Università degli Studi di Firenze

INTERNATIONAL DOCTORATE IN

"Mechanistic and Structural Systems Biology"

CYCLE XXV

COORDINATOR Prof. Roberta Pierattelli

"Mechanistic, structural and dynamic characterization of biomolecules: from DNA to membrane proteins"

S.S.D. CHIM/03

PhD student

Cerofolini Linda

Tutor

Prof. *Luchinat Claudio*

2010-2012

This thesis has been approved by the University of Florence,

the University of Frankfurt and the Utrecht University

Table of contents

1. Introduction

One of the present challenge in life sciences is the investigation of the roles of biomolecules in the molecular machines. Structural biology contributes to address this relevant biological puzzle by deciphering the unique tridimensional structure of biomolecules since it is directly linked with their function in the cell.

However, a complete illustration of the activity of biomolecules requires a combination of the knowledge of the tridimensional structure at the atomic level with the understanding of the nature and role of the conformational dynamics[1]. Therefore, a description of the dynamics must be associated to the snapshots of biomolecules frozen in their static structures to solve the structure-function paradigm.

The three-dimensional structure of a biomolecule only provides the lowest energy average atomic positions originated from motions of sizable amplitudes experienced by the atoms[2]. Such motions can occur on a small timescale (fast motions of the order of picosecond to nanosecond), involving atomic fluctuations around the average structure, or can involve large-scale reorganization (slow motions up to seconds), occurring in many biological processes, such as folding processes, enzyme catalysis, signal transduction and protein-protein interactions (Figure 1).

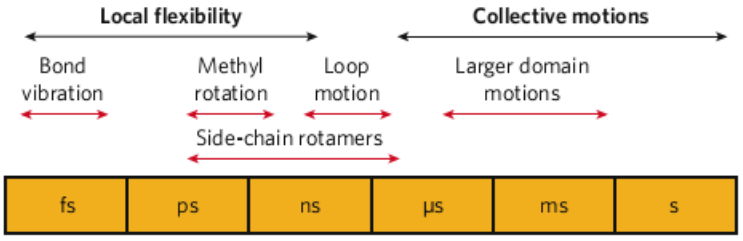


Figure 1. Timescale range for different biomolecules motions.

Biomolecules must be able to move in order to function and also the fluctuations on small timescale observed at the equilibrium govern the biological function itself. In fact, the conformational sub-states, sampled at the equilibrium in the fast-timescale, disclose the functional conformational states covered by collective domain motions on the slow-timescale. Therefore, DNA replication, enzyme catalysis, protein–ligand interactions and signal transduction occur as a result of the binding of specific ligands to complementary pre-existing states of a biomolecule and the consequent shifts in the equilibria[3].

Since the last decades two different mechanisms have been invoked for the description and coupling of the processes of ligands binding and conformational changes occurring in macromolecules: *induced fit* and *conformational selection*. The *induced fit* mechanism concerns with the previous binding of the ligand to the biomolecule and the occurring of a subsequent conformational change. Instead, the *conformational selection* mechanism regards the pre-existence of conformational states in the biomolecule independently of the ligand presence and their selective stabilization by the ligand binding itself. The issue of distinction between the two mechanism has proven to be of fundamental importance in understanding the mechanism of biological processes, such as folding mechanism, action of enzymes, protein-protein interactions and signal transduction[4]. However, when dealing with conformational ensembles, where multiple conformations are involved, the distinction between the two mechanisms is confusing and both *conformational selection* and *induced fit* should be recalled to have a more accurate description of the biological process.[5]

Nuclear Magnetic Resonance (NMR) spectroscopy has emerged as the method of choice for studying both biomolecules structure and dynamics in solution.

NMR-based structure determination relies on the collection of a set of experimental structural parameters, such as interatomic distances and dihedral angles, that are implemented as restraints in a molecular modeling algorithm to obtain a representation of the tridimensional structure of the biomolecule[6].

NMR can be exploited also to determine the rates of interconversion between different biomolecule states and have a detailed picture of the dynamics, meant as any time-dependent change in atomic coordinates. Various timescales (from picosecond up to seconds) can be investigated by NMR, with experiments measuring the nuclear relaxation rates. The values of longitudinal and transverse relaxation rates provide, indeed, indications about internal motions occurring on the fast timescale, shorter than the rotational correlation time (τ_c), while relaxation dispersion measures dynamics in the timescale between microseconds to milliseconds.

The presence of a paramagnetic metal ion in a system affords the possibility to extract additional distance, orientational and dynamics information for a more accurate description of the biomolecule. Paramagnetism-based constraints, such as pseudo-contact shifts (PCS) and residual dipolar couplings (RDC), can be exploited as long-range distance restraints and as orientational restraints, respectively, for the structural refinement of proteins in solution[7].

Moreover, the measurement of the modulation of the values of residual dipolar couplings (RDC)[8] provides important indications about the dynamics in the timescale window between nanoseconds and microseconds (blind zone), otherwise hidden to NMR relaxation measurements[9].

The fundamental challenge of NMR remains to combine and reconcile all the available structural and dynamic information. In this way, a complete, more accurate and realistic representation of the conformational space sampled by biomolecular systems can be obtained, and the relationship between structure, dynamics, and function can be resolved[6].

Experimental data obtained by NMR represent average values over the entire ensemble of conformations that coexist and rapidly interconvert in dynamic equilibrium. Therefore, a system can be better described by ensembles of conformations with respect to a single static structure[10]. For this reason, solution structures obtained by NMR are provided as families of structures displaying higher heterogeneity and lower definition in the regions characterized by higher mobility. Moreover, the description of the activity of multidomain proteins can be provided only by ensembles of conformations with different reciprocal interdomain orientations and positions. The accurate identification of these ensembles represent an "ill-defined inverse problem" whose resolution requires the development of robust statistical approaches to determine the existence probability of any conformation[11].

Changes in global orientation of protein domains are more difficult to characterize using NMR alone, and can be determined more accurately by combining NMR with small-angle X-ray scattering (SAXS)[12]. SAXS is particularly suitable for the study of less structured systems, such as intrinsically disordered proteins and multidomain proteins with flexible linkers. It provides information about the global shape of the systems, that is handled as a mixture of conformations in solution, and accounts for complementary information to NMR[12].

Conformational changes always requires hopping from one allowed state to another that involves the crossing of a barrier of high energy. A complete description of dynamics thus requires a multidimensional energy landscape that defines the relative probabilities of the conformational states and the energy barriers between them. As consequence, the timescale of the transition depends on the energy barriers themselves (Figure 2).

Although NMR experiments can determine with atomic resolution the dynamics of biomolecules, only molecular-dynamics simulations can speculate about the rationale behind the biomolecule motions and provide a complete description of the dynamics. In molecular-

dynamics simulations the precise position of each single atom at any instant in time with the associated energy can be followed, provided that one high-resolution structure is known as starting point.

An accurate energy description of the system can be, however, achieved only if the computational prediction are then experimentally validated.

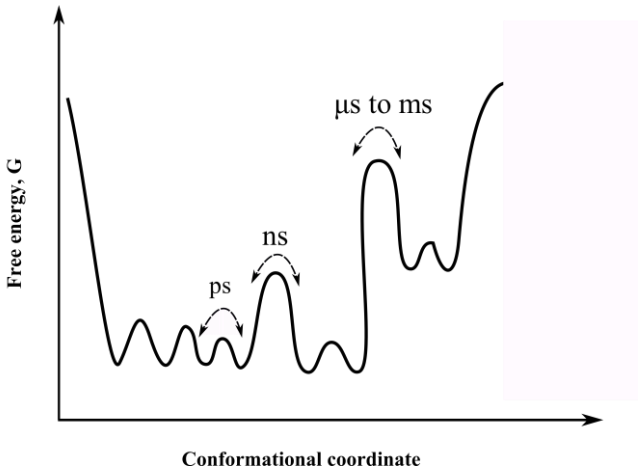


Figure 2 Free-energy landscape describing the energy related to possible motions of biomolecules.

However, conventional molecular dynamics can access only the timescale of the motions of hundreds of nanoseconds and biomolecules dynamics on the microsecond-to-millisecond timescale, that mostly occur during biological processes, are out of reach for these simulations[3,13].

Nevertheless advances in the study of long timescale processes have been recently reported with the aid of enhanced sampling techniques, such as metadynamics[14]. Metadynamics allows the exploration of the free energy surface (FES) of the process of interest combining the idea of coarse-grained dynamics in a space defined by a few collective coordinates with the introduction of a history-dependent (adaptive) bias. The chosen collective variables (CVs) must describe in the best way all the slow modes relevant for the process under study. Metadynamics simulations can be applied for the investigation of ligand-protein interaction process[13], protein conformational rearrangement[15] and folding/unfolding processes of nucleic acids.

In the present research project, an extensive NMR analysis has been carried out on four different biomolecules to clarify the structure/dynamics-function paradigm. NMR data have been integrated with structural and dynamical information provided by other biophysical and computational techniques such as SAXS and metadynamics to better characterize the structural and dynamical features of the investigated biomolecules in solution. NMR has thus been used i) for the identification of a new DNA structural folding, ii) for the speculation of the mechanism of catalysis of an enzyme, highlighting the most probable conformations in solution, iii) to underline the functional states of proteins in solution, and iv) to shed light on the mechanism of signal transduction of the Receptor for Advanced Glycation Endproducts.

NMR is, indeed, the most suitable technique for the structural characterization of nucleic acids. A high percentage of DNA and RNA structures, present in structural databases (Protein Data Bank (PDB) and Nucleic Acid Database (NDB)), has, actually, been solved by NMR. In the present research project, NMR has been used for the experimental validation in solution of a new structural folding of DNA identified during metadynamics simulations on the folding/unfolding process of a particular DNA construct. The integration of NMR and metadynamics techniques has provided a complete description of the properties of this new motif and hints for the speculation about the folding/unfolding processes of DNA occurring during cell replication.

The assessment of the conformational space sampled by multidomain enzymes, required for the explanation of their mechanism of catalysis, can also be accessed using NMR. The description of the ensemble of conformations spanned by a multidomain protein in solution can be achieved only with the implementation of sophisticated computational approaches. In the present research project, a new computational method (the *maximum occurrence*[16] approach) has been exploited for the inspection of the conformational heterogeneity of the enzyme matrix metalloprotease-1. Using a combination of paramagnetic NMR constraints and SAXS experimental data the conformations that can exist for the maximum percent of time in solution have been identified; those conformations account for the description of the antecedent step of the mechanism of catalysis.

Detailed knowledge of proteins structural characteristics is necessary for understanding their biological function. NMR can provide a high-resolution description of the structure acquired by a protein in solution and information about its dynamical properties. In this respect, in the present research, the structure, dynamics and metal-binding properties of one of the members of S100 proteins family (S100A14) have been solved using NMR techniques. The new conformation discovered for this member and its sizable dynamical properties account for its functional properties and assess the high heterogeneity of the EF-hand protein superfamily.

Membrane proteins and receptors can also be investigated by NMR. Speculation about the signal transduction mechanism can be provided by the dynamics analysis of the cytoplasmic domain of those proteins. In fact, the extracellular activation of a receptor triggers the interactions of its cytoplasmic tail with intracellular partners. In the present

research project, the cytoplasmic tail of the multiligand receptor RAGE has been investigated by NMR as part of the full-length protein in the presence of a membrane-mimicking environment and for comparison purpose as isolated domain. The highlighted versatile properties of this domain shed light on the mechanism of the signal transduction.

1.1 A New DNA Structural Motif: the G-Triplex

Nucleic acids represent one of the main class of the biomolecules of interest for structural biology because they regulate functions at the basis of cell life.

Despite DNA oligonucleotides exhibit lower conformational and chemical diversity than RNA oligonucleotides, since the discovery of DNA double helix conformation in 1953[17], several folding have been revealed for DNA. Beside the canonical Watson-Crick hydrogen bonding scheme present in DNA double-stranded structures (A-DNA, B-DNA, C-DNA, Z-DNA) a large number of base paring schemes that allow DNA molecules to form high-order structures have been identified. Hydrogen-bond complementarity is at the basis of the recognition and stabilization of the strands of DNA and among the nonstandard DNA base pairing the Hoogsteen pairings[18] has been found in many tertiary structures. These complex tertiary structures of DNA, such as hairpins[19], cruciforms[20], parallel-stranded duplexes[21], triplexes[22], G-quadruplexes[23] and the i-motif[24] are involved in many important processes like DNA packaging, replication, transcription and recombination. Moreover a strict relationship between the sequence composition, the geometry and the function of these oligonucleotides has been highlighted. Among those peculiar organizations of DNA, G-quadruplex structures have been widely studied.

1.1.1 G-quadruplex DNA structures

During the last decades unusual four-stranded DNA arrangements, named G-quadruplexes, has emerged as three-dimensional structures of large interest.

Guanine-rich sequences exhibit G-quadruplex formation tendency and are widely distributed in the eukaryotic genome because of their fundamental role in protecting the cells from recombination and degradation. Prevalently, these motives can be found at the telomeric ends of chromosome[25], as non-coding repeat sequences associated with the chromosomal maintenance, and in other regions with various roles, such as in the promoters of many important genes and oncogenes[26], at the immunoglobulin-switch regions[27] and in the regulation of insulin gene[28]. Telomeres, in particular, are composed of double-stranded (TTAGGG)n repeat sequences with protruding 3’ single-stranded overhang that can fold back into a G-quadruplex structure. The maintenance of telomeres length is ensured by telomerase

activity that adds repeats to their ends. In human somatic cells, telomerase activity is gradually lost at each mitotic cycle causing the shortening of telomeres with the associated genomic instability, growth arrest and eventual cell death[29]. Conversely, transformed cells, such as cancer cells, can bypass the progressive telomere loss and the limitations on proliferation by activating telomerase and thus becoming immortal cells. Many telomerase inhibitors have therefore been proposed as anticancer drugs. Furthermore, G-quadruplexes have been shown to inhibit telomerase activity and drugs that stabilize these tetraplex structures can interfere with telomere replication[30,31]. In this respect, the interest in the study of the structure and folding/unfolding process of this particular DNA folding has increased.

G-quadruplexes are constituted by a core of squares of guanine tetrads or G-tetrads stacked one upon the other. A G-tetrad consists of a planar arrangement of four guanine bases associated through a cyclic array of Hoogsteen-like hydrogen-bonds in which each guanine accepts and donates two hydrogen-bonds (Figure 3).

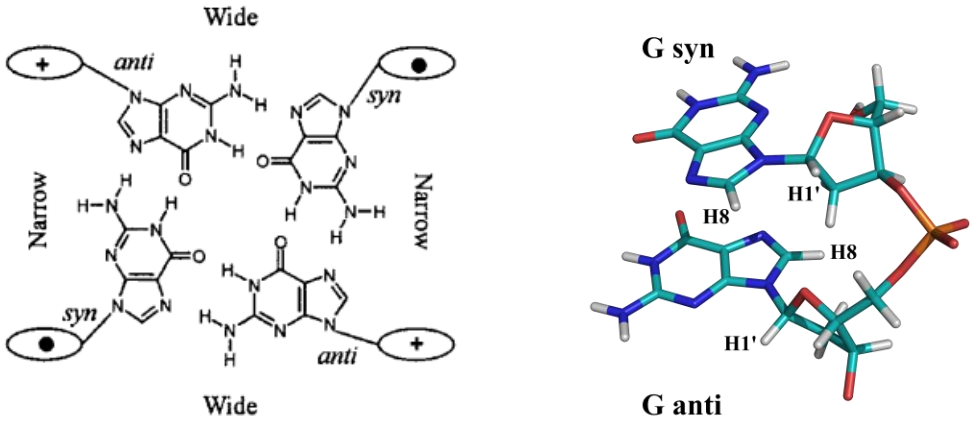


Figure 3. G-tetrad representation. In this case, the strands are alternating antiparallel and the Guanines have opposite glycosidic torsion angle conformation. Two wide and two narrow grooves are produced. Representation of the *syn* and *anti* conformations.

G-quadruplexes are characterized by extensive polimorfism depending on the strand stoichiometry and polarity, on the glycosidic torsion angle variation and on the conformation of the connecting loops[32]. In fact, G-quadruplexes can be formed by the association of one up to four strands and irrespective of the stoichiometry, the strands can come together in four different ways: all parallel, three parallel and one antiparallel, adjacent parallel, or alternating antiparallel. Moreover, unlike canonical B-DNA tertiary structures, that present exclusively a *anti* conformation for the glycosidic torsion angle, in G-quadruplexes organization both the *syn* and the *anti* conformation are observed (Figure 3). The glycosidic conformation

influences indeed the relative orientation of bases in the G-tetrad and also affects the stacking energy. For example, if the guanines forming the G-tetrads are located in antiparallel strands they must have opposite torsion angles to be in the correct orientation to form Hoogsteen-like hydrogen-bonds. Furthermore the stacking of guanine tetrads produces four grooves that are not necessarily identical but can be wide, narrow and medium; i.e. in the case of alternating antiparallel strands exclusively two wide and two narrow grooves are produced (Figure 3).

Those grooves are the sites of specific recognition of partners such as proteins and of therapeutic agents that acts through electrostatic interactions and intercalation with π-π stacking, exploiting the particular features of this structure.

The sugar puckering usually adopted by bases in B-DNA is the *C2'endo*. In the case of G-quadruplexes the bases in the *anti* conformation usually adopt a *C2'endo* sugar puckering, instead the bases in the *syn* conformation can adopt also the *C3'endo* organization.

The loops that connect the G-tetrads run on the outside of the guanines core and form either diagonal or edgewise tracts. Most of those loops are not loose or flexible but form stacking interactions with the guanine tetrads close to them. The conformation and composition of the loops can vary among the different G-quadruplexes and are also responsible of the specific interaction with partners[23].

In addition to the eight Hoogsteen hydrogen bonds the G-quadruplex tertiary structures are further stabilized by the presence of metal ions, such as sodium, potassium and strontium that can fit well in the cavities formed by the stacking of guanine tetrads. The cavity between two planes of guanine tetrads is lined by eight carbonyl oxygen atoms that can all participate in the precise coordination of cations, reducing the repulsions and further promoting the stacking itself of the G-quartets. Most of G-quadruplexes greatly prefer potassium as coordinating ion, whose concentration is also higher than other metal ions in live cells nuclei (around 150 mM). In this respect, in living cells the presence of other ions can hinder the G-quadruplexes formation.

New G-quadruplex structures still continue to emerge and provide insight into their biological roles in the telomere function. NMR is a unique method to screen for the formation of nonstandard structural elements and establish local details and stability as well as dynamic behavior of noncanonical helical-type structures with single residue resolution.

1.1.2 The Thrombin-binding aptamer (TBA)

The thrombin-binding aptamer (15-TBA) is an intriguing example of a DNA assuming a G-quadruplex structure and for this reason it has been widely characterized and studied. Aptamers are synthetic DNA or RNA oligonucleotides that specifically bind with high affinity a wide range of targets, from small molecules to whole cells.

15-TBA was screened from a randomly generated population of sequences for its ability to bind a desired molecular target, such as the serine protease thrombin, which plays a key role in the blood coagulation cascade. This aptamer was, indeed, able to inhibit the thrombin-catalyzed fibrin-clot formation in vitro via thrombin binding. In addition to its intrinsic affinity for thrombin and potential medicinal value, 15-TBA also represents an important system to study the structure and the basic physical chemistry of G-quadruplex DNA folding. It contains the minimum number of G-quartets (just two) formed by 15 nucleotides (15-mer), d(GGTTGGTGTGGTTGG). The oligonucleotide folds into a unimolecular quadruplex[33] where the two G-quartets are linked by two TT loops at one end and a TGT loop at the other end. The nucleotides in the two G-quartets are 5'-syn-anti-3' along each strand. Adjacent strands are anti-parallel, resulting in G-quartets in which the bases are alternately *syn* and *anti* around the quartet and interact with head-to-tail faces; two wide and two narrow grooves are formed (as illustrated in the scheme above, figure 3). The two TT loops span the narrow grooves at one end and the TGT loop spans one of the wide grooves at the other end of the quadruplex[34] (Figure 4).

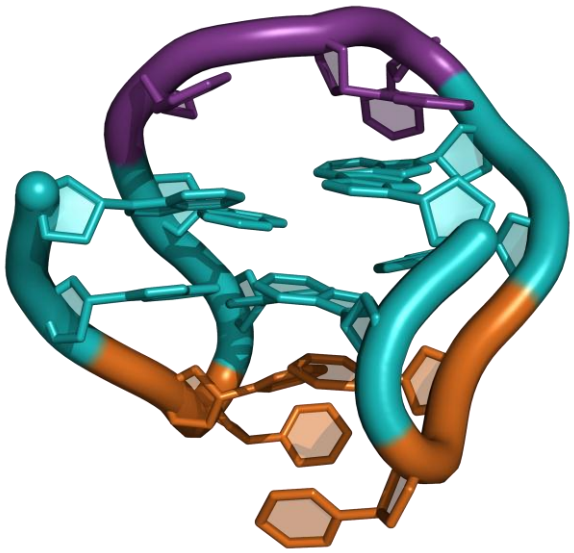


Figure 4 Thrombin binding aptamer. In cyan the G-tedrads core, in orange the TT loops spanning from one side and in violet the TGT loop spanning on the other side of the G-tetrad.

X-ray and NMR structures have

been deposited for the isolated 15-TBA[34,35] and for 15-TBA in complex with thrombin[36]. The two models are mutually inconsistent differing both in chain direction and loop geometry. Molecular dynamics simulations performed using these models as starting structures, respectively, strongly support the NMR-based model of 15-TBA[34]. Moreover, those simulation highlight the importance of the loops in the stabilization of the G-stem, being the whole 15-TBA NMR-structure more stable than the G-quartet stem alone, and at least two nucleotides (G8 and T9) stacked with the upper G-quartet are needed for the molecule's viability. Conversely, the X-ray structure resulted less stable than both the NMR structure and the G-quartet stem[37]. The TT loops have a destabilizing effect and are essential for the interaction with thrombin. A further stabilization for the structure is provided by potassium ion coordinated in the middle of the two G-quartets.

1.1.3 Folding/unfolding process of G-quadruplex formation

The understanding of quadruplex conformation and dynamics is crucial to elucidate its function. Although, the G-quadruplex conformation has been described in many studies, the details of the dynamics occurring during the folding/unfolding are not clearly understood.

Single-molecule fluorescence resonance energy transfer (FRET) was previously used to investigate conformational heterogeneity and dynamics fluctuations in nucleic acids[38]. Other conditions and constructs were then exploited by another group to study the process of G-quadruplex formation[39]. Conformations different from G-quadruplex were highlighted in these studies as important intermediates for the folding process.

Independent studies performed recently using molecular dynamics simulations[40], have proposed that those intermediates could assume a hairpin and a triplex structure. These predicted G-triplex conformations diverge from the already known triplex[41] structures because constituted by G:G:G triad planes stabilized by an array of Hoogsteen-like hydrogen-bonds.

The existence of a G-triplex structure has been evidenced by NMR in the (3+1) G-quadruplex structure, where it has been proposed that constructs displaying only three tandem guanine repeats can form an asymmetric heterodimeric complex between the three-repeats and a single-repeat of the human telomeric sequence[42]. Nevertheless, stand-alone G-triplex structures were not directly observed in three tandem guanine repeats.

The only experimental evidence of an intramolecular folding in three tandem guanine repeats of human telomeric DNA has been recently proposed using optical-tweezers, MD simulation and circular dichroism[43]. A mechanically and thermodynamically stable species in this sequence has shown a structure consistent with a triplex conformation. The G-triplex was found to be less stable than the G-quadruplex due to the greater contribution from stronger base stacking and higher number of Hoogsteen hydrogen bonds in the latter. However, with this experimental evidence the existence of an intramolecular G-triplex folding can only be hypothesize.

Currently, the G-triplex conformation has thus not been experimentally demonstrated yet with a model at atomic resolution.

1.1.4 Aim of the project

Only advanced state-of-the-art computations, such as metadynamics, can allow the identification of short-lived and less stable states of biomolecules and can thus be used for the description of the intermediates of a pathway. Therefore, in the present research project a combination of metadynamics simulations with NMR, circular dicronism (CD) and differential scanning calorimetry (DSC) experiments was carried out for the identification of an intramolecular G-triplex DNA structural motif.

The folding/unfolding process of the DNA G-quadruplex, thrombin binding aptamer (15-TBA), has been described through well-tempered metadynamics simulations and the formation of a very stable intermediate, displaying G:G:G triad planes stabilized by an array of Hoogsteen-like hydrogen-bonds, has been highlighted.

The present project aims at providing an experimental evidence to this new DNA structural motif. A truncation of 15-TBA, involving the last strand at the 3' ends (d(GGTTGGTGTGG)), has been investigated and its intramolecular folding with G-triads planes has been confirmed using NMR distance constraints.

1.2 Towards the characterization of the mechanism of collagenolysis by matrix-metalloprotease-1

a) The family of Matrix Metalloproteases

Matrix metalloproteases (MMPs) are a family of extracellular hydrolytic enzymes in charge of the degradation of the components of the extracellular matrix (ECM).

MMPs belong to the metzincin superfamily of metalloproteinases, characterized by the presence of a catalytic zinc atom in the active center and the HEXXHXXGXXH/D zinc-binding consensus sequence, followed by a conserved methionine residue. The family of human matrix metalloproteinases includes 24 members, displaying different substrate specificity for the components of the extracellular matrix. The latter is a heterogeneous mixture of well-organized and well-structured proteins, such as collagens, elastines, laminins, fibronectins and proteoglycans, that provides the scaffold on which cells and tissues are anchored. However, extracellular matrix, is not a passive support for cells but acts also as reservoir for embedded cytokines and growth factors. Moreover, the ECM conceals "cryptic" information within the proteins that constitute itself[44].

An old classification of MMPs[45] was performed on the basis of substrate preferences and MMPs were subgrouped as: i) stromelysins (MMP-3, MMP-10, MMP-11), ii) gelatinases (MMP-2, MMP-9), iii) collagenases (MMP-1, MMP-8, MMP-13), iv) elastases (MMP-12), v) matrylisins (MMP-7, MMP-26) and vi) membrane proteins (MMP-14, MMP-15, MMP-16, MMP-24). However, their selectivity towards these substrates is not high[46] and most of them hydrolyze also other extracellular components[47], such as extracellular domains of membrane receptors[48], as well as other proteases.

Recent classifications are based on bioinformatics and structural analysis, which through sequence and structural alignment allow to identify five main groups: i) non-furine regulated MMPs, ii) gelatinases, iii) transmembrane MMPs, iv) glycosylphosphatidylinositol (GPI)-anchored MMPs and v) others[49].

MMPs are involved in several physiological functions such as embryogenesis, tissues growth, development[50,51], wound healing, and cell migration. However, MMPs function does not depend only on the direct effect of the hydrolysis of the extracellular matrix[52,53] but is also related to the release and activation of cytokines and growth factors stored inside and to the

disclosure of the "cryptic" information by generating bioactive fragments. In this way, MMPs can modulate the activities of a wide range of extracellular but also intracellular proteins and regulate cell proliferation, adhesion, migration, growth factor bioavailability, chemotaxis, and cell-signaling[54]. Recently, an intracellular localization has been reported for some members and a role in the proteolysis of intracellular substrates proposed[55].

Because of their potential destructive effect, MMPs activity is strictly regulated in healthy tissues (Figure 5) by a tight control of expression, secretion and clearance and by the presence of endogenous inhibitors, tissue inhibitors of metalloproteinases (TIMPs)[56]. Moreover, MMPs are produced by the cell as inactive with a pro-domain preventing the enzymatic activity that is then removed by proteolytic cleavage. Some of the members of the MMPs family themselves participate in the activation of other members, for example the subgroup of stromelysins regulates collagenases function[57] and the membrane-type metalloproteinases, such as MMP-14, activate MMP-2[58].

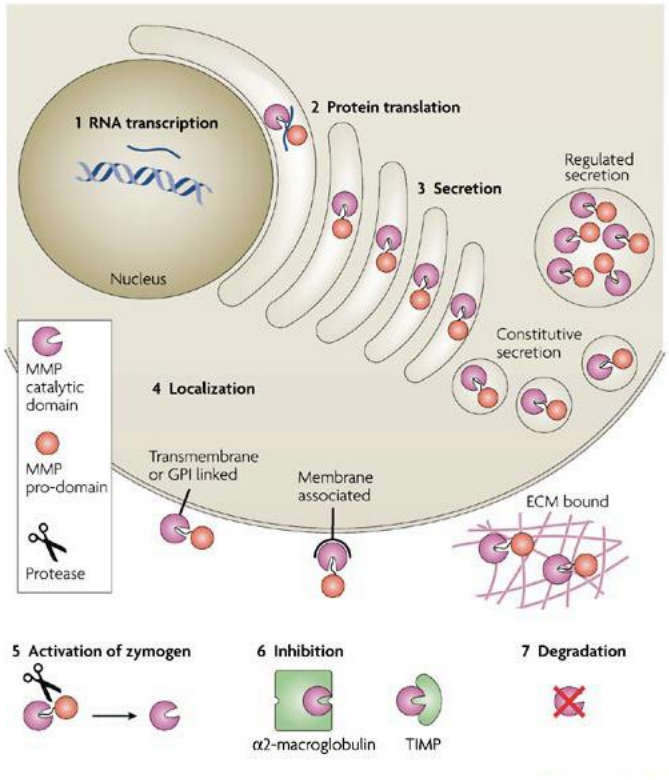


Figure 5. Regulation of MMPs function at various levels.[49]

When the balance among the inhibition and activation of MMPs is disrupted, pathological states occurs. In fact, the dysregulated activity of these enzymes is associated to some diseases, such as inflammatory and auto-immune disorders, rheumatoid arthritis, cancer and metastasis[59]. Some synthetic inhibitors of MMPs were thus designed as potential anti-cancer drugs[60]. Unfortunately, these molecules failed the clinical trials due to side effects and a poor pharmacological activity. The main reasons of these unsuccessful results are related to the no proper MMPs subgroup selection by these inhibitors[61] and to the different roles exerted by various MMPs in distinct stages of cancer[62], to such extent that the inhibition of the activity of some MMPs can exacerbate the pathology. Therefore, the future development of MMPs inhibitors as anticancer drugs comprise the identification of more selective molecules

towards the mechanism of action of specific MMPs[63] or the identification of particular exosites (see later).

b) Structural and dynamic features

MMPs are synthesized as zymogens with a signal peptide which leads them to the secretory pathway. Then, they can be secreted outside the cell or can be anchored to the plasma membrane, thereby confining their catalytic activity to the extracellular space or to the cell surface, respectively (Figure 5). However, several MMP family members[52,55,64] can be found also as intracellular proteins, although their functions at this subcellular location are still unclear.

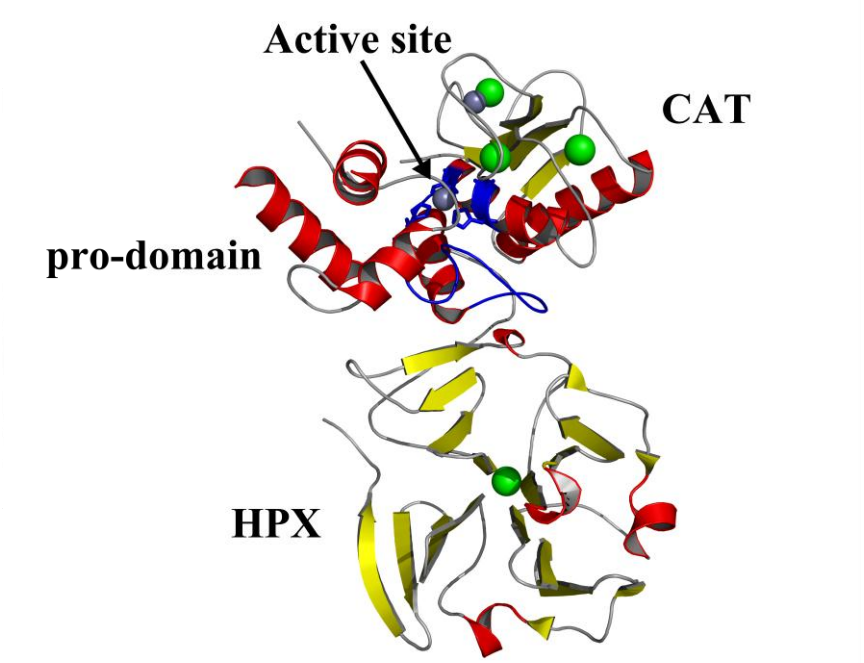


Figure 6. General structural representation of MMPs. The pro-domain, CAT and HPX domains are highlighted in the figure. The active site with the zinc(II) ligands and the loop L8 that flanks the S1' pocket are colored in blue.

Most of MMPs are constituted by a prodomain (from 66 to 80 amino acids), a catalytic metalloproteinase domain (of about 160 amino acids) and a hemopexin domain (of about 210 amino acids) connected to the catalytic domain by a linker peptide of variable length (Figure 6).

The prodomain maintains the active site inaccessible to substrates through a cysteine switch PRCXXPD consensus sequence, until the proteolytic activation occurs.

The catalytic (CAT) domain is largely conserved in all MMPs and consists of three α-helices, a twisted five-stranded β-sheet and eight intervening loops. In almost all MMPs the CAT domain contains two zinc(II) ions, one responsible for the catalytic activity, and the

second with a structural function. Furthermore, in the CAT domain from one to four calcium(II) ions with structural role are present.

The catalytic zinc(II) ion is coordinated by three conserved histidines and by one water molecule that is hydrogen bonded to the catalytically relevant glutamate and is, in this way, activated for a nucleophilic attack towards the peptide bond of the substrate, allowing its hydrolysis also at neutral pH[65]. The substrate binding groove is constituted by the catalytic zinc ion and several binding pockets. The hydrophobic S1' pocket, delimited by the last loop (L8), represent together with the zinc(II) ion, the preferred site for synthetic inhibitors. In fact, this loop is a region of relatively large variability among MMPs and can be targeted in order to have selectivity[66]. The inhibitors bind the active site in a way that resembles the substrate in the transition state, fitting the deep S1' pocket with a lipophilic moiety[67].

The hemopexin-like (HPX) domain has the same structural features in all the members of the family and it is constituted by four β-sheets organized in a symmetric four-blade propeller, forming a deep tunnel closed by a calcium ion at the bottom. The folding of HPX domain is stabilized by a disulfide bridge that links the beginning of the first blade with the end of the fourth blade of the propeller.

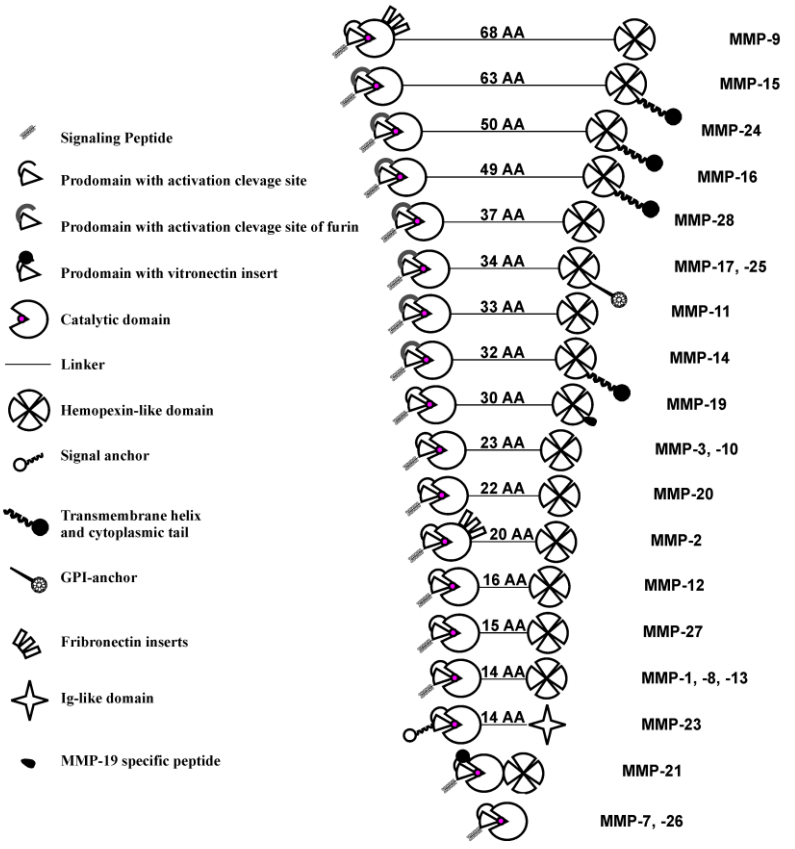


Figure 7 Schematic domain structure of MMPs with the corresponding linker length[68]

These two domains experience large amplitude motions in correspondence of the loops. Especially the loop flanking the S1' pocket in the CAT domain, mainly involved in the substrate recognition, is characterized by high flexibility in all the members of the

family, indicating the importance of conformational heterogeneity in these domains for the hydrolysis of the substrates[68].

The CAT and the HPX domain are connected together by a proline-rich linker whose length vary among the members of the family (Figure 7) ranging from 14 amino acids in collagenases (MMP-1, MMP-8 and MMP-13) to 68 AA in MMP-9.

The linker allows large interdomain flexibility and binding to structurally unrelated substrates with a variety of molecular conformations. It, thus, permits interdomain reorientation during the explication of the mechanism of these enzymes.

Interdomain flexibility of two of the members of the family with the shortest linkers (MMP-1 and MMP-12)[69,70] has been recently investigated by a combination of solution NMR and SAXS measurements. NMR relaxation data indicated that the two domains are not rigidly held to each other but they can experience independent motions. Moreover, the analysis of SAXS data indicated that a good description of the behavior of these proteins in solution can be provided only when ensembles of heterogeneous conformations are taken into account. In fact, the crystallographic structures displaying a closed conformations, when taken alone, fail in the description of experimental data. Instead, when these structures are flanked by a significant amount of extended conformations in equilibrium with the closed ones, a good agreement with the experimental data is achieved.

From these observations it was shown that the two domains can freely reorient among each other. This data can, however, hardly provide a quantitative and detailed description of the conformational heterogeneity of the protein and highlight the most probable conformations sampled in solution, that are necessary to afford hints for the mechanism of collagenolysis.

c) MMP-1 and the collagenolytic mechanism

Collagens are the major structural proteins of connective tissues and can be classified in three different types, I, II and III. Interstitial collagens have a unique molecular structure that consist of three α-chains staggered by one residue and constituted by repeating Gly–X–Y triplets, where X and Y are often proline and hydroxyproline, respectively[71]. Each α-chain, for the high content of proline and hydroxyproline in the sequence, forms a left-handed poly-Pro II-like helix. The three α-chains are coiled in overall right-handed superhelical conformation[72], stabilized by hydrogen bonds among the backbones of three α-chains, that

make the structure resistant to many proteases. In vertebrates, the only enzymes that can cleave this very complex triple-helical structure are collagenases[56] of the MMPs family and cathepsin K[73] produced by osteoclasts.

The preferential cleavage site of MMPs on the three α-chains of native triple-helical collagens is after a Gly residue. In particular, the sequence of choice is (Gln/Leu)–Gly#(Ile/Leu)–(Ala/Leu) (# indicates the scissile bond), that is located approximately three quarters away from the N-terminus of the entire collagen molecule. Once collagens are cleaved into 3/4 and 1/4 length fragments[74], they denature at physiological temperature and can be degraded by gelatinases[75] or by other non-specific proteases.

Among the collagenases, MMP-1 can cleave collagen type I, that is constituted by two α_1 chains and one α_2 and is 3000 Å in length and 15 Å in diameter. Although the CAT domain of MMP-1 by itself can cleave a number of non-collagenous proteins, its activity on native triple-helical collagen is negligible. In particular, it has been demonstrated that the presence of the HPX domain is required for the degradation of collagen type I by MMP-1[76,77,78,79,80,81,82], although the molecular details of the collagenolysis have not been completely elucidated. HPX domain has been proven to locally unwind the triple-helix in correspondence of the cleavage site, that is also more susceptible to change in conformation at high temperature. In this way, the accommodation of a single filament of collagen, otherwise inaccessible, in the CAT domain cleft, is allowed. The CAT domain partially participate itself in the local perturbation of the triple helical structure of collagen and is responsible of the sequential cleavage of the chains one by one (Figure 8). The two domains, joint together by the linker, cooperate for the interaction with collagen and the global affinity of the full-length protein is increased because of the benefit in the entropic contribution[81].

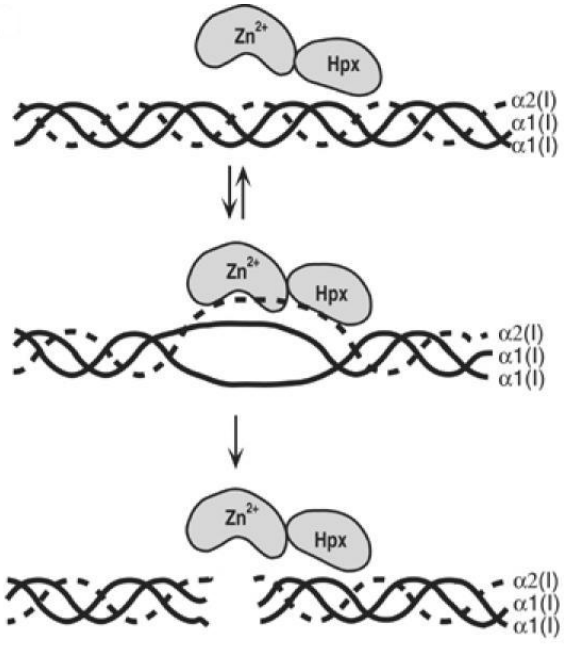


Figure 8 MMP-1 binds to and locally unwinds collagenbefore it cleaves the triple-helical interstitial collagen.

The overall reinforcement of the interaction due to cooperativity in the full-length protein with respect to the isolated

domains has been recently demonstrated also by NMR[83]. The surface of HPX domain, comprising blades 1 and 2 of the four-bladed propeller, has been identified[83,84] as responsible for the binding of collagen and for its destabilization and unwinding. Intriguingly, the HPX domain exosites are concealed in the interface between CAT and HPX domains in the crystal structure of full-length MMP-1[85]. This further suggests the presence of an equilibrium in solution among the closed conformation and more open/extended ones, that are also more poised for the interaction with collagen.

Also the region on collagen, where the binding of HPX domain occurs, has been

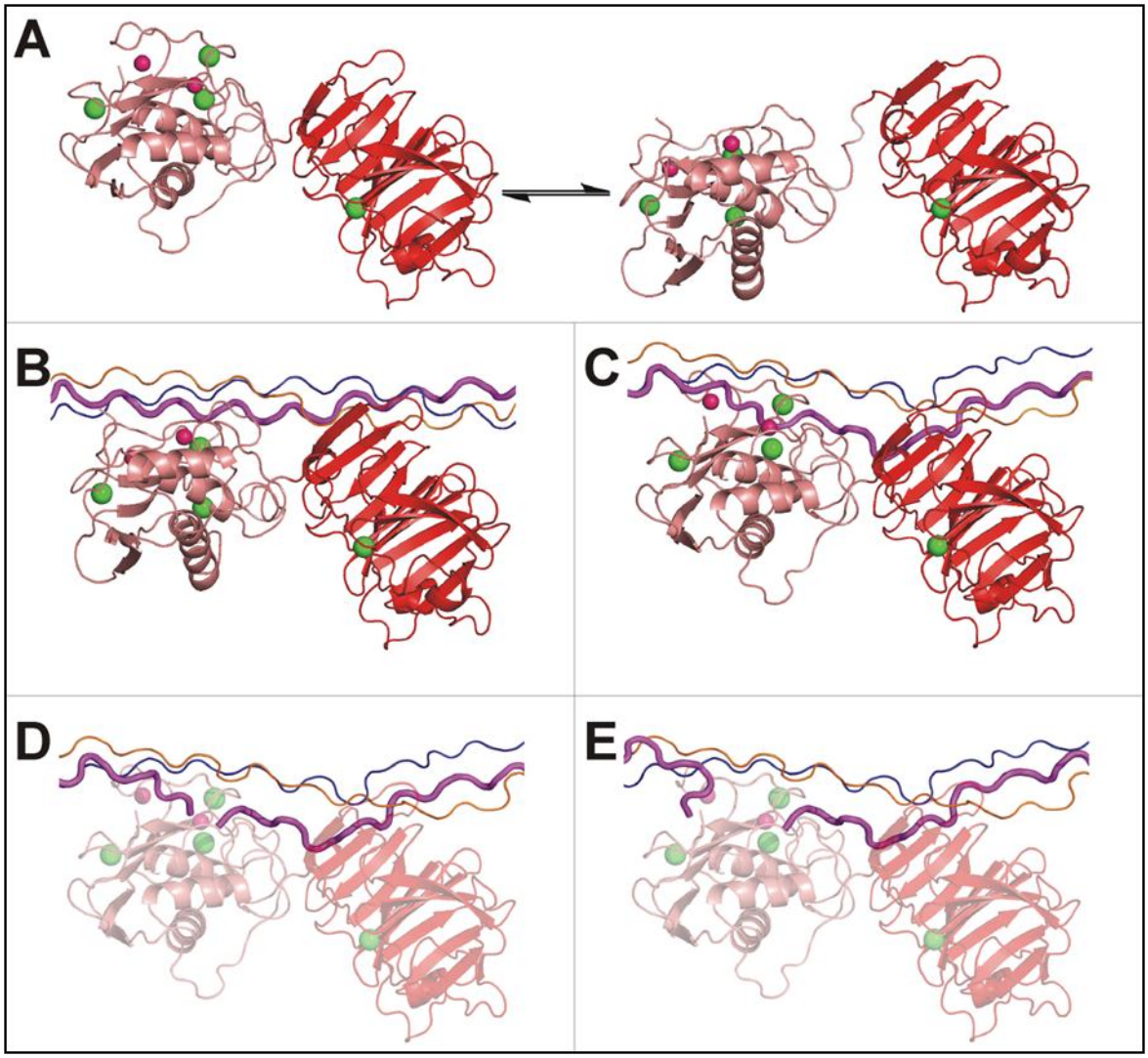


Figure 9. The initial steps of collagenolysis. (A) Equilibrium between the closed (left) and open/extended (right) forms of FL-MMP-1 in solution. (B) The extended protein domain binds the first two chains of THP with the HPX in a position downstream the cleavage site and with the CAT domain near the cleavage site. The THP is still in a compact conformation. (C) Closed FL-MMP-1 interacting with one of the released chain (in magenta). (D) After hydrolysis, both peptide fragments (C- and N-terminal) are initially bound to the active site. (E) The C-terminal region of the N-terminal peptide fragment is release[84,65]

highlighted using a collagen-like peptide and a plausible mechanism of collagenolysis has been proposed (Figure 9). The HPX domain drives the mechanism, interacting with the first two chains of THP in a position downstream to the cleavage site. Then the correct positioning of the CAT domain in front of the cleavage site is facilitated by interdomain flexibility and reorientation.

The linker flexibility appears thus crucial for the mechanism of collagenolysis by allowing the proper reorientation of the two domains during the catalysis. In particular, the residue Gly-271 in the linker has been reported to be critical for this mechanism and its substitution with bulkier amino acids reduces the efficiency of catalysis because of alteration of the flexibility of the linker[86]. In fact, once bound to collagen, a back-rotation of the HPX and CAT domains, to achieve the energetically favored closed conformation of the crystallographic structure, needs to occur, to induce the conformational change in the collagen. In this way, the overall favorable energetics, associated with the mechanism of collagenolysis, can readily account for collagen catabolism without the necessity of external energy source (such as ATP)[83].

Recently the complex between the full-length MMP-1 and a collagen-like peptide has been crystalized and the extensive interactions between CAT and HPX domains and collagen highlighted[87]. MMP-1 binds collagen in an extended manner rather than surrounding it and involves for the interaction many recognition sites distant from the catalytic site itself (exosites). This findings also agrees with a proposed "inchworm" mechanism[88] for the progressive movement and processing of collagen fibrils by matrix metalloproteases.

Although all aspects of collagenolysis have not been clarified yet, most of the experimental evidences suggest that the interdomain flexibility and conformational reorganization is fundamental for this enzymatic mechanism. Therefore, a detailed inspection in the internal dynamics of this multidomain enzyme is a prerequisite to enlarge the knowledge about collagenolysis.

d) Use of paramagnetic NMR for the study of dynamical properties

Paramagnetic NMR has been widely exploited for the study of the structure and dynamics of globular and multidomain proteins and for protein-protein interactions.

Biological systems can be naturally paramagnetic because of the presence of constitutive paramagnetic metal ion (such as Fe(III) or Cu(II)) or can be made artificially

paramagnetic by replacement of structural diamagnetic metal ions with paramagnetic ones [89]. Moreover, the implementation of tags chelating lanthanides with different paramagnetic strength has been largely employed as alternative[90], since this allows to select whatever position on the surface of any protein and monitor specific properties. For the use of paramagnetic experimental data as structural constraints, at least three different paramagnetic metal ions have to be exploited. In fact, only in this way it is possible to remove the symmetric, not real (or *ghost*) solution deriving from the degeneracy in the PCS and RDC equation (see Methodological Aspects).

Important prerogatives for these tags are the absence of isomerization and the rigid functionalization to the protein with well-defined position of the metal ion with respect to the protein. In fact, only in these cases single signals can be detected for the protein and no averaging of paramagnetic constraints due to motions of the tag with respect to the protein occurs, respectively.

Caged Lanthanide NMR Probe 5 (CLaNP-5)[91] has been widely used to introduce paramagnetic centers in diamagnetic proteins because it does not give isomerization, for its symmetry, and it is rigidly attached[92] to the protein via two disulphide bridge provided by two cysteins engineered on the protein surface.

Pseudo-contact shift (PCS) and residual dipolar coupling (RDC) can be investigated as distance and orientation constraints. Their combined use allows to highlight many important characteristics of a system. First of all it is possible to derive with high accuracy the orientation and magnitude of the tensors of the paramagnetic metal ions from PCS since they are less sensitive to local structural disorder and mobility. Second, the simultaneous use of PCS and RDC provides indications about local and global flexibility and about conformational freedom. Moreover, it is possible to exploit these paramagnetic constraints to determine the relative position and orientation of secondary structure elements (α-helices or β-sheet) and of entire domains with respect to the metal ion frame and to refine crystallographic structures accordingly to their values.

1.2.1 The catalytic domain of MMP-1 studied through tagged lanthanides

1.2.1.1. Structure refinement

Detailed information about the structure in solution of soluble proteins is required to better understand their function. However, NMR solution structures are not very precise because of the relatively few and loose experimental restraints and the precision of an "atomic resolution" crystal structure can never be obtained by NMR. On the other hand, crystallographic structures are more reliable, and are often taken as models of solution structures, although they may be inaccurate for the representation of the structure in solution because affected by crystal packing forces and evaluated in static conditions.

A possible strategy to obtain accurate and precise structures in solution is that of using a relatively good crystal structure as starting model and refining it by applying well-selected NMR restraints sensitive to local and/or global changes. An example of this strategy is provided by the use of PCS from a paramagnetic ion in a metalloprotein[93]. Also RDC measured with external orienting media have been proposed and used with high success[94]. More recently, a strategy for the refinement of protein structures or domains assumed to be rigid, based on the measurement of paramagnetism-based PCS and RDC deriving from a paramagnetic center, was proposed[7].

PCS and RDC, in fact, can act as reporters of structural information because they depend on the position of the observed nuclei (for PCS) or on the orientation of the vector connecting dipole–dipole coupled nuclei (for RDC) with respect to the paramagnetic susceptibility anisotropy tensor of the metal ion. PCS and RDC are, thus, differently sensitive to local and global motions.

PCS are mostly sensitive to large global protein conformational changes, and are scarcely affected by local mobility and structural inaccuracies. A robust estimation of the tensor of magnetic susceptibility anisotropy and of the position of the metal ion with respect to the protein can thus be obtained using PCS, once a structural model of the protein is available.

RDC, on the other hand, are sensitive also to small local reorientations and their agreement with crystal structures can be very poor. This can be caused both by local structural

inaccuracies and by global conformational differences and therefore RDC can be used for the adjustment of the orientation of protein regions.

Hence both PCS and RDC can be used to refine the structure through the PCS tensor. Moreover, the independent availability of an accurate estimate of the orientation tensor from PCS, actually permits a more quantitative use of RDC themselves for both structural and dynamical considerations.

1.2.1.2. Aim of the project

Previous protocols for structural refinement concerned with the use of dihedral angles, resulting from a crystal structure to fix the model, and the implementation of paramagnetic constraints (PCS and RDC), to drive local adjustments. The drawback of such approach is that even small changes in the values of backbone dihedral angles may result in relatively large structural changes after propagation on a number of residues, so that the protein structure may result distorted if the experimental restraints are few.

In the present research project, a new protocol has been proposed with the aim of obtaining structures with increased accuracy by anchoring the structure itself to the position of the atoms in the crystal model and adjusting only locally the orientation of the plane of bond vectors according to the solution paramagnetic constraints. With this protocol, PCS and RDC have been used to access the quality of X-ray structures and to refine them in solution performing only small variation. Two distinct cases have been considered and analyzed. In particular, if the solution and solid state structures are similar, a good agreement with paramagnetic experimental constraints can be achieved and the structure can be refined. Conversely, when the solution and solid state structures present different orientation for one or more secondary structural elements, the good agreement cannot be obtained.

1.2.2 Analysis of conformational heterogeneity in multidomain enzymes: the prologue of MMP-1 collagenolysis

1.2.2.1 Study of the conformational heterogeneity of multidomain proteins using paramagnetic NMR: the maximum occurrence approach.

Flexible systems, such as intrinsically disordered proteins and multidomain proteins with flexible linkers, can only be described by ensembles of structures in equilibrium with each other that take into account all the possible orientations and conformations acquired by the system. Usually multidomain proteins and unfolded proteins are involved in relevant biological functions and understanding their conformational heterogeneity is crucial for the investigation of their mechanism of action. However, the identification of ensembles of conformations sampled by these proteins from the solution average data represents an "ill-defined inverse problem"[11], because it has not a unique solution and instead infinite number of ensembles, not necessarily endowed of physical significance, can reproduce the experimental data. Various approaches have thus been proposed to solve this paradigm.

A combination of paramagnetic NMR with theoretical modeling through ensemble simulations is generally used to describe the conformational space covered by flexible systems in spite of NMR alone that can only provide a single, unique and averaged conformation.

The general approach adopted for globular proteins of limited mobility has been that of generating heterogeneous ensembles of conformations including the dynamics features of the protein to better fit the experimental parameters[1,95,96,97].

The algorithm flexible meccano[98] has been developed in the case of intrinsically disordered proteins to sample the conformational space on the basis of residue-specific propensity and side chains volume[99]. In this way, with the implementation of paramagnetic NMR restraints and SAXS data, it is possible to study the structural propensity of these proteins by selecting the best ensembles of conformers that better fits the experimental data

by averaging. Non-uniform weights can be given to the conformers with the inclusion of different numbers of conformers with similar shapes[12].

Towards the characterization of flexible systems, paramagnetic relaxation enhancement (PRE) has been widely used to detect transient sparsely populated states of multidomain proteins[100] as well as transient intermediates of macromolecular dynamic complexes[101], such as encounter ensembles[102,103]. With this approach, it is possible to determine the minimum degree of mobility of the system, necessary to recover an agreement with the experimental data and to define a probability distribution map for the lower populated states using ensemble simulated annealing refinement against paramagnetic relaxation enhancement data[101,102].

In the laboratory where I have performed my doctorate, for the structural and dynamics characterization of systems displaying conformational heterogeneity the idea of the maximum allowed probability (MAP)[104,105,106], then extended to the *maximum occurrence* (MO)[16,107,108,109,110] of a conformation, has been proposed. With the MAP approach, the conformations, defined by orientation plus translation, with the maximum probability to be sampled in solution by a protein with two rigid domains relatively free to move with respect to each other, could be searched and highlighted. The combined use of three set of PCS and RDC allowed to discriminate between real and symmetric *ghost* solutions.

With the evolution of the concept of the *maximum occurrence,* moreover, it is possible to assess the entire conformational space experienced by a multidomain protein and provide for each sterically allowed conformation a MO value, combining together paramagnetic NMR and SAXS constraints as averaged experimental data. The MO of any given conformation can be thus calculated and the regions with largest MO, as well as the regions which cannot be significantly sampled by the system, are identified. The *maximum occurrence* is defined as the maximum weight that a conformation can have, when taken together with any ensemble of conformations with variable weights, and still be in agreement with the experimental data. Since many different ensembles can equally well describe the experimentally averaged data, without necessarily be really represented in solution, no reliability can be given to the structures of the ensemble surrounding the selected conformation. Therefore, unlike ensemble averaging approaches, which can recover the conformational variability but give only plausible ensembles of conformations, with the *maximum occurrence* approach specific conformations sampled by the systems with their upper limit probability can be identified. In

this way, a more rigorous description of the conformational space of multidomain proteins can be given[111].

Moreover, the *maximum occurrence* method, scoring the conformations according to the maximum percent of time they can exist in solution, provides a distinct, even if not fully complementary, information to that obtained with the other described approaches[100]. Those approaches, in fact, search for the minimum lifetime of lower populated conformational states of multidomain proteins and complexes using paramagnetic relaxation enhancement alone.

The possibility of using together independent constraints, that are averaged in different ways in solution, such as paramagnetic constraints and SAXS data, allows to obtain MO values closer to the actual probability of the conformations in solution. In fact, increasing the number of variables used in the simulated annealing calculation, the MO values of less probable conformations decrease more than those of more probable conformations allowing an accurate mapping of the conformational space[108,110]. Moreover, with a higher number of independent constraints some possible *ghost* solutions due to degeneracy of both PCS and RDC can be better removed.

1.2.2.2 Aim of the project

The mechanism of collagenolysis by matrix metalloprotease-1 is at the basis of many physiological and pathological processes. Many speculations about this mechanism based on experimental data have been carried out since many years. The identification of the interaction areas of the two domains of MMP-1 and of a collagen-like peptide by NMR allowed to hypothesized a plausible mechanism for the collagenolysis in this laboratory[83].

Interdomain flexibility has been also largely invoked for the occurrence of the hydrolysis of structurally well-organized components of ECM, such as collagens.

In the present research project, the most probable conformations to be sample in solution by this multidomain enzyme have been highlighted using a recently developed computational method, the *maximum occurrence* approach. Moreover, with this method it is possible to obtain a complete characterization of the conformational space of the active MMP-1 in the absence of the substrate. The maximum occurrence value, defined as the maximum percent of time a conformer can exist in solution together with any ensemble of

conformations and be in agreement with experimental data, has been evaluated for more than 1000 random conformations covering all the possible conformational space. In this way, both the regions with high occupancy and those with low probability to be sampled have been identified.

The conformations with the highest maximum occurrence value are more extended and a have different interdomain orientation with respect to crystallographic structures. Moreover, the HPX binding region in these high MO conformations is solvent exposed.

This analysis sheds light on the elusive step preceding the collagenolysis and better elucidate the mechanism itself.

1.3. Solution structure and dynamics of human S100A14

1.3.1. The S100 proteins family

The family of the S100 proteins includes more than twenty calcium(II)-binding proteins, that belong to the EF-hand superfamily[112]. The members of the EF-hand superfamily are involved in the regulation of the intracellular calcium(II) homeostasis and/or in the transduction of the intracellular calcium(II)-mediated signals. These proteins are thus important for the regulation of numerous cellular functions, including the cell cycle, muscle contraction and apoptosis. Moreover, they have been implicated in various disease states, including cancer, Alzheimer's disease, and rheumatoid arthritis.

S100 proteins are calcium(II) sensor proteins and constitute one of the main protein families implicated in the aforementioned activities. They are expressed in a cell and tissue specific manner and have a variety of intracellular biological roles, such as the regulation of cytoskeletal protein assembly[113], protein phosphorylation[114], the activities of enzymes[115], the cycle of contraction–relaxation, calcium(II) homeostasis, cell proliferation and differentiation[116]. Among the intracellular partners of S100 proteins, over than 90 potential protein targets have been identified, including annexins (A1, A2, A5, A6, A11)[117,118], the ubiquitination protein CacyBP[119], p53[120], and tau[121].

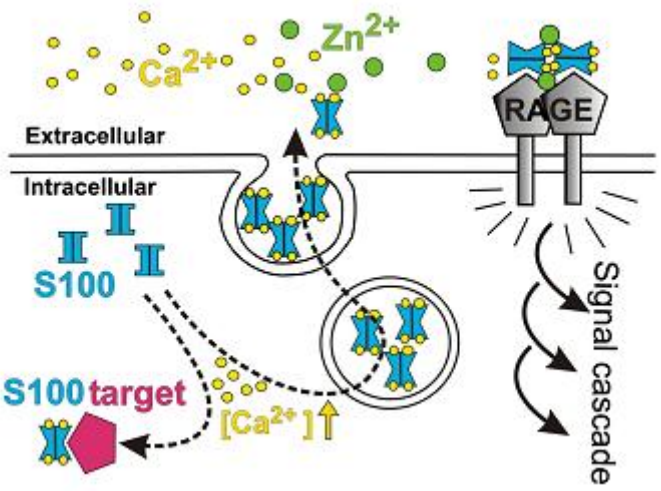


Figure 10 Scheme of the biological activity of S100 proteins in the intracellular and extracellular spaces[113].

S100 proteins can be also released or secreted into the extracellular space and exert their function in a cytokine-like manner through the Receptor for Advanced Glycation End-products (RAGE) (Figure 10) and participate in tissue organization during development, in

the inflammatory response, and/or in tissue remodeling in specific pathological states[122]. S100 family members can also perform their extracellular functions by interacting with other receptors besides RAGE.

The altered expression of several S100 proteins is associated with tumor development and progression to metastatic phenotype. By the way, while S100A4, S100A6, S100A7, and S100B are up-regulated in human tumors[123], S100A1, S100A2[124] and S100A11[125] are down-regulated in neoplastic tissues and have been postulated to have a tumor suppressor function. Overexpression of specific S100 proteins is also associated with other diseases, such as neurodegenerative diseases (S100B), cardiomyopathies (S100A1), and several inflammatory disorders, such as rheumatoid arthritis, chronic bronchitis, cystic fibrosis, psoriasis and allergy (S100A8, S100A9, S100A12, S100A7, S100A15).

1.3.2 Structural features

All EF-hand proteins contain two paired EF-hand motifs (helix-loop-helix), connected by an hinge region or linker (Figure 11), which constitute the so-called EF-hand domain (EFhD).

S100 proteins are small acidic 10-12 kDa proteins, displaying high sequence homology (aminoacid identity ranging from 25% to 65%) and a typical structural architecture for the members of the family. The highest sequence identity is found in the loop of each EF-hand motif where are present the ligands for the calcium(II) binding. Conversely, the

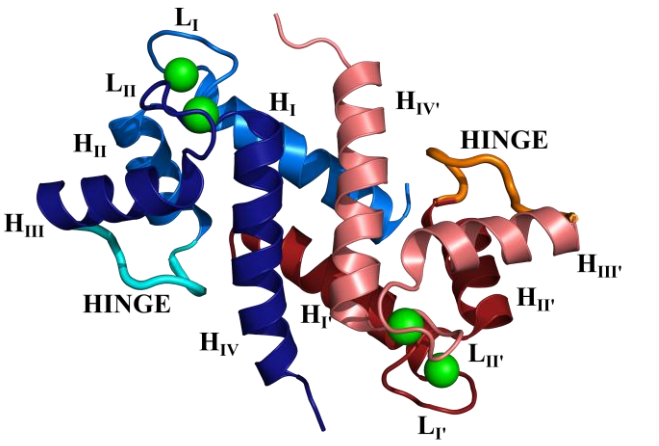


Figure 11. The S100 subunits (EF-hand domains) are shown in blue and red, respectively. Each subunit is composed of two EF-hands motives, shown in light and dark colors, connected by a hinge region or linker. The nomenclature of helices (H) and loops (L) is indicated. Calcium(II) ions are displayed as green spheres.

hinge region and the C-terminal extension, usually present in these proteins, exhibit high sequence diversification, suggesting that these two regions might have a role in their individual and specific biological activity.

S100 proteins usually exist as homo- and hetero-dimers in which the two monomers are related by a two-fold axis of rotation and held together by Van der Waals interactions between strictly conserved hydrophobic residues placed at the interface of the two monomers. In each monomer, the two EF-hand motifs, a modified S100-specific EF-hand, located at the N-terminus, and a canonical calcium(II)-binding EF-hand, interplay in a cooperative manner to bind calcium(II) ions with the N-terminal EF-hand displaying a reduced binding affinity (up to 100 times lower) with respect to the C-terminal one. The N-terminal EF-hand consists of 14 amino acids consensus sequence motif and the carbonyl oxygen of the residues placed in position 1, 4, 6, 9 and the side chain of the residue in position 14 represent the ligands. In this metal binding site only the residue in position 14 is highly conserved and can be a glutamate or an aspartate. The canonical calcium(II)-binding loop, instead consists of 12 amino acids, that form the well characterized binding motif D-x-N-x-D-x-(E/K/R/Q/A)-x-x-x-x-E, where the calcium(II) ion is bound to Asp(1), Asn(3), Asp(5) and Glu(12) side-chains and to the carbonyl oxygen of the less conserved amino acid in position 7. The replacement of one or more amino acids at positions 1, 3, and 12 affects dramatically the metal binding affinity of the C-terminal EF-hand motif and can disrupt the global affinity of the protein towards calcium(II).

In the apo state these proteins can adopt a wide range of conformations especially in the second EF-hand binding loop related to the positioning of helix III, that is the most loosely packed helix and undergoes the most dramatic conformational change after the binding of calcium(II). In fact, the presence of this metal ion causes a structural reorganization that places helix III from antiparallel, in the apo-form, to almost perpendicular, in the calcium(II)-bound form, with respect to helix IV. This “opens” the structure of the protein and determines the exposure of a hydrophobic cleft delimited by helix III and IV and the hinge region that regulates most of the S100 proteins function and is responsible for the binding of substrates. In this way, the conformational heterogeneity experienced in the apo form is directly correlated to the function of S100 proteins because it discloses the protein active states and the calcium(II)-induced conformational change results in this way facilitated[126].

S100 proteins, therefore, usually exhibit a so-called "closed" conformation in their apo state and an "open" conformation in the calcium(II) bound state. However, some exceptions have also been described, such as S100A10[127], that is already in the "open" conformation in the apo state, does not undergo calcium(II)-dependent conformational changes and it is in a permanently activated state. Conversely, S100A16[128] assumes a "closed" conformation in both the apo and calcium(II)-bound forms.

The structural features of S100 proteins have been widely analyzed in this laboratory using the Principal Component Analysis (PCA) performed on the six interhelical angles of the EFhD[129]. In this way, it is possible to obtain a complete description of the conformational space spanned by the EFhDs in the S100 protein family in comparison with in the other members of the EF-hand superfamily. The PCA analysis provides a quantitative parameter to classify in a reliable way the "open" and "closed" conformations and to highlight eventual outliers to this partition. Distinct packing of the helices in the EFhD with respect to the "open" and "closed" conformations have been identified for other members of the EF-hand superfamily. For example, a "semi-open" conformation has been described for the C-terminal domain of myosin light chains[130], while it has not been reported for the S100 proteins family.

In addition to the well characterized affinity for calcium(II) ion, many S100 proteins display high affinities towards zinc(II) and copper(II) ions[131-132], which are suggested to influence the biological activity of these proteins in the extracellular space where these metals are more abundant. Moreover, the binding of these metal ions can impair or enhance the relative affinity towards calcium(II).

1.3.3 S100A14: a novel member of the S100 family.

S100A14 is one of the latest characterized members of the S100 proteins family, whose biological function is largely unknown, and presents many peculiar and unique features.

The expression of S100A14 gene has been reported only in mammals and represent a phylogenetically young protein among the S100 proteins and, more in general, the EF-hand proteins. This protein shares the highest sequence homology with S100A13[133] (38% of sequence identity) and has the S100-specific N-terminal EF-hand formed by only 13 amino acids, in contrast with the 14-amino-acid loop of the other family members, that, however, preserves the critical residue (Glu) in the last position of the consensus motif (see above). The

C-terminal canonical EF-hand instead has only two out of the six conserved ligands of calcium(II) and thus appears not fully functional. In this way, the activity of S100A14 seems not to be regulated by the calcium(II) level as it occurs also for other members, such as S100A10. These mutations in the C-terminal EF-hand loop are reported for all the species expressing S100A14 except for of the evolutionary predecessor Meleagris gallopavo, where one of the four calcium binding residues is still in place as in S100A13.

S100A14 presents in its sequence some cysteins and histidines that could have a role in the binding of zinc(II) and copper(II). Moreover, two of the residues responsible of the binding of cooper(II) in S100A13 are conserved in S100A14.

S100A14 contains with respect to the other S100 family members an extended hydrophilic loop at the N-terminus, including the N-myristoylation site, that suggests the possibility of an interaction with a membrane receptor or the lipid bilayer itself. This long N-terminal loop probably accounts for the specific function of this protein[134].

S100A14 present a heterogenic expression in tumors. In fact, it is preferentially over-expressed in ovary, breast, and uterus, and mainly under-expressed in kidney, rectum, and colon tumors, unlike the other S100 proteins that usually present a homogeneous expression and are either under- or over-expressed in tumor cells[134]. In addition to the role in the regulation of cell proliferation S100A14 has been recently proven to have a function in the regulation of the invasiveness in the oral squamous cell carcinomas by influencing the expression of MMP-1 and MMP-9. In particular, in these cancer cells it has been found that the under-expression of S100A14 is associated with the over-expression of these matrix hydrolytic enzymes[135]. The metastatic potential of S100A14 has been demonstrated also in colorectal cancer as correlated with high-expression of S100A4[136].

Furthermore, it has been demonstrated that S100A14 promotes cell proliferation and survival at relatively low doses and induces apoptosis at high doses in a RAGE-dependent manner on esophageal squamous cell carcinoma[137]. S100A14 thus triggers the RAGE intracellular cascade, activating ERK1/2 and NF-kB signaling.

1.3.4 Aim of the project.

A detailed description of the structure, dynamics and metal binding properties of S100A14 is required to shed light on its largely unknown biological function. NMR is the

most suitable technique to provide simultaneously those information at the atomic level and to highlight the unique features of this member of S100 proteins family.

In the present project, classical NMR experiments for structure determination together with programs for structure calculation, refinements and validation have been exploited to solve the structure of S100A14 protein.

Longitudinal and transverse relaxation measurements have allowed to elucidate the dimeric state and the large flexibility occurring at the N- and C-terminal regions of this protein.

NMR allowed to prove that this protein is in the apo state and is in a permanently activated conformation at physiological temperature.

1.4. NMR characterization of the C-terminal tail of full-length RAGE in a membrane mimicking environment

1.4.1 The RAGE receptor

RAGE is a multiligand receptor of the immunoglobulin superfamily present both in normal tissues and in vascular cells. It is constitutively expressed during embryonic development and is under-regulated in adult life in physiological states, whereas in inflammatory states it becomes up-regulated in the sites where its ligands accumulate.

RAGE is able to recognize tertiary structures rather than amino acid sequences, thus it engages multiple families of molecules instead of individual ligands and is considered a pattern recognition receptor (PRR)[138]. In particular it can bind β-sheet fibrillar structures, such as amyloid components characteristic of Alzheimer's disease, the members of the S100 protein family, already described above, the advanced glycation end products (AGEs), that derive from nonenzymatic glycoxidation reaction of extracellular proteins and accumulate especially in diabetic states and neurodegenerative diseases, and amphoterin. Therefore, the RAGE receptor plays a key role in many inflammation-related pathological states such as diabetes, vascular diseases, neurodegenerative diseases, autoimmune diseases and in cancer[139].

Ligations to RAGE determine a sustained period of cellular activation mediated by receptor-dependent signaling. RAGE activation, indeed, triggers multiple signaling pathways such as ERK1/2 MAP kinases, SAPK/JNK MAP kinases, phosphoinositol-3 kinase and the JAK/STAT pathway and regulates important cellular functions through the transcription factors AP-1, NF-κB, etc.

1.4.2 Structural features

RAGE is a 45 kDa receptor constituted by three immunoglobulin-like extracellular domain, a variable portion, "V type" domain (23-116), and two constant portions, "C type" domains ("C1" (124-221) and "C2" (227-317)), by a transmembrane helix (343-363), and by a short cytoplasmic tail (364-404). The "V" and "C1" domains form an integrated structural unit connected with a flexible linker to "C2"[140], which is fully independent and the angle

between the long axes of "VC1" and "C2" units is around 80°[141]. Free rotation around the C1-C2 linker allows the "VC1" domains to utilize different surfaces to interact with different ligands. The "V" domain has a positively charged surface[142] and is responsible for most of the extracellular ligand recognitions. It is characterized by large ms-μs fluctuations in both the secondary structures and loops that ensure high plasticity. Nevertheless, in the case of S100 proteins, all the three domains can be involved in the interaction and the differences in RAGE binding sites may account for the diverse cellular responses caused by these proteins[143].

RAGE is expressed as both full-length, membrane-bound form, and as various soluble forms, lacking both the transmembrane domain and the cytosolic domain, which appears to be essential for the intracellular signaling. The soluble forms can derive from alternative mRNA splicing and from the proteolytic cleavage of the RAGE full-length receptor by the membrane metalloprotease ADAM-1[144]. Soluble RAGE acts both as extracellular (decoy) sink for the ligands, antagonizing their binding to the active receptor and has also a pro-inflammatory effect.

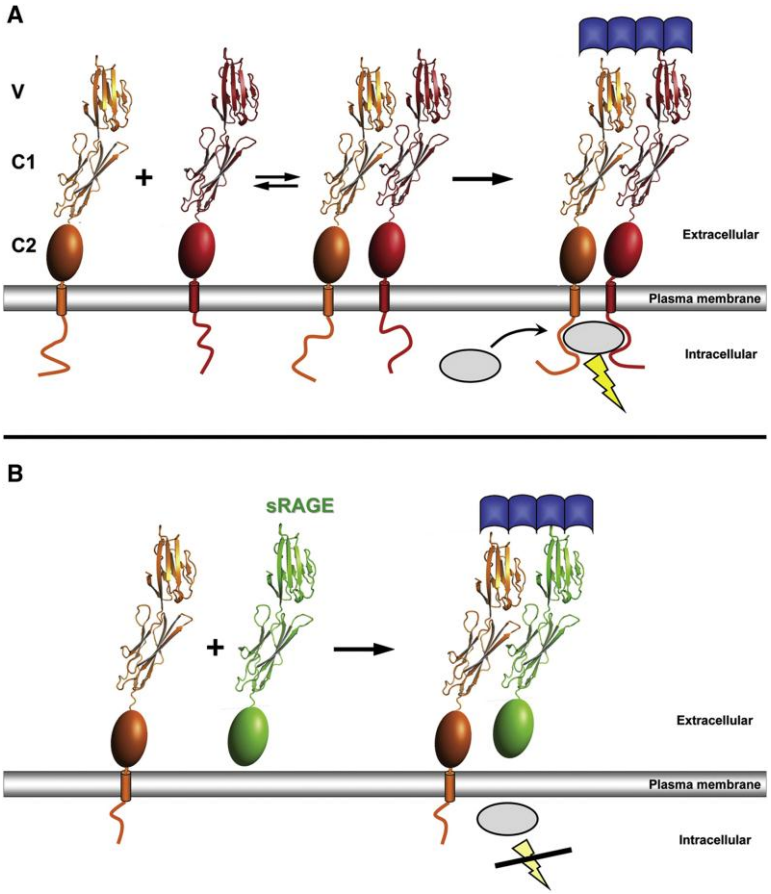


Figure 12. (A) RAGE preassembles in the plasma membrane and ligand binding to RAGE stabilizes oligomers, which then can bind a signaling adaptor protein to the cytoplasmic region of RAGE.
(B) Action of soluble RAGE. The interaction with intact RAGE form a hetero-oligomer that limits the binding of intracellular adaptors and blocks signal transduction[148].

Recently, it was shown that RAGE forms homodimers and also oligomers on the plasma membrane and that this is an important step in receptor signaling. Constitutive oligomerization of RAGE that brings together several "V" domains, increases the number of sites available for binding and thus enhances the affinity of the receptor towards structurally

heterogeneous ligands[145]. Moreover, oligomers of S100 proteins, such as S100B tetramers, that binds RAGE with higher affinity than the dimeric protein, can induce RAGE oligomerizzation at the "V" domain[146].

The effect of oligomerization might potentiate RAGE signaling not only by increasing the affinity of the receptor towards its ligands but also triggering signal transduction through the association of RAGE cytosolic tail to its interacting proteins[147], e.g.Dia-1 (Figure 12).

In this way, the soluble forms of RAGE, since they lack the cytoplasmic domain, can prevent RAGE signaling not only through competitive ligand binding but also through blockade of intrinsic RAGE homodimerization on the cell surface (Figure 12).

1.4.3 The cytosolic domain of RAGE receptor and its adaptor proteins.

The cytosolic domain of the RAGE receptor is essential for the intracellular signal transduction consequent to the receptor activation. In fact, many evidences have revealed that the deletion of the cytoplasmic tail exerts a "dominant-negative" effect and blunts the signal transduction downstream to the ectodomains interaction with ligands[148,149].

It has been highlighted that proteins and domains involved in cell-signaling in eukaryotic systems are usually characterized by an increased propensity in intrinsic disorder due to the functional advantages deriving from molecular recognition involving disordered structures. In fact, intrinsic disorder enables binding with high specificity and low affinity and promotes the interaction with multiple different partners[150].

The cytosolic domain of RAGE receptor is characterized by low sequence complexity with low content of bulky hydrophobic aminoacids and high portions of polar and charged aminoacids. These features are familiar to intrinsically unstructured proteins[151] and suggest the presence of a heterogeneous and not well-defined structure also for this domain, that is essential for the execution of its function. In the case of intrinsically disordered proteins the resonance assignment through NMR allows site-specific characterization of average properties over the ensemble of conformers.

The unique primary sequence of this domain is conserved in various species and can be divided into three distinct regions, one proximal to the cell membrane, rich of basic amino acids, a central fragment rich of acidic amino acids, especially glutamic acid, and a low-conserved C-terminal region[152]. Moreover, this sequence lacks any known motifs involved in receptor signaling and any enzymatic activity such as the endogenous tyrosine kinase activity.

Thus it necessarily has to interact with endogenous multiple partners or adaptor to trigger the recruitment of various downstream effector pathways.

Among the endogenous signaling molecules associated with the cytosolic domain the extracellular signal-regulated protein kinase-1 and -2 (ERK-1/2) have been identified. It has been proposed that the cytosolic RAGE stabilizes ERK under the proximal region of the plasma membrane and favors the interaction between ERK and its substrates[152].

Moreover, it has been reported that the cytosolic domain of the RAGE receptor is able to bind, after phosphorylation by PKCζ, TIRAP and MyD88, two adaptor proteins for the Toll-like receptor-4 (TLR-4), and to transduce signal to downstream molecules. In this way, a possible coordinated regulation of inflammation and immune response has been postulated for the two receptors[153].

Another adaptor protein for RAGE receptor is the diaphanous-1 (Dia1). Recently, it has been demonstrated that the cytosolic domain stimulate fundamental signaling networks, in particular the activation of Rac-1 and Cdc42, and the cellular migration through the interaction with the formin homology (FH1) domain of diaphanous-1 (Dia1) [154], that induces actin and microtubule polimerization.

The solution structure of cytosolic domain has been recently solved and reveals the presence of a α-turn structure located nearby the membrane and a long flexible C-terminal tail. The interaction area of the cytosolic domain with the FH1 of Dia-1 has also been identify[155]. A novel mechanism of action has been proposed for this receptor and implicate a constitutive presence of FH1 bound to the cytosolic domain in an inactive state and its consequent activation in the presence of ligands for the ectodomain that cause RAGE oligomerization.

1.4.4 Aim of the project

Investigation of the signaling pathway occurring through the RAGE receptor is critically important in understanding its function and its role in the pathologies correlated to it. High resolution structure of the RAGE receptor and knowledge of the way of interaction with its ligands is necessary for the elucidation of the molecular basis for RAGE function. Although, several extracellular and intracellular partners of the RAGE receptor have been identified and possible mechanism of signal transduction have been proposed, the molecular details of the signal transduction are still unknown.

Since the binding capability of cytosolic domain are related to its dynamics and accessibility to the intracellular partner, the present research project aims at investigating the structural and dynamics features of the cytosolic domain when it is linked to the full length receptor.

The cytosolic RAGE has been thus investigated as isolated domain and as tethered to the full-length receptor in a membrane-mimicking environment and the dynamics have been compared in the two different conditions.

1.5 Bibliography:

1. Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128–132 (2005).
2. Cui, Q. & Karplus, M. Allostery and cooperativity revisited. *Protein Sci* **17**, 1295–1307 (2008).
3. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
4. Changeux, J.-P. & Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol Rep* **3**, 19 (2011).
5. Hammes, G. G., Chang, Y.-C. & Oas, T. G. Conformational selection or induced fit: A flux description of reaction mechanism. *PNAS* **106**, 13737–13741 (2009).
6. Kay, L. E. NMR studies of protein structure and dynamics. *J. Magn. Reson.* **173**, 193–207 (2005).
7. Bertini, I. *et al.* Accurate Solution Structures of Proteins from X-ray Data and a Minimal Set of NMR Data: Calmodulin−Peptide Complexes As Examples. *Journal of the American Chemical Society* **131**, 5134–5144 (2009).
8. Lipsitz, R. S. & Tjandra, N. Residual Dipolar Couplings in Nmr Structure Analysis*1. *Annual Review of Biophysics and Biomolecular Structure* **33**, 387–413 (2004).
9. Lakomek, N.-A. *et al.* Residual dipolar couplings as a tool to study molecular recognition of ubiquitin. *Biochem. Soc. Trans* **36**, 1433 (2008).
10. Bernadó, P. & Blackledge, M. Structural biology: Proteins in dynamic equilibrium. *Nature* **468**, 1046–1048 (2010).
11. Rieping, W., Habeck, M. & Nilges, M. Inferential Structure Determination. *Science* **309**, 303–306 (2005).
12. Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **129**, 5656–5664 (2007).
13. Limongelli, V. *et al.* Sampling protein motion and solvent effect during ligand binding. *PNAS* **109**, 1467–1472 (2012).
14. Laio, A. & Parrinello, M. Escaping free-energy minima. *PNAS* **99**, 12562–12566 (2002).

15. Branduardi, D., Gervasio, F. L. & Parrinello, M. From A to B in free energy space. *J Chem Phys* **126**, 054103 (2007).
16. Bertini, I. *et al.* Conformational space of flexible biological macromolecules from average data. *J. Am. Chem. Soc.* **132**, 13553–13558 (2010).
17. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. , *Nature* **171**, 737–738 (1953).
18. Hoogsteen, K. The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallographica* **12**, 822–823 (1959).
19. Varani, G. Exceptionally stable nucleic acid hairpins. *Annu Rev Biophys Biomol Struct* **24**, 379–404 (1995).
20. Lilley, D. M. J. All change at Holliday junction. *PNAS* **94**, 9513–9515 (1997).
21. van de Sande, J. H. *et al.* Parallel stranded DNA. *Science* **241**, 551–557 (1988).
22. Sklenár, V. & Feigon, J. Formation of a stable triplex from a single DNA strand. *Nature* **345**, 836–838 (1990).
23. Parkinson, G. N., Lee, M. P. H. & Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **417**, 876–880 (2002).
24. Guéron, M. & Leroy, J. L. The i-motif in nucleic acids. *Curr. Opin. Struct. Biol.* **10**, 326–331 (2000).
25. Blackburn, E. H. Structure and function of telomeres. , *Nature* **350**, 569–573 (1991).
26. Simonsson, T., Pecinka, P. & Kubista, M. DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.* **26**, 1167–1172 (1998).
27. Sen, D. & Gilbert, W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334**, 364–366 (1988).
28. Hammond-Kosack, M. C. & Docherty, K. A consensus repeat sequence from the human insulin gene linked polymorphic region adopts multiple quadriplex DNA structures in vitro. *FEBS Lett.* **301**, 79–82 (1992).
29. Hackett, J. A., Feldser, D. M. & Greider, C. W. Telomere Dysfunction Increases Mutation Rate and Genomic Instability. *Cell* **106**, 275–286 (2001).
30. Mergny, J.-L. & Hélène, C. G-quadruplex DNA: A target for drug design. *Nature Medicine* **4**, 1366–1367 (1998).
31. Read, M. *et al.* Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors. *PNAS* **98**, 4844–4849 (2001).

32. Simonsson, T. G-quadruplex DNA structures--variations on a theme. *Biol. Chem.* **382**, 621–628 (2001).
33. Macaya, R. F., Schultze, P., Smith, F. W., Roe, J. A. & Feigon, J. Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution. *Proc Natl Acad Sci U S A* **90**, 3745–3749 (1993).
34. Schultze, P., Macaya, R. F. & Feigon, J. Three-dimensional solution structure of the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). *J. Mol. Biol* **235**, 1532–1547 (1994).
35. Padmanabhan, K., Padmanabhan, K. P., Ferrara, J. D., Sadler, J. E. & Tulinsky, A. The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer. *J. Biol. Chem.* **268**, 17651–17654 (1993).
36. Padmanabhan, K. & Tulinsky, A. An ambiguous structure of a DNA 15-mer thrombin complex. *Acta Crystallogr. D Biol. Crystallogr* **52**, 272–282 (1996).
37. Reshetnikov, R., Golovin, A., Spiridonova, V., Kopylov, A. & Šponer, J. Structural Dynamics of Thrombin-Binding DNA Aptamer d(GGTTGGTGTGGTTGG) Quadruplex DNA Studied by Large-Scale Explicit Solvent Simulations. *J. Chem. Theory Comput.* **6**, 3003–3014 (2010).
38. Lee, J. Y., Okumus, B., Kim, D. S. & Ha, T. Extreme conformational diversity in human telomeric DNA. *PNAS* **102**, 18938–18943 (2005).
39. Shirude, P. S., Okumus, B., Ying, L., Ha, T. & Balasubramanian, S. Single-Molecule Conformational Analysis of G-Quadruplex Formation in the Promoter DNA Duplex of the Proto-Oncogene C-Kit. *J. Am. Chem. Soc.* **129**, 7484–7485 (2007).
40. Mashimo, T., Yagi, H., Sannohe, Y., Rajendran, A. & Sugiyama, H. Folding Pathways of Human Telomeric Type-1 and Type-2 G-Quadruplex Structures. *Journal of the American Chemical Society* **132**, 14910–14918 (2010).
41. Rhee, S., Han, Z. j, Liu, K., Miles, H. T. & Davies, D. R. Structure of a triple helical DNA with a triplex-duplex junction. *Biochemistry* **38**, 16810–16815 (1999).
42. Zhang, N., Phan, A. T. & Patel, D. J. (3 + 1) Assembly of Three Human Telomeric Repeats into an Asymmetric Dimeric G-Quadruplex. *J. Am. Chem. Soc.* **127**, 17277–17285 (2005).
43. Koirala, D. *et al.* Intramolecular folding in three tandem guanine repeats of human telomeric DNA. *Chem. Commun.* **48**, 2006–2008 (2012).

44. Mott, J. D. & Werb, Z. Regulation of matrix biology by matrix metalloproteinases. *Curr. Opin. Cell Biol.* **16**, 558–564 (2004).
45. Nagase, H., Barrett, A. J. & Woessner, J. F., Jr Nomenclature and glossary of the matrix metalloproteinases. *Matrix Suppl* **1**, 421–424 (1992).
46. Massova, I., Kotra, L. P., Fridman, R. & Mobashery, S. Matrix metalloproteinases: structures, evolution, and diversification. *FASEB J* **12**, 1075–1095 (1998).
47. Fanjul-Fernández, M., Folgueras, A. R., Cabrera, S. & López-Otín, C. Matrix metalloproteinases: evolution, gene regulation and functional analysis in mouse models. *Biochim. Biophys. Acta* **1803**, 3–19 (2010).
48. Nesi, A. & Fragai, M. Substrate Specificities of Matrix Metalloproteinase 1 in PAR-1 Exodomain Proteolysis. *ChemBioChem* **8**, 1367–1369 (2007).
49. Andreini, C., Banci, L., Bertini, I., Luchinat, C. & Rosato, A. Bioinformatic Comparison of Structures and Homology-Models of Matrix Metalloproteinases. *J. Proteome Res.* **3**, 21–31 (2004).
50. Aiken, A. & Khokha, R. Unraveling metalloproteinase function in skeletal biology and disease using genetically altered mice. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1803**, 121–132 (2010).
51. Page-McCaw, A., Ewald, A. J. & Werb, Z. Matrix metalloproteinases and the regulation of tissue remodelling. *Nature Reviews Molecular Cell Biology* **8**, 221–233 (2007).
52. Butler, G. S. & Overall, C. M. Updated Biological Roles for Matrix Metalloproteinases and New 'Intracellular' Substrates Revealed by Degradomics. *Biochemistry* **48**, 10830–10845 (2009).
53. Rodríguez, D., Morrison, C. J. & Overall, C. M. Matrix metalloproteinases: What do they not do? New substrates and biological roles identified by murine models and proteomics. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1803**, 39–54 (2010).
54. McCawley, L. J. & Matrisian, L. M. Matrix metalloproteinases: they're not just for matrix anymore! *Curr. Opin. Cell Biol.* **13**, 534–540 (2001).
55. Limb, G. A. *et al.* Matrix metalloproteinase-1 associates with intracellular organelles and confers resistance to lamin A/C degradation during apoptosis. *Am. J. Pathol.* **166**, 1555–1563 (2005).
56. Visse, R. & Nagase, H. Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry. *Circ. Res* **92**, 827–839 (2003).

57. Barksby, H. E. *et al.* Matrix metalloproteinase 10 promotion of collagenolysis via procollagenase activation: implications for cartilage degradation in arthritis. *Arthritis Rheum.* **54**, 3244–3253 (2006).
58. Sounni, N. E. & Noel, A. Membrane type-matrix metalloproteinases and tumor progression. *Biochimie* **87**, 329–342 (2005).
59. Ala-aho, R. & Kähäri, V.-M. Collagenases in cancer. *Biochimie* **87**, 273–286 (2005).
60. Puerta, D. T., Lewis, J. A. & Cohen, S. M. New Beginnings for Matrix Metalloproteinase Inhibitors: Identification of High-Affinity Zinc-Binding Groups. *J. Am. Chem. Soc.* **126**, 8388–8389 (2011).
61. Konstantinopoulos, P. A., Karamouzis, M. V., Papatsoris, A. G. & Papavassiliou, A. G. Matrix metalloproteinase inhibitors as anticancer agents. *Int. J. Biochem. Cell Biol.* **40**, 1156–1168 (2008).
62. Overall, C. M. & Kleifeld, O. Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. *Nature Reviews Cancer* **6**, 227–239 (2006).
63. Lauer-Fields, J. *et al.* Triple-Helical Transition State Analogs: A New Class of Selective Matrix Metalloproteinase Inhibitors. *J Am Chem Soc* **129**, 10408–10417 (2007).
64. Luo, D., Mari, B., Stoll, I. & Anglard, P. Alternative splicing and promoter usage generates an intracellular stromelysin 3 isoform directly translated as an active matrix metalloproteinase. *J. Biol. Chem.* **277**, 25527–25536 (2002).
65. Bertini, I. *et al.* Snapshots of the reaction mechanism of matrix metalloproteinases. *Angew. Chem. Int. Ed. Engl.* **45**, 7952–7955 (2006).
66. Maskos, K. & Bode, W. Structural basis of matrix metalloproteinases and tissue inhibitors of metalloproteinases. *Mol. Biotechnol.* **25**, 241–266 (2003).
67. Grams, F. *et al.* X-ray structures of human neutrophil collagenase complexed with peptide hydroxamate and peptide thiol inhibitors. Implications for substrate binding and rational drug design. *Eur. J. Biochem.* **228**, 830–841 (1995).
68. Bertini, I. *et al.* Conformational variability of matrix metalloproteinases: Beyond a single 3D structure. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5334 –5339 (2005).
69. Bertini, I. *et al.* Interdomain Flexibility in Full-length Matrix Metalloproteinase-1 (MMP-1). *Journal of Biological Chemistry* **284**, 12821 –12828 (2009).

70. Bertini, I. *et al.* Evidence of Reciprocal Reorientation of the Catalytic and Hemopexin-Like Domains of Full-Length MMP-12. *Journal of the American Chemical Society* **130**, 7011–7021 (2008).
71. Ramachandran, G. N. Structure of Collagen. , *Nature* **177**, 710–711 (1956).
72. Orgel, J. P. R. O., Irving, T. C., Miller, A. & Wess, T. J. Microfibrillar structure of type I collagen in situ. *PNAS* **103**, 9001–9005 (2006).
73. Garnero, P. *et al.* The collagenolytic activity of cathepsin K is unique among mammalian proteinases. *J. Biol. Chem.* **273**, 32347–32352 (1998).
74. Aimes, R. T. & Quigley, J. P. Matrix metalloproteinase-2 is an interstitial collagenase. Inhibitor-free enzyme catalyzes the cleavage of collagen fibrils and soluble native type I collagen generating the specific 3/4- and 1/4-length fragments. *J. Biol. Chem.* **270**, 5872–5876 (1995).
75. Rosenblum, G. *et al.* Direct Visualization of Protease Action on Collagen Triple Helical Structure. *PLoS ONE* **5**, e11043 (2010).
76. Clark, I. M. & Cawston, T. E. Fragments of human fibroblast collagenase. Purification and characterization. *Biochem. J.* **263**, 201–206 (1989).
77. Bode, W. A helping hand for collagenases: the haemopexin-like domain. *Structure* **3**, 527–530 (1995).
78. Li, J. *et al.* Structure of full-length porcine synovial collagenase reveals a C-terminal domain containing a calcium-linked, four-bladed β-propeller. *Structure* **3**, 541–549 (1995).
79. Ottl, J. *et al.* Recognition and catabolism of synthetic heterotrimeric collagen peptides by matrix metalloproteinases. *Chemistry & Biology* **7**, 119–132 (2000).
80. Overall, C. M. Molecular determinants of metalloproteinase substrate specificity: matrix metalloproteinase substrate binding domains, modules, and exosites. *Mol. Biotechnol.* **22**, 51–86 (2002).
81. Chung, L. *et al.* Collagenase unwinds triple-helical collagen prior to peptide bond hydrolysis. *EMBO J* **23**, 3020–3030 (2004).
82. Piccard, H., Van den Steen, P. E. & Opdenakker, G. Hemopexin domains as multifunctional liganding modules in matrix metalloproteinases and other proteins. *J. Leukoc. Biol.* **81**, 870–892 (2007).
83. Bertini, I. *et al.* Structural Basis for Matrix Metalloproteinase 1-Catalyzed Collagenolysis. *J. Am. Chem. Soc.* **134**, 2100–2110 (2011).

84. Lauer-Fields, J. L. *et al.* Identification of specific hemopexin-like domain residues that facilitate matrix metalloproteinase collagenolytic activity. *J. Biol. Chem.* **284**, 24017–24024 (2009).
85. Arnold, L. H. *et al.* The Interface Between Catalytic and Hemopexin Domains in Matrix Metalloproteinase-1 Conceals a Collagen Binding Exosite. *J. Biol. Chem.* **286**, 45073–45082 (2011).
86. Tsukada, H. & Pourmotabbed, T. Unexpected Crucial Role of Residue 272 in Substrate Specificity of Fibroblast Collagenase. *J. Biol. Chem.* **277**, 27378–27384 (2002).
87. Manka, S. W. *et al.* Structural insights into triple-helical collagen cleavage by matrix metalloproteinase 1. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 12461-6 (2012).
88. Overall, C. M. & Butler, G. S. Protease Yoga: Extreme Flexibility of a Matrix Metalloproteinase. *Structure* **15**, 1159–1161 (2007).
89. Bertini, I., Fragai, M., Lee, Y.-M., Luchinat, C. & Terni, B. Paramagnetic Metal Ions in Ligand Screening: The CoII Matrix Metalloproteinase 12. *Angewandte Chemie* **116**, 2304–2306 (2004).
90. Keizers, P. H. J. & Ubbink, M. Paramagnetic tagging for protein structure and dynamics analysis. *Progress in Nuclear Magnetic Resonance Spectroscopy* **58**, 88–96 (2011).
91. Keizers, P. H. J., Saragliadis, A., Hiruma, Y., Overhand, M. & Ubbink, M. Design, Synthesis, and Evaluation of a Lanthanide Chelating Protein Probe: CLaNP-5 Yields Predictable Paramagnetic Effects Independent of Environment. *Journal of the American Chemical Society* **130**, 14802–14812 (2008).
92. Keizers, P. H. J., Desreux, J. F., Overhand, M. & Ubbink, M. Increased Paramagnetic Effect of a Lanthanide Protein Probe by Two-Point Attachment. *Journal of the American Chemical Society* **129**, 9292–9293 (2007).
93. Gochin, M. & Roder, H. Protein structure refinement based on paramagnetic NMR shifts: applications to wild-type and mutant forms of cytochrome c. *Protein Sci.* **4**, 296–305 (1995).
94. Chou, J. J., Li, S., Klee, C. B. & Bax, A. Solution structure of Ca2+-calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Mol Biol* **8**, 990–997 (2001).
95. Clore, G. M. & Schwieters, C. D. Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small alpha/beta protein: a unified

picture of high probability, fast atomic motions in proteins. *J. Mol. Biol.* **355**, 879–886 (2006).

96. Tang, C., Schwieters, C. D. & Clore, G. M. Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature* **449**, 1078–1082 (2007).
97. Lange, O. F. *et al.* Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471–1475 (2008).
98. Ozenne, V. *et al.* Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* **28**, 1463–1470 (2012).
99. Bernadó, P. *et al.* A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *PNAS* **102**, 17002–17007 (2005).
100. Anthis, N. J., Doucleff, M. & Clore, G. M. Transient, sparsely populated compact states of apo and calcium-loaded calmodulin probed by paramagnetic relaxation enhancement: interplay of conformational selection and induced fit. *J. Am. Chem. Soc.* **133**, 18966–18974 (2011).
101. Iwahara, J. & Clore, G. M. Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature* **440**, 1227–1230 (2006).
102. Tang, C., Iwahara, J. & Clore, G. M. Visualization of transient encounter complexes in protein-protein association. *Nature* **444**, 383–386 (2006).
103. Bashir, Q., Volkov, A. N., Ullmann, G. M. & Ubbink, M. Visualization of the Encounter Ensemble of the Transient Electron Transfer Complex of Cytochrome c and Cytochrome c Peroxidase. *Journal of the American Chemical Society* **132**, 241–247 (2010).
104. Bertini, I. *et al.* Experimentally exploring the conformational space sampled by domain reorientation in calmodulin. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6841–6846 (2004).
105. Bertini, I. *et al.* Paramagnetism-Based NMR Restraints Provide Maximum Allowed Probabilities for the Different Conformations of Partially Independent Protein Domains. *Journal of the American Chemical Society* **129**, 12786–12794 (2007).
106. Longinetti, M., Luchinat, C., Parigi, G. & Sgheri, L. Efficient determination of the most favoured orientations of protein domains from paramagnetic NMR data. *Inverse Problems* **22**, 1485–1502 (2006).
107. Bertini, I. *et al.* MaxOcc: a web portal for maximum occurrence analysis. *J. Biomol. NMR* **53**, 271–280 (2012).

108. Dasgupta, S. *et al.* Narrowing the conformational space sampled by two-domain proteins with paramagnetic probes in both domains. *J. Biomol. NMR* **51**, 253–263 (2011).
109. Nagulapalli, M. *et al.* Recognition Pliability Is Coupled to Structural Heterogeneity: A Calmodulin Intrinsically Disordered Binding Region Complex. *Structure* **20**, 522–533 (2012).
110. Bertini, I., Luchinat, C., Nagulapalli, M., Parigi, G. & Ravera, E. Paramagnetic relaxation enhancement for the characterization of the conformational heterogeneity in two-domain proteins. *Phys Chem Chem Phys* **14**, 9149–9156 (2012).
111. Bertini, I., Luchinat, C. & Parigi, G. Moving the frontiers in solution and solid-state bioNMR. *Coordination Chemistry Reviews* **255**, 649–663 (2011).
112. Marenholz, I., Heizmann, C. W. & Fritz, G. S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochemical and Biophysical Research Communications* **322**, 1111–1122 (2004).
113. Frizzo, J. K. *et al.* Involvement of the S100B in cAMP-induced cytoskeleton remodeling in astrocytes: a study using TRTK-12 in digitonin-permeabilized cells. *Cell. Mol. Neurobiol.* **24**, 833–840 (2004).
114. Deora, A. B., Kreitzer, G., Jacovina, A. T. & Hajjar, K. A. An annexin 2 phosphorylation switch mediates p11-dependent translocation of annexin 2 to the cell surface. *J. Biol. Chem.* **279**, 43411–43418 (2004).
115. Hatakeyama, T., Okada, M., Shimamoto, S., Kubota, Y. & Kobayashi, R. Identification of intracellular target proteins of the calcium-signaling protein S100A12. *Eur. J. Biochem.* **271**, 3765–3775 (2004).
116. Donato, R. S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. *Int. J. Biochem. Cell Biol.* **33**, 637–668 (2001).
117. Lewit-Bentley, A., Réty, S., Sopkova-de Oliveira Santos, J. & Gerke, V. S100-annexin complexes: some insights from structural studies. *Cell Biol. Int.* **24**, 799–802 (2000).
118. Rezvanpour, A. & Shaw, G. S. Unique S100 target protein interactions. *Gen. Physiol. Biophys.* **28 Spec No Focus**, F39–46 (2009).
119. Filipek, A., Jastrzebska, B., Nowotny, M. & Kuznicki, J. CacyBP/SIP, a calcyclin and Siah-1-interacting protein, binds EF-hand proteins of the S100 family. *J. Biol. Chem.* **277**, 28848–28852 (2002).

120. Fernandez-Fernandez, M. R., Rutherford, T. J. & Fersht, A. R. Members of the S100 family bind p53 in two distinct ways. *Protein Sci* **17**, 1663–1670 (2008).
121. Yu, W. H. & Fraser, P. E. S100β Interaction with Tau Is Promoted by Zinc and Inhibited by Hyperphosphorylation in Alzheimer's Disease. *J. Neurosci.* **21**, 2240–2246 (2001).
122. Leclerc, E., Fritz, G., Vetter, S. W. & Heizmann, C. W. Binding of S100 proteins to RAGE: An update. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1793**, 993–1007 (2009).
123. Ilg, E. C., Schäfer, B. W. & Heizmann, C. W. Expression pattern of S100 calcium-binding proteins in human tumors. *Int. J. Cancer* **68**, 325–332 (1996).
124. Wicki, R., Franz, C., Scholl, F. A., Heizmann, C. W. & Schäfer, B. W. Repression of the candidate tumor suppressor gene S100A2 in breast cancer is mediated by site-specific hypermethylation. *Cell Calcium* **22**, 243–254 (1997).
125. Sakaguchi, M. *et al.* Relationship between contact inhibition and intranuclear S100C of normal human fibroblasts. *J. Cell Biol.* **149**, 1193–1206 (2000).
126. Malik, S., Revington, M., Smith, S. P. & Shaw, G. S. Analysis of the structure of human apo-S100B at low temperature indicates a unimodal conformational distribution is adopted by calcium-free S100 proteins. *Proteins* **73**, 28–42 (2008).
127. Rety, S. *et al.* The crystal structure of a complex of p11 with the annexin II N-terminal peptide. *Nat Struct Mol Biol* **6**, 89–95 (1999).
128. Babini, E. *et al.* Structural characterization of human S100A16, a low-affinity calcium binder. *JBIC Journal of Biological Inorganic Chemistry* **16**, 243–256 (2010).
129. Babini, E. *et al.* Principal component analysis of the conformational freedom within the EF-hand superfamily. *J. Proteome Res.* **4**, 1961–1971 (2005).
130. Nelson, M. R. & Chazin, W. J. An interaction-based analysis of calcium-induced conformational changes in Ca2+ sensor proteins. *Protein Sci* **7**, 270–282 (1998).
131. Heizmann, C. W. & Cox, J. A. New perspectives on S100 proteins: a multi-functional Ca(2+)-, Zn(2+)- and Cu(2+)-binding protein family. *Biometals* **11**, 383–397 (1998).
132. Kordowska, J., Stafford, W. F. & Wang, C. L. Ca2+ and Zn2+ bind to different sites and induce different conformational changes in human calcyclin. *Eur. J. Biochem.* **253**, 57–66 (1998).
133. Taimei Zhou, Xueying Zheng, Deqing Yi & Qingying Zhang Molecular Modeling and Structure Analysis of S100 Calcium Binding Protein A14: Molecular Modeling and

Structure Analysis of S100A14. in *2nd International Conference on Biomedical Engineering and Informatics, 2009. BMEI '09* 1–4 (IEEE, 2009).

134. Pietas, A. *et al.* Molecular cloning and characterization of the human S100A14 gene encoding a novel member of the S100 family. *Genomics* **79**, 513–522 (2002).
135. Sapkota, D. *et al.* S100A14 regulates the invasive potential of oral squamous cell carcinoma derived cell-lines in vitro by modulating expression of matrix metalloproteinases, MMP1 and MMP9. *European Journal of Cancer* **47**, 600–610 (2011).
136. Wang, H.-Y. *et al.* Expression status of S100A14 and S100A4 correlates with metastatic potential and clinical outcome in colorectal cancer after surgery. *Oncol. Rep.* **23**, 45–52 (2010).
137. Jin, Q., Chen, H., Luo, A., Ding, F. & Liu, Z. S100A14 stimulates cell proliferation and induces cell apoptosis at different concentrations via receptor for advanced glycation end products (RAGE). *PLoS ONE* **6**, e19375 (2011).
138. Liliensiek, B. *et al.* Receptor for advanced glycation end products (RAGE) regulates sepsis but not the adaptive immune response. *J. Clin. Invest.* **113**, 1641–1650 (2004).
139. Chavakis, T., Bierhaus, A. & Nawroth, P. P. RAGE (receptor for advanced glycation end products): a central player in the inflammatory response. *Microbes Infect* **6**, 1219–1225 (2004).
140. Dattilo, B. M. *et al.* The extracellular region of the receptor for advanced glycation end products is composed of two independent structural units. *Biochemistry* **46**, 6957–6970 (2007).
141. Sárkány, Z. *et al.* Solution structure of the soluble receptor for advanced glycation end products (sRAGE). *J. Biol. Chem.* **286**, 37525–37534 (2011).
142. Park, H. & Boyington, J. C. The 1.5 A crystal structure of human receptor for advanced glycation endproducts (RAGE) ectodomains reveals unique features determining ligand binding. *J.Biol.Chem.* **285**, 40762–40770 (2010).
143. Leclerc, E., Fritz, G., Weibel, M., Heizmann, C. W. & Galichet, A. S100B and S100A6 differentially modulate cell survival by interacting with distinct RAGE (receptor for advanced glycation end products) immunoglobulin domains. *J. Biol. Chem.* **282**, 31317–31331 (2007).
144. Sparvero, L. J. *et al.* RAGE (Receptor for Advanced Glycation Endproducts), RAGE ligands, and their role in cancer and inflammation. *J Transl Med* **7**, 17 (2009).

145. Xie, J. *et al.* Structural basis for pattern recognition by the receptor for advanced glycation end products (RAGE). *J. Biol. Chem.* **283**, 27255–27269 (2008).
146. Ostendorp, T. *et al.* Structural and functional insights into RAGE activation by multimeric S100B. *EMBO J* **26**, 3868–3878 (2007).
147. Zong, H. *et al.* Homodimerization is essential for the receptor for advanced glycation end products (RAGE)-mediated signal transduction. *J. Biol. Chem.* **285**, 23137–23146 (2010).
148. Sakaguchi, T. *et al.* Central role of RAGE-dependent neointimal expansion in arterial restenosis. *J. Clin. Invest.* **111**, 959–972 (2003).
149. Taguchi, A. *et al.* Blockade of RAGE-amphoterin signalling suppresses tumour growth and metastases. *Nature* **405**, 354–360 (2000).
150. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584 (2002).
151. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197–208 (2005).
152. Ishihara, K., Tsutsumi, K., Kawane, S., Nakajima, M. & Kasaoka, T. The receptor for advanced glycation end-products (RAGE) directly binds to ERK by a D-domain-like docking site. *FEBS Lett.* **550**, 107–113 (2003).
153. Sakaguchi, M. *et al.* TIRAP, an Adaptor Protein for TLR2/4, Transduces a Signal from RAGE Phosphorylated upon Ligand Binding. *PLoS One* **6**, (2011).
154. Hudson, B. I. *et al.* Interaction of the RAGE cytoplasmic domain with diaphanous-1 is required for ligand-stimulated cellular migration through activation of Rac1 and Cdc42. *J. Biol. Chem.* **283**, 34457–34468 (2008).
155. Rai, V. *et al.* Signal transduction in RAGE: solution structure of C-terminal RAGE (ctRAGE) and its binding to mDia1. *The Journal of Biological Chemistry* **278**, 5133-5144 (2011).

2. Methodological Aspects

2.1 NMR and solution structure calculation

2.1.1 NMR experiments for assignment and structure calculation

Two different strategies can be followed for the assignment of NMR spectral resonances and for the solution structure determination of biomolecules. In the case of biomolecules with low molecular weight (up to 7 kDa) the NMR analysis is based on the use of two-dimensional ^{1}H–^{1}H NMR experiments for both, spectral assignment and collection of structural restraints. Conversely, for biomolecules with molecular weights larger than 10 kDa the isotopic enrichment with ^{15}N and ^{13}C is required to overcome the resolution problems related to the small chemical shift range of protons compared to the high number of protons. In particular, carbon and nitrogen labeling allows to increase the spectral resolution and spread the signals overlap on a higher number of nuclear dimensions, correlating together three different nuclei (^{1}H, ^{13}C, and ^{15}N) through heteronuclear scalar couplings (triple resonance NMR experiments).

In the frame of this research project, the first approach has been used to characterize the relatively small DNA construct (11-mer, 3.5 kDa) and to confirm the G-triplex conformation. Both coherence transfer experiments based on scalar coupling and dipole–dipole experiments, acquired at 274 K, have been implemented in the assignment procedure.

2D COSY[1,2] (*CO*rrelation *S*pectroscop*Y*), for the correlation of geminal or vicinal protons, and 2D TOCSY[3] (*TO*tal *C*orrelation *S*pectroscop*Y*), for the spin system assignment, collected at 700 MHz allowed the assignment of all protons but the imino and amino ones. 2D ^{1}H-^{31}P COSY experiment was acquired at 600MHz to confirm the sequential connectivities. 2D JR-HMBC[4] (*J*ump and *R*eturn *H*eteronuclear *M*ultiple *B*ond *C*orrelation) experiment, performed at 600 MHz and 950 MHz on isotopic natural-abundance, allowed the assignment of imino protons through bond connectivities with the aromatic protons of the same base.

2D ^{1}H-^{1}H NOESY[5] (using mixing times ranging between 50 and 300 ms) and 2D ^{1}H-^{1}H ROESY acquired in H_2O at 900 and 800 MHz allowed the identification of spatially near nuclei through dipolar interactions.

On the contrary, triple-resonance NMR experiments have been used for the characterization of MMP-1, S100A14 and RAGE. These experiments rely on the transfer of magnetization through ^{1}J or ^{2}J heteronuclear scalar couplings, which are relatively large and

allow to get high sensitivity and selectivity (Figure 1). In particular, the 3D experiments used for the complete backbone resonance assignment in the present research projects are[6]: 3D HNCA[7], 3D CBCA(CO)NH, 3D HNCACB[8], 3D HNCO[7], 3D HN(CA)CO. For side-chain assignment, instead, the 3D spectra HBHA(CO)NH and (H)CCH-TOCSY[9] have been performed. For the evaluation of dipolar interactions between nuclei 3D NOESY experiments, such as 3D ^{15}N-NOESY[10] and 3D^{13}C-NOESY, have been implemented.

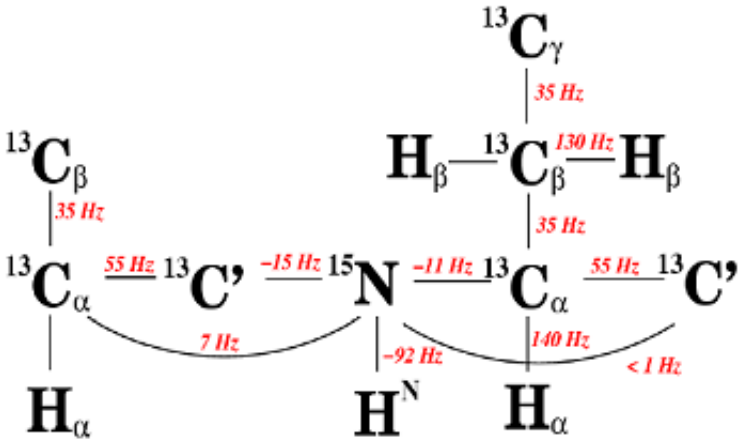


Figure 1 Size of the ^{1}J and ^{2}J coupling constants along the spin system that are used for magnetization transfer in ^{13}C-, ^{15}N-labelled proteins.

2D ^{1}H-^{1}H experiments (see S100A14 project) have been associated to the triple-resonance experiments for the assignment of aromatic side-chains and for the identification of further spatial connectivities among aliphatic and aromatic side-chains.

NMR experiments relying on transfer of magnetization through scalar coupling have been acquired at magnetic fields of 500 and 700 MHz; instead the experiments relying on dipolar interactions have been acquired at higher magnetic fields (800-900 MHz).

In the case of very large biomolecules (such as in the case of the full-length RAGE receptor), experiments based on *T*ransverse *R*elaxation-*O*ptimized *S*pectroscop*Y* (TROSY)[11], which makes use of interference effects between different relaxation mechanisms, have been exploited to reduce broadening and to achieve satisfactory linewidths.

Proton-less experiments[12] (2D CON, 2D CACO, 2D CBCACO) have used instead for a better characterization of unfolded proteins or less-structured regions. In fact C-direct detected experiments allow a better spreading of the frequency resonances of overlapped signals based on amino acids resonances specificity.

All NMR experiments were acquired on Bruker AVANCE 900, AVANCE 800, AVANCE 700 AVANCE 600 and DRX 500 spectrometers equipped with triple resonance CRYO-probes.

2.1.2 Assignment strategy

All the acquired spectra were then processed with the Bruker TOPSPIN software Packages and analyzed with the program CARA[13] (Computer Aided Resonance Assignment, ETH Zürich) for the resonance assignment.

In the assignment procedure for unlabeled biomolecules, at first instance, the spin system is assigned through the implementation of 2D COSY and 2D TOCSY and then the sequential assignment is obtained through the identification of NOEs among nearby bases.

In the case of DNA, while with 2D COSY/TOCSY the sugar spin systems can be assigned, with the 2D ^{1}H-^{1}H NOESY experiment it is possible to establish the conformation of the glycosidic torsion angle, according to the relative intensity of the NOEs between the aromatic protons of the base and the anomeric proton of the sugar of the same nucleotide (higher for bases in the *syn* conformation with respect to those in the *anti*). Moreover, from the 2D ^{1}H-^{1}H NOESY the sequential assignment can be achieved from aromatic-anomeric backwards connectivities for bases adopting the *anti* conformation. The topology of DNA construct can be evinced from the NOEs among non-adjacent bases and also exploiting the effect of spin diffusion (with long NOESY mixing times) among paired base. In this way, it is easier to identify H-bonded coupled bases.

The application of multidimensional NMR spectroscopy allowed the development of general strategies for the assignment of nuclei resonances in proteins. All procedures use the known protein sequence to connect nuclei of amino acid residues which are neighbors in the sequence.

The assignment procedure can be divided into three main steps. The first step consists in the sequential assignment of the backbone and in the association of the protein residues to the signals of the 2D ^{1}H-^{15}N HSQC spectrum, for which are required the abovementioned experiments. Once the backbone assignment is obtained it is possible to monitor the residue-specific behavior during a titration experiment, dynamics properties and to exploit paramagnetic constraints for the protein structural and dynamics characterization.

The complete set of backbone chemical shifts for the Hα, Cα, Cβ, CO and N nuclei resonances can be used to predict the secondary structure of the protein and to derive the backbone torsion angles (φ, ψ), with the program TALOS+[14], that are then implemented in the structure calculation procedure.

The second step of the assignment procedure, instead, regards with the side-chains assignment in order to determine carbons and protons belonging to the same spin system, starting from already assigned Cα and Cβ. Interestingly, in multidimensional NMR spectroscopy, the sequence-specific assignment of the backbone resonances precedes the identification of side-chains spin systems.

The third step concerns with the interpretation of the NOESY spectra to find connections between protons close in space, that are then used as distance constraints for the structure calculation.

2.1.3 Structure calculation, refinement and validation

Protein structure calculation based on NMR data typically involves two steps: structure generation and structure refinement.

In the first step, all the conformational restraints are assembled as input to suitable programs, whose output is a bundle of conformers that are all equally consistent with the restraints.

For the structure generation, NMR measured parameters need then to be converted into structural restraints. This process is based on a known physical relationship between the NMR observable and the geometrical property they refer to, that derives from the analysis of the already known structures.

Structure calculation mostly rely on the identification of a dense network of distance restraints from NOEs between nearby hydrogen atoms, permitting the identification of pairs close in space, although distant in the protein sequence. Long-range NOEs are, in fact, crucial for defining the overall protein folding.

NOEs are converted into distance restraints according to the following relation, that correlate together the NOESY cross-peak volume (V) to the distance (r) between the two hydrogen atoms:

$$V = \langle r^{-6} \rangle f(\tau c)$$

The program CYANA 2.1[15] implements an automated calibration routine of the NOESY peak volume on the basis of the expected average distance observable in the spectra,

that can be adjusted and NOEs are usually treated as upper interatomic distances rather than as precise distance restraints.

Supplementary constraints can be derived from dihedral angles that restrict the local conformation of a residue to the α-helical or β-sheet region of the Ramachandran plot.

In addition to the experimental constraints, information about the covalent structure of the protein, such as the amino acid sequence, bond lengths, bond angles, chiralities, planar groups, as well as steric repulsion between non-bonded atom pairs, should be implemented in the program used for the structure calculation. The program CYANA-2.1[15] uses a torsion angle dynamics algorithm, where torsion angles are used as degree of freedom in the molecular dynamics simulation, while bond lengths, angles, and backbone peptide plane angles are fixed, according to the values of the library.

At the end of the structure calculation two important parameters have to be considered: the agreement of the calculated structure with the experimental constraints, expressed by the target function, and the convergence of the calculation, evaluated by the root mean square deviation (RMSD) of the calculated models with the lowest target function. The average values of target function for a considered good bundle of structures, should be smaller than 1 $Å^2$ and each single violation should be smaller than 0.3 $Å^2$. Instead the RMSD evaluated for the family of structures should be close or less than 1 Å.

The second step instead involve structure refinement. Programs using force field, such as AMBER[16], are particularly suitable for this purpose. The structure refinement is carried out with a molecular dynamics simulation in explicit water box to further improve the quality of the structure[17].

The refined structure is then validated according to specific programs, such as iCING (http://nmr.cmbi.ru.nl/cing/Home.html) and PSVS[18] servers that evaluate the geometry of the protein and the Ramachandran plot appearance of backbone torsion angles. The solution structure is considered acceptable if more than 90% residues fall into the allowed region of Ramachandran plot and less than 1% residues in the disallowed region.

2.1.4 Evaluation and analysis of paramagnetic constraints

Paramagnetic constraints can be used both for structure calculation and refinement because they provide long-range distance and orientational information. Since, paramagnetic constraints are consistent with the diamagnetic ones (NOEs and the dihedral angles), they can

be profitably used to improve both the precision and the accuracy of the structure of the investigated protein.

The chemical shifts and ^{1}J splittings are largely affected by the presence of a paramagnetic metal ion and therefore the comparison of these parameters in the presence and in the absence of the metal ion itself allow the measurement of paramagnetic constraints.

The paramagnetism-based constraints that have been addressed to, in the present research project, are pseudo-contact shift (PCS) and residual dipolar coupling (RDC).

a) Pseudo-conctact shifts

The pseudo-contact shift arises from the dipolar interaction between the nuclear magnetic moment and the average induced magnetic moment of the paramagnetic metal ion unpaired electron[19]. In fact, when the electron magnetic moment is anisotropic, such as in most of lanthanides (Ce^{3+}, Pr^{3+}, Nd^{3+}, Pm^{3+}, Sm^{3+}, Eu^{3+}, Tb^{3+}, Dy^{3+}, Ho^{3+}, Er^{3+}, Tm^{3+}, Yb^{3+}), the dipolar interaction with the nearby nuclear spins magnetic moment is not completely quenched by molecular rotation. The nuclear spins thus sense the sum of the external magnetic field and of the field generated by the electron magnetic moment, which causes a difference in chemical shift of the paramagnetic molecule with respect to the diamagnetic one.

The pseudo-contact shift depends on the distance (r) and the angular position (expressed by the angular coordinates θ and φ) of the nucleus with respect to the metal ion and on the anisotropy of the magnetic susceptibility tensor of the paramagnetic metal ion itself (expressed by the axial ($\Delta\chi_{ax}$) and rhombic ($\Delta\chi_{rh}$) anisotropy parameters) (Figure 2):

$$\delta^{pcs} = \frac{1}{12\pi r^3}\left[\Delta\chi_{ax}(3cos^2\theta - 1) + \frac{3}{2}\,\Delta\chi_{rh}\,sin^2\theta\,\cos 2\varphi\right]$$

PCSs are calculated from the difference between the chemical shifts of the nuclei in the paramagnetic system and in a diamagnetic analog. The latter is obtained by substituting the paramagnetic metal ion with a diamagnetic metal. The acquisition and the complete assignment of the 2D ^{1}H–^{15}N HSQC spectra in both conditions is thus needed.

The assignment of the spectrum of the paramagnetic system can be obtained by comparison with the correspondent one of the diamagnetic system, considering that in 2D ^{1}H–

^{15}N HSQC spectrum the shift must be the same for both dimensions (^{1}H and ^{15}N), since PCSs do not depend on the observed nucleus. However, the re-assignment of the paramagnetic spectra is usually not complete because the interaction with the electron spin enhances the relaxation rates of the nearby nuclear spins. This increased relaxation rate causes the loss of resonances in the NMR spectrum due to severe line broadening.

From a first subset of PCS (at least eight), if a structural model is available, a first estimate of the susceptibility anisotropy tensor can be obtained using the program FANTASIAN[20], that, according to the PCS equation, provides also the calculated PCS for the other nuclei. A much larger number of PCS can then be found from the comparison of observed values in the 2D ^{1}H-^{15}N HSQC spectrum with the calculated ones.

b) <u>Residual dipolar couplings</u>

The same magnetic susceptibility anisotropy of the paramagnetic metal ion responsible for PCS also determines a perturbation on the ^{1}J splittings of coupled nuclei. In fact, the average magnetic moment of the unpaired electron induces partial orientation of the molecule in high magnetic fields, as a result of the different values of the energy for the different orientations with respect to the magnetic field. In this way, the nuclei dipolar interactions are not averaged to zero by isotropic molecular tumbling in solution and the residual dipolar coupling can be measured.

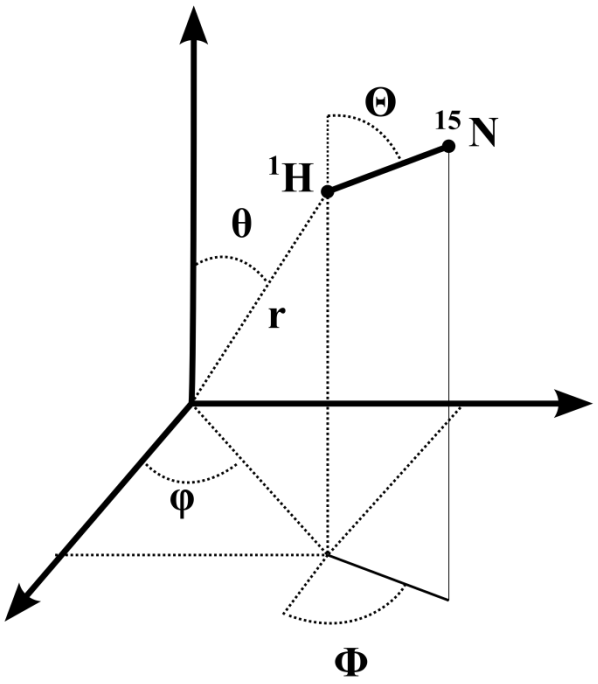


Figure 2 Structural parameters determining the H^N PCS (r,θ,φ) and PRDC values (Θ,Φ). The angles are provided in the frame established by the principal axes of the magnetic susceptibility anisotropy tensor.

The self-orientation RDC due to paramagnetism can thus be expressed as a function of $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$:

$$\Delta\nu^{rdc} = -\frac{1}{4\pi}\frac{B_0^2}{15kT}\frac{\gamma_A\gamma_B\hbar}{2\pi r_{AB}^3}[\Delta\chi_{ax}(3cos^2\Theta - 1) + \frac{3}{2}\Delta\chi_{rh}\,sin^2\Theta\cos 2\Phi]$$

where the Θ and Φ angles are defined in the principal frame of the paramagnetic susceptibility tensor (Figure 2). These angles define the orientation of the coupled nuclei independently of their position with respect to the metal ion. Therefore paramagnetic RDC depends only on the orientation of the nuclear pair, on the applied magnetic field and on the gyromagnetic ratio of the coupled nuclei (γ_A and γ_B).

RDC can be obtained from the difference between paramagnetic system and the diamagnetic analog in the measured ^{1}J splitting (in Hz) on the ^{15}N dimension (for a given N–H^N pair) in the 2D ^{1}H–^{15}N HSQC IPAP[21] (in-phase/antiphase) experiments.

RDC can be implemented in program for structural refinement, such as *paramagnetic-cyana*. Solution structure determination and refinement protocols are based on cycling between the simulated annealing structural calculation, performed using paramagnetism-based restraints (PCS and RDC), with fixed $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ values provided by PCS, and the new evaluation of the tensors based on the new refined coordinates of the protein structure, until convergence is reached.

However, in these simulated annealing calculation distinct sets of PCSs and RDCs are necessary to reach convergence. In fact, a large number of possible positions and orientation can equally fulfill the PCSs and RDCs equations. Therefore, to escape false minima, reach convergence and remove degeneracy of solutions, different alignment provided by substitution of different paramagnetic ions in the metal binding site, are needed.

2.2 NMR and protein dynamics

2.2.1 NMR Relaxation experiments

Dynamics information of proteins can be easily obtained from the measurement of relaxation rates. Relaxation refers to the return of a system to equilibrium after a perturbation and the relaxation time/rate is the time/rate constant for such a process. Relaxation is linked with the dynamics of a system because this effect is caused by the molecular motions that induce local magnetic field fluctuations.

The dynamics of a protein can be regarded as a superposition of global reorientational dynamics of the whole particle and internal dynamics at a more local level. However, the two classes of dynamics take place in separate time scales and can be readily separated.

In relaxation experiments it is monitored how the orientation of a particular bond vector changes with time. The ^{15}N relaxation experiments are the most common and have been extensively used to characterize backbone dynamics. The relaxation of this nucleus is due to the ^{15}N chemical shift anisotropy (CSA) and to the dipolar interaction with the covalently attached hydrogen. Since the reorientation of this bond vector, that causes the relaxation, with respect to the static magnetic field is faster than the molecular tumbling, information about pico- to nanosecond motions can be obtained by this analysis.

The return of the magnetization to the equilibrium depends on two components, parallel to magnetic field (longitudinal relaxation), that relaxes with the rate-constant R_1, and perpendicular to the magnetic field (transverse relaxation) that relaxes with the rate-constant R_2.

The ^{15}N longitudinal relaxation rates (R_1) have been measured using a sequence modified to remove cross correlation effects between the dipolar and ^{15}N CSA relaxation mechanisms during the relaxation delay[22]. Different inversion recovery times have been used for the evaluation of the decay of the signals in the experiments.

The ^{15}N transverse relaxation rates (R_2) were measured using a Carr-Purcell-Meiboom-Gill (CPMG) sequence[23] with different delays.

Heteronuclear NOE[24] can be also measured as reporter of the internal dynamics of the system. The measurement is performed by saturating the proton signal and observing changes in the ^{15}N signal intensities.

The longitudinal (R_1) and transverse (R_2) relaxation rates were determined by fitting the cross-peak volumes as a function of the different used delays to a single-exponential decay using the program Origin. The equation used for the fitting is written below:

$$f(t) = I_0 \times e^{(-Rt)} + A_0$$

where f(t) is the calculated volume, t is the variable time used in seconds, R is the relaxation constant R_1 or R_2 in (s-1) which has to be determined and I_0 and A_0 are the constants used for the fitting of the curve.

The experimental data are usually compared with predicted data considering a rigid bead hydrodynamic model of the protein with the program HydroNMR[25]. Deviation of the experimental values from calculated ones is an indication of protein motion.

R_1 and R_2 experimental values contain both the information about the overall rotational dynamics, linked to the protein dimension and expressed by the correlation time, and about the local bond vectors dynamics. The correlation time is derivable from the ratio of longitudinal and transversal relaxation times of rigid structures of the protein, in the case of the isotropic approximation, using the program TENSOR2[26]. R_1, R_2, and NOE values can be then used in the anisotropic approximation to map out the spectral density functions describing protein motions at a number of frequencies faster than the correlation time.

2.2.2 The protocol of the maximum occurrence for the evaluation of conformational heterogeneity of multidomain proteins

Multidomain proteins exist in solution in different conformations with similar energy and PCSs and RDCs are quite informative restraints of this conformational heterogeneity. In fact, the experimentally measured values are the averages of the values corresponding to the different conformations assumed by the proteins.

RDCs can be used to assess the presence of interdomain mobility in proteins. In fact, RDCs do not depend on the distance of the observed nuclear pairs from the metal ion and in case of rigid systems their distribution should be similar in both the domain bearing the paramagnetic metal ion and the other one. On the contrary, in case of motion, the spreading of the observed RDC values is reduced and it can collapse to zero for cases of overall isotropic reorientation of one domain with respect to the metal-bearing domain. Conformational heterogeneity should be thus invoked in the case of significantly reduced RDC for domain without the paramagnetic metal ion with respect to the domain with the coordinated paramagnetic ion.

The analysis of the experimental RDCs and PCSs also provides information on the conformations actually experienced by the system. With the protocol of the maximum occurrence[27], it is possible to score any possible conformation with a value indicating its maximum probability to be sampled in solution. The protocol is based on the availability of

the magnetic susceptibility anisotropy tensors from the PCS data measured for the domain hosting the paramagnetic metal.

Random conformations with all possible relative domain positions and orientations are generated (with the program RANCH[28]) and PCSs and RDCs are calculated (with the program CALCALL) for all the generated conformations. Then a simulated annealing minimization is carried out for each selected conformation, that is so included with a fixed weight in a ensemble of 50 other conformers with variable weights, in order to provide weighted average of calculated PCSs and RDCs that are in agreement with the experimental data. During the minimization, the weight of the fixed conformation is then gradually increased. The MO of such a conformer is defined as the largest weight for which the TF is smaller than a given threshold. The minimum for the TF is calculated by generating structural ensembles without any fixed conformation and giving 10% of tolerance.

These calculations were performed through the GRID infrastructure that allow big computational sources.

2.3 Bibliography:

1. Derome, A. & Williamson, M. COSY,2D homonuclear shift correlation ;with double quantum filter ;phase sensitive ;water suppression using excitation sculpting with gradients ;allowing for presaturation during relaxation delay in cases ; of radiation damping ; ;phasecycle: *J. of Magn Resonance* **88**, 177–185 (1990).
2. Hwang, T.-L. & Shaka, A. J. COSY, 2D homonuclear shift correlation ;with double quantum filter ;phase sensitive ;water suppression using excitation sculpting with gradients ;allowing for presaturation during relaxation delay in cases ; of radiation damping ; ;phasecycle: *J. of Magn Resonance* **112**, 275–279 (1995).
3. BAX, A. & Davis, D. G. Homonuclear Hartman-Hahn transfer using MLEV17 sequence for mixing using two power levels for excitation and spinlock phase sensitive. *J. of Magn Resonance* **112**, 275–279 (1995).
4. Phan, A. T. Long-range imino proton-13C J-couplings and the through-bond correlation of imino and non-exchangeable protons in unlabeled DNA. *J. Biomol. NMR* **16**, 175–178 (2000).
5. Piotto, M., Saudek, V. & Sklenář, V. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *Journal of Biomolecular NMR* **2**, 661–665 (1992).
6. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* **89**, 496–514 (1990).
7. Grzesiek, S. & Bax, A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *Journal of Magnetic Resonance (1969)* **96**, 432–440 (1992).
8. Grzesiek, S. & Bax, A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* **99**, 201–207 (1992).
9. Olejniczak, E., Xu, R. & Fesik, S. A 4D HCCH-TOCSY experiment for assigning the side chain ^{1}H and ^{13}C resonances of proteins. *Journal of Biomolecular NMR* **2**, 655–659 (1992).
10. Zhang, O., Kay, L. E., Olivier, J. P. & Forman-Kay, J. D. Backbone ^{1}H and ^{15}N resonance assignments of the N-terminal SH3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient NMR techniques. *Journal of Biomolecular NMR* **4**, 845–858 (1994).

11. Fernández, C. & Wider, G. TROSY in NMR studies of the structure and function of large biological macromolecules. *Curr. Opin. Struct. Biol.* **13**, 570–580 (2003).
12. Bermel, W. *et al.* Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J. Am. Chem. Soc.* **128**, 3918–3919 (2006).
13. Keller, R. *CARA*. (2003).
14. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44**, 213–223 (2009).
15. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA. *Journal of Molecular Biology* **319**, 209–227 (2002).
16. Bertini, I., Case, D. A., Ferella, L., Giachetti, A. & Rosato, A. A Grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics* **27**, 2384–2390 (2011).
17. Ponder, J. W. & Case, D. A. Force fields for protein simulations. *Adv. Protein Chem.* **66**, 27–85 (2003).
18. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778–795 (2007).
19. Bertini, I., Luchinat, C. & Parigi, G. Magnetic susceptibility in paramagnetic NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy* **40**, 249–273 (2002).
20. Banci, L. *et al.* The use of pseudocontact shifts to refine solution structures of paramagnetic metalloproteins: Met80Ala cyano-cytochrome c as an example. *Journal of Biological Inorganic Chemistry* **1**, 117–126 (1996).
21. Ottiger, M., Delaglio, F. & Bax, A. Measurement ofJand Dipolar Couplings from Simplified Two-Dimensional NMR Spectra. *Journal of Magnetic Resonance* **131**, 373–378 (1998).
22. Kay, L. E., Nicholson, L. K., Delaglio, F., Bax, A. & Torchia, D. . Pulse sequences for removal of the effects of cross correlation between dipolar and chemical-shift anisotropy relaxation mechanisms on the measurement of heteronuclear T1 and T2 values in proteins. *Journal of Magnetic Resonance (1969)* **97**, 359–375 (1992).
23. Peng, J. W. & Wagner, G. Investigation of protein motions via relaxation measurements. *Meth. Enzymol.* **239**, 563–596 (1994).
24. Zhu, G., Xia, Y., Nicholson, L. K. & Sze, K. H. Protein dynamics measurements by TROSY-based NMR experiments. *J. Magn. Reson.* **143**, 423–426 (2000).

25. García de la Torre, J., Huertas, M. L. & Carrasco, B. HYDRONMR: Prediction of NMR Relaxation of Globular Proteins from Atomic-Level Structures and Hydrodynamic Calculations. *Journal of Magnetic Resonance* **147**, 138–146 (2000).

26. Korchuganov, D. S. *et al.* Determination of protein rotational correlation time from NMR relaxation data at various solvent viscosities. *J Biomol NMR* **30**, 431–442 (2004).

27. Bertini, I. *et al.* Conformational space of flexible biological macromolecules from average data. *J. Am. Chem. Soc.* **132**, 13553–13558 (2010).

28. Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **129**, 5656–5664 (2007).

3. Results

3.1 A New DNA Structural Motif: the G-Triplex

Vittorio Limongelli,[1,2] Stefano De Tito,[1] Linda Cerofolini,[3] Marco Fragai,[3] Bruno Pagano,[1] Roberta Trotta,[1] Sandro Cosconati,[1] Luciana Marinelli,[1] Ettore Novellino,[1] Ivano Bertini,[3] Antonio Randazzo,[1,±] Claudio Luchinat[3,±] and Michele Parrinello[2,±]

[1] Dipartimento di Chimica Farmaceutica e Tossicologica, Università di Napoli "Federico II", Via D. Montesano, 49, I-80131 Napoli, Italy

[2] Department of Chemistry and Applied Biosciences, ETH Zurich, and Facoltà di Informatica, Istituto di Scienze Computazionali (ICS), Università della Svizzera Italiana, Via Giuseppe Buffi 13, CH-6900 Lugano, Switzerland

[3] Magnetic Resonance Center (CERM), University of Florence, 50019 Sesto Fiorentino, Florence, Italy and Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Florencùe, Italy

Accepted, Angewandte Chemie

A New DNA Structural Motif: the G-Triplex

Vittorio Limongelli,[1,2] Stefano De Tito,[1] Linda Cerofolini,[3] Marco Fragai,[3] Bruno Pagano,[1] Roberta Trotta,[1] Sandro Cosconati,[1] Luciana Marinelli,[1] Ettore Novellino,[1] Ivano Bertini,[3] Antonio Randazzo,[1,±] Claudio Luchinat[3,±] and Michele Parrinello[2,±]

[1] Dipartimento di Chimica Farmaceutica e Tossicologica, Università di Napoli "Federico II", Via D. Montesano, 49, I-80131 Napoli, Italy

[2] Department of Chemistry and Applied Biosciences, ETH Zurich, and Facoltà di Informatica, Istituto di Scienze Computazionali (ICS), Università della Svizzera Italiana, Via Giuseppe Buffi 13, CH-6900 Lugano, Switzerland

[3] Magnetic Resonance Center (CERM), University of Florence, 50019 Sesto Fiorentino, Florence, Italy and Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Florence, Italy

In memory of Prof. Ivano Bertini

[±]Corrisponding Authors: parrinello@phys.chem.ethz.ch; luchinat@cerm.unifi.it; antonio.randazzo@unina.it

Nucleic acids represent the alphabet of the cellular language and through their sequence and their topology regulate vital cell functions. In recent years it has been found that many variations from the Watson-Crick duplex structure play key roles in many cellular processes. Examples are hairpins,[2] cruciforms,[3] parallel-stranded duplexes,[4] triplexes,[5] G-quadruplexes[6] and the i-motif.[7] These structures can be formed by nucleotide sequences distributed throughout the whole human genome, their location is not random and often associated with human diseases.[8] These complexes are formed from one up to four strands, stabilized by base stacking and hydrogen bond interactions with a variety of non-standard pairings. For instance, DNA triplexes can present G:G-C, A:A-T, C^+:G-C and T:A-T pairings, with two strands in the standard Watson-Crick duplex structure (i.e. G-C and A-T) and the third one lying in the major groove of the duplex. Instead, G-quadruplexes are four stranded structures stabilized by stacking of two or more guanine tetrads (Figure 1).

These examples highlight the high structural polymorphism of DNA and suggest that other DNA structures might exist, perhaps with specific cellular functions that are to date unknown. In this scenario, the discovery of novel DNA structural motifs is more than ever an important breakthrough. Here, by using metadynamics simulations,[9] we have identified a stable folding intermediate of the Thrombin Binding Aptamer (TBA) quadruplex.[10] This intermediate is characterized by a "G-triplex" structure, having G:G:G triad planes stabilized by an array of Hoogsteen-like hydrogen-bonds (Figure 1). This kind of structure has been already hypothesized in other investigations on different DNA sequences,[11] but never experimentally proven. Here, for the first time, we have structurally and thermodynamically characterized this new DNA structural motif, through a combination of a number of biophysical experiments.

The metadynamics calculations have been used to study the unfolding/folding of TBA, which is a 15-mer oligonucleotide (5'-dGGTTGGTGTGGTTGG-3') organized in an anti-parallel

monomolecular G-quadruplex with a chair-like structure (Figure 2a). This structure consists of two G-tetrads, able to coordinate a metal ion at the center, connected by two TT loops and a single TGT loop.

Metadynamics accelerates the sampling adding a bias on few degrees of freedom of the system, called collective variables (CVs). In such a way, long time scale events, such as ligand/protein docking[12] or protein/DNA folding, can be sampled in an affordable computational time and the free energy profile of the whole process can be computed. In the present case, the Free Energy Surface (FES) was calculated as a function of two CVs, the radius of gyration CV defined by the oxygen atoms of the guanines forming the G-tetrads and a second CV that counts the number of hydrogen bonds between these guanines (see Experimental Section in Supporting Information). Looking at the FES obtained after approximately 80 ns of metadynamics simulation, three main energy minima can be identified (Figure 2b). The deepest one, basin A, corresponds to the experimental G-quadruplex structure of TBA[13] with two G-tetrad planes and eight guanines involved in standard Hoogsteen hydrogen bonds. In the second minimum, basin B, TBA shows a partial opening of the 3' end with G15 oriented towards the solvent. In this basin G14 moves slightly from its original position, with the oxygen of the base pointing towards the center of the planes formed by G1:G6:G10 and G2:G5:G11 (Figure 2b). In such a way, G14 conserves the hydrogen bond interactions with G2 and G11 and at the same time it partially fills the vacancy in the coordination shell of the metal present in the core. This step can be considered the very first event in the unfolding process of TBA. In basin C, the 3' end opens completely with G14 and G15 leaving the G-tetrad planes and pointing towards the solvent (Figure 2b). This minimum is approximately 6.5 kcal/mol higher in energy than basin A and here TBA assumes a number of different conformations due to the conformational flexibility of the 3' overhang formed by T13-G14-G15. In fact, these bases rearrange to form a single helix that assumes a number of

conformations, all of them stabilised by stacking interactions. While the 3' overhang is flexible, the rest of TBA is rather stable forming two G:G:G planes, namely G-triads, composed by G1:G6:G10 and G2:G5:G11, that form an array of Hoogsteen-like hydrogen bonds (Figure 1 and 2b). In this conformation the metal ion is coordinated at the center of the two triads in a way similar to that of the G-quadruplex structure. This structure, named "G-triplex", differs from the known triplex structures not only for the base pairing but also for the structure (Figure 1).

It can be observed that the 3' end of TBA opens and closes several times, passing from basin A to basin B and basin C, and then folded again in the G-quadruplex structure, basin A (Figure 2c). Thanks to these recrossing events, the calculated FES is accurate and quantitatively well characterized. The stability of the TBA conformation in the three energy minima A, B and C, has been further assessed through molecular dynamics calculations (see Supporting Information). A movie showing the metadynamics simulation on TBA is provided in the Supporting Information.

To establish the intrinsic stability of the G-triplex, we removed from the 3' end of TBA the last four residues that are highly mobile in basin C. The resulting structure was found stable in an ordinary MD simulation that lasted over 100 ns (see Supporting Information for details). Its coordinates are provided in the Supporting Information.

The prediction of a stable G-triplex structure needed to be validated by experiments. To this effect, we performed a number of experiments on the same truncated form of TBA (5'-GGTTGGTGTGG-3', **I**). First, **I** was studied through ^{1}H-NMR experiments at 25 °C showing the presence of a predominant well defined hydrogen bonded structure in solution. This is supported by the presence of four well-defined exchangeable proton signals in the 11.0-12.5 ppm region of the 1D ^{1}H-NMR spectrum (Figure 3a). These signals are typical of DNA structures with Hoogsteen hydrogen bonds.[14] On the other hand, the region of aromatic

protons (6.5-8.5 ppm) is characterized by the presence of eleven intense signals that can be attributed to the seven guanines H8 and four thymines H6 protons, and by a number of minor signals (Figure 3a). The latter might be due to the presence of unstructured DNA in equilibrium with the structured one. This equilibrium is very sensitive to temperature with the structured form highly favored at low temperature, as shown by the spectrum obtained at 1 °C (Figure 3b). The structure adopted by **I** turned out to be very stable over the weeks.

The non-exchangeable base and sugar protons of **I** were assigned through the analysis of 2D NOESY, 2D TOCSY and 2D COSY NMR spectra (see Table S2 and Supporting Information). Interestingly, the presence of three intense cross-peaks between the H8 proton bases and sugar H1' resonances for residues G1, G5 and G10 in the NOESY spectrum (900 MHz, T = 1 °C, mixing time 100 ms), along with the presence of weak cross-peaks between the same aromatic protons and the H2' and H2" protons of their own ribose moiety, indicated that these three residues adopt a *syn* glycosidic angle conformation (Figure 3c). On the contrary, G2, G6, G8 and G11 turned out to have an anti glycosidic conformation, being inverted the intensity of the cross-peaks mentioned before (Figure 3c). The three H8 peaks of syn G residues are downfield shifted with respect to those of the anti ones, exactly as reported for TBA.[15] Furthermore, the three anti-Gs (G2, G6 and G11) show H8/H2'-H2'' sequential connectivities with the 5' neighboring syn-Gs (G1, G5 and G10). This suggests the presence of the tracks G1–G2, G5–G6, G10–G11 (underlined residues have a syn conformation) and the formation of a helical structure. The presence of unusual NOE connectivities between a number of Gs and Ts indicates that 5'-TG-3' and 5'-GT-3' do not adopt a helical winding, and that the TT and TGT tracts realistically form loops. Finally, the alternation of syn and anti G residues implies that **I**, as TBA, folds into an antiparallel structure.

The assignment of the exchangeable imino protons was instead obtained by JR-HMBC correlation experiment. The jump-return version of HMBC[16] allowed us to assign for each

guanine the imino proton starting from its H8 proton (see Figure 4a and Supporting Information). Once the imino protons have been assigned, we could determine the folding topology of **I** using the jump-return version of the 2D NOESY experiment. This allows to identify NOEs representing the interaction between the imino proton of one base and the H8 proton of the other. Generally, these NOEs are diagnostic of the presence of Hoogsteen base pairings. In our case, we observed NOEs between G11-NH (12.08 ppm)/ G5-H8 (7.44 ppm) and G5-NH (12.23 ppm)/G2-H8 (8.04 ppm). These two correlations indicate that, as in TBA, G5 is involved in the formation of Hoogsteen hydrogen-bonds with both G11 and G2 and this is in agreement with the calculated structure (Fig. 4b-c). We have also observed a NOE between G1-NH (12.00 ppm) and G6-H8 (8.26 ppm), indicating that G1 and G6 are also paired. All of this indicates that the structure of **I** could be characterized also by a second G-triad formed by G10, G6 and G1. Due to unfavorable T1 noise, no NOE between G6-NH and G10-H8 could be unambiguously detected. Nevertheless, the fact that the subunit G10-G11 adopts a helical winding strongly supports the idea that actually also G10 does take part in the second G-triad. The formation of the G-triplex is also supported by the presence of further NOEs between NH protons and other exchangeable protons identified by ^{1}H-^{15}N-HSQC experiments (see Supporting Information). In particular, we observed strong NOEs between G1-NH/G5-NH and G6-NH/G11-NH; medium intensity NOEs between G1-NH/G6-NH and G5-NH/G11-NH; and weak NOEs between G5-NH/G6-NH, that once again support the formation of the G-triplex, having two G-triads: G1-G6-G10 and G2-G5-G11 characterized by syn-anti-syn and anti-syn-anti arrangement of the residues, respectively. Finally, a number of other NOEs definitely confirmed the structure of the G-triplex. Particularly, T7-Me shows correlations with G1-NH, G6-NH and G10-NH, confirming that the loop TGT is spatially very close to the G-triad G1-G6-G10. Then, G11-NH shows correlations with T4-

H2',H2",H4',H5',H6 and T3-H4',H5',H5",H6. Furthermore, G5-NH is also correlated with T3-H3', indicating that the TT loop is very close to the G-triad G2-G5-G11.

All these experimental evidences are in fully agreement with the G-triplex structure found in our calculations (see Figure S3 and Supporting Information).

The thermodynamic stability and the molecularity of **I** was investigated by Circular Dichroism (CD) and Differential Scanning Calorimetry (DSC) experiments. The CD spectrum of **I** shows two positive bands at 289 and 253 nm, and two negative at 235 and 265 nm, indicative of an homopolar stacking of the nucleobases,[17] in agreement with the proposed G-triplex structure (see Supporting Information). Such spectrum closely resembles that of TBA (Figure 5a). However, both positive and negative bands are slightly shifted, thus suggesting a similar, but not identical, stacking of the bases. The CD melting profile of **I**, recorded at the wavelength of maximum absorbance variation upon folding (λ = 289 nm), shows almost superimposable heating and cooling curves (Figure 5b), *i.e.* the melting profile was reversible, with no significant hysteresis, indicating a *quasi*-equilibrium process. From these measurements, a melting temperature, T_m, of 33.5 (± 1.0) °C and a van't Hoff enthalpy change, $\Delta H°_{vH}$, of 145 (± 15) kJ mol^{-1} are derived (see Experimental Section in the Supporting Information).

DSC experiments were carried out to characterize the denaturation thermodynamics of **I** with a model-independent analysis method.[18] DSC thermograms for denaturation of **I** (Figure 5c) were obtained at two different heating rates, 0.5 and 1.0 °C min^{-1}. The different heating rate does not alter the thermodynamic parameters significantly, thereby demonstrating that the investigated process is not kinetically controlled.[19] Furthermore, the unfolding of **I** is a highly reversible process, since the original signal is recovered by rescanning the same sample. The DSC curves show a symmetric shape with a maximum centered at T_m of 34.0 (± 0.5) °C, in good agreement with that obtained by CD melting. The melting temperature was

almost concentration-independent, consistent with a structure resulting from unimolecular folding. The integration of the denaturation peak gives a $\Delta H°_{cal}$ of 135 (± 5) kJ mol^{-1}, which is almost identical to the van't Hoff enthalpy calculated from DSC curves and very close to that calculated from CD. This indicates that the transition of **I** is a two-states process. Finally, the calculated Gibbs energy value, $\Delta G°$, at 298 K is 4 (± 1) kJ mol^{-1}, and results from the characteristic compensation of the favorable enthalpy term with an unfavorable entropy contribution ($\Delta S°$ = 0.44 ± 0.02 kJ $mol^{-1}K^{-1}$). The whole set of thermodynamic parameters show that, as expected, **I** is less stable than TBA.[20] In particular, the enthalpy term is lower than the one derived for TBA, probably due to the greater contribution resulting from the stronger base stacking and larger number of hydrogen-bonds involved in the quadruplex structure. This is consistent with the lack of G-tetrads in the structure adopted by **I**.

In summary, during the folding process of the G-quadruplex DNA aptamer TBA, we have observed the formation of the G-triplex structural motif. The existence of this structure has been proven in an 11-mer oligonucleotide, whose structural and thermodynamics properties have been characterized by advanced computations, NMR, CD and DSC experiments. At variance with the already known triplex structures characterized by triads having standard Watson-Crick base pairings, G-triplex presents G:G:G triad planes stabilized by an array of Hoogsteen-like hydrogen bonds (Figure 1). Although this kind of structure was already hypothesized as intermediate of the folding process of other quadruplex forming sequences,[11] this is the first time that DNA has been unambiguously isolated and structurally characterized in this conformation.

This discovery is an important breakthrough that paves the way to new horizons in biology. In fact, guanine-rich regions, potentially able to form G-triplex structures, are very abundant in the genome and our study provides the molecular bases and the tools to investigate the presence of these structures in the genome, their biological role and the way to interact with

them. Furthermore, our study raises new questions: are G-triplex structures present in the folding process of other DNA or RNA structures? Are there small molecules able to bind and interact with G-triplex? Are G-triplex structures exploitable to design new aptamers? These are only some of the issues that will be addressed in the near future.

References.

[1] J. D. Watson, F. H. Crick, Nature 1953, 171, 737-738.

[2] G. Varani, Annu. Rev. Biophys. Biomol. Struct. 1995, 24, 379-404.

[3] D. M. J. Lilley, Proc. Natl Acad. Sci. USA 1997, 94, 9513-9515.

[4] J. H. van de Sande, N. B. Rmasing, M. W. Germann, W. Elhorst, B. W. Kalish, E. von Kitzing, R. T. Pon, R. C. Clegg, T. M. Jovin Science 1988, 241, 551-557.

[5] V. Sklená, J. Felgon, Nature 1990, 345, 836-838.

[6] G. N. Parkinson, M. P. H. Lee, S. Neidle, Nature 2002, 417, 876-880.

[7] M. Guéron, J. L. Leroy, Curr. Opin. Struct. Biol. 2000, 10, 326-331.

[8] R. D. Wells, Trends Biochem Sci. 2007, 32, 271-278.

[9] (a) A. Laio, M. Parrinello, Proc. Natl. Acad. Sci., USA 2002, 99, 12562-12566. (b) A. Barducci, G. Bussi, M. Parrinello, Phys. Rev. Lett. 2008, 100, 020603.

[10] O. C. Bock, L. C. Griffin, J. A. Latham, E. H. Vermaas, J. J. Toole, Nature 1992, 355, 564-566.

[11] (a) R. D. Gray, R. Buscaglia, J. B. Chaires, J. Am. Chem. Soc. 2012, 134, 16834-16844. (b) M. Boncĭna, J. Lah, I. Prislan, G. Vesnaver, J. Am. Chem. Soc. 2012, 134, 9657-9663. (c) T. Mashimo, H. Yagi, Y. Sannohe, A. Rajendran, H. Sugiyama, J. Am. Chem. Soc. 2010, 132, 14910-14918. (d) R. Stefl, T. E. Cheatham 3rd, N. Spacková, E. Fadrná, I. Berger, J. Koca, J. Sponer, *Biophys J.* **2003**, 85, 1787-1804.

[12] (a) V. Limongelli, L. Marinelli, S. Cosconati, C. La Motta, S. Sartini, L. Mugnaini, F. Da Settimo, E. Novellino, M. Parrinello, Proc. Natl. Acad. Sci. USA 2012, 109, 1467-1472. (b) G. Grazioso, V. Limongelli, D. Branduardi, E. Novellino, C. De Micheli, A. Cavalli, M. Parrinello, J. Am. Chem. Soc. 2012, 134, 453-463. (c) V. Limongelli, M. Bonomi, L. Marinelli, F. L. Gervasio, A. Cavalli, E. Novellino, M. Parrinello, Proc. Natl. Acad. Sci. USA 2010, 107, 5411-5416.

[13] V. M. Marathias, K. Y. Wang, S. Kumar, T. Q. Pham, S. Swaminathan, P. H. Bolton J.

Mol. Biol. 1996, 260, 378-94.
[14] J. Feigon, F. W. Smith, Nature 1992, 356, 164-168.
[15] P. Schultze, R. F. Macaya, J. Feigon, J. Mol. Biol. 1994, 235, 1532-1547.
[16] A. T. Phan, J. Biomol. NMR. 2000, 16, 175-178.
[17] (a) S. Masiero, R. Trotta, S. Pieraccini, S. De Tito, R. Perone, A. Randazzo, G. P. Spada, Org Biomol Chem. 2010, 8, 2683-2692. (b) A. I. Karsisiotis, N. M. Hessari, E. Novellino, G. P. Spada, A. Randazzo, M. Webba da Silva, Angew. Chem. Int. Ed. Engl. 2011, 50, 10645-10648. (c) A. Randazzo, G. P. Spada, M. Webba da Silva, *Top. Curr. Chem.* **2012** DOI: 10.1007/128_2012_331
[18] G. E. Plum, K. J. Breslauer, Curr. Opin. Struct. Biol. 1995, 5, 682-690.
[19] L. Petraccone, B. Pagano, V. Esposito, A. Randazzo, G. Piccialli, G. Barone, C. A. Mattia, C. Giancola, J. Am. Chem. Soc. 2005, 127, 16215-16223.
[20] I. Smirnov, R. H. Shafer, Biochemistry 2000, 39, 1462-1468.

Acknowledgments

The authors acknowledge Anh Tuân Phan for providing an optimized JR HMBC sequence and Janez Plavec for helpful suggestions. Furthermore, the authors acknowledge that the results of this research have been achieved using the PRACE Research Infrastructure resource JUGENE based in Germany at Forschungszentrum Juelich. This work is supported by the Gabriele Charitable Foundation, Italian Institute of Technology (IIT), Italian M.U.R.S.T., P.R.I.N. 2009, Biologia Strutturale Meccanicistica, Italian Association for Cancer Research (A.I.R.C.) and EU FP7, ERC Advanced Grant No 247075.

Correspondence and requests for materials should be addressed to parrinello@phys.chem.ethz.ch; luchinat@cerm.unifi.it; antonio.randazzo@unina.it

Figures:

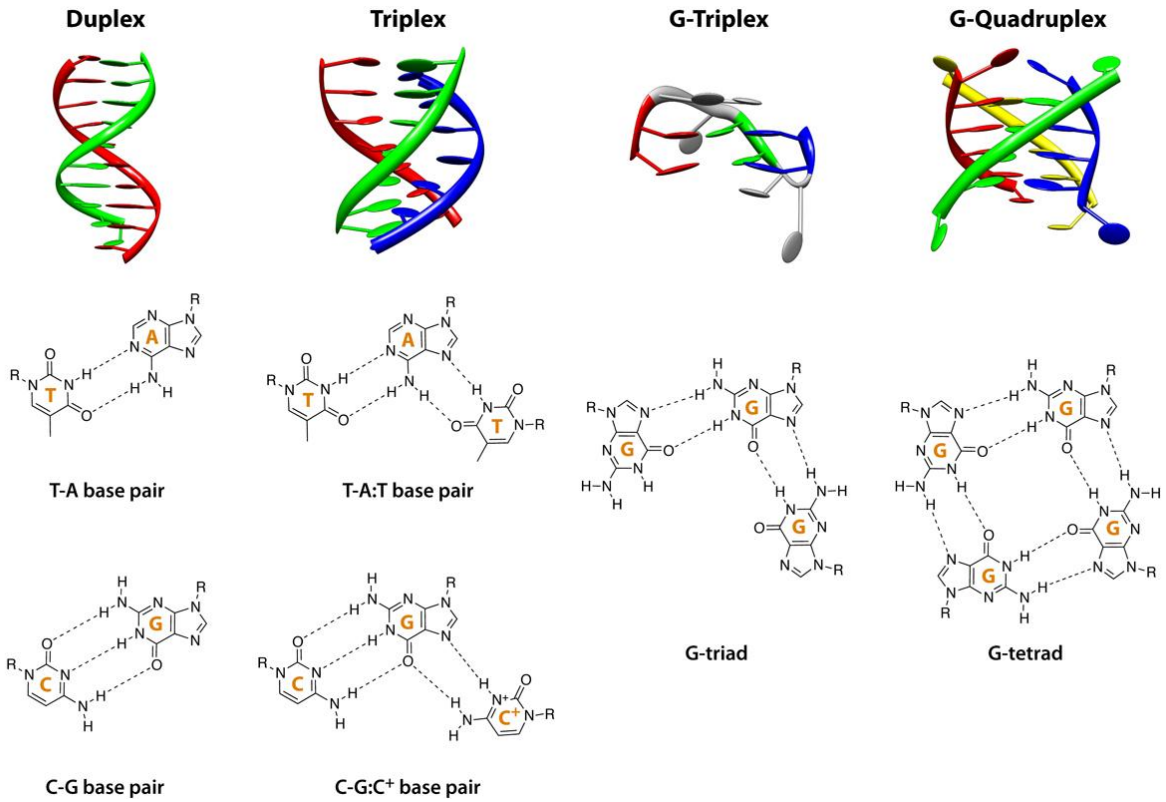


Figure 1. DNA structural motifs. (Top panel) Schematic illustration of duplex, triplex, G-triplex and G-quadruplex. (Bottom panel) Examples of base pairing: T-A and C-G for duplex, T-A:T and C-G:C^+ for triplex (where C-G and T-A form the standard Watson-Crick duplex structure, whereas the colon signs indicate the pairing with the third strand, that lies in the major groove of the duplex), G-triad for G-triplex, and G-tetrad for G-quadruplex. G-triad and G-tetrad consist of a planar arrangement of three and four guanines, respectively, held together by Hoogsteen-like hydrogen bonds.

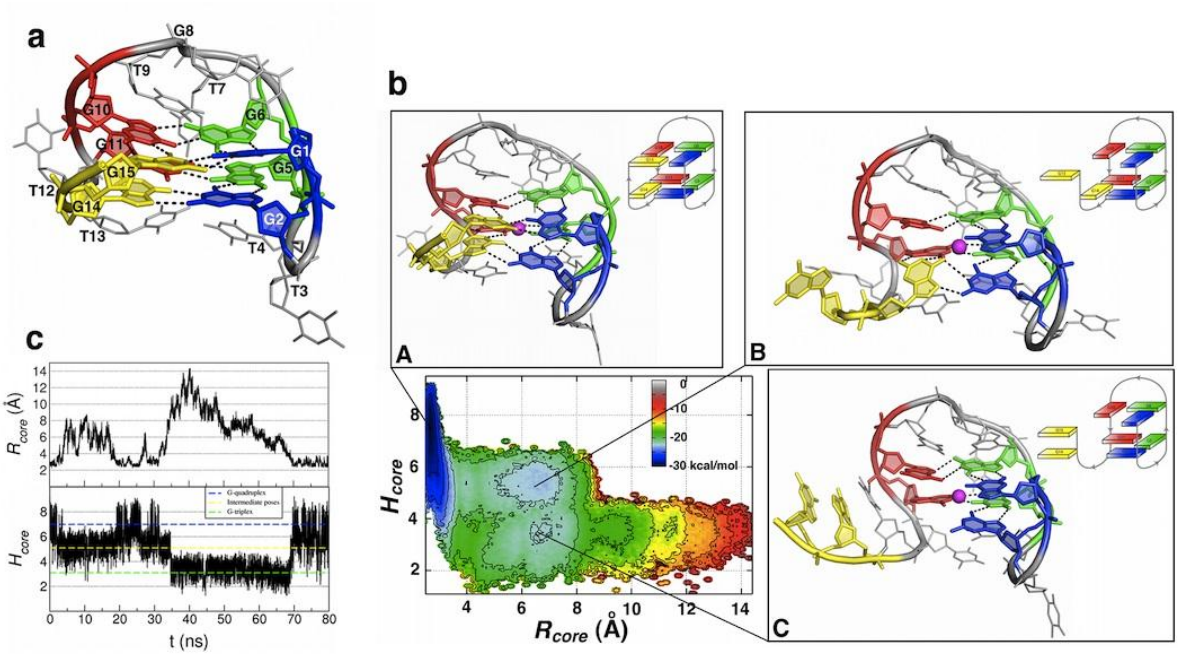


Figure 2. The 3' end opening of TBA. a) Tridimensional Representation of the NMR structure of TBA (PDB ID code 1qdf).[13] b) The Free Energy Surface (FES) of the 3' end opening of TBA shows three main energy minima are shown: one deep and narrow, basin A, which represents TBA in the G-quadruplex structure; the second one, basin B, which represents an intermediate state; and the last one, basin C, which shows the G-triplex structure formed by the triads G1:G6:G10 and G2:G5:G11. Residue labels are the same of (a). c) Plots showing the phase space represented as the H_{core} and R_{core} CV, explored during the metadynamics simulation.

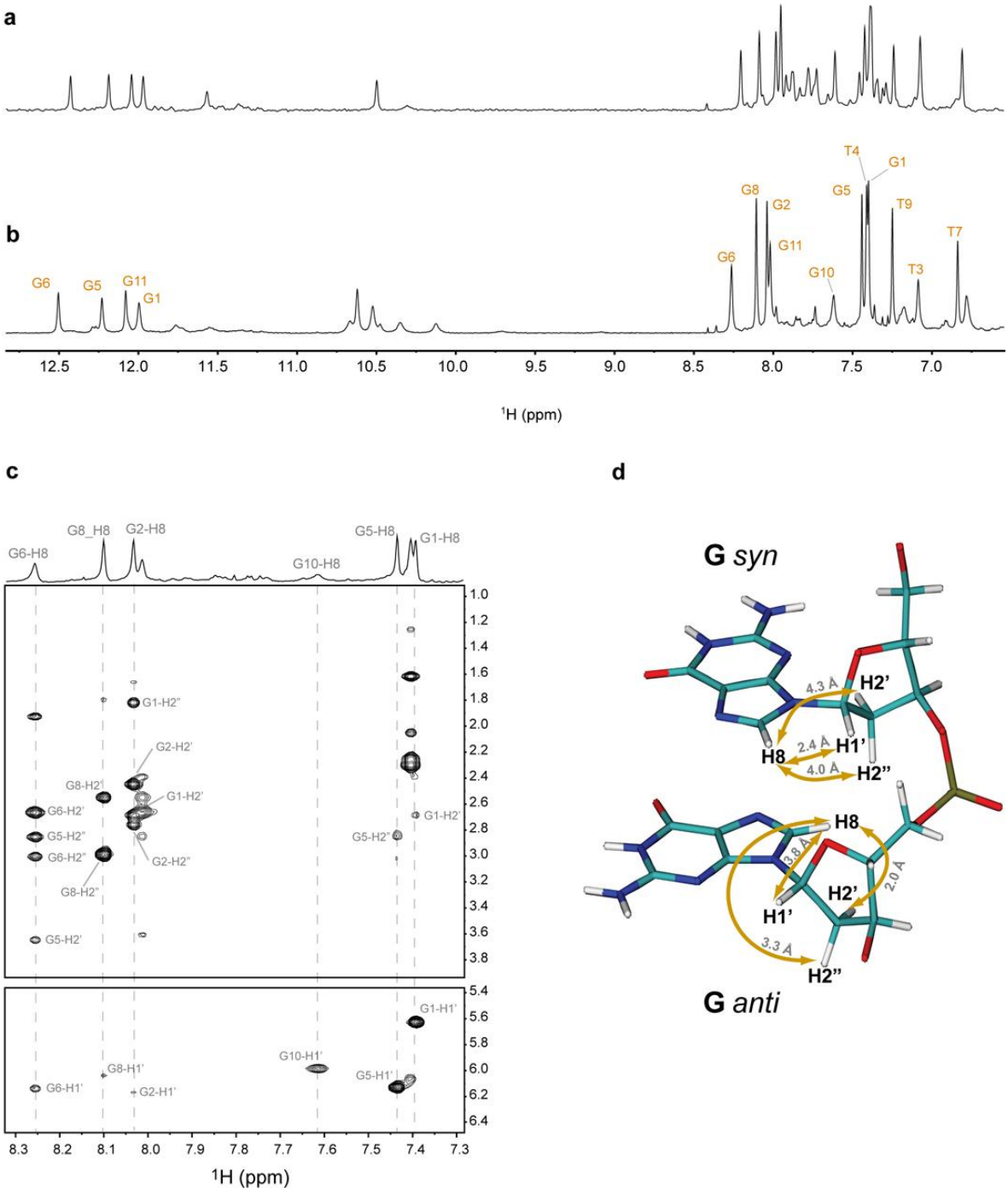


Figure 3. 1D ^{1}H-NMR and 2D NOESY spetra of 5'-GGTTGGTGTGG-3' (I) (70 mM KCl, 10 mM KH_2PO_4, 0.2 mM EDTA, pH 7). Imino, amino and aromatic regions of the 1D ^{1}H-NMR spectra of **I** acquired at 25°C (a) and 1 °C (b). In the latter spectrum (b) the signal at 7.63 ppm, attributed to G10-H8, is slightly broad, suggesting that its conformation can vary on the NMR timescale. c) Two expanded regions of the NOESY spectrum (900 MHz, T = 1 °C, mixing time 100 ms). H8 protons of the syn residues G1, G5 and G10 show intense cross-peaks with H1' proton and weak connectivities with H2' and H2" protons of their own sugar. On the contrary, anti residues G2, G6, G8 have opposite relative intensity. d) Distances of the correlated protons in the NOESY spectrum.

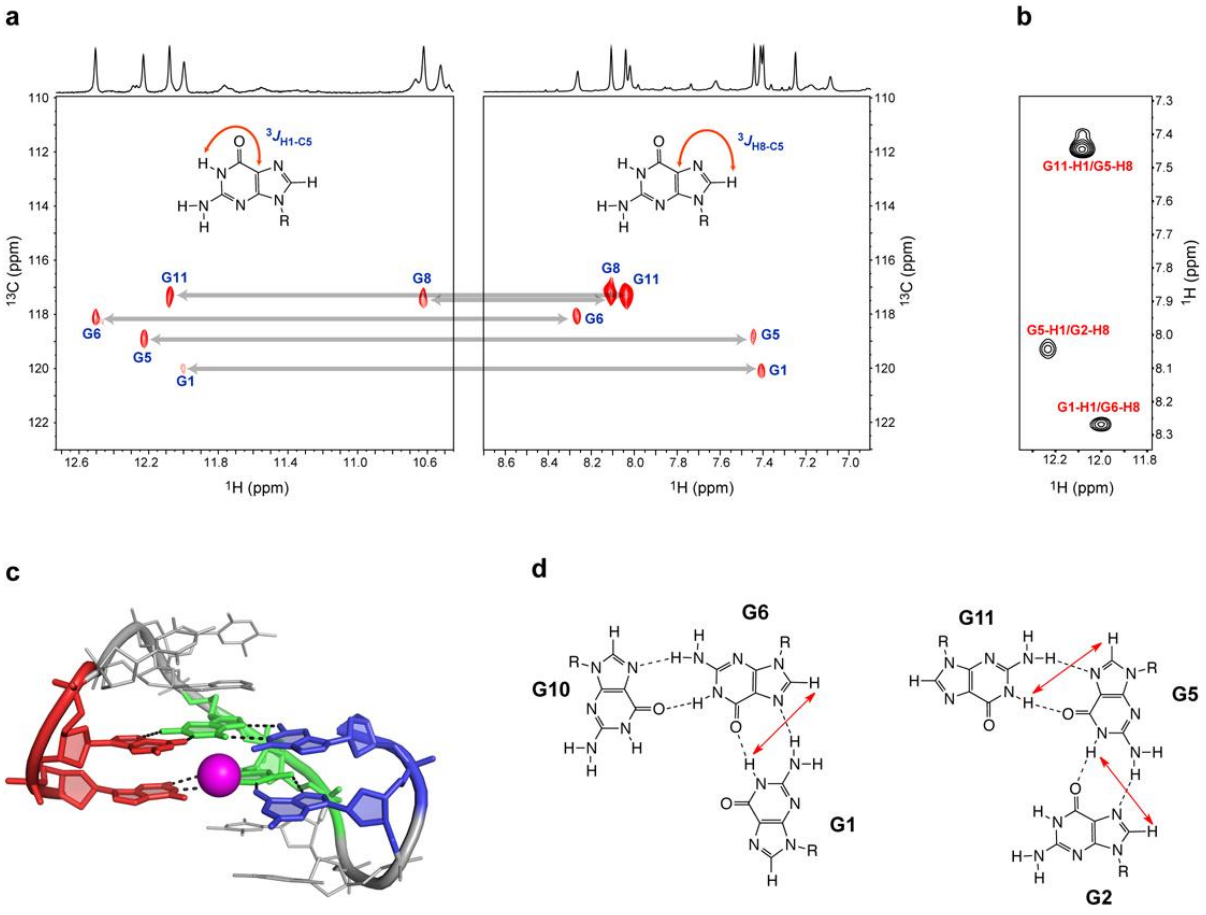


Figure 4. JR-HMBC and JR-NOESY spectra of 5'-GGTTGGTGTGG-3' (**I**) (70 mM KCl, 10 mM KH_2PO_4, 0.2 mM EDTA, pH 7). a) Expanded regions of the JR-HMBC (600 MHz, T = 1 °C) showing correlations between H8 and H1 guanine protons with C5 carbons. b) Expanded region of the 2D JR-NOESY (900 MHz, T = 1 °C) showing diagnostic correlations between H8 and H1 protons of paired bases. c) Tridimensional representation of the G-triplex structure adopted by **I** (Supporting Information). d) G-triads involved in the formation of the G-triplex. Red arrows show NOE correlations between H8 and H1 protons of paired bases.

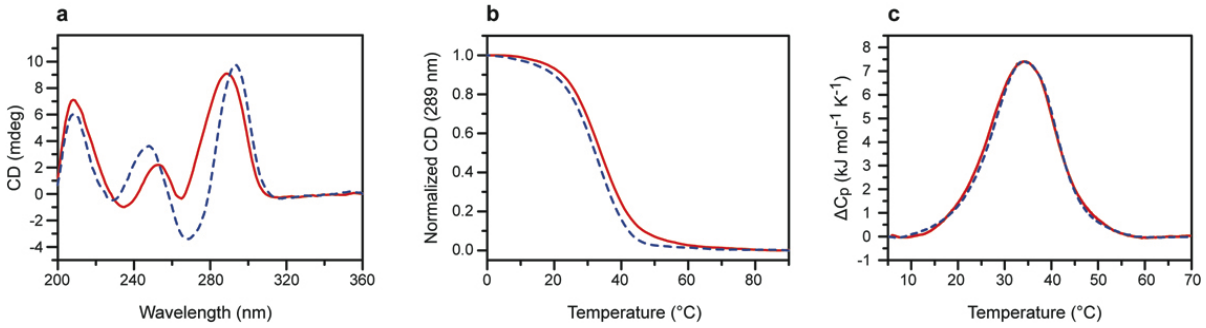


Figure 5. Biophysical characterization of 5'-GGTTGGTGTGG-3' (**I**). a) CD spectra of **I** (solid line) and TBA (dashed line) at 1 °C. b) Normalized CD melting curves of **I** recorded at 289 nm at a scan rate of 0.5 °C min^{-1}. Heating and cooling curves are shown with solid and dashed line, respectively. c) DSC profiles for **I** at 0.5 (solid line) and 1.0 (dashed line) °C min^{-1} heating rate. All the experiments were performed in a buffer solution containing 10 mM potassium phosphate, 70 mM KCl, 0.2 mM EDTA (pH 7.0).

SUPPORTING INFORMATION

Supporting Information Available:

- The results of the standard MD simulations carried out on the basin A, B and C conformation.
- The results of the standard MD simulation carried out on the 11-mer oligonucleotide I.
- NMR assignment of other exchangeable protons.
- NOEs supporting the formation of the G-triplex.
- Other investigated truncations of TBA.
- Further consideration on CD experiments.
- Experimental Section.
- The atoms used to define the radius of gyration and hydrogen bond CV (Table S1).
- ^{1}H-NMR assignment of I at 1°C (900 MHz) (Table S2).
- The movie of the folding/unfolding process of the 3' end of TBA (tba_movie.m4v).
- The atomic coordinates of I in the G-triplex structure (G-triplex.pdb). The reported structure is representative of the conformations ensemble obtained from the over 100-ns-long MD simulation.

Molecular dynamics simulations on TBA in the conformations representing the energy basins A, B and C.

We assessed the stability of the conformations representing the energy basins A, B and C using 50 ns long molecular dynamics simulations. Different behaviors have been observed for the different poses considered. As expected, basin A resulted very stable throughout the whole simulation with the two G-tetrads planes formed by G1:G6:G10:G15 and G2:G5:G11:G14, constantly formed (Figure S2A). As seen in the metadynamics calculation, even in the MD simulation the T7-G8-T9 loop is rather stable while T12 in the T12-T13 loop, is very flexible and points away towards the solvent. Furthermore, T3, T4 and T13 present a good conformational stability favored by the engagement of stacking interactions with the nucleobases placed above.

At variance with basin A, basin B is rather unstable. At this basin the partial opening of the 3' end of TBA has been observed during the metadynamics simulation, with the G15 base pointing towards the solvent (Figure 2 in the main text). As shown in Figure S2B, after approximately 18 ns of MD simulation TBA leaves its starting conformation, basin B, transforming in that of basin A. In particular, G15 slowly moves towards the center of the G-tetrads forming again the G-tetrads planes and the Hoogsteen hydrogen bonds first with G1 and then with G10. This finding is not totally surprising since, looking at the FES of the metadynamics simulation (Figure 2 in the main text), the energy barrier that separates basin B from basin A, is rather small, approximately 2 kcal/mol, and thus it can be overcome in the standard MD time-scale. This motion can be appreciated looking at Figure S2B where the r.m.s.d. of the guanine bases with respect to their position in the G-quadruplex conformation has been calculated.

A different behavior has been instead observed for basin C. In fact, during the MD simulation the TBA conformation of this energy minimum is stable conserving all the interactions previously found in the metadynamics calculation. In particular, at this pose, the 3' end of TBA is completely open with T13, G14 and G15 forming a single helix overhang that points away from the core of the TBA structure (Figure 2 in the main text). The other nucleobases are arranged to form a well-organized structure where G1:G6:G10 and G2:G5:G11 form two G-triad planes stabilized by an array of Hoogsteen-like hydrogen bonds. At variance with the other two minima A and B, in basin C, T12 does not point out towards the solvent, on the contrary it is rotated towards the core of TBA and is involved in stacking interactions with T4 of the TT loop at the 5' end (Figure 2b in the main text). This position favors the formation of hydrogen bond interactions between the imide hydrogens of T13 and G14 and the backbone phosphates of G11. These interactions further contribute to the energy stability of this pose.
The G-triplex structure is very stable along the whole MD simulation with a low average r.m.s.d. of 0.63 Å for the guanine bases forming the G-triad planes (Figure S2C).

Molecular dynamics simulation on the 11-mer oligonucleotide I.

We have carried out an extensive MD simulation, more than 100 ns, on the oligonucleotide I in the G-triplex conformation at the ionic concentration of 70 mM KCl used in the experiments. During this simulation the G-triplex structure is very stable with a low average r.m.s.d. of 0.88 Å for the guanines forming the triad planes. As found in the metadynamics simulation, the triads formed by G1:G6:G10 and G2:G5:G11 are stabilized by Hoogsteen-like hydrogen bonds (Figure S3). At the center of the two triad planes the potassium ion is coordinated by the six oxygen atoms of the guanines. Analogously to the metadynamics results, T4 has a good conformational stability thanks to the stacking interactions engaged with G5, while T3 is more flexible. However, during the simulation this base often occupies a

position under G2 where favorable stacking interactions can be formed. On the other hand, the T7-G8-T9 loop is stable with G8 stacked above the plane formed by the G1:G6:G10 triad engaging favorable interactions with G6. It is interesting to note that the T7-G8-T9 loop in oligonucleotide I occupies a conformation different from that assumed in TBA. However, these two different conformations are equally stable in all the simulations suggesting a considerable conformational freedom for the bases forming this loop.

All the simulations, on TBA and I, have shown a great stability of the G-triplex structure suggesting to investigate in the near future the presence of G-triplex structures even in other poliguanine sequences of DNA and RNA.

NMR assignment of other exchangeable protons

In the 1D ^{1}H-NMR spectrum, along with the intense and narrow signals of the assigned imino protons (δ_H 12.00, 12.08, 12.23, 12.51) a number of other exchangeable protons could be observed. In order to clarify whether these signals could be attributable to other imino or amino protons, a 2D ^{1}H-^{15}N HSQC experiment was acquired at 25 °C. At this temperature, the already assigned signals with δ_H (T=1°C) of 12.00, 12.08, 12.23, 12.51 and 10.65, were shifted respectively at δ_H 11.92, 11.99, 12.14, 12.37 and 10.45. These signals turned out to be correlated to nitrogens resonating at δ_N 143.37, 144.36, 144.66, 145.74 and 144.67 respectively. All these values are characteristic of N1 guanine bases. It is interesting to note that also the broad signal at δ_H 11.52 (25 °C) turned out to be correlated to similar frequencies (143.72 ppm), indicating that also this resonance is attributable to an imino proton of guanine. Unfortunately, no correlation could be observed in the JR-HMBC spectrum for this signal and it could not be unambiguously assigned to the pertinent base. Nevertheless, this signal is correlated with T9-H6,H1',H2",Me; T7-Me; G10-H2" and G8-H4'. All of this can be interpreted assuming that the imino proton resonating at δ_H 11.52 is assigned to G10.

Probably, the signals of this NH proton is observable because involved in hydrogen bond. In fact, taking into account that the strand G14-G15 originally present in TBA is now missing, it is reasonable to assume that the strands G1-G2 and G10-G11 tend to bring themselves closer to each other, so that the formation of hydrogen bonds between G1 and G2 and G10 and G11, respectively, cannot be ruled out.

NOEs supporting the formation of the G-triplex.

The assignment of the exchangeable imino protons was obtained by JR-HMBC correlation experiment. The jump-return version of HMBC (ref. 16 in the main text) allowed us to assign for each guanine the imino proton starting from its H8 proton (Figure 4a in the main text). In particular, H8 signals at δ_H 7.40 (G1), 8.02 (G11), 7.44 (G5) and 8.26 (G6) turned out to be correlated to the NH protons of their own base resonating at δ_H 12.00, 12.08, 12.23 and 12.51, respectively. As discussed in the main text, only in the case of G8, H8 (δ_H 8.11) is correlated with the imino proton at δ_H 10.65. This low value for an imino proton suggests that G8 is not involved in classical Hoogsteen hydrogen bonds.

Other investigated truncations of TBA.

Three additional oligodeoxynucleotides (5'-GGTTGGTGTGGTTG-3' (**A**), 5'-GGTTGGTGTGGTT-3' (**B**) and 5'-GGTTGGTGTGGT-3' (**C**)) have been studied by NMR to investigate the formation of the G-triplex structure in other truncations of TBA. Samples **A**, **B** and **C** were analyzed by ^{1}H-NMR at 1 °C in 80 mM K^+-containing buffer (see experimental section). The spectra of the three samples exhibited conformational heterogeneity. This does

not depend on the nature of the buffer solution, be it a potassium or sodium containing buffer, or on its concentration and temperature.

Further consideration on CD experiments.

Circular Dichroism (CD) spectroscopy has been exploited to further investigate the structure adopted by **I** in solution. CD spectra were recorded at 1 °C, a temperature at which the folded-unfolded equilibrium of **I** is strongly shifted towards the folded form, as shown by NMR, CD melting and DSC experiments. For DNA, a CD spectrum is generated by a chiral orientation of the chromophores, i.e. the nucleobases, and it is strictly related to the base-stacking pattern. Since the two faces of a base are heterotopic, when two bases are stacked together, they can interact through the same (head-to-head or tail-to-tail) or the opposite (head-to-tail) faces, leading to a heteropolar or homopolar stacking, respectively. When the glycosyl bonds of the guanines alternate in syn and anti conformations along the strand, head-to-head and/or tail-to-tail interactions are realized, leading to a CD signal characterized by two positive bands at approx. 295 and 245 nm and two negative bands at approx. 230 and 265 nm (Ref. 17 in the main text). The experimental CD spectrum of **I** is characterized by two positive CD bands at 289 and 253 nm, and two negative bands at 235 and 265 nm (Figure 5a in the main text), values that are consistent with the structure proposed by metadynamics calculations and NMR.

Experimental Section.

Metadynamics simulations. The starting conformation coordinates for TBA were obtained in its G-quadruplex conformation (PDB ID code 1qdf).[1] The system was solvated using the TIP3P water model[2] and neutralized adding Na^+ ions with one of these ions placed at the center of the G-tetrad planes. All the simulations were carried out using periodic boundary

conditions and Particle Mesh Ewald to treat long range electrostatic. Before doing metadynamics simulations the system was equilibrated through 25 ns MD under NPT conditions at 1 atm and 300 K using the parmbsc0 parameters, which is a modified version of the Amber force field adapted for nucleic acids.[3-5] The Amber charges were applied to the DNA and waters atoms. The PLUMED plugin[6] was used to carry out metadynamics calculations with the NAMD code.[7]

The estimation F(s,t) at time t of the free-energy surfaces F(s) as a function of the CVs was determined by metadynamics in its new well-tempered variant, using the following formula:

$$F(s,t) = -\frac{T+\Delta T}{\Delta T}\, V(s,t)\ ,$$

where V(s,t) is the bias potential added to the system and T is the temperature of the simulation. ΔT is the difference between the fictitious temperature of the CV and the temperature of the simulation. The bias potential is made up by the sum of the Gaussians deposited along the trajectories of the CVs. Thanks to this new formalism, one can increase barrier crossing and facilitate the exploration in the CVs space by tuning ΔT. A Gaussians deposition rate of 0.2 kcal/mol per picosecond was initially used and gradually decreased on the basis of the adaptive bias with a ΔT of 2,700 K.

We used two CVs. The first one is the radius of gyration (R_{core}) calculated on the oxygen atoms of the guanines forming the G-tetrad planes (Table S1). This collective variable is calculated using the following formula:

$$R_{core} = \left(\frac{\sum_i^n |r_i - r_{com}|^2}{\sum_i^n m_i} \right)^{1/2},$$

where the sums are over the n atoms and the center of mass is defined by

$$r_{com} = \frac{\sum_i^n r_i\, m_i}{\sum_i^n m_i}.$$

The gyration radius is a common descriptor in protein folding studies because it is able to discriminate between completely unfolded and globule states. A Gaussian width of 0.02 Å was used for this CV. In order to distinguish between globule and folded states, we used as second CV the number of intramolecular hydrogen bonds engaged by the guanine bases forming the G-tetrad planes (H_{core}). The number of hydrogen bonds is evaluated using the switching function

$$H_{core} = \sum_{ij} \frac{1 - (\frac{d_{ij}}{r_0})^n}{1 - (\frac{d_{ij}}{r_0})^m},$$

where r_0 is set to 2.0 Å, n and m are set to 6 and 12, respectively, i and j are the donor and acceptor hydrogen bond atoms of the guanines used to calculate the number of hydrogen bonds (Table S1).

The H_{core} CV has been used to compute the two-dimensional FES by means of a reweighting algorithm.[8] In fact, this algorithm is able to reconstruct the Boltzmann distribution of CVs different from those used originally in the metadynamics run. Once the free energy of the metadynamics simulation is converged, using the newly computed probability distribution, the reweighting method allows to build the FES on these CVs.

NMR spectroscopy. All the experiments were collected on samples of the TBA and of its truncated construct I at concentrations ranging from 0.7 mM up to 1.5 mM in potassium phosphate buffer (70 mM KCl, 10 mM KH_2PO_4, 0.2 mM EDTA, pH 7). NMR spectra were performed at 1 °C and 25 °C on Bruker spectrometers operating at 900, 700 and 600 MHz, equipped with triple resonance cryo-probes. Different pulse sequence schemes (presaturation, excitation sculpting[9,10] and watergate[11]) were used to suppress the water signal. 2D ^{1}H-^{1}H TOCSY and 2D ^{1}H-^{1}H COSY experiments were performed at 1 °C operating at 700 MHz for the assignment of the spin systems. Proton–proton distance restraints were derived from the

analysis of 2D ^{1}H-^{1}H NOESY and 2D ^{1}H-^{1}H ROESY acquired in H_2O at 900 and 800 MHz. 2D ^{1}H-^{1}H NOESY was acquired using several mixing times ranging between 50 and 300 ms. 2D ^{1}H-^{15}N HSQC spectrum was acquired at 800 MHz to identify imino protons. Intranucleotide connectivities between imino and aromatic protons were obtained by a jump-and-return HMBC spectrum acquired at 600 MHz and 950 MHz on an isotopically natural-abundance sample using 5120 transients. All spectra were processed with the Bruker TOPSPIN software packages and analyzed by the program CARA (Computer Aided Resonance Assignment, ETH Zurich).[12]

CD and DSC experiments. Oligonucleotide samples were prepared by using the same buffer solution used for NMR experiments. CD spectra and CD melting curves of oligonucleotide samples were recorded by using a Jasco J-715 spectropolarimeter equipped with a Jasco JPT-423-S temperature controller using 1 mm path-length cuvettes. CD spectral scans were accumulated over the wavelength range 200−360 nm at 1 °C. The spectra were recorded at a scan rate of 100 nm/min with a response of 1 s, at 2.0 nm bandwidth and were averaged over 5 scans. Buffer baseline was subtracted from each spectrum. CD melting and annealing curves were recorded as a function of temperature in the range 0-90 °C at 289 nm with a scan rate of 0.5 °C/min. The CD melting curve of I was analyzed with a two-state model, using a theoretical equation for an intramolecular association, according to the van't Hoff analysis.[13] The T_m and $\Delta H°_{vH}$ values provide the best fit of the experimental melting data.

Differential scanning calorimetry (DSC) measurements were carried out using a Nano DSC III (TA Instruments, New Castle, DE). The experiments were performed at oligonucleotide concentration of 0.2 mM. Scans were carried out at 0.5 and 1.0 °C/min scan rate in the temperature range 0-90 °C. Reversibility and repeatability were proven by multiple scans. A buffer-buffer scan was subtracted from the sample-buffer scans and linear-polynomial

baselines were drawn for each scan. Baseline corrected thermograms were then normalized with respect to the oligonucleotide concentration to obtain the corresponding molar heat capacity curves. We note that the initial and final states of the transition have similar heat capacity values, indicating that the unfolding of I is accompanied by a negligible heat capacity change. The model-independent transition enthalpies were obtained by integrating the area under the heat capacity versus temperature curves.[13] The melting temperatures were estimated as the temperatures corresponding to the maximum of each thermogram peak. Entropy values were obtained by integrating the $\Delta C_p/T$ versus T curves (where ΔC_p is the molar heat capacity and T is the temperature in kelvin) and the Gibbs energy values were computed by the equation $\Delta G° = \Delta H° - T\Delta S°$. The thermodynamic parameters reported are the averages of at least three different heating experiments. The reported errors are the standard deviations of the mean from the multiple determinations.

References

[1] V. M. Marathias, K. Y. Wang, S. Kumar, T. Q. Pham, S. Swaminathan, P. H. Bolton J. Mol. Biol. 1996, 260, 378-94.

[2] W. L. Jorgensen, J. D. Madura, J. Am. Chem. Soc. 1983, 105, 1407-1413.

[3] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, J. Am. Chem. Soc. 1995, 117, 5179-5197.

[4] T. E. 3rd Cheatham, P. Cieplak, P. A. Kollman, J. Biomol. Struct. Dyn. 1999, 16, 845-862.

[5] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham III, C. A. Laughton, M. Orozco Biophys. J. 2007, 92, 3817-3829.

[6] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Petrucci, R. A. Broglia, M. Parrinello Comp. Phys. Comm. 2009, 180, 1961-1972.

[7] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, J. Comput. Chem. 2005, 26, 1781-1802.

[8] M. Bonomi, A. Barducci, M. Parrinello, J. Comput. Chem. 2009, 30, 1615-1621.

[9] T. L. Hwang, A. J. Shaka, J. Magn. Reson. 1995, A112, 275–279.

[10] C. Dalvit, J. Biomol. NMR 1998, 11, 437-444.

[11] M. Piotto, V. Saudek, V. J. Sklenar, J. Biomol. NMR 1992, 2, 661-665.

[12] R. Keller, The Computer Aided Resonance Assignment Tutorial (CANTINA Verlag, Goldau) (2004).

[13] L. A. Marky, K. J. Breslauer, Biopolymers 1987, 26, 1601-1620.

Supporting Information Figures

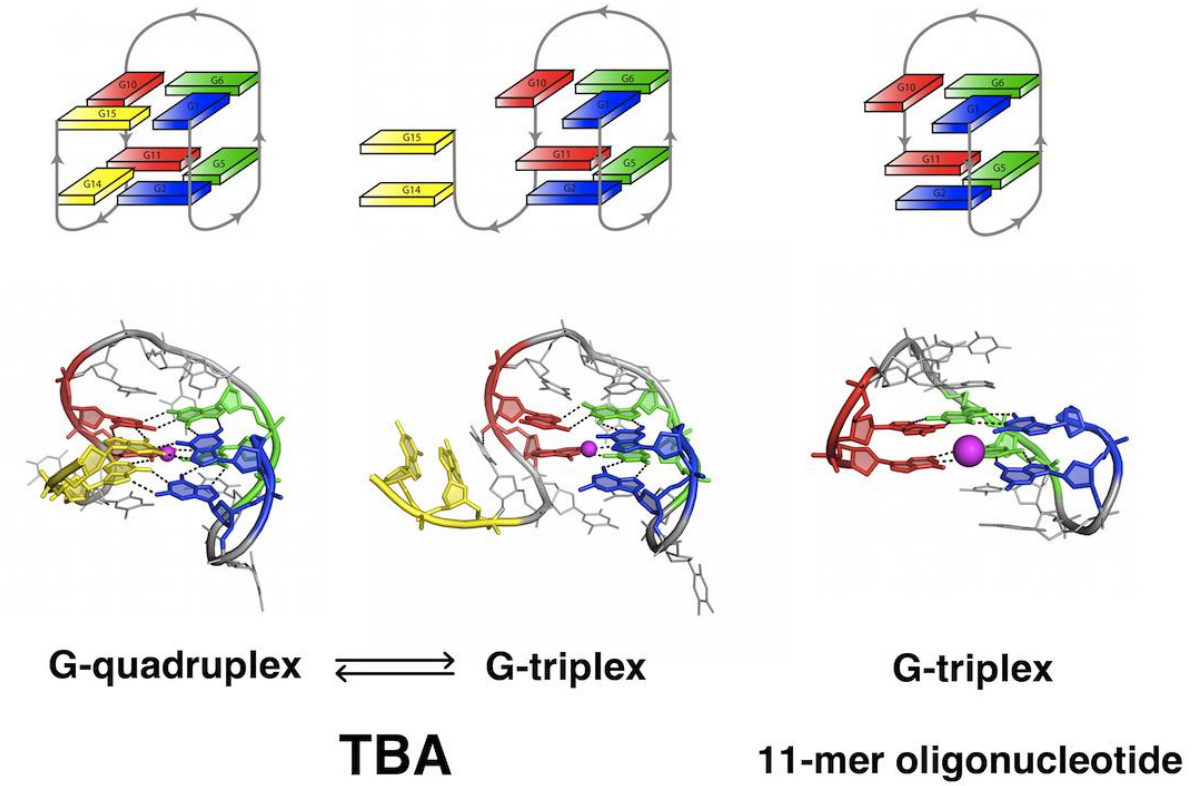


Figure S1. Schematic representation of the main results of the study. Through metadynamics calculations we have found that the DNA aptamer, TBA, is able to assume the G-quadruplex conformation (left) and a totally new structure named G-triplex (middle). This new DNA structural motif has been also found in an 11-mer oligonucleotide, called I in the main text, combining advanced computations and experiments.

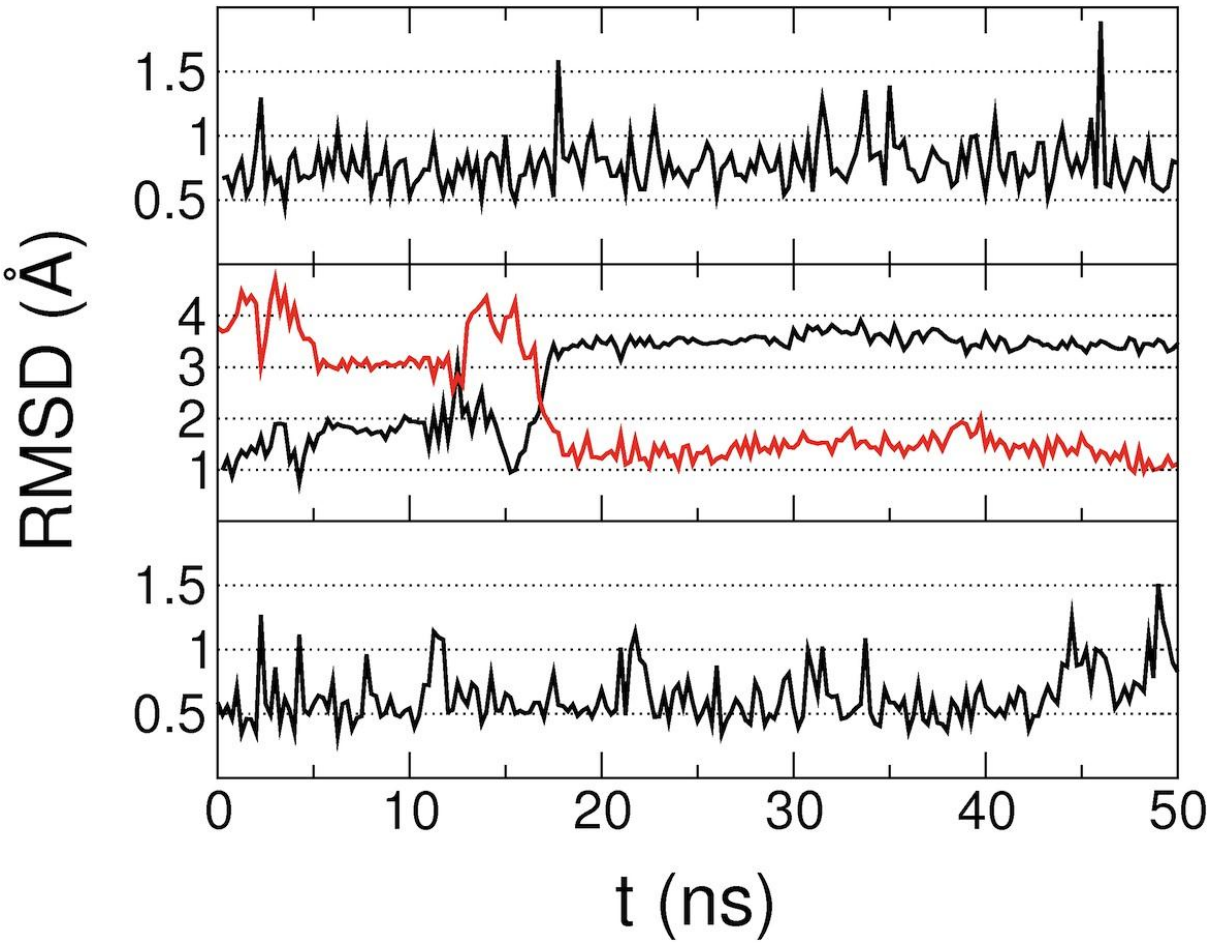


Figure S2. Plots regarding the MD simulations on basin A, B and C conformation. (Top graph) Plot of the rmsd of the heavy atoms of guanines foming the G-tetrad planes, G1:G6:G10:G15 and G2:G5:G11:G14, during over 50 ns of MD simulation. The very low rmsd values reflect the good stability of this pose. (Middle graph) Plot of the rmsd of the heavy atoms of guanines forming the G-tetrad planes, G1:G6:G10:G15 and G2:G5:G11:G14, during over 50 ns of MD simulation using the basin B conformation as starting pose. The rmsd are calculated relatively to the basin B (black lines) and basin A (red lines) conformation. The red plots clearly show that the basin B conformation transforms in the basin A one after approximately 17 ns. As discussed in the Supporting Information, this change is possible thanks to the relatively low energy barrier that separates basin B from basin A. (Lower graph) Plot of the rmsd of the heavy atoms of guanines forming the G-triad planes, G1:G6:G10 and G2:G5:G11, during over 50 ns of MD. The very low average rmsd value of 0.63 Å reflects the high stability of the G-triplex structure.

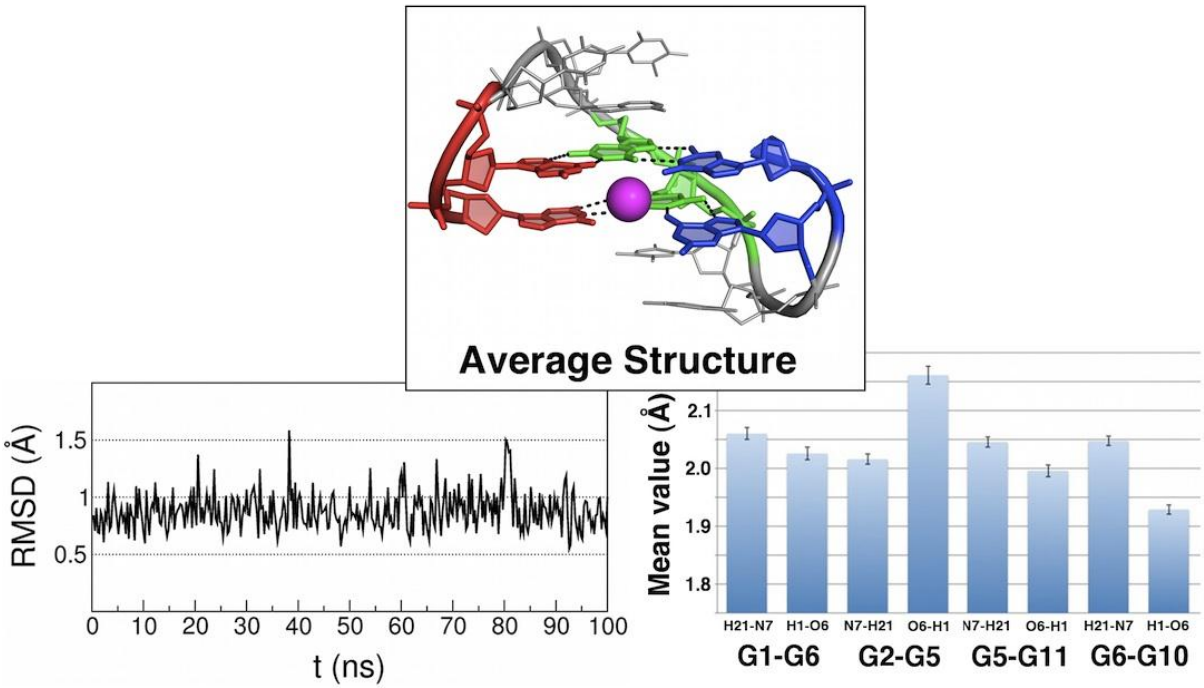


Figure S3. Plots regarding the MD simulation of the oligonucleotide I. (Left graph) Plot of the rmsd of the heavy atoms of the guanines forming the G-triad planes, namely G1:G6:G10 and G2:G5:G11, in I. The very low average rmsd value of 0.88 Å reflects the high stability of the G-triplex structure that is represented in the inset picture. (Right graph) Representation as histograms of the mean values of the distances between the hydrogen bond acceptor and hydrogen bond donor groups of the guanines forming G-triplex. The error bars show the SEM (standard error of the mean). Data for both the graphs are obtained from the over 100-ns-long MD simulation at the ionic concentration of 70 mM KCl.

Table S1. List of the atoms used to define the R_{core} and H_{core} CV.

CV type	Atoms	
R_{core}	O6 (G1); O6 (G2); O6 (G5); O6 (G6); O6 (10); O6 (G11); O6 (G14); O6 (G15)	
	H-bond acceptor	H-bond donor
H_{core}	O6 (G1)	H1 (G15)
	N7 (G1)	H21 (G15)
	O6 (G2)	H1 (G5)
	N7 (G2)	H21 (G5)
	O6 (G5)	H1 (G11)
	N7 (G5)	H21 (G11)
	O6 (G6)	H1 (G1)
	N7 (G6)	H21 (G1)
	O6 (G10)	H1 (G6)
	N7 (G10)	H21 (G6)
	O6 (G11)	H1 (G14)
	N7 (G11)	H21 (G14)
	O6 (G14)	H1 (G2)
	N7 (G14)	H21 (G2)
	O6 (G15)	H1 (G10)
	N7 (G15)	H21 (G10)

Table S2. ^{1}H-NMR assignment of I at 1°C (900 MHz).

	H8/H6	CH_3	H1'	H2'	H2"	H3'	H4'	H5'/H5"	H1/H3	H2 (H-bonded)	H2 (non H-bonded)
G1	7.40	--	5.62	2.68	1.81	4.67	3.60	3.40, 2.86	12.00	9.72	6.15
G2	8.04	--	6.16	2.44	2.75	4.92	4.41	4.80, 4.02	12.42	6.90	6.90
T3	7.09	1.65	5.04	1.30	2.04	4.41	4.29	4.23, 3.83	--	--	--
T4	7.41	1.61	6.05	2.23	2.29	4.45	3.01	2.81, 2.59	10.34	--	--
G5	7.44	6.12	6.12	3.64	2.84	4.87	4.51	4.37, 3.96	12.23	7.17	7.17
G6	8.26	--	6.13	2.66	3.00	5.02	4.57	4.25	12.51	6.91	6.91
T7	6.86	1.92	5.50	0.50	1.79	4.65	4.17	4.00, 3.99	11.76	--	--
G8	8.11	--	6.03	2.54	2.98	5.01	4.18	4.06, 4.00	10.65	--	5.90
T9	7.25	1.54	5.91	1.72	2.22	4.52	3.62	3.44, 2.54	12.42	--	--
G10	7.63	--	5.98	2.84	2.53	4.72	4.31	3.95	11.55	6.48	6.48
G11	8.02	--	5.74	2.38	2.65	4.78	4.15	4.10, 3.60	12.08	10.12	6.91

3.2 The catalytic domain of MMP-1 studied through tagged lanthanides

Ivano Bertini [a,b], Vito Calderone [a], Linda Cerofolini [a], Marco Fragai [a,b], Carlos F.G.C. Geraldes [c], Petr Hermann [d], Claudio Luchinat [a,b], Giacomo Parigi [a,b], João M.C. Teixeira [a,c]

[a] Magnetic Resonance Center (CERM), University of Florence, via Sacconi 6, 50019 Sesto Fiorentino, Italy

[b] Department of Chemistry, University of Florence, via Sacconi 6, 50019 Sesto Fiorentino, Italy

[c] Department of Life Sciences and Center of Neurosciences and Cell Biology, Faculty of Science and Technology, University of Coimbra, P.O. Box 3046, 3001-401 Coimbra, Portugal

[d] Department of Inorganic Chemistry, Faculty of Science, Univerzita Karlova (Charles University), Hlavova 2030, 12840 Prague 2, Czech Republic.

*FEBS Letters, (2012), **586**, 557-567*

FEBS Letters 586 (2012) 557–567

journal homepage: www.FEBSLetters.org

Review

The catalytic domain of MMP-1 studied through tagged lanthanides

Ivano Bertini [a,b,*], Vito Calderone [a], Linda Cerofolini [a], Marco Fragai [a,b], Carlos F.G.C. Geraldes [c], Petr Hermann [d], Claudio Luchinat [a,b], Giacomo Parigi [a,b], João M.C. Teixeira [a,c]

[a] Magnetic Resonance Center (CERM), University of Florence, via Sacconi 6, 50019 Sesto Fiorentino, Italy
[b] Department of Chemistry, University of Florence, via Sacconi 6, 50019 Sesto Fiorentino, Italy
[c] Department of Life Sciences and Center of Neurosciences and Cell Biology, Faculty of Science and Technology, University of Coimbra, P.O. Box 3046, 3001-401 Coimbra, Portugal
[d] Department of Inorganic Chemistry, Faculty of Science, Univerzita Karlova (Charles University), Hlavova 2030, 12840 Prague 2, Czech Republic

ARTICLE INFO

Article history:
Received 30 June 2011
Revised 12 September 2011
Accepted 12 September 2011
Available online 19 September 2011

Edited by Miguel Teixeira and Ricardo O. Louro

Dedicated to António V. Xavier who pioneered the use of lanthanides in NMR in the early seventies. He exploited the potential of lanthanides as NMR probes for obtaining structural restraints for the conformational analysis of mononucleotides. His results are still inspiring nowadays researches, fulfilling the prediction that "this method will be capable of determining the structural properties in solution of a large number of biophysically interesting molecules" (Barry, C.D., North, A.C.T., Glasel, J.A., Williams, R.J.P. and Xavier, A.V. (1971) Quantitative determination of mononucleotide conformations in solution using lanthanide ion shift and broadening NMR probes. Nature 232, 236). Along these lines, the structural analysis of a biomolecule in solution is here performed through the use of the paramagnetic lanthanides.

Keywords:
Paramagnetic restraint
Paramagnetic tag
Lanthanide
Matrix metalloproteinase
Residual dipolar coupling

ABSTRACT

Pseudocontact shifts (pcs) and paramagnetic residual dipolar couplings (rdc) provide structural information that can be used to assess the adequacy of a crystallographic structure to represent the solution structure of a protein. This can be done by attaching a lanthanide binding tag to the protein. There are cases in which only local rearrangements are sufficient to match the NMR data and cases where significant secondary structure or domain rearrangements from the solid state to the solution state are needed. We show that the two cases are easily distinguishable. Whereas the use of solution restraints in the latter case is described in the literature, here we deal with how to obtain a better model of the solution structure in a case (the catalytic domain of the matrix metalloproteinase MMP-1) of the former class.

1. Introduction

A strategy proposed for monitoring and possibly improving the accuracy of a protein structure in solution is that of taking a crystal structure as a starting model and to validate or "correct" it by applying few, highly informative, long range NMR restraints [1–5]. Solution NMR data which are easily accessible are residual dipolar

* Corresponding author at: Magnetic Resonance Center (CERM), University of Florence, via Sacconi 6, 50019 Sesto Fiorentino, Italy.
E-mail address: ivanobertini@cerm.unifi.it (I. Bertini).

0014-5793/$36.00
doi:10.1016/j.febslet.2011.09.020

couplings, chemical shifts, and pseudocontact shifts in paramagnetic systems. By their own, these data are hardly able to determine the protein structure, and in any case the precision of the latter would not be comparable with that of the crystal structures. However, they can be used to monitor whether they are consistent within their experimental error with a crystal model: in positive cases, the crystal structure can be assumed to be accurate also in solution. Differently, these NMR data can be used as restraints to modify the model and obtain a more accurate structure in solution.

An early example of this strategy is provided by the use of pseudocontact shifts (pcs) from a paramagnetic ion in a metalloprotein [1]. Residual dipolar couplings (rdc) originating from external orienting media have also been proposed and highly successfully used [2,3]. More recently, a strategy based on the occurrence of a metal ion binding site – either natural or artificial – and the measurement of the paramagnetism-based pcs and rdc was proposed [5]. Protein structures were also calculated by modeling on other proteins with some homology in the primary sequence and refining the model through the paramagnetism-based restraints [6]. The paramagnetism-induced NMR restraints can actually provide valuable information for the structural and dynamic characterization of proteins [5,7–10] and, being long-range solution restraints, are particularly valuable for validating and refining a protein model.

The extent of the disagreement of the NMR restraints with the available model can be checked with a novel protocol that we have here implemented, based on anchoring the protein atoms to the coordinates of the model. In this way, the structural changes possibly needed to reproduce the experimental solution data are restricted to the minimum. Two cases can be encountered: (i) a good fit of the experimental data can be obtained with minor and uniformly distributed changes in the protein structure, or (ii) the experimental data remain in disagreement with the protein structure unless the latter is allowed to have sizable global conformational changes. In the first case, it is possible to conclude that the starting model is indeed close to the protein structure in solution. This information is highly relevant because even when the model is a crystal structure solved at high resolution, in many cases structural differences can arise in solution due to the absence of crystal packing forces and/or the possible presence of inhibitors bound to the protein. In the second case, i.e., when sizable global changes, like those involving secondary structural element rearrangements, are needed for obtaining a structure in agreement with the NMR data, different previously described approaches [2,4,5] should be applied. They are based on the use of force fields to keep the structure as close as possible to the original model, without restricting the possibility of sizable rearrangements under the driving force of the NMR restraints. Of course, the accuracy of the structure, as dictated by a few NMR restraints, is limited, but in any case better than that of the solid state model.

In this manuscript we show that the proposed protocol can discriminate between the two cases, by using the catalytic domain of the matrix metalloproteinase MMP-1 [11] and the protein calmodulin as representative examples. The structural analysis of the catalytic domain of MMP-1 was performed by solving the crystal structure of the protein, and by comparing the latter to the several structures deposited in the PDB with and without inhibitors. The paramagnetic NMR restraints (pcs and rdc) were acquired after the binding of a rigid lanthanide binding tag [12–21]. A number of paramagnetic lanthanide tags have been designed in the last years, to be rigidly attached to proteins, in order to take advantage of paramagnetism-assisted NMR for the structural and dynamic characterization of proteins and protein complexes. CLaNP-5 represents one of the last advances, showing no mobility nor multiple conformations [13,20,22]. This tag has been previously used for the study of the conformational heterogeneity of the complex formed by adrenodoxin reductase and its electron receiving partner adrenodoxin [23], and by cytochrome c and adrenodoxin [17]. Pcs and rdc resulted in agreement with the crystallographic model of the catalytic domain of MMP-1 after minor changes in the protein structure. On the opposite, in the case of calmodulin, we show that the same protocol can detect the presence of sizable structural changes occurring on passing from solid state to solution, when secondary structural elements assume different relative orientations [4,5].

2. Materials and methods

2.1. Sample preparation

2.1.1. CLaNP-5 (1,4,7,10-tetraazacyclododecane-1,7-[bis(N-oxido-pyridine-2-yl)methyl]-4,10-diyl-bis(2-(acetylamino)ethylmethane-sulfonothioate)) synthesis

The tag was synthesized as previously described [22,24], except that the last step consisted of functionalizing the compound precursor, 1,4,7,10-tetraazacyclododecane-1,7-[bis(N-oxido-pyridine-2-yl)methyl]-4,10-diacetic acid with (2-aminoethyl) methanethiosulfonate (MTS). The functionalization was achieved through reaction of the precursor with 2.3 equivalents MTS in the presence of 2 equivalents DMAP (4-dimethylaminopyridine), 1 equivalent HOBT (1-hydroxybenzotriazole), 4 equivalents TBTU O-(benzotriazol-1-yl)-*N,N,N′,N′*-tetramethyluronium tetrafluoroborate), and 5–6 equivalents DIPEA (*N*-ethyldiisopropylamin), and by stirring overnight at room temperature in dried acetonitrile. A HPLC purification step was then performed by using a Phenomenex C8 semi-preparative column (0.1% TFA, gradient 10–20% MeCN in H_2O). After lyophilisation of pure fraction (confirmed by analytical HPLC), a thick, slightly green colored solid was obtained, and the tag (CLaNP-5) stored at −20 °C.

2.1.2. Lanthanide binding to CLaNP-5

The tag was dissolved in DMF, incubated with an excess of a lanthanide(III) acetate salt (Ln^{3+} = Lu^{3+}, Yb^{3+}, Tm^{3+} or Tb^{3+}) at 32 °C, and shaken for 3 days. The unsoluble excess of lanthanide(III) salt was removed by centrifugation.

2.1.3. Protein expression and purification

The cDNA encoding sequence (Asn-106–Gly- 261) of the catalytic domain of MMP-1 (CAT) was generated by the polymerase chain reaction (PCR), using the inactive full length MMP-1 gene (E219A) as template [11], and was cloned into pET21a by using NdeI and XhoI as restriction enzymes. The double mutation H132C/K136C, performed for the attachment of the CLaNP-5 tag, was engineered during a single PCR step using the QuickChange Site Directed Mutagenesis Kit (Stratagene): 5′-GCC AAG AGC AGA TGT GGA CTG TGC CAT TGA GTG TGC CTT CCA ACT CTG GAG-3′; 5′-CTC CAG AGT TGG AAG GCA CAC TCA ATG GCA CAG TCC ACA TCT GCT CTT GGC-3′. The mutations were confirmed by nucleotide sequencing. The expression vector was inserted into competent *Escherichia coli* BL21(DE3) CodonPlus RIPL strain cells, and the colonies were selected for ampicillin and chloramphenicol resistance. Single-labelled ^{15}N and doubly-labeled $^{13}C/^{15}N$ protein were expressed using minimal medium containing ^{15}N-enriched $(NH_4)_2SO_4$ and ^{13}C-enriched glucose (Cambridge Isotope Laboratories). Cell growth occurred at 37 °C, with induction at 0.6 O.D. with 500 μM of IPTG and harvested after 5 h expression. CAT MMP-1, precipitated as inclusion bodies, was solubilized, after lysis of the cells, in a solution of 8 M urea, 20 mM dithiothreitol, and 20 mM Sigma–Aldrich Trizma-base (pH 8), and stored at −20 °C. The refolding of CAT MMP-1 involved decreasing the urea gradient at 4 °C. The desired amount of protein was diluted into a 500 ml solution containing 6 M urea, 50 mM Trizma-base, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 20 mM Cysteamine, at pH 8.0. The solution was then dialyzed against 1) 4 l of 4 M urea, 50 mM Trizma-base pH 8.0, 10 mM $CaCl_2$

and 0.1 mM $ZnCl_2$; 2) 4 l of 2 M urea, 50 mM Trizma-base pH 7.2, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 0.3 M NaCl; 3) three steps of 20 mM Trizma-base pH 7.2, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 0.3 M NaCl. The resulting 500 ml protein sample was concentrated down to 100 ml using MiniKros Modules (Spectrumlabs). CAT MMP-1 was purified by using HiLoad 26/60 Superdex 75 pg (Amersham Biosciences). Protein pure stocks were stored at 4 °C.

2.1.4. CLaNP-5 attachment

An aliquot of 5 ml of 10 μM protein was dialyzed against anaerobic buffer (20 mM Trizma-base pH 7.2, 10 mM $CaCl_2$ and 0.1 mM $ZnCl_2$, 0.3 M NaCl) under anaerobic conditions. The protein was then treated with 20 mM of Dithiothreitol (DTT) to reduce the engineered cysteines. Still in anaerobic conditions the DTT was removed by using PD10 desalting columns, and the protein was concentrated down to 1 ml. 10 equivalents of tag CLaNP-5 were added to the sample; reaction proceeded overnight at 4 °C (20 mM Trizma-base pH 7.2, 10 mM $CaCl_2$ and 0.1 mM $ZnCl_2$, 0.3 M NaCl). Precipitation was observed after overnight incubation. Final buffer conditions were obtained by washing the sample with PD10 desalting columns against 20 mM Trizma-base pH 7.2, 10 mM $CaCl_2$ and 0.1 mM $ZnCl_2$, 0.15 M NaCl. The final yield was around 60% due to protein precipitation. The NMR spectra showed that the protein is fully reacted with the metal complex.

2.1.5. Protein preparation for crystal structure determination

The wild type CAT MMP-1 was prepared as previously described [25].

2.2. Protein crystallization and structure determination

The protein samples were concentrated to 0.7 mM. Crystals of CAT MMP-1 were obtained under aerobic conditions by using the vapour diffusion technique at 289 K from solutions containing 0.1 M Tris–HCl pH 8.5 and 30% PEG 8000. The dataset was collected by using synchrotron radiation at DESY (EMBL, Hamburg) on beamline BW7A, at 100 K; the crystal used for data collection was cryo-cooled by using 20% ethylene glycol in the mother liquor.

The data were processed as monoclinic C2 by using the program MOSFLM [26] and scaled by using the program SCALA [27] with the TAILS and SECONDARY corrections on (the latter restrained with a TIE SURFACE command), to achieve an empirical absorption correction.

Table 1 shows the data collection and processing statistics for all datasets. The structure was solved by using the molecular replacement technique; the structure of collagenase-1 (PDB 966C) was used as a model, where water molecules and ions were omitted. The correct orientation and translation of the three molecules present in the asymmetric unit were determined with standard Patterson search techniques [28,29] as implemented in the program MOLREP [30,31]. The isotropic refinement was carried out by using REFMAC5 [32]. REFMAC5 default weights for the crystallographic and the geometrical term have been used in all cases.

In between the refinement cycles the models were subjected to manual rebuilding by using XtalView [33]. Water molecules were added by using the standard procedures within the ARP/WARP suite [34]. The stereochemical quality of the refined model was assessed by using the program Procheck [35].

2.3. NMR measurements

All experiments were performed on samples of a mutant (E219A, H132C, K136C) of CAT MMP-1 functionalized with the tag CLaNP-5 coordinated to a lanthanide ion (Lu^{3+}, Yb^{3+}, Tm^{3+} or Tb^{3+}), at concentrations ranging between 0.15 and 0.2 mM (20 mM Tris, pH = 7.2, 0.15 M NaCl, 0.1 mM $ZnCl_2$, 10 mM $CaCl_2$).

Table 1
Data collection and refinement statistics.

Space group	C2
Cell dimensions (Å, °)	a = 147.69, b = 54.53, c = 94.90, β = 120.69°
Resolution (Å)	51.4–2.2
Unique reflections	32923 (4785)[a]
Overall completeness (%)	99.2 (99.2)
R_{sym} (%)	6.6 (49.1)
Multiplicity	3.0 (3.0)
$I/(\sigma I)$	9.1 (1.7)
Wilson plot B-factor (Å^2)	40.10
R_{cryst}/R_{free} (%)	22.4/28.8
Protein atoms	3730
Water molecules	196
Ions	15
RMSD bond lengths (Å)	0.020
RMSD bond angles (°)	2.40
Mean B-factor (Å^2)	36.34

[a] Numbers in parenthesis refer to high resolution shells.

All NMR experiments were performed at 310 K and acquired on Bruker AVANCE 700 and DRX 500 spectrometers, equipped with triple resonance cryo-probes. All spectra were processed with the Bruker TOPSPIN software packages and analyzed by the program CARA (Computer Aided Resonance Assignment, ETH Zurich) [36]. ^1H-^{15}N HSQC spectra were recorded at 500 MHz. The assignment of the protein functionalized with Lu-CLaNP-5 was obtained by the comparison of the ^1H-^{15}N HSQC spectrum with the assignment reported on BMRB [37] and the analysis of the 3D HNCA experiment performed at 700 MHz. The ^1H-^{15}N HSQC spectrum of Yb-CLaNP-5-CAT MMP-1 was assigned with the help of the 3D HNCA and CBCA(CO)NH experiments performed at 500 MHz. The ^1H-^{15}N HSQC of Tm-CLaNP-5- and Tb-CLaNP-5-CAT MMP-1 were assigned by using the other assigned spectra and pseudocontact shift predictions.

^1H-^{15}N residual dipolar couplings were measured at 310 K and 700 MHz for the CAT MMP-1 functionalized with the Tm-CLaNP-5 and Tb-CLaNP-5 tags by using the IPAP method [38]. In the case of the Yb-CLaNP-5 tag, due to the smaller magnetic susceptibility anisotropy of the metal, the measurements were performed on an Avance 900 MHz Bruker spectrometer to achieve a higher alignment of the protein, and thus larger rdc values.

2.4. Paramagnetism-based restraints

The electron-nucleus dipolar coupling does not average to zero upon rotation in the presence of anisotropy in the paramagnetic susceptibility tensor. A contribution to the hyperfine shift, which is called *pseudocontact shift* (pcs) thus arises, which is described by Eq. (1) [39]

$$pcs = \frac{1}{12\pi r^3}\left[\Delta\chi_{ax}(3\cos^2\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\theta\cos 2\varphi\right] \quad (1)$$

where r is the distance between observed nuclei and metal ion, $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ are the axial and rhombic anisotropy parameters of the magnetic susceptibility tensor of the metal, and θ and φ are the spherical angles defining the position of the nucleus in the frame of the paramagnetic susceptibility tensor. Therefore, pcs values depend only on the position of the nuclei with respect to the frame defined by the magnetic susceptibility tensor, with origin on the metal ion, and on the anisotropy values.

Rdc due to partial self-orientation of the paramagnetic protein in the magnetic field is described by Eq. (2) [39,40]

$$rdc\,(Hz) = -\frac{S_{LS}}{4\pi}\frac{B_0^2}{15kT} \times \frac{\gamma_N\gamma_H\hbar}{2\pi r_{HN}^3}\left[\Delta\chi_{ax}(3\cos^2\alpha - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\alpha\cos 2\beta\right] \quad (2)$$

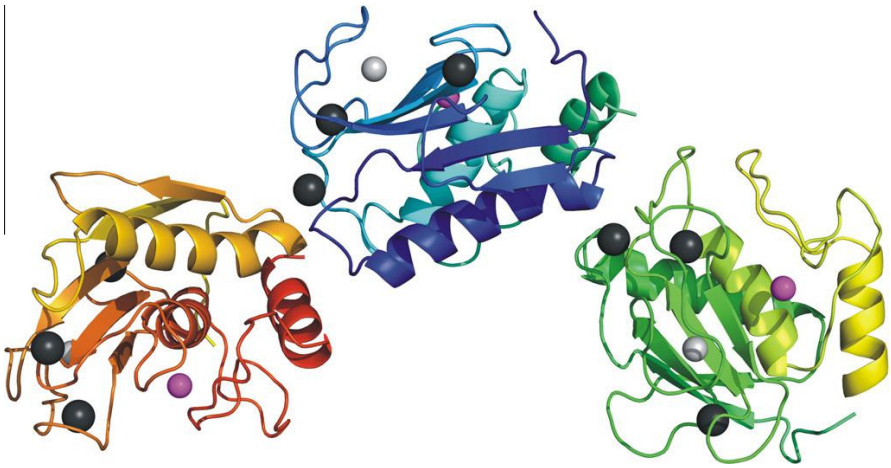


Fig. 1. Crystal structure of the catalytic domain of MMP-1. Structural zinc ions are in light grey, catalytic zinc ions are in magenta and calcium ions are in dark grey.

where r_{HN} is the distance between the two coupled nuclei N and ^{N}H (set equal to 1.02 Å) and the spherical angles α and β define the orientation of the vector connecting the coupled nuclei in the frame of the magnetic susceptibility tensor. S_{LS} is the model-free order parameter, introduced to take into account some local mobility of the nitrogen-amide proton vector. Other symbols have the usual meaning. Therefore, rdc values depend on the orientation of the vector connecting the coupled nuclei in the reference frame of the magnetic susceptibility tensor axes, on the values of the anisotropies of the latter, on the applied magnetic field and on the gyromagnetic ratio of the coupled nuclei [41–44]. They are not related at all to the position of the coupled nuclei with respect to both the metal ion and the magnetic susceptibility tensor.

3. Results and discussion

3.1. Crystallographic structure

The structure of the catalytic domain of wild type MMP-1 was solved at 2.2 Å resolution, and deposited in the PDB with code 3SHI. The space group is C2 and three molecules are present in the asymmetric unit (Fig. 1). The backbone RMSD between the different protein molecules in the unit cell is 0.25 Å. There are no regions showing significant lack of density involving the main chain of the three molecules in the asymmetric unit. The Ramachandran plot is of good quality (90.8% core, 8.4% allowed, 0.8% generously allowed and 0.0% disallowed residues). CAT MMP-1 is a "spherical" molecule that contains a twisted five-stranded β-sheet (I, II, III, IV and V) and three α-helices (A, B and C). The β-sheet contains four parallel strands and one antiparallel strand. The active site cleft is bordered by β-strand IV, helix B, and a stretch of random coil adjacent to the COOH terminus of helix B. The catalytic zinc is at the bottom of the cleft and is ligated by His 218, His 222, and His 228. In addition to the catalytic zinc, there is a second zinc ion that interacts with an extended loop between β-strand III and IV, and also three calcium atoms. Although the model used for molecular replacement contained all residues in trans conformation, the refinement provided two residues (Tyr A210 and Arg G108) in cis conformation.

The calculated structure 3SHI resulted similar to both the 1CGE crystal structure of the same protein without inhibitor already deposited in the PDB (the BB RMSD between the two structures is 0.44 Å in the residue range 108-261) and to the crystal structure at highest resolution (1HFC, resolution 1.56 Å), crystallized with a bulk hydroxamate inhibitor (with BB RMSD with 3SHI of 0.38 Å). The space group of these two structures is different from the C2 space group of 3SHI, as a result of different conditions of crystallization. The BB RMSD of 3SHI with the available solution structure (2AYK) is 1.34 Å. Also in 1HFC the residue Tyr A210 is in cis conformation, whereas in the other two structures all residues are in trans conformation.

In 1HFC, the space group is $P2_12_12_1$, with cell parameters different from those obtained in 3SHI. There are two small regions with significantly high RMSD between the 3SHI and the 1HFC structures (Fig. 2): the first one involves residues 188–190 and shows an RMSD of about 1 Å and the second one involves residues 242–245 with an RMSD which reaches 2.3 Å for residue 243 only. Both regions belong to protein loops; the second region belongs to the long loop forming the S_1' cavity. The regions which might be affected by the binding of the inhibitor in 1HFC are 178–182 and 238–240. These residues are not those with high RMSD, although they are adjacent. The high RMSD values can be determined by the presence of the inhibitor in 1HFC and by the different crystal contacts with symmetry related molecules present in the two space groups.

In the ligand-free 1CGE structure, the space group is $P4_12_12$, with one molecule in the asymmetric unit and different cell parameters with respect to the other mentioned structures. The superposition between the 3SHI and the 1CGE structures shows that residues 155–156 deviate by 1–2 Å, residues 243–244 deviate by 1–1.5 Å and residues 208 and 238 by slightly more than 1 Å. All other deviations are well below 1 Å. It is worth noticing that the region involving residues 178–182 does not show significant deviations (BB RMSD 0.3–0.4 Å), and that the deviations in the region 242–245 are much smaller than for the ligand-bound 1HFC structure.

3.2. Translating the protein model into a CYANA structure

Pcs and rdc provide structural information because they depend on the position of the observed nuclei and on the orientation of the vectors connecting coupled nuclei, respectively, in the frame defined by the paramagnetic susceptibility anisotropy tensor (see Eqs. (1) and (2)) [45]. They can thus be used for assessing whether an available structure is in agreement with these data, and possibly for calculating a solution structure in better agreement.

In order to estimate the extent of the structural changes needed to reproduce the paramagnetism-based restraints, the initial model was first adapted to fulfill all chemical bond constraints (in

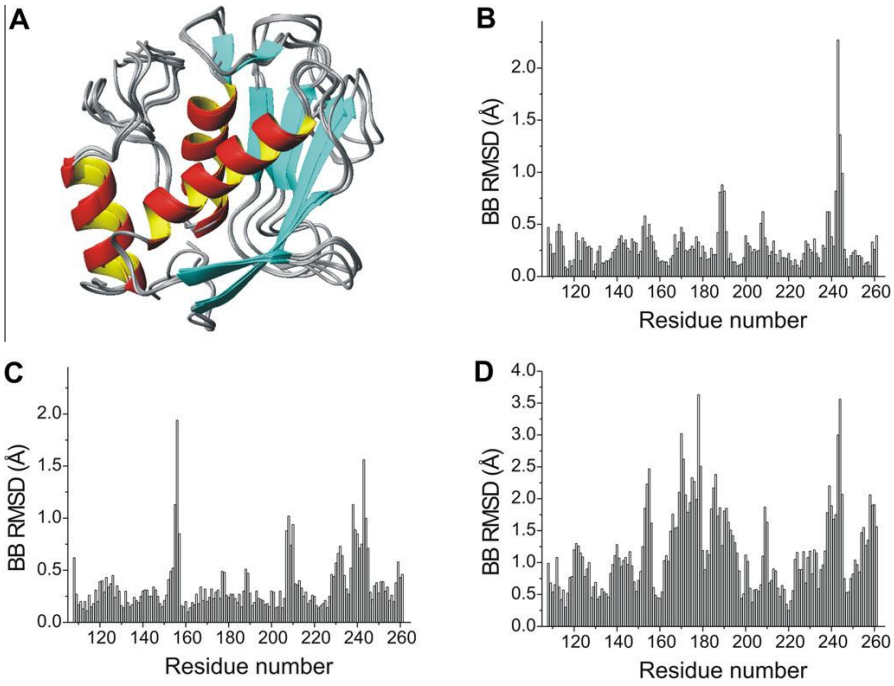


Fig. 2. (A) Superimposition of the four analyzed structures (the crystal structure 3SHI, here calculated, and the 1HFC, 1CGE and 2AYK structures). The bar plots represent the RMSD between the 3SHI and the (B) 1HFC, (C) 1CGE and (D) 2AYK structures.

terms of bond angles and lengths) of the library of the program CYANA [46]. This was done by including in the protein sequence a "pseudoprotein residue", composed by as many pseudoatoms as the number of atoms of the protein, each of them labeled according to its residue number and atom name. These pseudoatoms have coordinates equal to the coordinates of the corresponding atoms in the model structure and no van der Waals radius. The pseudoprotein residue was linked to the protein sequence through dummy residues, which have the function of allowing the pseudoprotein residue to freely move with respect to the protein residues. A simulated annealing calculation was performed with CYANA with upper distance limits of 0.1 Å (with weight 0.1) between all the heteroatoms of the protein and the corresponding atoms of the pseudoprotein residue. The dihedral ϕ and ψ angles were also restrained to vary within ±90° around the value in the model structure. A further conjugate gradient minimization was then performed with the same restraints and with the weight of the upper distance limits reduced to 0.01. In this way, the protein atoms are positioned as close as possible to the starting structure, being at the same time constrained to the bond lengths and angles defined in the internal library. This determined a slight rearrangement with respect to the four protein structures of CAT MMP-1 (3SHI, 1HFC, 2AYK, 1CGE), with backbone RMSD values to the starting structures of 0.25–0.30 Å.

3.3. Experimental NMR restraints

Pcs of ^{N}H nuclei and rdc of the ^{N}H–N pairs for the three paramagnetic Ln-CLaNP-5-CAT MMP-1 (Ln = Yb, Tm, Tb) forms were measured. Pcs restraints were first introduced into the CYANA calculation by using the routines dealing with the paramagnetic restraints available in PARAMAGNETIC CYANA [47]. They provided the metal position and the orientations of the magnetic anisotropy susceptibility tensors of the three metals. The metal ions were found at distances of 7.5 and 8 Å from the C^{α} nuclei of the tag-binding residues 132 and 136, respectively, in agreement with expectations based on previous results obtained with the same tag [22]. The magnitude of the anisotropy tensors (i.e., the $\Delta\chi ax$ and $\Delta\chi rh$ parameters) were obtained by cycling between PARAMAGNETIC CYANA and a minimization program performing a fit of the pcs data to the metal-containing protein structure. Pcs are indeed quite robust for providing the magnetic susceptibility anisotropy tensors (reported in Table 2), because they are not very sensitive to small local protein conformational changes, when the tag is rigidly bound to the protein (see later). The agreement between experimental and back-calculated pcs values is quite satisfactory for all four considered structures (see Fig. 3) with Q factors of 0.07–0.09. The fact that all pcs of the three metals agree simultaneously with a single protein structure, and that the anisotropy values are as high as expected for these metals, is an indication of both small internal mobility and rigidity of the attached tag. In fact, if the tag were largely mobile with respect to the protein, the metal-nucleus distances would average differently for the different nuclei, depending on their motion with respect to the magnetic susceptibility anisotropy tensors, and no average tensors could in principle be defined from the pcs values [5,48].

The agreement between experimental rdc values and rdc calculated with the pcs-derived tensors and either the initial structures or the structure calculated with CYANA was however unsatisfactory (Q factor = 0.72 for 3SHI). This may be due to the local mobility of each residue, to extensive mobility of the tag (which seems however excluded by the pcs, as discussed above), or to inaccuracies of the bond vector orientations in the available structure. A good estimation of the amide proton mobility can be obtained from ^{15}N relaxation rate measurements. R_1, R_2 and NOE measurements for the catalytic domain of MMP-1 are available [11] and actually indicate a sizable mobility for residues 108, 111, 116, 134, 135, 137, 145, 154, 157, 184, 191, 217, 227, 244, 245, 246, 249, 250, 260,

Table 2
Pcs-derived magnetic susceptibility anisotropy values for the three lanthanides and for the 3SHI, 1HFC, 1CGE and 2AYK structures, respectively.

	Yb^{3+}	Tm^{3+}	Tb^{3+}
$\Delta\chi ax$ (10^{32} m^3)	8.6, 8.0, 8.2 and 9.0	49.3, 47.4, 45.9 and 50.1	−44.9, −45.3, 43.4 and −44.6
$\Delta\chi rh$ (10^{32} m^3)	−2.1, −2.2, −2.8 and −3.0	−8.7, −8.6, −8.0 and −8.7	−20.1, −14.8, −14.5 and −11.5

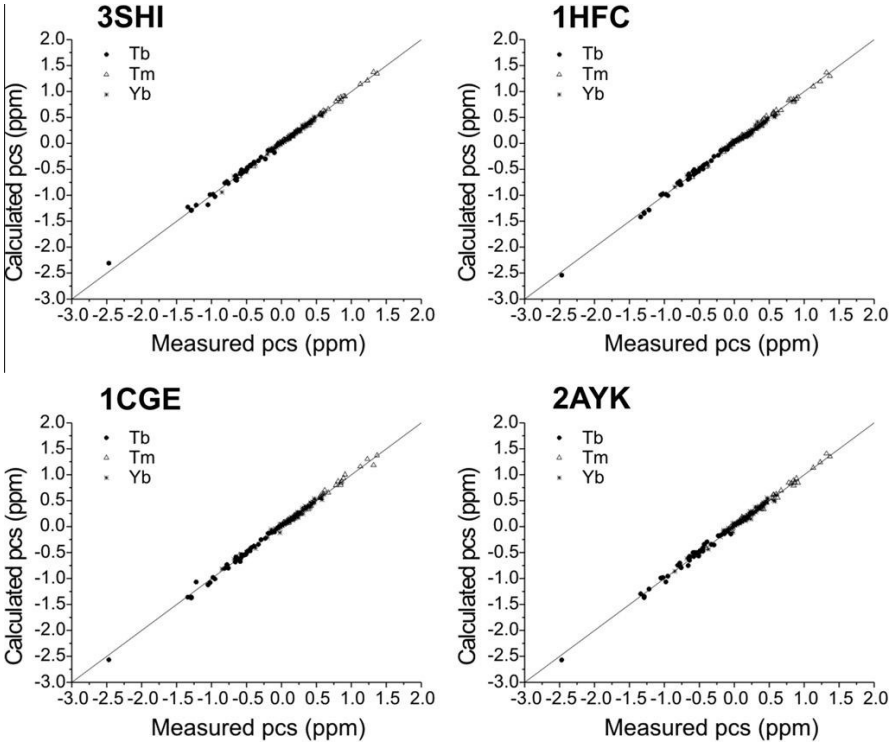


Fig. 3. Agreement between measured and calculated pcs obtained by using the magnetic susceptibility anisotropy parameters reported in Table 2 and the structures determined with PARAMAGNETIC CYANA, after the introduction of pcs and the upper distance limits for anchoring the protein nuclei to the nuclear coordinates of the 3SHI, 1HFC, 1CGE and 2AYK structures.

261. The rdc of these residues were thus excluded from all subsequent calculations. The *Q* factor of the remaining rdc is 0.67, 0.52, 0.69 and 0.83 for the 3SHI, 1HFC, 1CGE and 2AYK structures, respectively (Fig. 4A). The presence of a small mobility for the other residues can then be taken into account by including an order parameter S_{LS} (see Eq. (2)). The lowest *Q* factor was obtained for the rdc of the 1HFC structure, suggesting that this structure represents the best model in solution.

The rdc of non-mobile residues were fit to the 1HFC and 3SHI structures (see Fig. 4B) according to Eq. (2), with an order parameter S_{LS} fixed to 0.9 (fits of similar quality are obtained also with $S_{LS} = 1$) for all the values. The $\Delta\chi ax/\Delta\chi rh$ values were 7.1/−0.8, 39.7/−1.5, and −39.4/−16.7 × 10^{-32} m^3 for Yb^{3+}, Tm^{3+} and Tb^{3+}, respectively, for the 1HFC structure, and 7.0/−1.2, 36.6/−10.2, and −38.7/−10.4 × 10^{-32} m^3 for the 3SHI structure. The similar size of pcs- and rdc-derived tensors indicates that there is not a sizable global motion of the metal-bearing tag with respect to the protein, pointing out to its rigidity. The disagreement between calculated and experimental rdc is however outside the error (2 Hz) for many residues. This suggests that the H-^{H}N vectors of these residues experience a somewhat different conformation in solution with respect to these solid state structural models, which needs to be quantified. The rdc-derived tensors are actually expected to be somewhat reduced with respect to the pcs-derived tensors in the presence of inaccuracies in the relative orientation of the individual H–^{H}N vectors.

Therefore, rdc can be used for checking the consistency of the crystal model in solution by using the $\Delta\chi ax/\Delta\chi rh$ values determined from the pcs data. Remarkably, much worse fits of the rdc data were calculated for the crystal 1CGE structure as well as for the solution 2AYK structure (see Fig. 4B), although these two structures were obtained in the absence of inhibitors, again pointing out that they are less accurate models for the protein.

3.4. The protein solution structure

All pcs and rdc data (except the rdc of residues affected by large mobility, as shown by relaxation measurements) were introduced as restraints in the assumption that a unique tensor for any metal, and precisely that calculated from the pcs [5,8,39,49,50], is responsible for all the observed pcs and rdc values, as occurring in the absence of motion. The usual local mobility of H–^{H}N vectors was considered by using an order parameter S_{LS} of 0.9 for rdc. The weights of the restraints were 0.20 for Yb-rdc, 0.04 for the larger Tm- and Tb-rdc and 100 for pcs, which are much smaller in absolute value. Upper distance limits between the protein heteronuclei

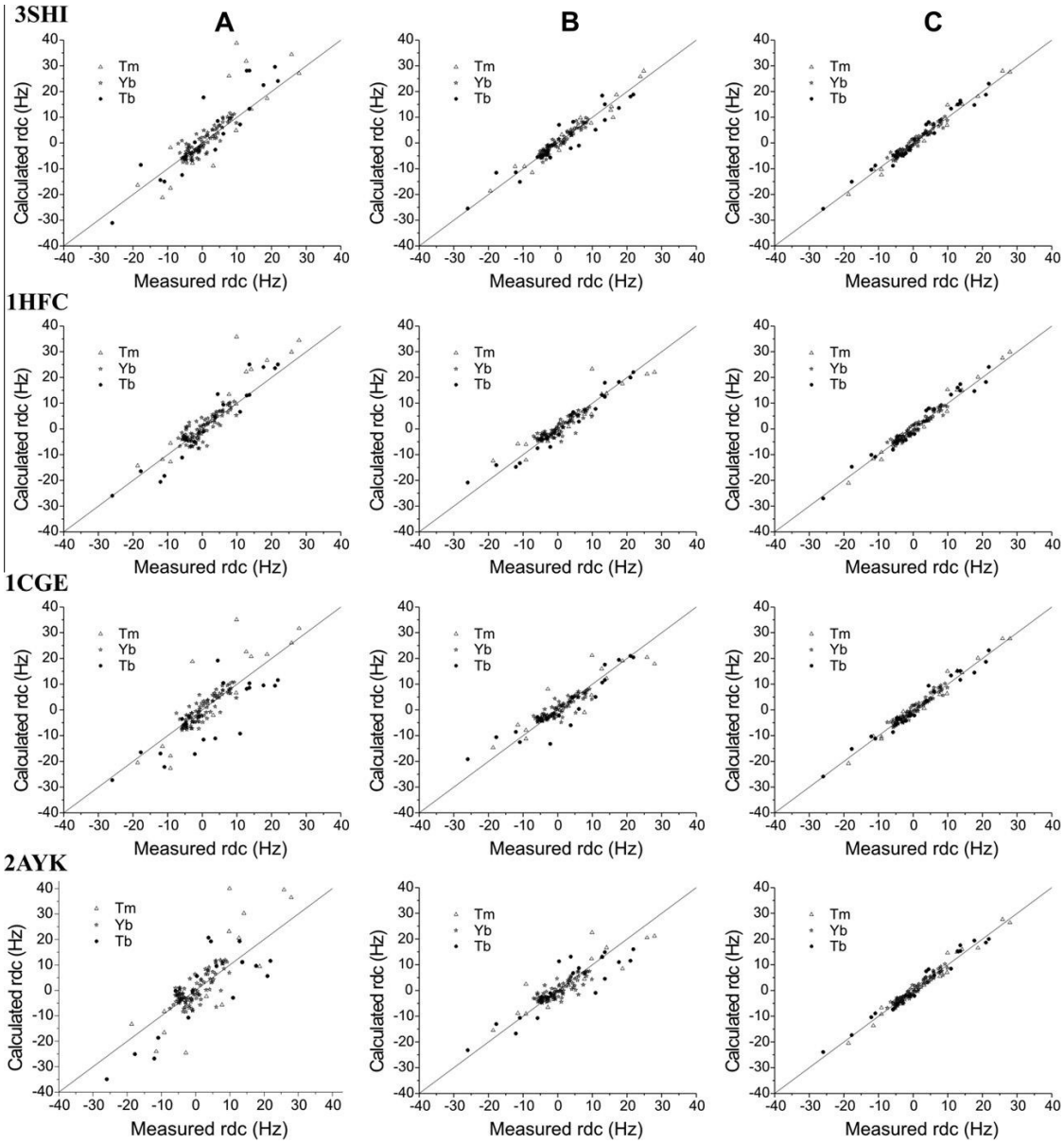


Fig. 4. (A) Agreement between experimental rdc and rdc calculated from the pcs-derived tensors and the 3SHI, 1HFC, 1CGE and 2AYK structures. (B) Best fit of the experimental rdc values to the 3SHI, 1HFC, 1CGE and 2AYK structures. (C) Agreement between experimental rdc and rdc calculated from the pcs-derived tensors and the solution structures.

and the pseudoatoms of the pseudoprotein residue were also included in the same way as described for the calculations in the absence of the paramagnetic restraints, for anchoring the position of the protein atoms to the coordinates of the selected model. A simulated annealing followed by a conjugated gradient minimization were performed with PARAMAGNETIC CYANA.

Because the 1HFC structure provides the best agreement with the paramagnetic data, as seen before, the calculation was first performed using this model. The resulting structure is shown in Fig. 5 together with the initial 1HFC crystal structure; the backbone RMSD between the two is 0.29 Å. The Ramachandran plot of the solution structure is still of very good quality (90.1% core, 9.9% allowed, 0.0% generously allowed and disallowed residues), without van der Waals contact violations, and with the bond length and angle parameters fixed to the library values of CYANA. Fig. 4C shows the good agreement between calculated and observed rdc values. The rdc Q factor decreases to 0.19 from 0.52.

Calculations were also performed using the 1CGE and 3SHI structures as models. For the structures so obtained, the agreement of the data is as good as that obtained for the 1HFC structure ($Q = 0.20$, versus 0.67 and 0.69), and the RMSD between the crystal and the corresponding solution structures is again lower than 0.4 Å.

The fact that single structures so similar to the crystal models are in very good agreement with the experimental data indicates that there is no need to invoke ensemble averaging approaches

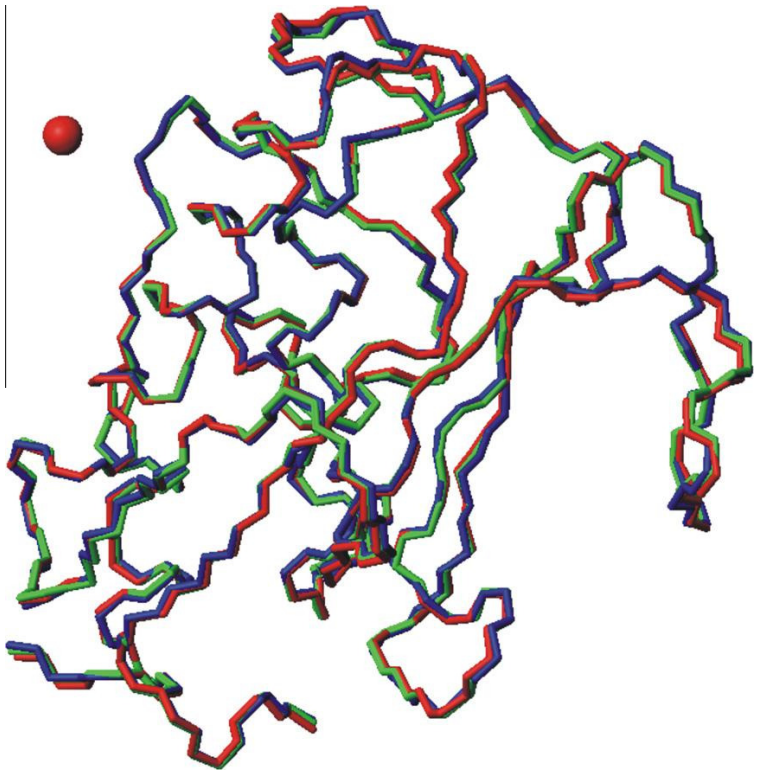


Fig. 5. Solution structure (red) superimposed to the crystal 1HFC structure (blue) and to the CYANA structure (green) calculated before the inclusion of the paramagnetic restraints. The red sphere shows the position of the paramagnetic metals.

[51,52], although a slight structural heterogeneity may likely be present. In fact, since all these solution models nicely satisfy the experimental restraints, any ensemble composed of these structures will also satisfy all restraints, as well as many other ensembles not containing any of them. However, it is clear that no physical meaning can be given to such ensembles. The concept that the number of structures needed to represent a molecule should be restricted to the minimum required by the experimental data is referred as an application of the Occam's razor [53].

The improvement in the accuracy of the solution structures can be monitored by comparing the dihedral angles of the structures before and after the inclusion of the paramagnetic restraints. Fig. 6 shows the differences in the ϕ dihedral angles among the three crystal structures (1HFC, 1CGE and 3SHI) as well as the differences with the corresponding solution structures. It can be seen that there are small but significant changes in these angles, which were necessary to adjust the rdc values to the experimental ones. These changes are in many cases consistently outside the range of the values of the crystal structures. This indicates that the paramagnetic data modify the ϕ dihedral angles in a way that is meaningful. The changes in the values of the ψ dihedral angles are minor and often within the variability of the different structures, as a result of the smaller effect of H–^{H}N vectors on these angles. The average difference of the ϕ dihedral angles between the starting structures and the corresponding solution ones is 9.1°, 10.8° and 10.8° for the 1HFC, 3SHI and 1CGE structures, respectively. As expected, the 1HFC structure, which was in best agreement with the rdc data, varied less to be in agreement with the experimental data.

The improvement in the accuracy of the solution structures was also verified by cross-validation with data not used in the calculations. In fact, if the latter are performed by removing the rdc measured for one metal of the residues for which rdc have been measured also for the other two metals (5 rdc values in total), the Q_{free} of these rdc is still small (0.23 versus 0.34 in the crystal structure). A similar decrease in the Q_{free} is observed when also the rdc measured for the residues of at least another metal, belonging to secondary structural elements (8 rdc values in total), are removed.

In conclusion, the paramagnetism-based restraints indicate that the solution structure of the catalytic domain of MMP-1 is in good agreement with the crystal structure and provide a tool for an improvement in solution of the orientation of the vectors for which rdc have been measured.

3.5. Proteins with solution structures different from the crystal structures

In the case of calmodulin, it was previously shown that the solution structure of the N-terminal domain differs from the crystal structure [4]. A crystal structure at high resolution (1 Å) is available (1EXR), as well as the rdc for H–^{H}N, H$^{\alpha}$–C$^{\alpha}$, C′–N, C$^{\alpha}$–C′, H$^{\alpha}$–C′ nuclear pairs, measured in liquid crystalline medium. The protocol described in the previous sections was applied to this system to check whether it correctly indicates that the NMR restraints are incompatible with the crystal structure unless large conformational changes are allowed. The best fit of the experimental rdc to the 1EXR structure is indeed quite unsatisfactory (Fig. 7A) and it does not improve sizably when the rdc data are introduced into the CYANA calculations (Fig. 7B): the Q factor, equal to 0.39 for the crystal structure, remains as high as 0.32 after the restrained minimization. Therefore, the approach itself indicates that the rdc are not compatible with any slightly modified crystal structure.

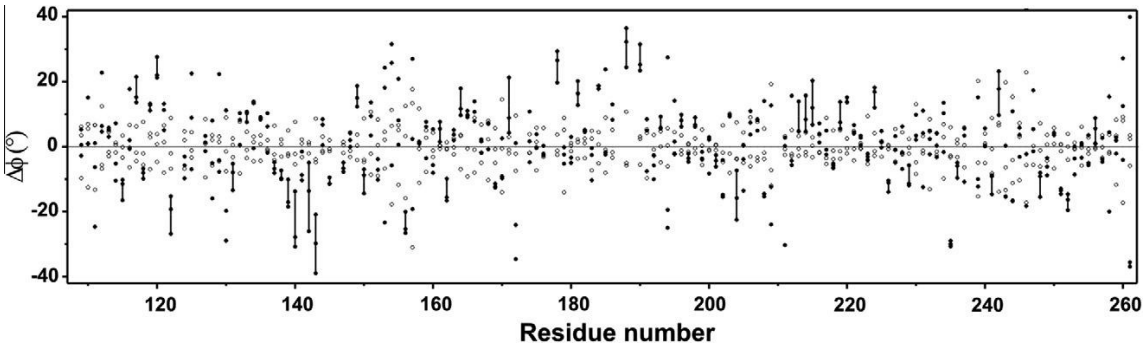


Fig. 6. Dihedral ϕ angles of the solution structures (solid symbols) and of the crystal structures (open symbols), after subtraction of the average of the angles calculated from the crystal structures. The bars indicate the residues whose dihedral angles in solution are consistently outside the range of the values observed at the solid state. The analyzed structures are the 1HFC, 3SHI and 1CGE structures.

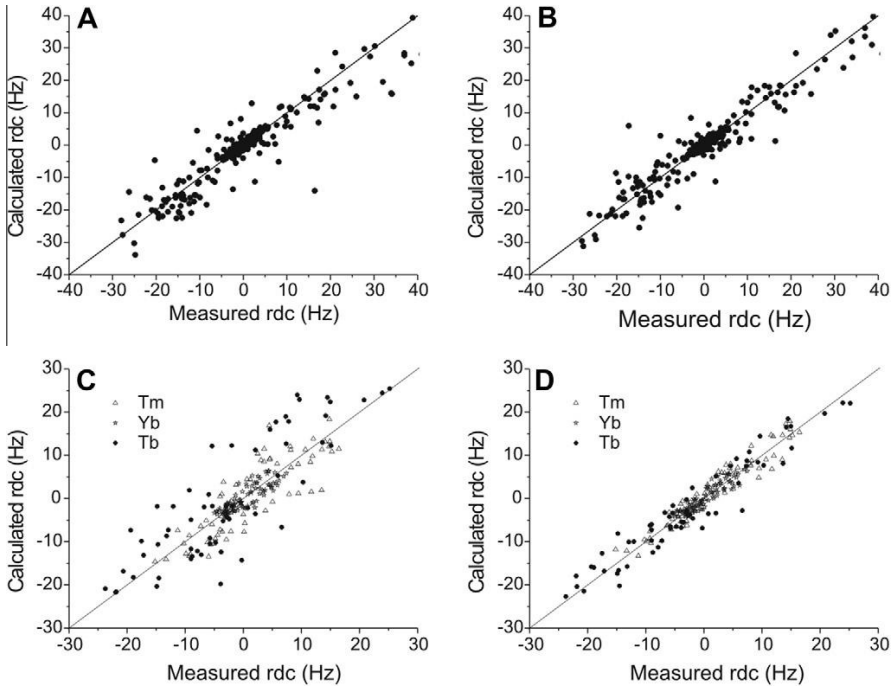


Fig. 7. (A) Best fit of the experimental rdc values to the crystal 1EXR structure of calmodulin and (B) to the structure calculated with the inclusion of the solution restraints. (C) Agreement between experimental rdc and rdc calculated from the pcs-derived tensors and the crystal 1YR5 structure of calmodulin bound to the binding peptide of DAPk or (D) the structure calculated with the inclusion of the paramagnetic restraints.

The agreement between the crystal structure and the paramagnetic restraints was also checked for the adduct of N60D calmodulin with the binding peptide of DAPk [5]. For this system, pcs and rdc were collected for the Yb^{3+}, Tb^{3+} and Tm^{3+} ions substituted to the second binding site of the N-terminal domain and used to show the occurrence of a conformational rearrangement involving the first helix of the N-terminal domain and the whole C-terminal domain with respect to the N-terminal domain [5]. Indeed, the agreement of the rdc data is modest (Fig. 7D, $Q = 0.28$; the Q factor of the pcs is 0.10), with differences up to 10 Hz between the calculated rdc and the experimental values, when the conformation of the protein is restrained to the coordinates of the crystal structure. Again, the protocol points to a sizable conformational rearrangement occurring on passing from crystal to solution. In these cases, previously described methods [2,4,5] should be applied for a structural refinement, based on the inclusion of dihedral angles restraints, which replace the distance restraints anchoring the nuclear coordinates to the model structure, and the inclusion of appropriate force fields for keeping the protein structure properly folded while allowing secondary structure or domain rearrangements.

4. Concluding remarks

Solid state structures are known to suffer from crystal packing forces so that they may not be accurate models for the structures in solution. Paramagnetism-assisted NMR can provide valuable restraints to assess the extent and the nature (local or global) of the deviations, and to produce better models for the solution

structures, without the need to assign all proton resonances and collect and analyze all the NMR spectra required for the collection of the classical NMR restraints. A protocol is presented for discriminating between the case of minor, local changes needed to reconcile the available model with the experimental data, and the case of major, global changes as those involving domain reorientations [54]. In the former case, the smallest structural changes that are needed to reproduce the experimental data can be determined and solution structures of improved accuracy are thus obtained. The tolerance and weight used for restraining the protein coordinates to the crystal model must be adjusted in order to allow a contribution to the target function of the NMR restraints low and comparable with the contribution of the upper distance limits anchoring the nuclear coordinates. If even with tolerances of few tenths of an Angstrom the NMR data cannot be reproduced within their error, the approaches developed for refining the proteins in the presence of major global changes, previously described [2,4,5], should be used.

These protocols are foreseen to be of large utility for the structural and dynamic characterization of protein complexes and of multidomain proteins, where the different domains have some degree of orientational freedom. In these cases, in fact, the paramagnetism-based restraints, collected thanks to the presence of paramagnetic metal ions rigidly positioned in one protein domain, can be used for determining the relative position of the protein domains with respect to one another in case of no motion, or for obtaining information on the conformational heterogeneity of the system in the presence of interdomain motion. Preliminary to this is, of course, the availability of the solution structures of each rigid protein domain. After checking the consistency of crystal and solution restraints, the simultaneous use of diffraction data and paramagnetic restraints for the structural calculation of a protein can be advantageous for obtaining structures of improved accuracy and precision, as also previously shown using classical NMR restraints [55].

Acknowledgements

This work has been supported by Fundação para a Ciência e Tecnologia (FCT), Portugal (grant SFRH/BD/45928/2008 to J.M.C.T.), MIUR-FIRB contracts RBLA032ZM7 and RBRN07BMCT, Ente Cassa di Risparmio di Firenze, and by the European Commission, contracts Bio-NMR n. 261863, East-NMR n. 228461, Strep-Sfmet n. 201640, and We-NMR 261572.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2011.09.020.

References

[1] Gochin, M. and Roder, H. (1995) Protein structure refinement based on paramagnetic NMR shifts. Applications to wild-type and mutants forms of Cytochrome c. Protein Sci. 4, 296–305.
[2] Chou, J.J., Li, S. and Bax, A. (2000) Study of conformational rearrangement and refinement of structural homology models by the use of heteronuclear dipolar couplings. J. Biomol. NMR 18, 217–227.
[3] Skrynnikov, N.R., Goto, N.K., Yang, D., Choy, W.-Y., Tolman, J.R., Mueller, G.A. and Kay, L.E. (2000) Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: Differences in solution and crystal forms of maltodextrin binding protein loaded with β-cyclodextrin. J. Mol. Biol. 295, 1265–1273.
[4] Chou, J.J., Li, S., Klee, C.B. and Bax, A. (2001) Solution structure of Ca^{2+} calmodulin reveals flexible hand-like properties of its domains. Nature Struct. Biol. 8, 990–997.
[5] Bertini, I., Kursula, P., Luchinat, C., Parigi, G., Vahokoski, J., Willmans, M. and Yuan, J. (2009) Accurate solution structures of proteins from X-ray data and minimal set of NMR data: calmodulin peptide complexes as examples. J. Am. Chem. Soc. 131, 5134–5144.
[6] Bertini, I., Faraone-Mennella, J., Gray, B.H., Luchinat, C., Parigi, G. and Winkler, J.R. (2004) NMR-validated structural model for oxidized Rhodopseudomonas palustris cytochrome c556. J. Biol. Inorg. Chem. 9, 224–230.
[7] Fragai, M., Luchinat, C. and Parigi, G. (2006) "Four-dimensional" protein structures: examples from metalloproteins. Acc. Chem. Res. 39, 909–917.
[8] Bertini, I., Luchinat, C., Parigi, G. and Pierattelli, R. (2008) Perspectives in NMR of paramagnetic proteins. Dalton Trans. 2008, 3782–3790.
[9] Bertini, I., Gupta, Y.K., Luchinat, C., Parigi, G., Peana, M., Sgheri, L. and Yuan, J. (2007) Paramagnetism-based NMR restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains. J. Am. Chem. Soc. 129, 12786–12794.
[10] Bertini, I., Giachetti, A., Luchinat, C., Parigi, G., Petoukhov, M.V., Pierattelli, R., Ravera, E. and Svergun, D.I. (2010) Conformational space of flexible biological macromolecules from average data. J. Am. Chem. Soc. 132, 13553–13558.
[11] Bertini, I., Fragai, M., Luchinat, C., Melikian, M., Mylonas, E., Sarti, N. and Svergun, D. (2009) Interdomain flexibility in full-lenght matrix metalloproteinase-1 (MMP-1). J. Biol. Chem. 284, 12821–12828.
[12] Su, X.C., Huber, T., Dixon, N.E. and Otting, G. (2006) Site-specific labelling of proteins with a rigid lanthanide-binding tag. ChemBioChem 7, 1599–1604.
[13] Keizers, P.H., Desreux, J.F., Overhand, M. and Ubbink, M. (2007) Increased paramagnetic effect of a lanthanide protein probe by two-point attachment. J. Am. Chem. Soc. 129, 9292–9293.
[14] Su, X.C., Man, B., Beeren, S., Liang, H., Simonsen, S., Schmitz, C., Huber, T., Messerle, B.A. and Otting, G. (2008) A dipicolinic acid tag for rigid lanthanide tagging of proteins and paramagnetic NMR spectroscopy. J. Am. Chem. Soc. 130, 10486–10487.
[15] Vlasie, M.D., Fernández-Busnadiego, R., Prudêncio, M. and Ubbink, M. (2008) Conformation of pseudoazurin in the 152 kDa electron transfer complex with nitrite reductase determined by paramagnetic NMR. J. Mol. Biol. 375, 1405–1415.
[16] Zhuang, T., Lee, H.S., Imperiali, B. and Prestegard, J.H. (2008) Structure determination of a Galectin-3-carbohydrate complex using paramagnetism-based NMR constraints. Protein Sci. 17, 1220–1231.
[17] Xu, X., Keizers, P.H.J., Reinle, W., Hannemann, F., Bernhardt, R. and Ubbink, M. (2009) Intermolecular dynamics studied by paramagnetic tagging. J. Biomol. NMR 43, 247–254.
[18] Saio, T., Ogura, K., Yokochi, M., Kobashigawa, Y. and Inagaki, F. (2009) Two-point anchoring of a lanthanide-binding peptide to a target protein enhances the paramagnetic anisotropic effect. J. Biomol. NMR 44, 157–166.
[19] Häussinger, D., Huang, J. and Grzesiek, S. (2009) DOTA-M8: an extremely rigid, high-affinity lanthanide chelating tag for PCS NMR spectroscopy. J. Am. Chem. Soc. 131, 14761–14767.
[20] Hass, M.A.S., Keizers, P.H.J., Blok, A., Hiruma, Y. and Ubbink, M. (2010) Validation of a lanthanide tag for the analysis of protein dynamics by paramagnetic NMR spectroscopy. J. Am. Chem. Soc. 132, 9952–9953.
[21] Man, B., Su, X.C., Liang, H., Simonsen, S., Huber, T., Messerle, B.A. and Otting, G. (2010) 3-Mercapto-2,6-pyridinedicarboxylic acid: a small lanthanide-binding tag for protein studies by NMR spectroscopy. Chem. Eur. J. 16, 3827–3832.
[22] Keizers, P.H.J., Saragliadis, A., Hiruma, Y., Overhand, M. and Ubbink, M. (2008) Design, synthesis, and evaluation of a lanthanide chelating protein probe: CLaNP-5 yields predictable paramagnetic effects independent of environment. J. Am. Chem. Soc. 130, 14802–14812.
[23] Keizers, P.H.J., Mersinli, B., Reinle, W., Donauer, J., Hiruma, Y., Hannemann, F., Overhand, M., Bernhardt, R. and Ubbink, M. (2010) A solution model of the complex formed by adrenodoxin and adrenodoxin reductase determined by paramagnetic NMR spectroscopy. Biochemistry 49, 6846–6855.
[24] Polášek, M., Šedinová, M., Kotek, J., Vander Elst, L., Muller, R.N., Hermann, P. and Lukeš, I. (2009) Pyridine-N-oxide analogues of DOTA and their Gadolinium(III) complexes endowed with a fast water exchange on the square-antiprismatic isomer. Inorg. Chem. 48, 455–465.
[25] Bertini, I., Fragai, M., Giachetti, A., Luchinat, C., Maletta, M., Parigi, G. and Yeo, K.J. (2005) Combining in silico tools and NMR data to validate protein-ligand structural models: application to matrix metalloproteinases. J. Med. Chem. 48, 7544–7559.
[26] Leslie, A.G.W. (1991) In crystallographic computing V in: Molecular Data Processing (Moras, D., Podjarny, A.D. and Thierry, J.-C., Eds.), pp. 50–61, Oxford University Press, Oxford.
[27] Evans PR. Data reduction. In: Proceedings of CCP4 Study Weekend. Data Collection & Processing, 1993. p. 114–122
[28] Rossmann, M.G. and Blow, D.M. (1962) The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst. D15, 24–31.
[29] Crowther, R.A. (1972) The Molecular Replacement Method. In: Rossmann, M.G., editor. Gordon & Breach, New York.
[30] Vagin, A.A. and Teplyakov, A. (1997) MOLREP: an automated program for molecular replacement. J. Appl. Crystallogr. 30, 1022–1025.
[31] Vagin, A.A. and Teplyakov, A. (2000) An approach to multi-copy search in molecular replacement. Acta Crystallogr. D: Biol. Crystallogr. 56, 1622–1624.
[32] Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D: Biol. Crystallogr. 53, 240–255.
[33] McRee, D.E. (1992) XtalView: a visual protein crystallographic software system for XII/XView. J. Mol. Graphics 10, 44–47.
[34] Lamzin, V.S. and Wilson, K.S. (1993) Automated refinement of protein models. Acta Crystallogr. D: Biol. Crystallogr. 49, 129–147.
[35] Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291.

[36] Keller R (2003) The CARA/Lua Programmers Manual. DATONAL AG.
[37] Moy, F.J., Pisano, M.R., Chanda, P.K., Urbano, C., Killar, L.M., Sung, M.L. and Powers, R. (1997) Assignments, secondary structure and dynamics of the inhibitor-free catalytic fragment of human fibroblast collagenase. J. Biomol. NMR 10, 9–19.
[38] Ottiger, M., Delaglio, F. and Bax, A. (1998) Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. J. Magn. Reson. 131, 373–378.
[39] Bertini, I., Luchinat, C. and Parigi, G. (2002) Magnetic susceptibility in paramagnetic NMR. Progr. NMR Spectrosc. 40, 249–273.
[40] Banci, L., Bertini, I., Huber, J.G., Luchinat, C. and Rosato, A. (1998) Partial orientation of oxidized and reduced cytochrome b_5 at high magnetic fields: Magnetic susceptibility anisotropy contributions and consequences for protein solution structure determination. J. Am. Chem. Soc. 120, 12903–12909.
[41] Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. Proc. Natl. Acad. Sci. USA 92, 9279–9283.
[42] Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1997) NMR evidence for slow collective motions in cyanometmyoglobin. Nature Struct. Biol. 4, 292–297.
[43] Bax, A. and Tjandra, N. (1997) Are proteins even floppier than we thought? Nature Struct. Biol. 4, 254–256.
[44] Bothner-By, A.A., Domaille, J.P. and Gayathri, C. (1981) Ultra high-field NMR spectroscopy: observation of proton-proton dipolar coupling in paramagnetic bis[tolyltris(pyrazolyl)borato]cobalt(II). J. Am. Chem. Soc. 103, 5602–5603.
[45] Bertini, I., Luchinat, C., Parigi, G. and Pierattelli, R. (2005) NMR of paramagnetic metalloproteins. ChemBioChem 6, 1536–1549.
[46] Guntert, P. (2004) Automated NMR structure calculation with CYANA. Methods Mol. Biol. 278, 353–378.
[47] Balayssac, S., Bertini, I., Luchinat, C., Parigi, G. and Piccioli, M. (2006) ^{13}C direct detected NMR increases the detectability of residual dipolar couplings. J. Am. Chem. Soc. 128, 15042–15043.
[48] Bertini, I., Luchinat, C. and Parigi, G. (2011) Moving the frontiers in solution solid state bioNMR. A celebration of Harry Gray's 75th birthday. Coord. Chem. Rev. 255, 649–663.
[49] Banci, L., Bertini, I., Cavallaro, G., Giachetti, A., Luchinat, C. and Parigi, G. (2004) Paramagnetism-based restraints for Xplor-NIH. J. Biomol. NMR 28, 249–261.
[50] Barbieri, R., Luchinat, C. and Parigi, G. (2004) Backbone-only protein solution structures with a combination of classical and paramagnetism-based constraints: a method that can be scaled to large molecules. ChemPhysChem 21, 797–806.
[51] Clore, G.M. and Schwieters, C.D. (2006) Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small α/β protein: a unified picture of high probability, fast motions in proteins. J. Mol. Biol. 355, 879–886.
[52] Lange, O.F., Lakomek, N.-A., Farès, C., Schröder, G.F., Walter, K.F.A., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C. and de Groot, B.L. (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320, 1471–1475.
[53] Clore, G.M. and Schwieters, C.D. (2004) How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? J. Am. Chem. Soc. 126, 2923–2938.
[54] Clore, G.M. and Kuszewski, J. (2003) Improving the accuracy of NMR structures of RNA by means of conformational database potentials of mean force as assessed by complete dipolar coupling cross-validation. J. Am. Chem. Soc. 125, 1518–1525.
[55] Shaanan, B., Gronenborn, A.M., Cohen, G.H., Gilliland, G.L., Veerapandian, B., Davies, D.R. and Clore, G.M. (1992) Combining experimental information from crystal and solution studies – joint X-ray and NMR refinement. Science 257, 961–964.

3.3 Analysis of conformational heterogeneity in multidomain enzymes: the prologue of MMP-1 collagenolysis

Ivano Bertini [1,2], Linda Cerofolini [2], Gregg B. Fields [3], Marco Fragai [1,2], Carlos F .G.C. Geraldes [4], Claudio Luchinat [1,2], Giacomo Parigi [1,2], Enrico Ravera [1,2], Dmitri I. Svergun [5], João M.C. Teixeira [2,4]

[1] Department of Chemistry "U. Schiff", University of Florence, via della Lastruccia 3, 50019, Sesto Fiorentino (FI), Italy

[2] CERM, University of Florence, Via Luigi Sacconi 6, 50019, Sesto Fiorentino (FI) Italy

[3] Torrey Pines Institute for Molecular Studies, Port St. Lucie, FL 34987, USA

[4] Department of Life Sciences, Center of Neurosciences and Cell Biology and Chemistry Center, Faculty of Science and Technology, University of Coimbra, P.O. Box 3046, 3001-401 Coimbra, Portugal

5 EMBL c/o DESY, Notkestrasse 85, Geb. 25 A, 22603 Hamburg, Germany

Submitted

Analysis of conformational heterogeneity in multidomain enzymes: the prologue of MMP-1 collagenolysis

Ivano Bertini (ivanobertini@cerm.unifi.it)[1,2,‡], Linda Cerofolini (cerofolini@cerm.unifi.it)[2], Gregg B. Fields (gfields@tpims.org)[3*], Marco Fragai (fragai@cerm.unifi.it)[1,2], Carlos F.G.C. Geraldes (geraldes@ci.uc.pt)[4], Claudio Luchinat (luchinat@cerm.unifi.it)[1,2*], Giacomo Parigi (parigi@cerm.unifi.it)[1,2], Enrico Ravera (ravera@cerm.unifi.it)[1,2], Dmitri I. Svergun (svergun@embl-hamburg.de)[5], João M.C. Teixeira (joao@ci.uc.pt)[2,4]

[1]Department of Chemistry "U. Schiff", University of Florence, via della Lastruccia 3, 50019, Sesto Fiorentino (FI), Italy
[2]CERM, University of Florence, Via Luigi Sacconi 6, 50019, Sesto Fiorentino (FI) Italy
[3]Torrey Pines Institute for Molecular Studies, Port St. Lucie, FL 34987, USA
[4]Department of Life Sciences, Center of Neurosciences and Cell Biology and Chemistry Center, Faculty of Science and Technology, University of Coimbra, P.O. Box 3046, 3001-401 Coimbra, Portugal
[5]EMBL c/o DESY, Notkestrasse 85, Geb. 25 A, 22603 Hamburg, Germany

Abstract

A significant number of proteases and kinases are multidomain enzymes. In matrix metalloproteinase 1 (MMP-1), interdomain flexibility is crucial for collagen degradation. A rigorous assessment of the most readily accessed conformations in solution is required to explain the onset of substrate recognition and collagenolysis. Recovering such information from experimental data that are averages over a conformational ensemble is a so-called "ill-posed inverse problem" that admits an infinite number of solutions. We address this issue by calculating the maximum occurrence (MO) of conformations, through paramagnetic NMR and small angle X-ray scattering (SAXS). Analysis of the highest MO conformations suggests that MMP-1 in solution is poised to interact with the substrate and then can easily proceed along the steps of collagenolysis. The MO approach described herein can evaluate the predominant domain conformations for numerous multidomain enzymes, providing insight into mechanistic action and novel inhibitor design.

[‡]Ivano Bertini passed away on July 7th, 2012.

Proteases and kinases represent ~6% of the human genome (1, 2). Within these enzyme superfamilies, a significant number are multidomain (2, 3). Unfortunately, there is often little knowledge as to how enzyme domains are oriented for substrate recognition or how they cooperate to enhance catalytic activity. Many proteases and kinases are important therapeutic targets, as their activities have been correlated or validated to disease progression. Documenting the predominant conformational states of multidomain enzymes represents an opportunity to better define mechanisms and to develop novel, selective inhibitors.

Matrix metalloproteinases (MMPs) are a family of proteases with the striking feature of hydrolyzing structurally unrelated substrates (4, 5). This broad proteolytic specificity, together with tight regulation of enzyme activation and localization, has been achieved by an evolutionary process where specialization of protein domains and protein flexibility interplay to facilitate recognition and hydrolysis of a variety of substrates (6). In particular, several active MMPs, including MMP-1, are two-domain [catalytic (CAT) and hemopexin-like (HPX)] enzymes capable of catalyzing the hydrolysis of highly structured substrates such as triple-helical, interstitial (types I-III) collagen. Interdomain flexibility appears particularly important for allowing movement of the MMP along collagen fibrils and for unwinding of the collagen and accommodation of a single, otherwise inaccessible, peptide chain into the active site (6-12). Assessment of the most easily accessed conformations within the ensemble of all sterically possible conformations in solution can be critical to understanding substrate recognition.

When a system rapidly samples multiple conformations, the experimental data are a weighted average of the data relative to each conformation. Recovering a conformational ensemble from averaged data is known as an "ill-defined inverse problem" (13) that has an infinite number of solutions. Various methods (14-23) have been proposed to reconstruct ensembles consistent with the experimental data. To advance from simply obtaining many "plausible" ensembles to identifying specific conformations within these ensembles that are more likely sampled by the system, we proposed the idea of maximum allowed probability (24), later extended to the concept of maximum occurrence (MO) (25). The MO of a given conformation is defined, and numerically calculated, as the maximum weight that this conformation can have in any suitable ensemble while still maintaining the ensemble's ability to reproduce the experimental data. Paramagnetic NMR spectroscopy and small angle X-ray scattering (SAXS) average data can be efficiently used as experimental restraints to calculate the MO of conformations of two-domain proteins, as

previously demonstrated for calmodulin (CaM) (25, 26) as well as for its complexes with target peptides (24, 27).

MMP-1 was analyzed herein using the MO approach. Many of the MMP-1 conformations with the highest MO value were found to have interdomain orientations and positions that can be clearly grouped into a cluster. Remarkably, in the conformations belonging to this cluster, (i) the collagen binding residues of the HPX domain are solvent exposed and (ii) the CAT domain is already correctly positioned for its subsequent interaction with the collagen. A modest structural rearrangement, that can be easily performed by a ~50° rotation around a single axis of the CAT domain with respect to the HPX domain, is sufficient to position the CAT domain right in front of the preferred cleavage site in triple-helical collagen. The conformations belonging to this cluster can thus be seen as the antecedent step of collagenolysis.

Results

Pseudocontact shifts (PCS) and self-orientation residual dipolar couplings (RDC) (13) from at least three metals ions with large paramagnetic susceptibility anisotropy are needed as paramagnetic NMR restraints to provide sufficient average data and minimize degeneracy (28). The introduction of covalently bound lanthanide chelators has been widely exploited to introduce paramagnetic centers in diamagnetic proteins (29-31); the rigid lanthanide chelator CLaNP-5 (32) does so by covalently binding two neighboring Cys residues in a rigid fashion[§].

Paramagnetism-Based NMR Data

The correspondence of the chemical shifts in the diamagnetic (Lu^{3+}) tagged and untagged protein indicates that the presence of (Ln^{3+})CLaNP-5 does not affect the CAT structure (35). The magnetic susceptibility anisotropy tensors were determined from the best fit of the PCSs of the CAT domain to the previously refined protein structure (Table 1). The averaged anisotropy tensors obtained from the best fit of the RDCs of the amide protons of the HPX domain to the available X-ray structure of full-length proMMP-1 (pdb entry: 1SU3) (36) were also evaluated

[§]It has been shown that the paramagnetic metal ion cobalt(II) can be introduced in the place of the catalytically active zinc(II) ion[33, 34], and this could complement paramagnetic NMR data. However, the magnetic susceptibility anisotropy of cobalt(II) in the CAT domain is not large enough to provide measurable effects as far away as needed to reach the HPX domain nuclei.

(Table 1). The two sets of tensors for the CAT and HPX domains should be similar to one another in the case of a rigid system. The much smaller values obtained for the HPX domain with respect to those of the CAT domain (to which (Ln^{3+})CLaNP-5 is attached) reveal the presence of sizable interdomain mobility, and the relative tensor magnitudes and orientations reflect the conformational heterogeneity experienced by the system in solution (see later). Because the RDCs induced by one paramagnetic center can always be described by a single averaged anisotropy tensor in the case of rigid domains, independently of the fact that they originate from a weighted average of a number of conformations, the good quality of the fits (Supplementary Figure 1) reflects good agreement of the data with the X-ray structure of the HPX domain. Thus, the HPX domain moves essentially as a rigid body with respect to the CAT domain.

To quantify the extent of the interdomain mobility we compare the RDCs back-calculated for the HPX domain from the susceptibility anisotropy tensors obtained for the CAT domain and the 1SU3 structure (which would correspond to the case of absence of internal motion) (Figure 1A left panel) with the RDCs back-calculated from the averaged anisotropy tensors, which report their real distribution in solution (Figure 1B left panel). For comparison purposes, the distribution of RDCs calculated from uniformly sampled sterically allowed conformations was determined (Figure 1C left panel). The spreading in the actual distribution (Figure 1B left panel) is sizably smaller (by a factor of 3-4) than expected for a rigid system (Figure 1A left panel), thus indicating considerable mobility, but also much larger than the uniform sampling case (Figure 1C left panel), thus indicating the occurrence of preferred conformations in solution. The ratio of the spreading between the real RDC distribution and the RDC distribution calculated in the assumption of no motion, which can be taken as a generalized order parameter reflecting the interdomain mobility (19), is sizably larger in MMP-1 with respect to what was observed for CaM (Figure 1 right panels). This generalized order parameter for MMP-1 is 0.28, 0.27, and 0.29 for Tb^{3+}, Dy^{3+}, and Tm^{3+}, respectively. Different generalized order parameters as well as different scaling factors of the components of the anisotropy tensor indicate that the HPX domain motion causes different motional averaging for the different metals, because of the different rhombicity and directions of the principal axes of the anisotropy tensors. SAXS data, previously measured for MMP-1 solutions under the same experimental conditions as utilized here (10, 37), also indicated that the structure of the protein cannot be described by the crystallographic

conformation alone, but that ensembles with closed and more extended conformations must be considered, further indicating that the protein experiences noticeable flexibility.

Maximum Occurrence (MO) Analysis

A MO analysis was performed using as restraints the motionally averaged PCSs and RDCs collected for the HPX domain, the metal anisotropy tensors determined from the PCSs of the CAT domain, and the SAXS data. The latter data provide restraints complementary to those of the NMR data, and were recently demonstrated to be very useful to make the overall dataset more stringent in characterizing the different conformations through their MO values (25).

If one examines the discrepancy of the calculated average data from the experimental dataset (expressed as Target Function values, TF) for each analyzed conformation as a function of their arbitrarily given weight in any suitable ensemble (see Supplementary Figure 2), the substantial differences in the weight at which the TF value starts increasing result in markedly different MO. Only 6% of the 1000 analyzed conformations were found to have a MO smaller than 5%, while most of the conformations (80%) have a MO smaller than 20%. Only 3% of the conformations have a MO larger than 30%, and only 0.3% have a MO larger than 40%. To visualize the results, the CAT domains of all the structures were superimposed and the position of the HPX domain was schematized by a triad of vectors pointing along the axes of a Cartesian coordinate system. Initial orientation is defined arbitrarily, pointing along the axes of the pdb file used for calculations and centered in the center of mass of the domain. The Cartesian axes system was color-coded with respect to the MO of the corresponding conformation, from blue (MO lower than 5%) to red (highest MO, 48%) (Figure 2A). Different orientations and positions of the Cartesian axes system thus reflect different orientations and positions of the HPX domain with respect to the CAT domain.

The conformations having the HPX domain in the region proximal to (Ln^{3+})CLaNP-5 (and distal to the catalytic site cleft) were found to have a negligible weight in solution, with MO values below 5% (blue tensors in Figure 2A). Thus, these conformations are not sampled significantly by the protein. A striking finding is that most of the conformations with the highest MO (orange-red tensors in Figure 2A-B) are clustered in a well-defined region of the distribution, corresponding to relatively elongated structures. A second region comprising high MO conformations with lower density of structures, and more spread in the conformational space, is

present. To increase the resolution of the regions populated by the structures with the highest MO, additional conformations near these high MO structures were selected from the pool of 50,000 generated by RanCh. In this way, the MO values of 281 additional conformations have been evaluated. All conformations with MO larger than 35% have been examined. It was confirmed that the largest share of high MO conformations was grouped into a well-defined cluster.

MO values can be represented as a function of the translational and rotational parameters of the corresponding structures with respect to the structure with highest MO (Figure 3). The translations are reported with respect to the center of mass of the reference structure. To simplify distance calculations, rotations are represented through the corresponding 4-components complex number (quaternion) and distances are calculated as the projection of one quaternion to the reference one (38).[**] There is continuity in the MO values as a function of these structural parameters, thus indicating a correlation between position/orientation and MO. The main regions with conformations possessing the highest MO values are clearly visible (Figure 3). A reasonably well-defined peak encompassing the conformations with the largest MO value was observed as well as another region with somewhat smaller MO values (up to 40%). From the shape of the 3D plot in Figure 3, it appears that the highest MO conformations can be clearly identified independently from the generation probability of RanCh.

All available X-ray structures of human full-length MMP-1 (pdb entries: 1SU3 (36) and 2CLT (39, 40)) display relatively closed conformations. It is crucial to understand how much these structures are represented in the ensemble sampled by the protein in solution. In order to calculate their MO values, these two structures were included in the pool of structures to be analyzed. The MO values obtained for the X-ray structures 1SU3 (proMMP-1) and 2CLT (active MMP-1) were 20% and 19%, respectively. 2CLT (active MMP-1) was highly similar to the X-ray crystallographic structure of porcine full-length MMP-1 (41), while the recently reported X-ray crystallographic structure of an MMP-1/triple-helical peptide (THP) complex (pdb entry: 4AUO) has a more closed structure than 2CLT (40) and has a MO of 18%.

[**]A quaternion is a complex number with 3 imaginary components. As 2D rotations can be represented by a complex number with norm 1 according to the Euler's rule, a 3D rotation can be represented with a quaternion with norm 1.

The radii of gyration (R_g) of the two crystallographic structures are 25.5 and 25.7 Å, respectively, whereas the structures with highest MO (>35%) have R_g of 28.9 ± 1.3 Å. This range of R_g is in better agreement with the experimentally determined values from the SAXS data alone, indicating that the X-ray structures are more compact than the average solution conformation. Furthermore, the relative orientations of the HPX and CAT domains in the structures with the highest MO are different from those in the X-ray crystallographic structures.

Discussion

Conformational selection or induced fit are often invoked to explain the mechanism used by proteins constituted of multiple domains and connected by flexible linkers to recognize partners or substrates. While a detailed structural characterization of the bound-state conformation is often possible, much more difficult is the analysis of the conformations sampled by multidomain proteins *before* the interaction. However, analysis of the conformational space experienced by the free protein is useful not only to investigate the mechanism of binding, but also to determine the role of the different domains in the identification of substrates or partners, to predict new possible substrates or partners, and to investigate natural and new mechanisms of inhibition (6, 16, 18, 23, 27, 42-44).

The full-length form of MMP-1 was observed by NMR and SAXS to experience a sizable interdomain flexibility and an open-closed equilibrium (10, 37). The compact arrangement of the two domains of MMP-1 observed in the crystal structure, in fact, was shown not to be fully representative of the conformations sampled by the protein in solution. It was demonstrated in previous work that for at least one third of the time, the enzyme exists with the CAT and HPX domains in a more extended arrangement (10). Moreover, it has been hypothesized that the interface of the CAT and HPX domains may conceal secondary binding sites (exosites) involved in the recognition of the substrate (37). While X-ray crystallographic analysis of an MMP-1/THP complex has revealed binding of the THP to a closed form of MMP-1, it has been noted that the mode of binding in the MMP-1/THP structure was unproductive (40). Interdomain mobility in solution appears essential for the protein to be poised to catalyze collagen hydrolysis.

More recently, the interaction of MMP-1 with a THP has been investigated utilizing NMR spectroscopy, leading to a plausible multistep mechanism for collagenolysis (11). In this model, the initial binding of the HPX domain to the THP is followed by the interaction of the CAT

domain with the THP in front of the cleavage site, and by a subsequent back rotation of the CAT and HPX domains toward the closed conformation that drives the unwinding of the triple-helix and causes the displacement of one peptide chain into the active site.

While there is experimental evidence for the formation of the initial MMP-1/THP complex, the mechanism that leads to the observed two-domain interaction is still unclear. The MO analysis performed for MMP-1 can shed light on this elusive step preceding catalysis. The MMP-1 conformations with large MO values (up to 48%) are restricted into a relatively small conformational region, much smaller than what observed for the previously investigated protein CaM, where all MO values were below 35% (24, 25). All conformations with high MO values largely differ from the closed MMP-1 structures obtained by X-ray crystallography. The MO of the latter is around 20%, which represents the upper limit for the presence of this conformation in the ensemble sampled by the protein in solution. In all the high MO conformations, the CAT and HPX domains are not in tight contact, and the residues of the HPX domain reported to be responsible for the binding to the THP are solvent exposed. The MMP-1 conformations that may be more relevant in solution can be examined by comparing the structures with the highest MO values (40-48%) among themselves, after having superimposed their HPX domains (Figure 4). The reciprocal orientation of the CAT domain has been evaluated by considering the differences in the orientation of the hA and hC helices of the CAT domain (defined by residues 130-141 and 250-258, respectively), which are almost perpendicular to one another (Figure 4B). The angles among these highest MO structures for the first and second helix change up to a maximum of 28° with respect to the mean orientation, with the exception of one structure with the longest helix pointing at about 180° with respect to the others. This indicates that all highest MO structures are characterized by an interdomain orientation and position that can be defined relatively well. The single structure differently oriented with respect to the others is most likely a "ghost" solution, arising from the quadratic form of the RDC equation (28), which in this case neither PCSs nor SAXS have been able to remove. A second, much less populated region with high MO values (up to 38%) is also present in the pool of conformations, but its importance for collagenolysis appears negligible, as the reciprocal orientation of CAT and HPX domain is distant from the MMP-1 structure in the first step of the collagenolytic mechanism and also from any MMP-1 crystallographic structure.

In the highest MO structures, the residues of the HPX domain essential for the binding to collagen are not buried between the CAT and HPX domains, and the open space between the two domains is wider than in the crystallographic structures. Furthermore, and more importantly, the secondary binding sites (exosites) of the HPX domain responsible for collagen interaction, and the active site of the CAT domain, face the same side. If triple helical collagen is modeled in its experimentally determined bound position to the HPX domain (11), the CAT domain faces closely the collagen cleavage site, and in about half of the highest MO structures even penetrates the triple helical substrate (Figure 5). Actually all the high MO conformations (MO >35%) of free MMP-1 fall along the boundary between the penetrating and non penetrating conformations.

Therefore, the largest MO conformations sampled by MMP-1 when free in solution, i.e. in the absence of collagen, appear to be much more poised for interaction with collagen than the compact X-ray crystallographic structures. Comparison of the non penetrating structures with high MO values with the structural models corresponding to the different steps of the catalytic mechanism (11) suggests that the protein in solution has a marked tendency to assume "catalytically prone" conformations: once the HPX domain is bound to triple helical collagen, the CAT domain can effectively search within a restricted and productive subset of binding modes that face the collagen hydrolysis site, and can start collagen unwinding and cleavage. Therefore, the high MO conformations that are not colliding can be seen as a possible antecedent step for the recently proposed mechanism of collagenolysis (11).

In order to confirm whether the protein can easily rearrange from the highest MO conformations to the conformation assumed when interacting with the substrate, a morphing between these two conformations has been performed with the programs Climber (45) and FATCAT (46). Rearrangement from one conformation to the other involves only one twist in the hinge region, and the angle that the CAT domain has to cover to reorient itself on the cleavage site of collagen, once the HPX domain is attached, is about 50° along one single axis (Figure 6). The transition seems to be feasible at physiological temperature as the difference in free energy between these steps in the pathway is favorable (-0.133 kcal/mol). Therefore, the conformational rearrangement reasonably occurs through a small energetic barrier, and the entropy loss is compensated by the enthalpy gain associated to the new interaction between the CAT domain and the triple-helix.

It has been previously reported that G271 is critical for the collagenolytic activity of MMP-1. In particular, it has been observed that the replacement of this Gly with bulkier amino acids such as Asp drastically reduce the catalytic efficiency of the enzyme (47). This effect has been explained as being due to an alteration of the linker mobility. Analysis performed with Climber on the G271D MMP-1 mutant showed that the conformational space sampled by the linker passing from the highest MO structures to the conformation in step 1 of the collagenolysis mechanism differs from that observed in the wild type protein, supporting previous results that G271 is largely involved in the hinge bending motion.

The flexibility of MMP-1 domains, and particularly the highly favored extended conformation, also has a critical role in enzyme movement on collagen fibrils that occurs during the proteolytic process. MMPs are known to bind to numerous regions within the collagen triple-helix (48). MMPs then progressively move on collagen fibrils (49). Elongated MMP structures have been observed upon binding to collagen (6), from which an "inchworm" mechanism for MMP movement has been proposed(50). The application of mechanical stress facilitates collagen hydrolysis in the fibril (51). Both the MMP movement and the mechanical stress could be derived from the closing of an open MMP-1 conformation.

The overall conclusion from the present study is that conformational selection followed by induced fit should be invoked to describe the MMP-1/collagen binding process. In fact, among the many conformations sampled by MMP-1 where the residues of the HPX domain essential for collagen binding are not buried between the protein domains, the largest MO conformations have the CAT domain in an orientation that can easy access the collagen once the latter binds to the HPX domain. The present study represents a striking example of the pathway followed by a multidomain protein with flexible linker(s) to perform its catalytic activity. In a broader context, the MO approach described here can evaluate the predominant domain conformations for numerous multidomain enzymes, including members of the protease and kinase superfamilies.

Methods

The MMP-1 mutations H132C and K136C were engineered to attach (Ln^{3+})CLaNP-5 to the protein through disulfide bonds. The residues mutated were on the rigid amphipathic helix (hA), far enough from the active site cleft and the HPX domain to avoid steric clashes that could affect

the conformational heterogeneity of the protein. The paramagnetic ions Tb^{3+}, Dy^{3+}, and Tm^{3+} were ligated to CLaNP-5(32), as previously described for the isolated CAT domain (35). Contrary to the procedure for the single MMP-1 CAT domain (35), no DTT or reductant of any kind was added to the protein at any stage of the (Ln^{3+})CLaNP-5-protein ligation (see Supplementary Text 1) to avoid reduction of the structurally important and solvent exposed disulfide bridge present in the HPX domain between C278 and C466.

Calculations of the MO values were performed for the first 1000 MMP-1 conformations obtained from a pool of 50,000 conformations that were randomly generated by RanCh (15, 25, 52), as representative of all possible conformations in solution. A flexible linker of 13 residues (from R262 to T274) was used to connect the rigid structures of the previously refined CAT(35) and HPX (1SU3) domains. Calculations were performed as described in the Supplementary Text 2, largely through the web server[††] at py-enmr.cerm.unifi.it/access/index/maxocc (53).

Acknowledgements

This work has been supported by Ente Cassa di Risparmio, MIUR-FIRB contracts RBLA032ZM7, RBRN07BMCT, the European Commission, contracts East-NMR no. 228461, WeNMR no. 261572, and Bio-NMR no. 261863, the National Institutes of Health (CA98799), and Fundação para a Ciência e Tecnologia (FCT), Portugal (grant SFRH/BD/45928/2008 to J.M.C.T.). We acknowledge the support and assistance at the X33 beamline of the EMBL (synchrotron DESY, Hamburg) for SAXS data collection. We acknowledge discussion of these studies with Hashim Al-Hashimi, David Fushman, and Marcellus Ubbink at the XII Chianti-Instruct meeting.

[††]Accessible to WeNMR registered users.

Table 1. Tensors obtained by FANTASIAN Software package implemented in the WeNMR Portal.

	Tb^{3+}		Dy^{3+}		Tm^{3+}	
	$\Delta\chi_{ax}$ (10^{-32} m^3)	$\Delta\chi_{rh}$ (10^{-32} m^3)	$\Delta\chi_{ax}$ (10^{-32} m^3)	$\Delta\chi_{rh}$ (10^{-32} m^3)	$\Delta\chi_{ax}$ (10^{-32} m^3)	$\Delta\chi_{rh}$ (10^{-32} m^3)
PCS Tensor of CAT	-45.4	16.5	-40.4	-13.2	51.9	-9.9
RDC Tensor of HPX	-12.7	7.7	-10.9	-2.3	15.0	-2.3

The PCS measured on the CAT domain and the RDC measured on the HPX domain for ^{1}H-^{15}N couplings were used.

Figure Legends

Figure 1. (Left) Distribution of the RDC values of the (Tm^{3+})CLaNP-5 MMP-1 HPX domain based on the magnetic susceptibility tensors obtained from (A) experimental PCS of the CAT domain, (B) experimental RDC of the HPX domain, and (C) average RDC of the HPX domain obtained from sterically allowed uniformly sampled conformations. (Right) Distribution of the RDC values of the Tm^{3+}CaM *C*-terminal Bax structure based on the magnetic susceptibility anisotropy values obtained from (A) experimental PCS of the *N*-terminal domain, (B) experimental RDC of the *C*-terminal domain, and (C) average RDC of the *C*-terminal domain obtained from sterically allowed uniformly sampled conformations(19).

Figure 2. Results of MO calculations for 1000 MMP-1 conformations. Two stereo views of MMP-1 are presented, one (**A**) with all conformations superimposed on the CAT domain, the other (**B**) with all conformations superimposed on the HPX domain. Within the CAT domain, the residues in pink color are those mutated to Cys to incorporate (Ln^{3+})CLaNP-5 and the gray spheres are the metals in the protein. The colored axes are positioned in the center of mass and indicate the orientation of 1000 different structures of the HPX and CAT domain, respectively, randomly generated in space. Colors from blue (<5%) to red (48%) represent the MO values of the various structures.

Figure 3. Representation of MO values as a function of translational and rotational parameters. Both plots are centered at the conformation with the highest MO value. (Left) 3D plot representation of the MO value as a function of the distance between the centers of mass of the different HPX domain conformations and the angle between their quaternion representations. (Right) The probability distribution in space of the conformations generated by RanCh. Plots were generated using Gnuplot, interpolating the data 30 times along each direction.

Figure 4. Results of MO calculations for 1281 MMP-1 conformations. In the stereo view representation, only conformations with MO higher than 40% are reported. One main cluster is distinguishable containing conformations with the highest MO (**A**). CAT domain □-helices hA

and hC are in blue and red, respectively (**B**). The complete structure of one of the CAT domains is also displayed in gray along with the colored helices for reference.

Figure 5. Stereo view representation of MO results for 1281 conformations superimposed on the HPX domain. The conformations where the CAT domain collides with collagen, when the HPX binds collagen as in step 1 of the proposed collagenolytic mechanism(11), are removed (**A**) and separately shown in panel **B**.

Figure 6. Climber calculations of MMP-1 conformations. From left to right: structure with the highest MO, two morphing intermediate steps, and the previously proposed first step of collagenolysis.(11) Structures in the bottom row are rotated 180° about the vertical axis with respect to the top row. The highest MO structure and morphing results were aligned to the HPX domain of the MMP-1/THP complex structure obtained previously.(11) In yellow is the surface representation of MMP-1, in blue is the MMP consensus sequence HEXXHXXGXXH, in orange is the catalytic Zn^{2+} ion, and in green is the surface of the THP. The blue and red arrows indicate the directions of the helices hA and hC, respectively, to facilitate visualizing the movement of the CAT domain with respect to the HPX domain and the THP.

Reference List

1. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002) *Science* **298**, 1912-1934.

2. *Handbook of Proteolytic Enzymes* Barrett,A.J.; Rawlings,N.D.; Woessner,J.F. (Eds.) (2012) (Accademic Press, New York).

3. Srinivasan, N. (2011) *Journal of Natural Science, Biology, Medicine* **2**, 10-11.

4. Overall, C. M. (2002) *Mol. Biotechnol.* **22**, 51-86.

5. Bode, W. (2003) *Proteases and the Regulation of Biological Processes* **70**, 1-14.

6. Rosenblum, G., Van den Steen, P. E., Cohen, S. R., Grossmann, J. G., Frenkel, J., Sertchook, R., Slack, N., Strange, R. W., Opdenakker, G. & Sagi, I. (2007) *Structure* **15**, 1227-1236.

7. Chung, L. D., Dinakarpandian, D., Yoshida, N., Lauer-Fields, J. L., Fields, G. B., Visse, R. & Nagase, H. (2004) *EMBO J.* **23**, 3020-3030.

8. Tam, E. M., Moore, T. R., Butler, G. S. & Overall, C. M. (2004) *J. Biol. Chem.* **279**, 43336-43344.

9. Bertini, I., Calderone, V., Fragai, M., Luchinat, C. & Maletta, M. (2006) *Angew. Chem. Int. Ed.* **45**, 7952-7955.

10. Bertini, I., Fragai, M., Luchinat, C., Melikian, M., Mylonas, E., Sarti, N. & Svergun, D. (2009) *J. Biol. Chem.* **284**, 12821-12828.

11. Bertini, I., Fragai, M., Luchinat, C., Melikian, M., Toccafondi, M., Lauer, J. L. & Fields, G. B. (2012) *J. Am. Chem. Soc.* **134**, 2100-2110.

12. Minond, D., Lauer-Fields, J. L., Cudic, M., Overall, C. M., Pei, D. Q., Brew, K., Moss, M. L. & Fields, G. B. (2007) *Biochemistry* **46**, 3724-3733.

13. Bertini, I., Luchinat, C. & Parigi, G. (2002) *Progr. NMR Spectrosc.* **40**, 249-273.

14. Bernadò, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R. W. H. & Blackledge, M. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 17002-17007.

15. Bernadò, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. (2007) *J. Am. Chem. Soc.* **129**, 5656-5664.

16. Anthis, N. J., Doucleff, M. & Clore, G. M. (2011) *J. Am. Chem. Soc.* **133**, 18966-18974.

17. Lange, O. F., Lakomek, N.-A., Farès, C., Schröder, G. F., Walter, K. F. A., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C. & de Groot, B. L. (2008) *Science* **320**, 1471-1475.

18. Bashir, Q., Volkov, A. N., Ullmann, G. M. & Ubbink, M. (2010) *J. Am. Chem. Soc.* **132**, 241-247.

19. Bertini, I., Del Bianco, C., Gelis, I., Katsaros, N., Luchinat, C., Parigi, G., Peana, M., Provenzani, A. & Zoroddu, M. A. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 6841-6846.

20. Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M. & Vendruscolo, M. (2005) *Nature* **433**, 128-132.

21. Iwahara, J. & Clore, G. M. (2006) *Nature* **440**, 1227-1230.

22. Volkov, A. N., Worrall, J. A. R., Holtzmann, E. & Ubbink, M. (2006) *Proc. Natl. Acad. Sci. USA* **103**, 18945-18950.

23. Zhang, Q., Stelzer, A. C., Fisher, C. K. & Al-Hashimi, H. M. (2007) *Nature* **450**, 1263-1267.

24. Bertini, I., Gupta, Y. K., Luchinat, C., Parigi, G., Peana, M., Sgheri, L. & Yuan, J. (2007) *J. Am. Chem. Soc.* **129**, 12786-12794.

25. Bertini, I., Giachetti, A., Luchinat, C., Parigi, G., Petoukhov, M. V., Pierattelli, R., Ravera, E. & Svergun, D. I. (2010) *J. Am. Chem. Soc.* **132**, 13553-13558.

26. Das Gupta, S., Hu, X., Keizers, P. H. J., Liu, W.-M., Luchinat, C., Nagulapalli, M., Overhand, M., Parigi, G., Sgheri, L. & Ubbink, M. (2011) *J. Biomol. NMR* **51**, 253-263.

27. Nagulapalli, M., Parigi, G., Yuan, J., Gsponer, J., Deraos, S., Bamm, V. V., Harauz, G., Matsoukas, J., de Planque, M., Gerothanassis, I. P. *et al.* (2012) *Structure* **20**, 522-533.

28. Longinetti, M., Parigi, G. & Sgheri, L. (2002) *J. Phys. A:Math. Gen.* **35**, 8153-8169.

29. Wohnert, J., Franz, K. J., Nitz, M., Imperiali, B. & Schwalbe, H. (2003) *J. Am. Chem. Soc.* **125**, 13338-13339.

30. Ikegami, T., Verdier, L., Sakhaii, P., Grimme, S., Pescatore, P., Saxena, K., Fiebig, K. M. & Griesinger, C. (2004) *J. Biomol. NMR* **29**, 339-349.

31. Su, X. C., Man, B., Beeren, S., Liang, H., Simonsen, S., Schmitz, C., Huber, T., Messerle, B. A. & Otting, G. (2008) *J. Am. Chem. Soc.* **130**, 10486-10487.

32. Keizers, P. H. J., Saragliadis, A., Hiruma, Y., Overhand, M. & Ubbink, M. (2008) *J. Am. Chem. Soc.* **130**, 14802-14812.

33. Bertini, I., Fragai, M., Lee, Y.-M., Luchinat, C. & Terni, B. (2004) *Angew. Chem. Int. Ed.* **43**, 2254-2256.

34. Bertini, I., Calderone, V., Fragai, M., Jaiswal, R., Luchinat, C., Melikian, M., Mylonas, E. & Svergun, D. (2008) *J. Am. Chem. Soc.* **130**, 7011-7021.

35. Bertini, I., Calderone, V., Cerofolini, L., Fragai, M., Geraldes, C. F. G. C., Hermann, P., Luchinat, C., Parigi, G. & Teixeira, J. M. C. (2012) *FEBS Lett.* **586**, 557-567.

36. Jozic, D., Bourenkov, G., Lim, N. H., Visse, R., Nagase, H., Bode, W. & Maskos, K. (2005) *J. Biol. Chem.* **280**, 9578-9585.

37. Arnold, L. H., Butt, L. E., Prior, S. H., Read, C. M., Fields, G. B. & Pickford, A. R. (2011) *Journal of Biological Chemistry* **286**, 45073-45082.

38. Kuffner, J. J. Effective sampling and distance metrics for 3D rigid body path planning. 4, 3993-3998. 2004. ICRA '04. 2004 IEEE International Conference on Robotics and Automation, 2004. Proceedings. 26-4-2004.

39. Iyer, S., Visse, R., Nagase, H. & Acharya, K. R. (2006) *J. Mol. Biol.* **362**, 78-88.

40. Manka, S. W., Carafoli, F., Visse, R., Bihan, D., Raynal, N., Farndale, R. W., Murphy, G., Enghild, J. J., Hohenester, E. & Nagase, H. (2012) *Proc. Natl. Acad. Sci. U. S. A* **109**, 12461-12466.

41. Li, J., Brick, P., Ohare, M. C., Skarzynski, T., Lloyd, L. F., Curry, V. A., Clark, I. M., Bigg, H. F., Hazleman, B. L., Cawston, T. E. *et al.* (1995) *Structure* **3**, 541-549.

42. Tang, C., Iwahara, J. & Clore, G. M. (2006) *Nature* **444**, 383-386.

43. Yuwen, T., Post, C. B. & Skrynnikov, N. R. (2011) *J Biomol NMR* **51**, 131-150.

44. Ryabov, Y. E. & Fushman, D. (2007) *J. Am. Chem. Soc.* **129**, 3315-3327.

45. Weiss, D. R. & Levitt, M. (2009) *J Mol Biol* **385**, 665-674.

46. Ye, Y. & Godzik, A. (2003) *Bioinformatics* **19**, ii246-ii255.

47. Tsukada, H. & Pourmotabbed, T. (2002) *J. Biol. Chem.* **277**, 27378-27384.

48. Sun, H. B., Smith, G. N., Hasty, K. A. & Yokota, H. (2000) *Analytical Biochemistry* **283**, 153-158.

49. Saffarian, S., Collier, I. E., Marmer, B. L., Elson, E. L. & Goldberg, G. (2004) *Science* **306**, 108-111.

50. Overall, C. M. & Butler, G. S. (2007) *Structure* **15**, 1159-1161.

51. Sarkar, S. K., Marmer, B., Goldberg, G. & Neuman, K. C. (2012) *Current Biology* **22**, 1047-1056.

52. Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikheney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012) *Journal Of Applied Crystallography* **45**, 342-350.

53. Bertini, I., Ferella, L., Luchinat, C., Parigi, G., Petoukhov, M. V., Ravera, E. & Rosato, A. (2012) *J. Biomol. NMR* **53**, 271-280.

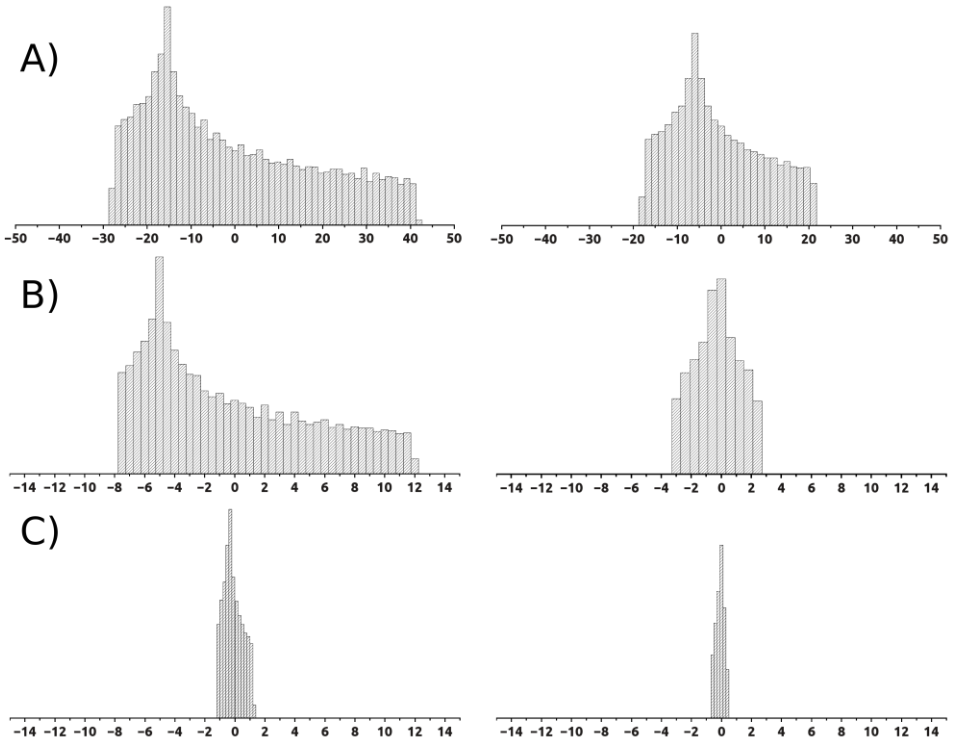


Figure 1

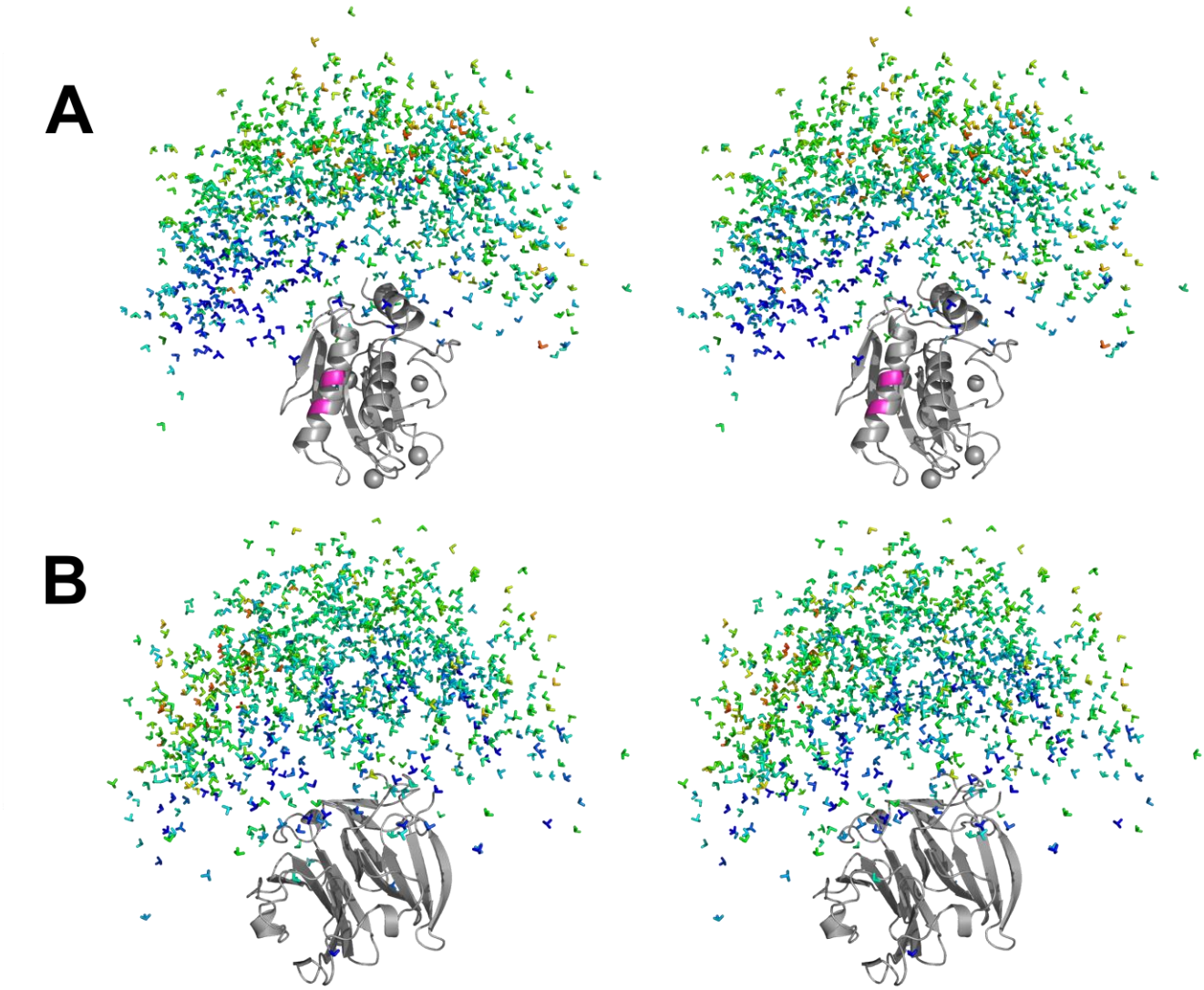


Figure 2

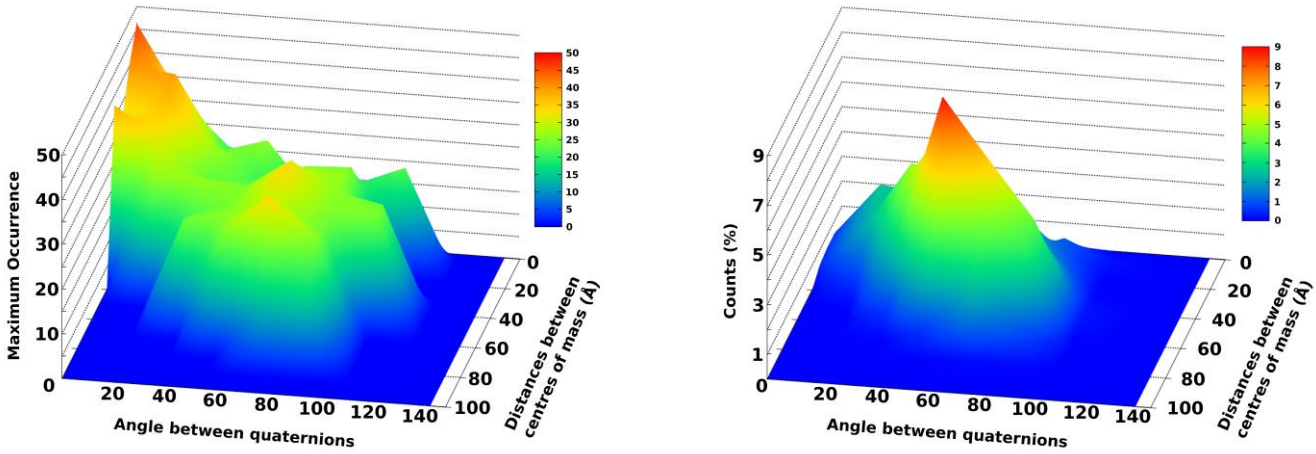

Maximum Occurrence
0
10
20
30
40
50
Angle between quaternions
0
20
40
60
80
100
120
140
Distances between centres of mass (Å)
0
20
40
60
80
100
0
5
10
15
20
25
30
35
40
45
50
Counts (%)
0
1
3
5
7
9
Angle between quaternions
0
20
40
60
80
100
120
140
Distances between centres of mass (Å)
0
20
40
60
80
100
0
1
2
3
4
5
6
7
8
9

Figure 3

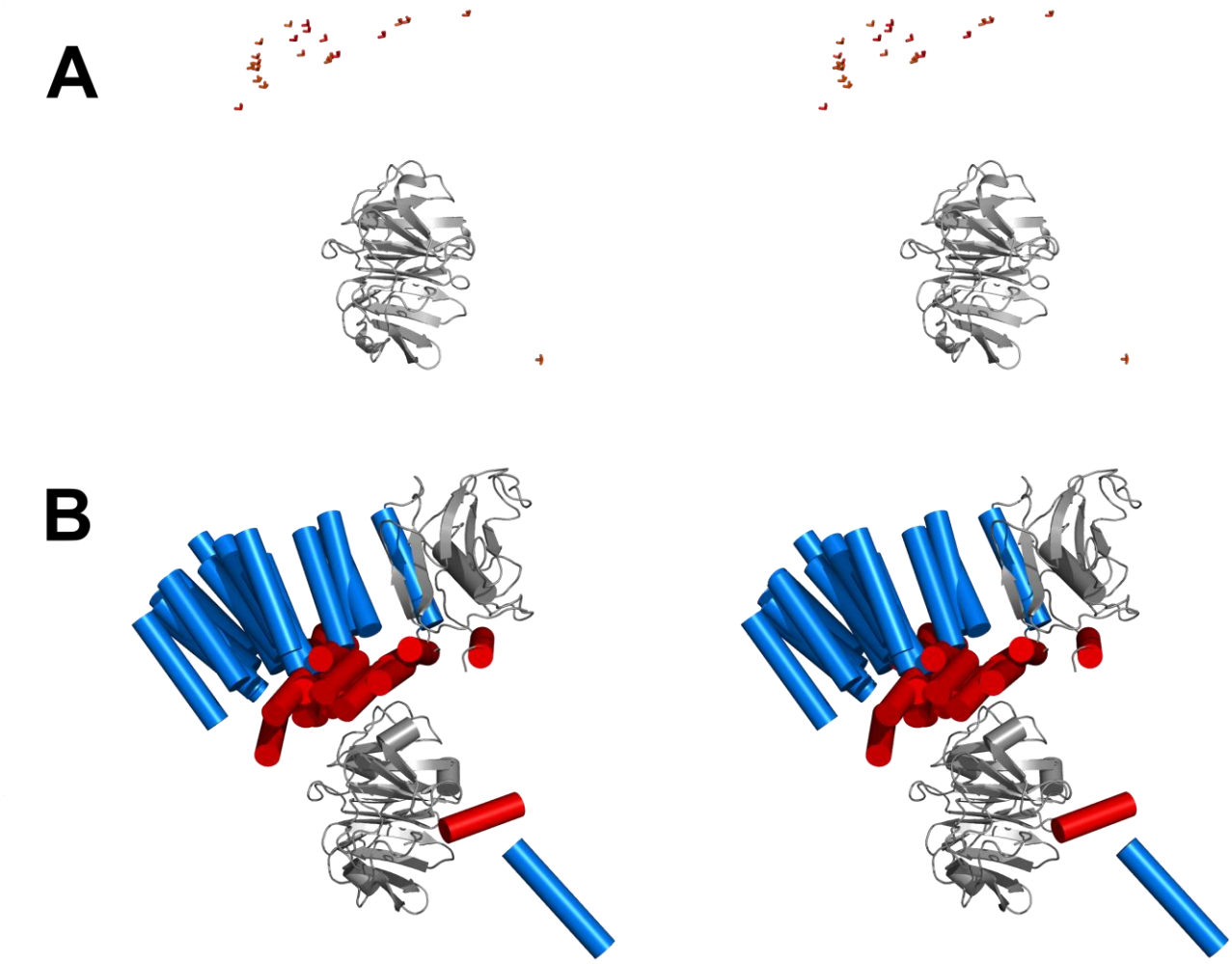


Figure 4

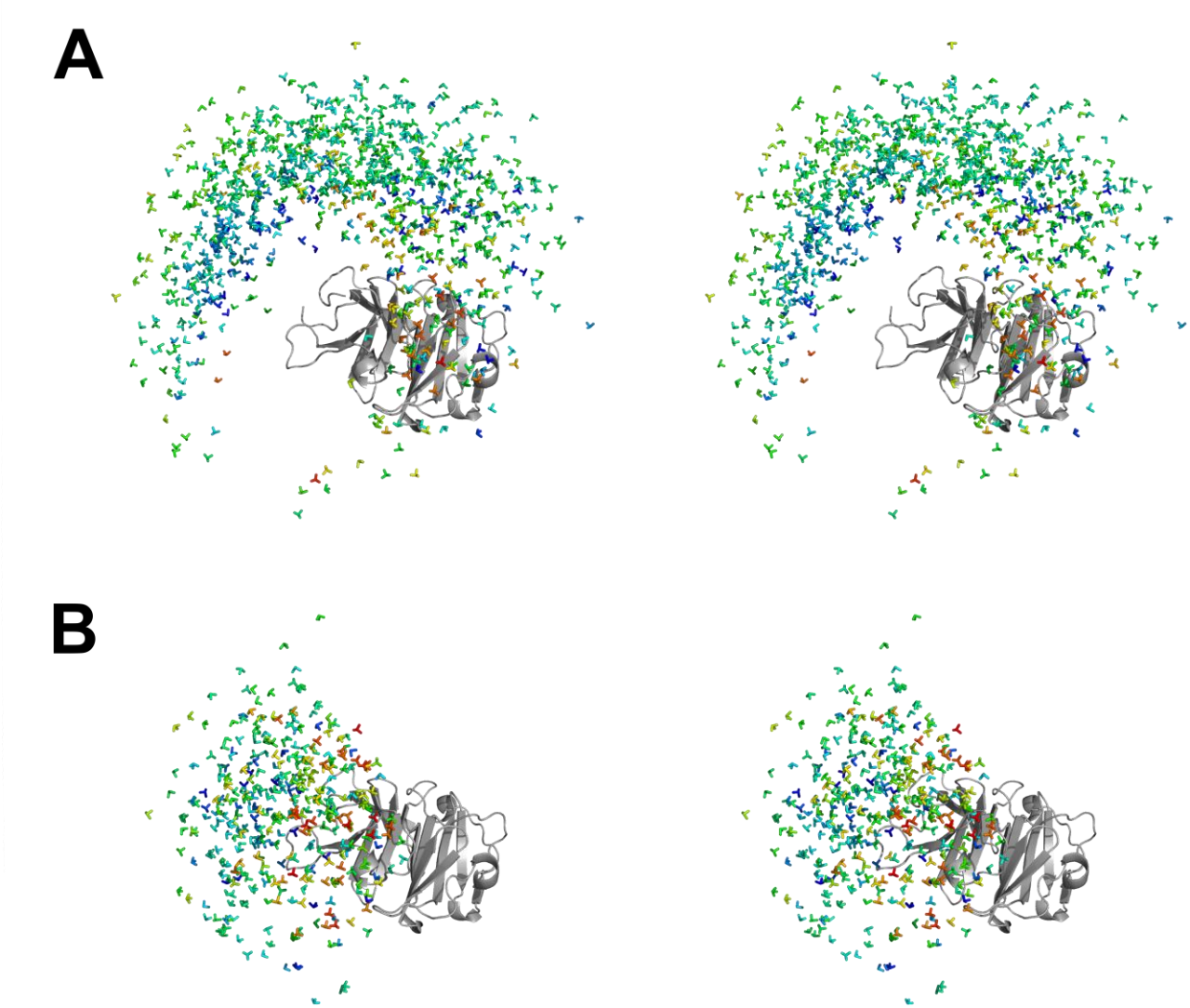


Figure 5

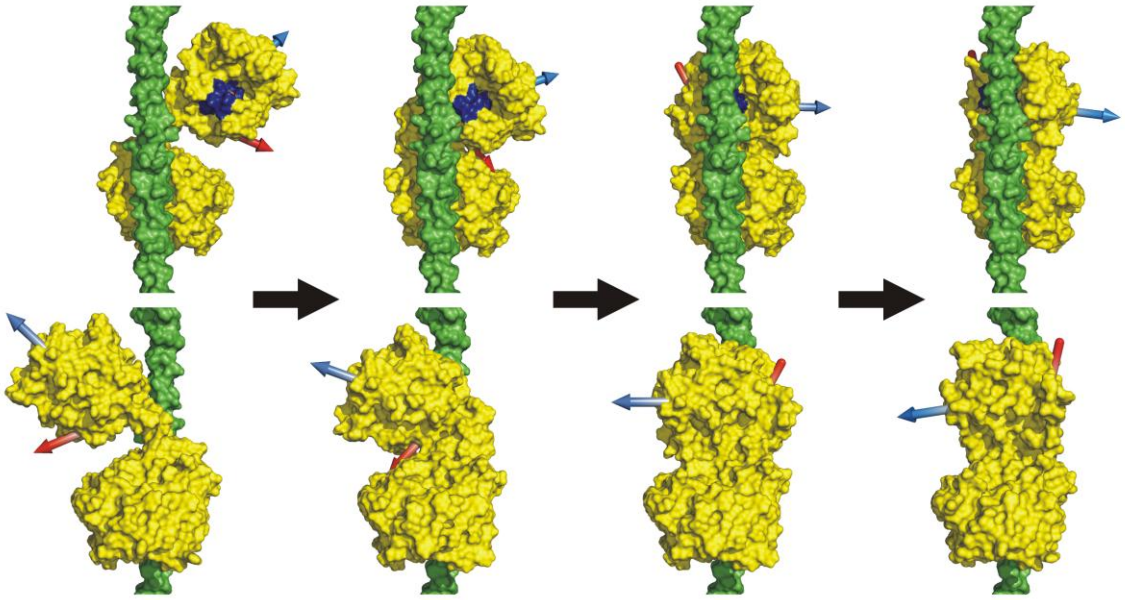


Figure 6

Supplementary Information

Analysis of conformational heterogeneity in multidomain enzymes: the prologue of MMP-1 collagenolysis

Ivano Bertini[1,2] , Linda Cerofolini[1], Gregg B. Fields[2*], Marco Fragai[1,2], Carlos F.C. Geraldes[4], Claudio Luchinat[1,2*], Giacomo Parigi[1,2], Enrico Ravera[1,2], Dmitri I. Svergun[5], João M.C. Teixeira[2,4]

[1]Department of Chemistry "U. Schiff", University of Florence, via della Lastruccia 3, 50019, Sesto Fiorentino (FI), Italy
[2]CERM, University of Florence, Via Luigi Sacconi 6, 50019, Sesto Fiorentino (FI) Italy
[3]Torrey Pines Institute for Molecular Studies, Port St. Lucie, FL 34987, USA
[4]Department of Life Sciences, Center of Neurosciences and Cell Biology and Chemistry Center, Faculty of Science and Technology, University of Coimbra, P.O. Box 3046, 3001-401 Coimbra, Portugal
[5]EMBL c/o DESY, Notkestrasse 85, Geb. 25 A, 22603 Hamburg, Germany

Supplementary Text 1: Protein Expression and Functionalization

Protein Preparation

MMP-1 E219A construct (residues N106 to N469) was prepared as described previously. The double mutation H132C/K136C, engineered for the attachment of the (Ln^{3+})CLaNP-5 tag, was obtained during a single PCR step using QuickChange Site Directed Mutagenesis Kit (Stratagene): 5' - GCC AAG AGC AGA TGT GGA CTG TGC CAT TGA GTG TGC CTT CCA ACT CTG GAG - 3'; 5' - CTC CAG AGT TGG AAG GCA CAC TCA ATG GCA CAG TCC ACA TCT GCT CTT GGC - 3'. The mutations were confirmed by nucleotide sequencing. The expression vector was inserted into the competent *E. coli* BL21(DE3) CodonPlus RIPL strain, and the colonies were selected for ampicillin and chloramphenicol resistance. Monolabeled ^{15}N protein was expressed using minimal medium containing ^{15}N-enriched $(NH_4)_2SO_4$. Cell growth occurred at 37 °C with induction at 0.6 O.D. with 500 □M of IPTG and harvesting after 5 h expression. Triple mutant MMP-1 (H132C/K136C/E219A) precipitated into inclusion bodies, and these were solubilized, after lysis of the cells, in a solution of 8 M urea, 20 mM dithiothreitol, and 20 mM Sigma-Aldrich Trizma-base (pH 8), stored at -20 °C. The refolding of triple mutant MMP-1 consisted of decreasing the urea concentration according to the following steps, performed at 4 °C. The desired amount of protein was diluted in a 500 mL solution containing 6 M urea, 50 mM Trizma-base, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 20 mM cysteamine, at pH 8.0. The solution was then dialyzed against (a) 4 L of 4 M urea, 50 mM Trizma-base, pH 8.0, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, 5 mM 2-mercaptoethanol, and 1 mM hydroxyethyl disulphide (overnight dialysis); (b) 4 L of 2 M urea, 50 mM Trizma-base, pH 7.2, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 0.3 M NaCl; and (c) three steps of 20 mM Trizma-base, pH 7.2, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 0.3 M NaCl. The resulting 500 mL protein sample was concentrated down to 100 mL using MiniKros Modules (Spectrumlabs). H132C/K136C/E219A MMP-1 was purified using HiLoad 26/60 Superdex 75 pg (Amersham Biosciences) in 20 mM Trizma-base, pH 7.2, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 0.3 M NaCl buffer. Protein pure stocks were stored at 4 °C.

(Ln^{3+})CLaNP-5-Protein Ligation. CLaNP-5 was synthesized and functionalized with the different lanthanides as previously described[1].

2 mg of purified H132C/K136C/E219A MMP-1 was concentrated down to 1 mL in 2 M Trizma-base, pH 7.2, 10 mM $CaCl_2$, 0.1 mM $ZnCl_2$, and 0.3 M NaCl buffer. 6-10 equivalents of (Ln^{3+})CLaNP-5 (where

the lanthanide ions were Lu^{3+}, Tb^{3+}, Dy^{3+}, and Tm^{3+}) from DMF stock (ca. 3-6 μL) were added to the protein solution. The triple mutant MMP-1/(Ln^{3+})CLaNP-5 mixture was left on mild stirring overnight. Some protein precipitation was observed after reaction. Contrary to the procedure for the MMP-1 CAT domain, here no DTT or reductant of any kind was added to the protein at any stage of the tag-protein ligation, so to avoid reduction of the structurally important and solvent exposed disulphide bridge present in the HPX domain between C278 and C466. After reaction with (Ln^{3+})CLaNP-5, approximately 10-20% of diamagnetic MMP-1 remained, as estimated from the 2D ^{1}H-^{15}N HSQC spectra acquired on these samples. The overall yield of obtained paramagnetic (Ln^{3+})CLaNP-5-MMP-1, considering precipitation occurring during CLaNP-5 reaction and efficiency of MMP-1 functionalization, was estimated to be ~60-70%.

NMR Measurements

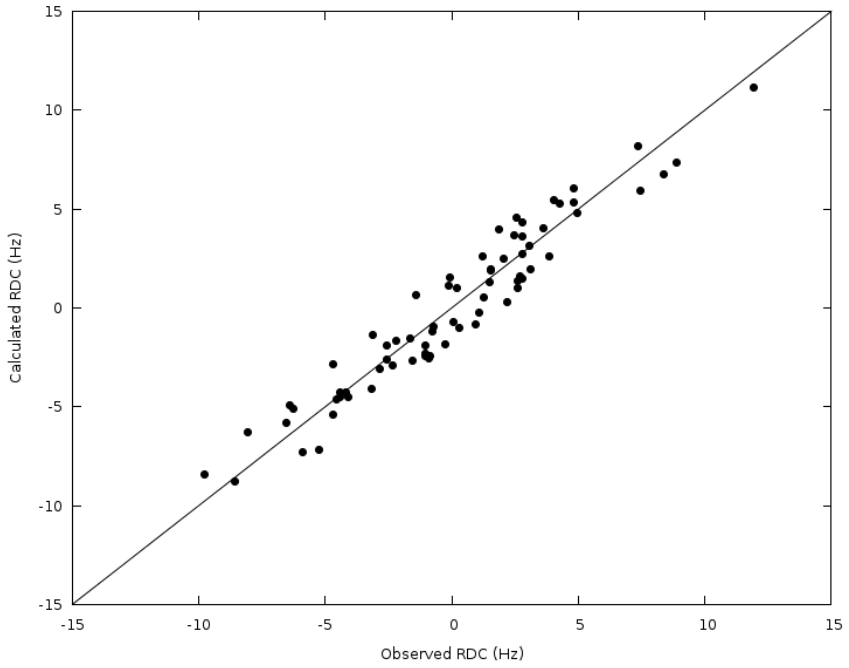
All experiments were performed on samples of triple mutant (H132C/K136C/E219A), full-length MMP-1 functionalized with the (Ln^{3+})CLaNP-5 (Ln = Lu^{3+}, $Tb^{3+,}$ Dy^{3+}, Tm^{3+}), at concentrations ranging between 0.10 and 0.20 mM in water buffer solution (20 mM Tris, pH 7.2, 0.15 M NaCl, 0.1 mM $ZnCl_2$, 10 mM $CaCl_2$, and 200 mM AHA). All NMR experiments were performed at 310 K and acquired on a Bruker AVANCE 700 spectrometer equipped with triple resonance cryo-probe. All spectra were processed with the Bruker TOPSPIN software packages and analyzed by the program CARA (Computer Aided Resonance Assignment, ETH Zurich). The 2D ^{1}H-^{15}N HSQC spectrum of (Ln^{3+})CLaNP-5-MMP-1 was recorded as the diamagnetic reference to evaluate the PCSs. The assignment of the protein functionalized with (Lu^{3+})CLaNP-5 was based on the assignment previously reported for MMP-1[2]; the spectrum was easily reassigned because no meaningful shifts with respect to the non-functionalized protein were observed, indicating that the presence of the CLaNP-5 does not alter the structure of the protein. The assignment of MMP-1 in the presence of the paramagnetic lanthanides was performed by comparison with the assigned spectra obtained for the isolated CAT domain in presence of the same metal ions. ^{1}H-^{15}N RDCs were measured for the MMP-1 functionalized with (Tb^{3+})CLaNP-5, (Dy^{3+})CLaNP-5, and (Tm^{3+})CLaNP-5, by using the IPAP-HSQC method.

Supplementary Text 2: Computational Details

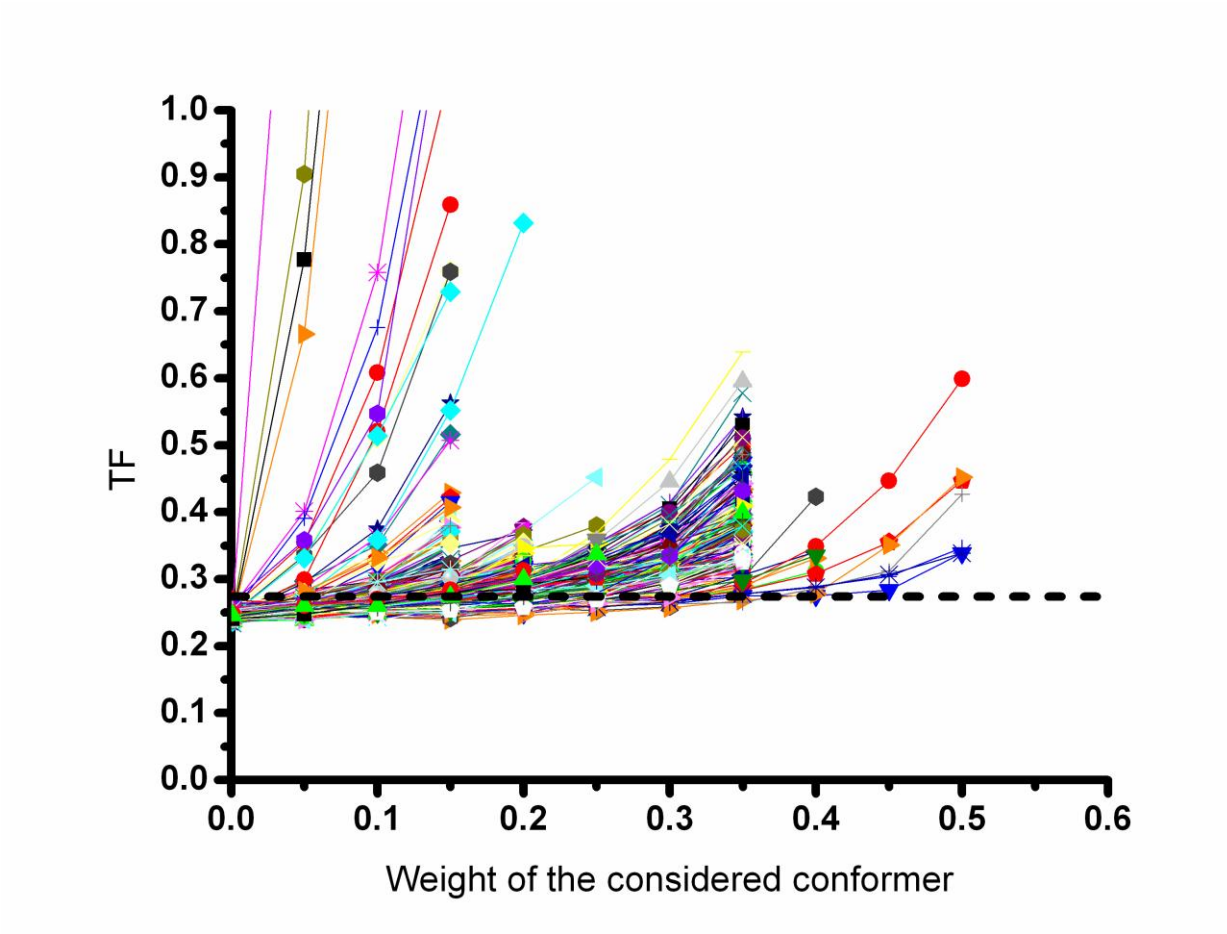
MO Calculations

The MO value of each of the 1000 conformations was obtained from the largest weight that the conformation can have when included in any best-fit ensemble together with 50 other conformations with different weights freely chosen from a pool of 50,000 conformations generated with the program RanCh. These best-fit ensembles were found as families of structures in best agreement with the experimental data by minimizing the target function (TF), defined as a measure of the disagreement from the experimental data of the weighted average PCS and RDC calculated according to the ensemble itself. During the minimization, the weight of the fixed conformation (one of the 1000 randomly generated conformations) was set to different values ranging from 0% to 50% in steps of 5%. The MO of such a conformation was defined as the largest weight for which the TF is smaller than a given threshold. A minimum for the TF was calculated by generating structural ensembles without any fixed conformation, and it was equal to ~0.253. The threshold was defined 10% larger than this lowest value (0.278). Most calculations were performed through the web portal MaxOcc (http://py-enmr.cerm.unifi.it/access/index/maxocc)[3]. The final weights of the restraints used in the calculations were 1.0 for RDC and PCS and 0.1 for SAXS.

Supplementary Figure 1: Quality of fit of calculated versus observed RDC values of the HPX domain.



Supplementary Figure 2: Progression of the target functions (TFs) as a function of the weight given to the selected structure within the minimized ensemble of 50 structures. The dashed line represents the threshold at which the MO value is assigned when intercepted by the TF. Only the TF for the first 500 calculated structures is represented here.



Supplementary References

1. Bertini,I. *et al.* The catalytic domain of MMP-1 studied through tagged lanthanides. *FEBS Lett.* **586**, 557-567 (2012).

2. Bertini,I. *et al.* Interdomain flexibility in full-length matrix metalloproteinase-1 (MMP-1). *J. Biol. Chem.* **284**, 12821-12828 (2009).

3. Bertini,I. *et al.* MaxOcc: a web portal for Maximum Occurence Analysis. *J. Biomol. NMR* **53**, 271-280 (2012).

3.4 Solution structure and dynamics of human S100A14

Ivano Bertini [1,2], Valentina Borsi [1], Linda Cerofolini [1], Soumyasri Das Gupta [1], Marco Fragai[1,2], Claudio Luchinat [1,2]

[1] Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy.

[2] Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Published online 30 November 2012, J. Biol. Inorg. Chem.

J Biol Inorg Chem
DOI 10.1007/s00775-012-0963-3

ORIGINAL PAPER

Solution structure and dynamics of human S100A14

Ivano Bertini · Valentina Borsi · Linda Cerofolini · Soumyasri Das Gupta · Marco Fragai · Claudio Luchinat

Received: 2 September 2012 / Accepted: 6 November 2012

Abstract Human S100A14 is a member of the EF-hand calcium-binding protein family that has only recently been described in terms of its functional and pathological properties. The protein is overexpressed in a variety of tumor cells and it has been shown to trigger receptor for advanced glycation end products (RAGE)-dependent signaling in cell cultures. The solution structure of homodimeric S100A14 in the apo state has been solved at physiological temperature. It is shown that the protein does not bind calcium(II) ions and exhibits a "semi-open" conformation that thus represents the physiological structure of the S100A14. The lack of two ligands in the canonical EF-hand calcium(II)-binding site explains the negligible affinity for calcium(II) in solution, and the exposed cysteines and histidine account for the observed precipitation in the presence of zinc(II) or copper(II) ions.

Ivano Bertini died on 7 July 2012.

An interactive 3D complement page in Proteopedia is available at http://proteopedia.org/w/Journal:JBIC:17

Electronic supplementary material The online version of this article (doi:10.1007/s00775-012-0963-3) contains supplementary material, which is available to authorized users.

I. Bertini · V. Borsi · L. Cerofolini · S. Das Gupta · M. Fragai · C. Luchinat (✉)
Magnetic Resonance Center (CERM),
University of Florence, Via L. Sacconi 6,
50019 Sesto Fiorentino, Italy
e-mail: luchinat@cerm.unifi.it

I. Bertini · M. Fragai · C. Luchinat
Department of Chemistry,
University of Florence,
Via della Lastruccia 3,
50019 Sesto Fiorentino, Italy

Keywords Nuclear magnetic resonance · Binding affinity · Protein folding

Introduction

The family of the S100 proteins includes more than 20 calcium(II)-binding proteins belonging to the EF-hand superfamily [1, 2]. All EF-hand proteins contain two paired EF-hand motifs (helix–loop–helix), which constitute the so-called EF-hand domain. All S100 proteins (but calbindin D9k [3]) are present as homodimers or heterodimers, with each monomer made by an EF-hand domain. The two domains are held together by van der Waals interactions occurring between strictly conserved hydrophobic residues placed at the interface of the two monomers [4, 5]. Some S100 family members can also form large oligomers with potential physiological functions [6]. The two EF-hand motifs in each monomer interplay in a cooperative manner to bind calcium(II) ions, with the N-terminal EF-hand domain showing reduced binding affinity (up to 100 times lower) with respect to the canonical C-terminal EF-hand domain [2]. Binding of calcium causes a conformational change that results in a large movement of helix III towards helix IV: from antiparallel in the apo form to almost perpendicular in the calcium(II)-bound form. This "opens" the structure of the protein and causes the exposure of a hydrophobic cleft delimited by helix III and helix IV [7]. In some family members, such as S100B, this hydrophobic cleft is the binding site for peptides and drug candidates [8–10]. Therefore, according to their conformational state, S100 proteins are termed "closed" and "open." Most of the apo S100 proteins are in the closed conformation, whereas the holo forms usually adopt the open structure [7]. However, in some cases the open conformation is already adopted by the apo form (i.e., for S100A10 [11]), and in other

Published online: 30 November 2012

cases both the apo form and the holo form display a closed conformation (i.e., S100A16 [12]). A "semi-open" conformation has also been described for EF-hand domains with a different packing of the helices, such as the C-terminal domain of myosin light chains, whereas it has not been reported for S100 proteins family [7]. The structural features of EF-hand proteins in general [13] and of S100 proteins in particular have been recapitulated using a principal component analysis (PCA) [14] of the six interhelical angles in order to obtain a complete description of the conformational space spanned by the EF-hand domains in the different members of the EF-hand superfamily. PCA provides a reliable tool to classify the conformations of the EF-hand domains, providing quantitative parameters to describe the open and closed conformations of this large protein superfamily, including the S100 subgroup.

In addition to the well-characterized affinity for calcium(II) ion, many S100 proteins display high affinities for some divalent ions, such as zinc(II) [15] and copper(II) [16] ions, which are suggested to influence the biological activity of these proteins in the extracellular space [17], where these metals are more abundant.

S100A14 is a distinct member of the family involved in several functional and pathological processes [18] and is predicted to be under tight transcription and posttranslational regulation [19]. This protein, which has been predicted to contain an N-glycosylation site, protein kinase phosphorylation sites, and an N-myristoylation site [18], has a high degree of sequence homology with S100A13, which is overexpressed in a variety of tumors, including ovarian, breast, lung, and uterine tumors, whereas it is downregulated in colon, kidney, and rectal tumors and in esophageal squamous cell carcinoma. The action of the exogenous S100A14 on esophageal squamous cell carcinoma cell lines has been investigated recently by Jin et al. [20]. They found that the protein can interact with the receptor for advanced glycation end products (RAGE), which binds to different ligands, among which are S100s [21], thus activating ERK1/2 and NF-κB signaling and stimulating cell proliferation or promoting cell survival at low dose. At high dose, binding of the protein to RAGE increases the production of reactive oxygen species, leading to cell apoptosis [22].

Here, we present the solution structure of S100A14 and the characterization of its dynamical and metal-binding properties through NMR investigations.

Materials and methods

Protein expression

The complementary DNA encoding human S100A14 was PCR-amplified from a human complementary DNA library (GenBank accession number NM_020672.1) with oligonucleotides containing 5′ *Nde*I and 3′ *Xho*I restriction sites. The construct was cloned into the pET21a vector (Novagen) by using *Nde*I and 3′ *Xho*I as restriction enzymes. The resulting plasmid, pET21a-S100A14, was transformed in *Escherichia coli* strain BL21(DE3)CodonPlus (Novagen) and cells were grown at 310 K until the optical density at 600 nm reached 0.6. Then, protein expression was induced with 0.5 mM isopropyl 1-thio-β-D-galactopyranoside and cells were allowed to grow overnight at 310 K. Singly labeled (^{15}N) protein and doubly labeled (^{13}C/^{15}N) protein were expressed using minimal medium containing ^{15}N-enriched $(NH_4)_2SO_4$ and ^{13}C-enriched glucose (Cambridge Isotope Laboratories).

Cells were harvested by centrifugation at 9,000*g*. After lysis of the cells, the protein, precipitated as inclusion body, was solubilized in 20 mM tris(hydroxymethyl)aminomethane (Tris; pH 8), 1 mM EDTA, 5 mM dithiothreitol (DTT), and 8 M urea. The supernatant was then loaded on a Q Sepharose column (GE Healthcare). The protein was eluted with the same buffer containing 800 mM NaCl. The protein was diluted with a buffer containing 20 mM Tris (pH 8), 1 mM EDTA, 5 mM DTT, and 4 M urea, and refolded by using a multistep dialysis against solutions containing 20 mM Tris (pH 8), 1 mM EDTA, 5 mM DTT, and 2 M urea and then against the same solution without urea.

The folded protein was further purified by size-exclusion chromatography on a Superdex 75 column (GE Healthcare) and was eluted with 30 mM 2-morpholinoethanesulfonic acid (pH 6.5), 100 mM NaCl, and 5 mM DTT.

NMR spectroscopy and structure calculation

The NMR experiments for the structure calculation of the homodimeric S100A14 were performed at 310 K on protein samples at concentrations ranging between 0.4 and 0.6 mM, at pH 6.5.

The NMR experiments were performed with Bruker AVANCE 900, 800, 700, and DRX 500 spectrometers, equipped with triple-resonance CryoProbes. All spectra were processed with the Bruker TopSpin 2.0 software package and were analyzed by the program CARA (ETH Zürich) [22].

The backbone resonance assignment was obtained by the analysis of 3D HNCA, 3D HNCO, 3D HN(CA)CO, 3D HNCACB, and 3D CBCA(CO)NH spectra performed at 500 MHz. ^{13}C-detected experiments (2D hCON, 2D hCACO, and 2D hCBCACO) were also performed at 700 MHz to solve some problems of signal overlap and improve the percentage of the assignment. The aliphatic side-chain ^{1}H and ^{13}C resonances were assigned through the analysis of the 3D (H)CCH total correlation spectroscopy spectrum at 500 MHz, the 3D HBHA(CO)NH

spectrum at 700 MHz, and the 3D ^{15}N and ^{13}C nuclear Overhauser effect (NOE) spectroscopy (NOESY) heteronuclear single quantum coherence (HSQC) spectra at 900 MHz. The assignments are reported in Tables S1 and S2 and in the Biological Magnetic Resonance Data Bank under accession code 18818.

Backbone dihedral angles were obtained from TALOS+ [23] from the chemical shifts of N, H^N, H^α, C′, C^α, and C^β nuclei. Proton–proton distance restraints between aliphatic and aromatic protons of side chains were derived from the analysis of a 2D NOESY spectrum acquired in D_2O at 900 MHz in addition to the 3D ^{15}N NOESY-HSQC and 3D ^{13}C NOESY-HSQC spectra. The program CYANA 2.1 [24] was used to calculate a family of 500 structures of the S100A14 protein starting from randomly generated conformers in 15,000 annealing steps. Since the chemical shifts of the atoms of both monomers are identical, it is not possible to discriminate between NOEs of each subunit. Therefore, structure calculations were performed by imposing the dimer symmetry constraint in the CYANA program. The solution structure statistics are reported in Table 1. The quality of the structures calculated by CYANA can be assessed by a properly defined energy function (target function) proportional to the squared deviations of the calculated constraints from the experimental ones, plus the standard covalent and nonbonded energy terms. The best 30 structures of the calculated structures of the CYANA family were then subjected to restrained energy minimization using AMBER 11.0. Energy minimization of the family was performed in a water box of 10 Å [25] and following the protocol available at the AMPS-NMR portal within the WeNMR gateway (http://py-enmr.cerm.unifi.it/access/index/amps-nmr), which includes a molecular dynamics step for the structure minimization [26]. NOE and torsion angle restraints were applied with force constants of 50 kcal mol^{-1} Å^{-2} and 32 kcal mol^{-1} rad^{-2}, respectively. Structure calculation statistics were obtained, and the structural quality was evaluated using PSVS [27] and iCing (http://nmr.cmbi.ru.nl/cing/iCing.html) (see Table 2). The coordinates of the solution structure of human S100A14 were deposited in the Protein Data Bank under accession code 2M0R.

Calcium(II) titration was monitored by NMR spectroscopy with a Bruker 900 MHz spectrometer at 298 K. Two-dimensional ^{1}H–^{15}N HSQC spectra were acquired for different calcium(II) concentrations in solution (0.04, 0.08, 0.16, 0.32, 0.64, 1.2, 2.5, 5.0, 10, 20, 40, 80 mM). Zinc(II) and copper(II) titrations were monitored by NMR spectroscopy with a Bruker 600 MHz spectrometer at 298 K. Two-dimensional ^{1}H–^{15}N HSQC spectra were acquired for different zinc(II) and copper(II) concentrations (0.05, 0.1, 0.2 mM) in a non-oxidizing environment.

Table 1 Structural restraints used in structure calculations

Total number of NOEs per subunit	943
Intrasubunit	811
Intraresidue ($i = j$)	339
Interresidue	
Sequential ($\|i–j\| = 1$)	209
Medium range ($1 < \|i–j\| < 4$)	150
Long range ($\|i–j\| > 5$)	113
Intersubunit	132
Dihedral angle restraints per subunit	
φ	69
ψ	71
Average RMSD from the mean (Å)	
Backbone	0.68 ± 0.17
Heavy atoms	1.08 ± 0.16
Residual CYANA target function	0.75 ± 0.17
Residual AMBER target function	0.44 ± 0.05

NOEs nuclear Overhauser effects, *RMSD* root mean square deviation

Table 2 Statistical analysis of the solution structure of S100A14 obtained with iCing

Structure *Z* scores[a,b]		
1st generation packing quality	1.187 ± 0.614	
2nd generation packing quality	6.132 ± 1.339	
Ramachandran plot appearance	−3.141 ± 0.566	
χ_1/χ_2 rotamer normality	−5.049 ± 0.525	
Backbone conformation	0.983 ± 0.487	
RMS *Z* scores[b,c]		
Bond lengths	1.198 ± 0.003	
Bond angles	0.791 ± 0.014	
ω angle restraints	1.290 ± 0.100	
Side-chain planarity	1.310 ± 0.183	
Improper dihedral distribution	1.472 ± 0.056	
Inside/outside distribution	0.935 ± 0.023	
Ramachandran plot		
	Secondary structures[b] (%)	All residues[d] (%)
Residues in core regions	98.7	88.4
Residues in allowed regions	1.3	9.5
Residues in generously allowed regions	0.0	1.2
Residues in disallowed regions	0.0	1.0

RMS root mean square

[a] Positive is better than average

[b] Ranges 17–32, 43–53, 64–71, and 81–94 for both monomers

[c] Should be close to 1.0

[d] Range 17–94 for both monomers

Relaxation data

The experiments for the determination of ^{15}N longitudinal and transverse relaxation rates and ^{1}H–^{15}N NOE [28] were performed at 298 K and 700 MHz on a ^{15}N-enriched sample of S100A14. The ^{15}N longitudinal relaxation rates (R_1) were measured using a sequence modified to remove cross-correlation effects during the relaxation delay [29]. Inversion recovery times ranging between 20 and 2,200 ms, with a recycle delay of 3.5 s, were used for the experiments. The ^{15}N transverse relaxation rates (R_2) were measured using a Carr–Purcell–Meiboom–Gill sequence [29, 30] with delays ranging between 8.48 and 203.52 ms and with a refocusing delay of 450 μs. The longitudinal and transverse relaxation rates were determined by fitting the cross-peak intensities as a function of the delay to a single-exponential decay using the program Origin. The heteronuclear NOE values were obtained from the ratio of the peak height for ^{1}H-saturated and unsaturated spectra. The relaxation data are reported in Table S3.

Estimates of the reorientation time were then obtained with the program TENSOR2 [31]. Theoretical predictions of NH R_1 and R_2 values for apo-S100A14 were made using the program HYDRONMR [32].

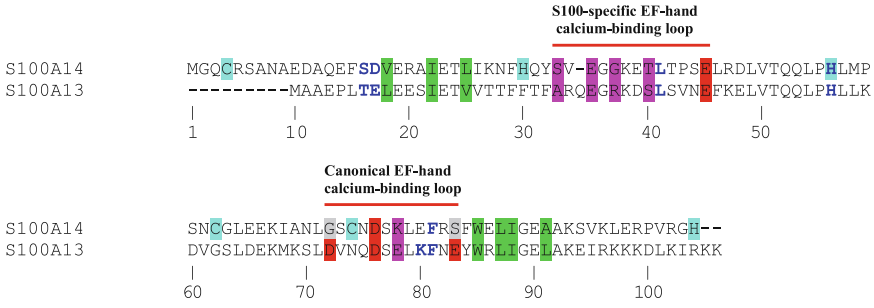
Dynamic light scattering

An online multiangle light scattering detector (DAWN EOS, Wyatt Technology, Santa Barbara, CA, USA) and differential refractive index detector (Optilab, Wyatt Technology) setup was used to measure the light scattered as a function of the angle and absolute protein concentration of fractions eluted from the size-exclusion chromatography column. The Zimm/Debye approximations were used in the program Astra (Wyatt Technology) to estimate the molar mass. Data were fit using a second-order polynomial. All measurements were done at 298 K. Human S100A14 in 30 mM 2-morpholinoethanesulfonic acid (pH 6.5), 100 mM NaCl, and 5 mM DTT, at a concentration of 0.1 mM, was filtered (0.1 μm) before the injection. The concentration of the S100A14 in solution was determined using a specific refractive index increment (dn/dc) of 0.180 at 690 nm. The analysis shows that S100A14 is dimeric in solution ($M_{avg} = 2.308 \times 10^4$ g/mol), although a small amount of tetrameric or higher-order-aggregation species is also present (see Fig. S1).

Results

The sequence alignment of S100A14 with its closest homolog S100A13 shows that in both N- and C-terminal domains some residues, responsible for the binding of calcium(II) in S100A13 as well as in other S100 proteins, are mutated. These mutations mostly occur at the C-terminal, on the canonical EF-hand calcium(II)-binding site, where two key residues for calcium(II) chelation (aspartate and glutamate placed at the 1 and 12 coordinating positions, respectively) are replaced by glycine and serine (Scheme 1). The replacement of the chelating residues at the canonical EF-hand calcium(II)-binding site accounts for the lack of affinity for calcium(II) ion observed in S100A14. Conversely, the hydrophobic core residues are strongly conserved in S100A14 [4, 33] (Scheme 1), and this suggests that the protein exists as a homodimer in solution. Moreover, S100A14 contains an extended positively charged N-terminal tail, which is uncommon in S100 proteins and could be involved in its specific biological activity.

Concerning the binding capability of S100A14 with regard to other metal ions, the analysis of the sequence shows the presence of three cysteines and three histidines that could be involved in the binding of metals, such as copper(II) and zinc(II) ions. In particular the sequence



Scheme 1 Sequence alignment of human S100A13 and S100A14 proteins. The Ca^{2+}-coordinating residues are highlighted in *red* (side-chain coordination) and *magenta* (backbone-oxygen coordination); the mutated residues of S100A14 in the place of the residues for calcium(II) binding are highlighted in *gray*. Hydrophobic residues that are essential for dimerization are highlighted in *green*. Residues that are putative copper(II)-binding sites are colored *blue* (Ser-16, Asp-17, Leu-41, His-56, and Phe-81 in S100A14) and the putative zinc(II) ligands of S100A14 are highlighted in *cyan* (Cys-4, His-30, His-56, Cys-62, Cys-74, and His-104)

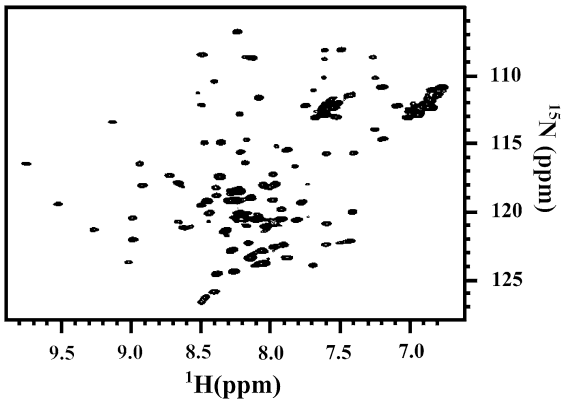


Fig. 1 Two-dimensional ^{1}H–^{15}N heteronuclear single quantum coherence spectrum of apo-S100A14 at 310 K

alignment shows that the two residues His-48 and Glu-10, which are supposed to be involved in the binding of copper(II) in S100A13 [16], are conserved (His-56 and Glu-19) in S100A14 (Scheme 1).

NMR resonance assignments

The 2D ^{1}H–^{15}N HSQC NMR experiments performed on S100A14 at 298 and 310 K show that the quality of the spectra improves with the temperature. A detailed analysis of the spectra acquired at 310 K (see Fig. 1) shows two different groups of cross-peaks, one with sharp and intense signals, characterized by low chemical shift dispersion, and a second characterized by broad and low-intensity, well-dispersed resonances in the proton and nitrogen dimensions. No splitting of signals was observed in the spectra. To assign the protein residues at 310 K, a combination of standard ^{1}H-detected and ^{13}C-detected experiments were performed on ^{13}C/^{15}N-enriched samples of S100A14. All H^N resonances but the first four residues (1–4) in the N-terminal region were easily assigned. Protein assignment was also obtained at 298 K starting from the data collected at 310 K. Three of the four unassigned H^N cross-peaks at 310 K (residues 2–4) were easily identified at low temperature. Conversely, at this temperature (298 K), some weak H^N peaks previously assigned at 310 K (Phe-81, Phe-84, Trp-85, Ala-91, Val-95, Val-98) further decrease in intensity beyond detection.

Secondary structure prediction was done with the program TALOS+ by using the chemical shifts of H^N, N, C′, C^{α}, H^{α}, and C^{β} as input data. The prediction suggests the presence of four helices (residues 17–32, 43–53, 64–70, 81–93) and three interconnecting loops (residues 33–42, 54–63, 71–80). Moreover, the program predicts high mobility for the N- and C-terminal regions (residues 1–16, 94–104) of S100A14.

Ca^{2+}, Zn^{2+}, and Cu^{2+} titration of apo-S100A14

The affinity of metals for S100A14 was evaluated by NMR investigations. The apoprotein (0.4 mM) was titrated with increasing amounts of calcium(II), up to a final concentration of 80 mM. The binding of calcium(II) ion to the protein was monitored by following the changes in the 2D ^{1}H–^{15}N HSQC NMR spectrum of ^{15}N-labeled apo-S100A14. The analysis of the chemical shift perturbation indicates that only very high concentrations of calcium(II) ions affect a few residues located in the EF-hand loops forming the putative calcium-binding sites at the N-terminal. The residues undergoing the largest changes in chemical shifts belong to the metal-binding site in the loop between helix I and helix II, in the so-called S100-specific EF-hand calcium-binding loop. In particular, residues Glu-35, Gly-36, Gly-37, Thr-40, and Thr-42 experience the largest effect, but also Gly-72 and Ser-77 in the canonical EF-hand calcium-binding loop are slightly affected. All these peaks continuously change their chemical shifts upon an increase in the concentration of calcium(II) ion, as for systems in the fast exchange regime, and the changes were not completed even in the presence of 80 mM calcium(II). The small chemical shift perturbations and the negligible affinity of S100A14 for calcium(II) ion indicate that only the apo form of the protein has a physiological relevance since even the extracellular concentration of calcium(II) ion is only about 1–2 mM. The chemical shift variation upon addition of increasing amounts of calcium(II) ions is shown in Fig. S2.

To prove the binding capability of S100A14 with regard to zinc(II) and copper(II), solutions of the apoprotein (0.2 mM) were treated with 5 mM DTT. The DTT was then removed and two samples of the protein were titrated with increasing amounts of zinc(II) and copper(II) ions, respectively, in the absence of oxygen. Protein precipitation occurred upon addition of the metals, with a sizable decrease of the intensity of all the cross-peaks in the 2D ^{1}H–^{15}N HSQC spectra. The observed precipitation and the general decrease in signal intensity is open to different explanations. In particular, the absence of high-affinity binding sites for both copper(II) and zinc(II) on S100A14 and the interaction of these metal ions with surface-exposed cysteines/histidines (Cys-4, Cys-62, Cys-74, His-30, His-56, His-104) could cause protein oligomerization. However, it is also possible that the resulting complex of S100A14 with copper(II) or zinc(II) is insoluble.

^{15}N relaxation measurements

The relaxation data for S100A14 are shown in Fig. S3 in a per-residue plot. The quality of the relaxation data is modest for the signals of the residues belonging to helices, owing to their low intensity and to the overlap with peaks

of the unfolded loops, but is still sufficient for the present purposes. The correlation time for molecular reorientation as estimated from the R_2/R_1 ratio for the secondary structure elements is 12.3 ± 0.3 ns, as expected for the molecular weight of the homodimeric protein. This value is also in agreement with the reorientation times observed for other S100 homodimeric proteins [12, 16, 34–38]. After S100A9 and S100A11, S100A14 exhibits the longest sequence among the members of the family. The relaxation data show that the first residues (1–16) at the N-terminal and the last residues (95–104) at the C-terminal are poorly structured as a consequence of their fast internal mobility, revealed by the small or negative $^{1}H–^{15}N$ NOE values, as well as by the large R_1 and the small R_2 values. Such values indicate that these residues experience local motions on the nanosecond to picosecond timescale (faster than the overall protein-tumbling rate) and behave essentially as unfolded random-coil peptides. Fast motion is also detected for some residues at the beginning of helix II (i.e., Thr-40), for the linker between the two EF-hand motifs (Asn-61, Cys-62, Gly-63), and for some residues in the loop before helix IV (Gly-72, Asp-76, Ser-77). Conversely, some residues located in helix IV (Arg-82, Ser-83, Leu-87), in the C-terminal of helix I (Tyr-32), and in helix II (Leu-49, Val-50) experience motions on a slower timescale (microsecond to millisecond timescale), as indicated by their significantly larger R_2 values (around 50 s^{-1}) compared with the average values observed for the residues belonging to the secondary structures (25 s^{-1}).

Solution structure of apo-S100A14

The solution structure of human S100A14 in the apo form was calculated from a total of 811 meaningful intrasubunit upper distance limits per monomer and 132 intersubunit upper distance limits. Only a few intraresidues and sequential NOEs were detected for loop L1, between helix I and helix II, for loop L2, between helix III and helix IV, and for the residues in the N- and C-terminal regions. Conversely, some NOEs were found in particular between the N-terminal part (53–57) of the linker connecting the two EF-hand domains (also named "hinge loop"), and helices I′ and IV.

The family of 30 structures, calculated with CYANA by imposing the dimer symmetry constraint, were minimized with AMBER. The 30 minimized structures have a root mean square deviation (RMSD) from the mean structure (taking into account only the structured regions of the dimeric protein) of 0.68 ± 0.17 Å for the backbone and 1.08 ± 0.16 Å for the heavy atoms.

PSVS and iCing were used to validate the structure. The Ramachandran plot shows φ–ψ torsion angles inside the core region for 98.7 % of the residues belonging to the secondary structures and inside the allowed areas for 1.3 % of the residues (the statistical analysis is reported in Table 2). The not excellent quality of the statistical analysis is common to many S100 proteins and is probably related to the conformational heterogeneity of these proteins. The experimental relaxation rates for the secondary structure elements are in overall agreement with the values calculated with the program HYDRONMR, starting from the minimized mean structures and under the assumption of no internal motions (see Fig. S3). The good agreement between the calculated and the experimental data again confirms that S100A14 is homodimeric in solution.

The calculated family of structures with the unstructured N- and C-terminal regions is shown in Fig. 2a. The four helices of the two EF-hand motifs of each subunit are well defined, although helices III and III′ and IV and IV′ exhibit a higher RMSD per residue. In particular, the RMSD per residue of the family of structures relative to the mean is 0.44, 0.53, 0.94, and 0.57 for helices I, II, III, and IV, respectively. These data suggest a higher conformational heterogeneity of helix III related to a small precession of this helix around its axis in solution even in the absence of calcium binding related to the slight opening of the structure (Fig. 3). Reorientation of helix III usually occurs in S100 proteins passing from the apo to the holo state which is characterized by a minor conformational heterogeneity [39, 40]. The long and unstructured N- and C-terminal tails that are present in S100A14 presumably account for our failure to obtain crystals for X-ray diffraction.

At this point, it is instructive to compare the plot of the RMSD per residue with the plot of R_1. The comparison shows that the region where a larger structural disorder is present (i.e., loops L1 and L2 and the hinge loop) both have high RMSDs and relatively high R_1 values. This indicates that the observed structural disorder is at least partly due to a real local mobility and not only to lack of NOEs. In fact, the residues that mostly contribute to the overall percentage of disallowed residues in both monomers of the homodimeric protein are Lys-78, located in loop L2, Asn-61 in the hinge loop, and Lys-38 in loop L1. Therefore, the observed deviations from Ramachandran ideality are most probably an artifact due to mobility.

Discussion

In homodimeric S100 proteins, the two subunits are related by a twofold axis of rotation. The major contributors to the dimer interface are helices I and IV of each subunit, which form an X-type four-helix bundle [12, 41]. These structural features are maintained also in S100A14. A detailed analysis of the solution structure shows that in each monomer of S100A14, the contacts between the aromatic

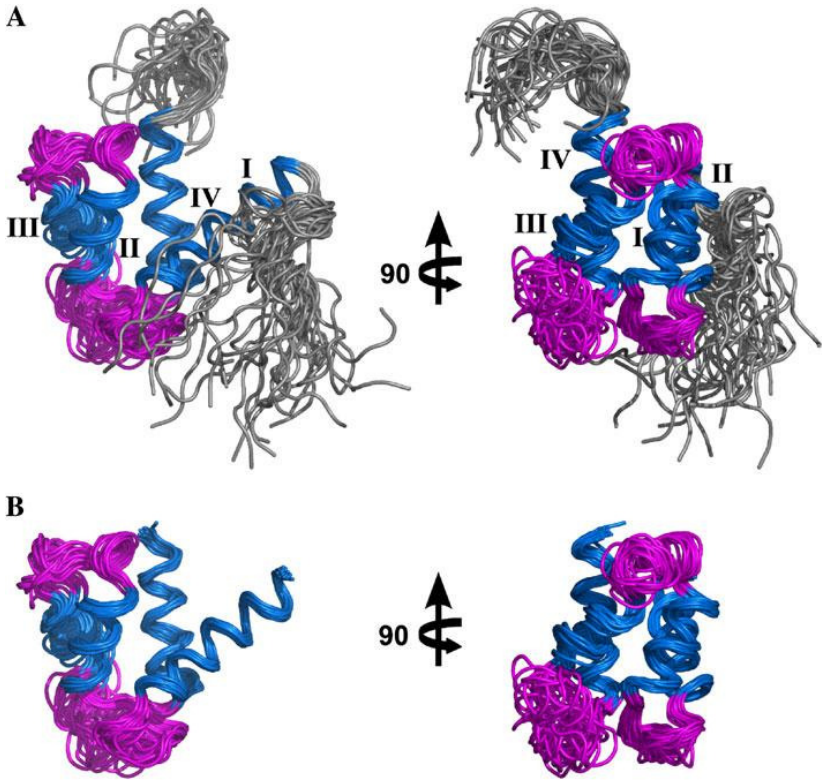


Fig. 2 Ribbon representation of the AMBER-minimized solution structures of a monomer of S100A14. The structures corresponding to the whole protein sequence and to protein stretch 17–94 are reported in **a** and **b**, respectively

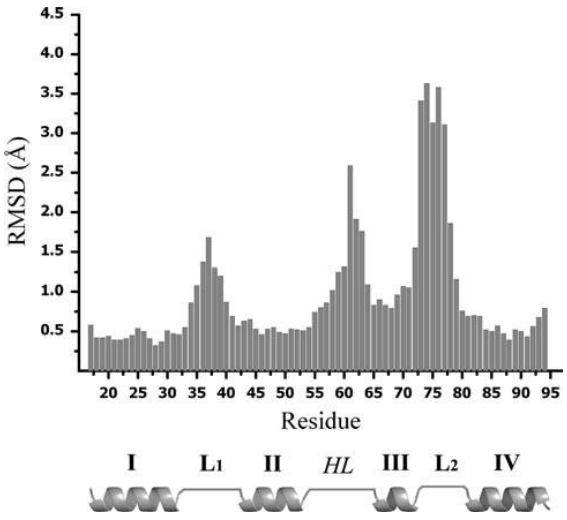


Fig. 3 Root mean square deviation (*RMSD*) per residue of the family of structures after AMBER minimization. Higher RMSDs occur in helices III and IV

residue Phe-29 in helix I and residues Phe-80, Phe-84, and Trp-85 in helix IV contribute to the orientation of the helices, and sustain the interaction between the two subunits (Fig. 4). The monomer–monomer adduct is stabilized by several interactions between the hydrophobic residues in helix I and those present in helix I′ and helix IV′ (Fig. 5). The adduct is further stabilized by a few interactions involving residues in helices IV and IV′. Finally, the interaction between the two monomers is reinforced by the contacts established by residues 53–57 in the hinge loop with the residues at the N-terminal of helix I′ and the C-terminal part of helix IV. The role of the hinge loop in the stabilization of the homodimeric structure is in agreement with the relaxation data and in particular with the $^{1}H–^{15}N$ NOE values, which indicate mobility on the nanosecond to picosecond timescale only for the last residues (60–63) of the loop. Collectively, this network of interactions orients helices I and IV in opposite directions to helices I′ and IV′, respectively.

It is thus apparent that S100A14 adopts a conformation that differs from that of both the canonical apo form and the calcium(II)-loaded form [16, 42–44] Helices II and III are almost antiparallel as in the canonical apo form of S100 proteins, but superimposition of the present structure on the structures of the apo and holo forms of S100A13 reveals a better structural agreement with the latter (Fig. 6a, b). In particular, the N-terminal part of helix III points towards helix II, assuming a conformation similar to that of calcium(II)-loaded S100A13 [45]. Similar structural features

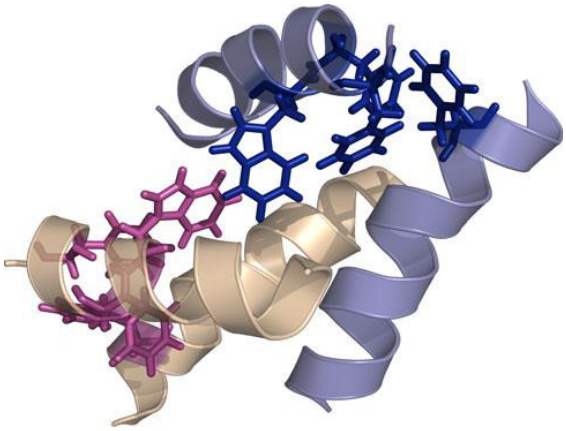


Fig. 4 The interacting aromatic residues in helix I (Phe-29) and helix IV (Phe-81, Phe-84, Trp-85) stabilize the interhelical orientation

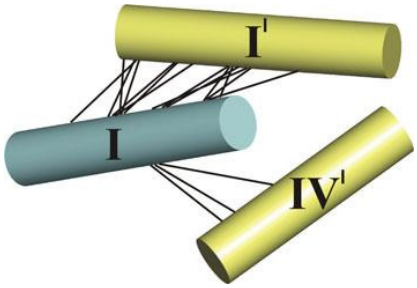


Fig. 5 Interdimer contacts between helices I and I′ and helices I and IV′

have been observed in S100A10, where the apo form is in the open conformation [11]. The comparison of the structures of S100A14 and S100A10 reveals that the first EF-hand motifs are completely superimposable on one another, with a backbone RMSD (residues 17–53 for S100A14 and residues 3–37 for S100A10) of 0.96 Å. Major differences are instead found in loop L1, which is two amino acids shorter in S100A10 (Fig. 6c). In the second EF-hand motif, helix IV is shorter in S100A14 than in S100A10, but the directions of the axis of the helices (III and IV) are very similar. A high sequence and structure similarity is also present between the two proteins in the canonical EF-hand calcium(II)-binding site, and highlights their common peculiarity of very weak affinity for calcium(II) ion.

The sequence alignment of S100 proteins shows details about the evolution of S100A14, and important structural features of the protein. The phylogenetic tree generated on the basis of multiple alignments of entire S100 sequences shows that S100A14 and S100A13 form a group on their own, well separated from the rest. The residues responsible for the binding of calcium(II) in S100 proteins, including S100A13, are mutated and highly conserved in all mammalian S100A14 proteins. This suggests that any physiological and/or pathological activity of S100A14 is not regulated by the concentration of calcium(II) ion as for most S100 proteins. The mutations involving the residues forming the C-terminal calcium-binding site provide new insight into the evolution of S100A14. In fact, although expression of the S100A14 gene has been reported only in mammals, an S100A14-like protein is present in the genome of the evolutionary predecessor *Meleagris gallopavo*. In such a protein, one of the four calcium-binding residues is still in place as in S100A13. This might suggest for S100A14 an evolutionary process characterized by a progressive loss of affinity for calcium(II) ions. In addition, human S100A14 exhibits the longest N-terminal tail among S100 proteins, and this remains largely unfolded and flexible in solution.

In S100 proteins, the C-terminal EF-hand motif binds calcium(II) with a pattern of highly conserved residues placed at positions 1, 3, 5, 7, and 12 that form the well-characterized binding motif D-x-N-x-D-x-(E/K/R/Q/A)-x-x-x-x-E [46]. In this metal-binding site, the calcium ion is bound to Asp-1, Asn-3, Asp-5, and Glu-12 side chains and to the carbonyl oxygen of the less conserved amino acid at position 7. The dissociation constants for the different mutants suggest that the replacement of one or more amino acids at positions 1, 3, and 12 affects dramatically the metal-binding affinity of the C-terminal EF-hand motif [47] (Table 3). In S100A14, the replacement of all three residues at positions 1, 3, and 12 with glycine, cysteine, and serine, respectively, resulting in the loss of three coordinating ligands, explains the negligible affinity for calcium. Apparently, the lack of affinity of the C-terminal EF-hand motif also prevents calcium binding at the N-terminal site, despite the residues forming the metal binding motif and the critical glutamate all being in place. This is further and clear proof of the cooperativity of calcium(II) binding within the S100 protein family.

A detailed structural comparison of S100A14 with all the S100 family members can be performed by a PCA of the angles among the α-helices [12, 13, 37]. In each EF-hand domain the six interhelical angles are measured from the directions of the four α-helices by considering the eight residues immediately preceding and following each EF-hand loop. For solution NMR structures, such values are calculated from the mean NMR structure and the corresponding errors are calculated from the standard deviation observed within the structures of the family (Table 4). As shown in Table 4, some angles of S100A14 are halfway between the values typical of apo and calcium(II)-loaded S100 proteins [13]. In fact, the intramotif angles I–II and I–III, which determine the so-called closed and open forms of the EF-hand domains, have values (133° for I–II and 82° for I–III) intermediate between those of the closed

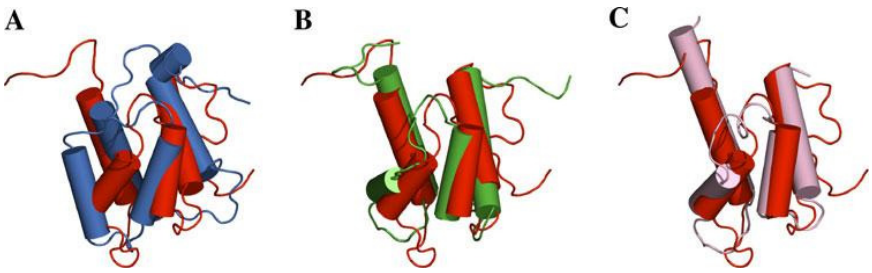


Fig. 6 Superimposition of S100A14 on apo-S100A13 (**a**), holo-S100A13 (**b**), and S100A10 (**c**). In *red* S100A14, in *blue* apo-S100A13, in *green* holo-S100A13, and in *pink* S100A10

Table 3 Calcium(II) affinity constants for the S100 protein family

S100	Ca^{2+} sites	N-terminal Ca K_{d1} (mM)	C-terminal Ca K_{d2} (mM)	Generic K_d (mM)	References
S100A1	4	>0.25	0.027		[52]
S100A2	4	0.25	0.060		[53]
S100A3	ND			4–35	[54]
S100A4	4	>0.570	0.003 ± 0.001		[55]
S100A5	4	0.16	0.0002		[56]
S100A6	4			0.30–0.60	[57]
S100A7	2		0.001		[56]
S100A10	None				[11]
S100A11	4			0.52	[58]
S100A12	2		0.05		[15]
S100A13	4	0.066 ± 0.001	0.008 ± 0.001		[59]
S100A14	None				This work
S100A16	4			0.27 ± 0.02	[12]
S100B	4	0.35	0.060		[60]
S100P	4	0.8	0.002		[61]
S100Z	4	>1	0.0002		[60]

ND not determined

(110°–130° for I–II and 35°–110° for I–III) and open (130°–140° for I–II and 90°–130° for I–III) conformations. Also the value of the angle between helix III and helix IV (130°), which provides important additional information for the classification of the structure, is intermediate between that of the two conformations [100–150° for the calcium(II)-loaded form, 120–160° for the apo form] even if a bit closer to that of the open conformation. Instead, the value of the angle between helices II and III (144°) falls in the range typical of the apo conformation (110–165°). Helices II and IV are almost parallel (33°), whereas helix III is neither strictly parallel nor perpendicular to them but tilted, with its N-terminal oriented towards helix II and its C-terminal oriented towards helix IV. These interhelical angle ranges show the broad conformational space sampled by helix III in both the apo structure and the calcium(II)-loaded structure as previously observed for other S100 proteins [40].

The principal component values [13] for S100A14 were plotted together with the values previously calculated for all the other S100 proteins [12, 37] (Fig. 7a) by using the same coefficients for the interhelical angles reported by Babini et al. [13]. The first principal component and second principal component values of S100A14 fall in the middle of the two clusters defined by the open and closed conformations of S100 proteins, although closer to the latter. This plot provides a visual representation of the extent to which the conformation of S100A14 in solution is intermediate between the open and the closed conformations. S100A14 can be thus considered as another example of the continuum of conformational states occurring within the whole EF-hand protein superfamily (Fig. 7b). This variety of structures reflects the different functions in which EF-hand proteins are involved, and underlines the difficulty of directly obtaining sequence–structure or sequence–function relationships.

The analysis of the interhelical angles is a good method for the classification of the EF-hand conformations and is free from user bias. However, it does not provide detailed information on the packing of the helices that constitute the protein core [7]. Such information can be obtained from the analysis of the contacts among the α-helices. The analysis of the NOEs shows that the N-terminal of helix III

Table 4 Angles between different helices for S100A14; the directions of the helices are defined by the eight residues immediately preceding or following each EF-hand loop and were calculated from the mean solution NMR structure (the errors were calculated from the standard deviations within the 30 structures of the family)

Helices	Angle (°)
I/II	133 ± 4
I/III	82 ± 6
I/IV	123 ± 4
II/III	144 ± 6
II/IV	33 ± 4
III/IV	130 ± 5
I/I′	158 ± 4
IV/IV′	148 ± 3

establishes contacts with the central and C-terminal part of helix II (NOEs of Leu-64 with Thr-51 and Arg-47 and of Lys-67 with Leu-46), whereas the C-terminal region faces the central part of helix IV (NOEs between Lys-67 and Leu-71 with Leu-87). This network of interactions stabilizes an offset X-shaped arrangement of helices III and IV that has been observed in the EF-hand domains with a semi-open conformation. This semi-open structural arrangement of the EF-hand domain is not an intermediate state between the closed and open conformations but represents a distinct conformational state. In fact, in the transition from the closed to the open conformation, helices III and IV form an inverted-V-shaped interface, whereas in the semi-open conformation the same helices adopt an offset X-shaped arrangement. At the same time, helices II and helix III create a V-shaped interface stabilized by contacts involving the residues at the C-terminal of helix II and the N-terminal of helix III (Val-50 and Thr-51 shows NOEs with Leu-64). To further characterize the semi-open conformation of S100A14, the number of interhelical NOEs should also be taken into account [39]. In particular, very few interhelical NOEs have been observed for the helix II/helix IV and helix III/helix IV pairs. Conversely, helices II and III are very close to each other. Collectively, these structural data indicate that only the interface between helices II and III is conserved in S100A14 as previously observed in the semi-open structure of the C-terminal domain of myosin light chains [48].

The electrostatic potential surface in the protein region corresponding to helix III, the hinge loop, and helix IV seems to play a pivotal role for the biological activity of S100 proteins. Indeed, in some members of the family such as S100B, protein activation occurs upon the transition from the close" to the open conformation with the exposure of hydrophobic patches and a wide negatively charged surface that promotes the interaction with the protein partners [21, 49, 50]. The analysis of the charge distribution shows that S100A14 exhibits a negatively charged surface on helix III and helix IV already in the semi-open conformation and suggests that the protein is "permanently" activated (Fig. 8).

The packing density provides important information on protein stability and functionality. The analysis of atomic-scale packing for the human S100 proteins has been performed using the Voronoia toolkit with the PyMOL plug-in [51], starting from the available structures. Most S100 proteins in the closed conformation have a higher average packing density with respect to their corresponding open form (Fig. 9). In fact, after calcium(II) binding, the exposure to the solvent of the hinge loop, the C-terminal of helix IV, and some residues in helix III sizably increases.

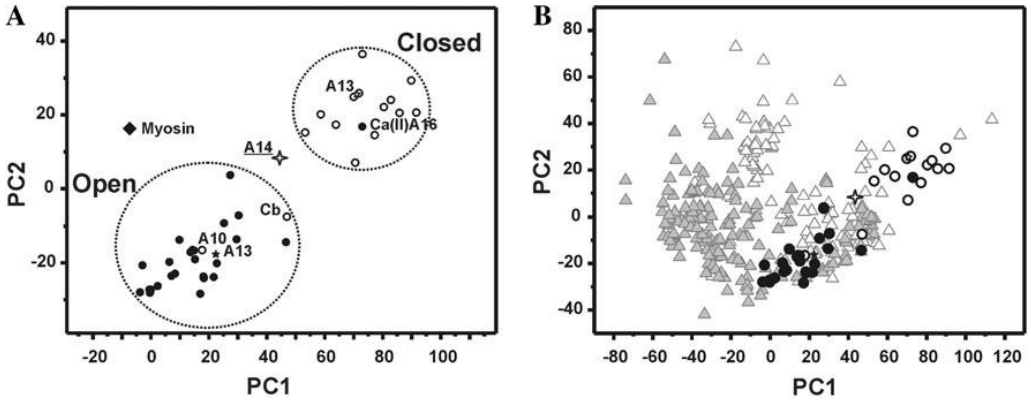


Fig. 7 **a** Principal component analysis performed on the six interhelical angles of the apo and calcium(II)-loaded forms of S100 proteins. Apo and holo proteins are represented as *open circles* and *filled circles*, respectively. The two open circles not regularly placed with respect to the others correspond to the apo form of calbindin D9k (*Cb*) and S100A10 (*A10*). The filled circle placed in the region of the protein in the closed conformation corresponds to holo-S100A16 (*A16*). S100A14 (*open star*; *A14*) is in a "semi-open" conformation in physiological conditions. Apo and holo forms of S100A13 (*A13*) are reported as an *open asterisk* and a *closed asterisk*, respectively. Myosin light chain is reported as a *diamond*. **b** Principal component plot of the whole EF-hand domain dataset derived from principal component analysis of the six interhelical angles. Apo and holo proteins are represented as *open triangles* and *closed triangles*, respectively. S100 proteins are represented by the same *symbols* as in **a**. *PC* principal component

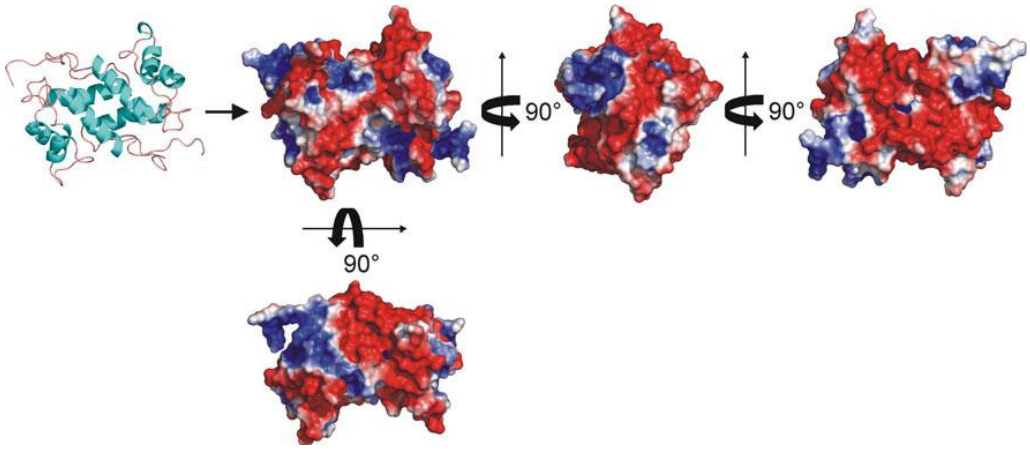


Fig. 8 Surface charge representation of S100A14, with *blue* and *red* representing regions of positive and negative electrostatic potential, respectively. All electrostatic surfaces were generated with the APBS plug-in of PyMOL

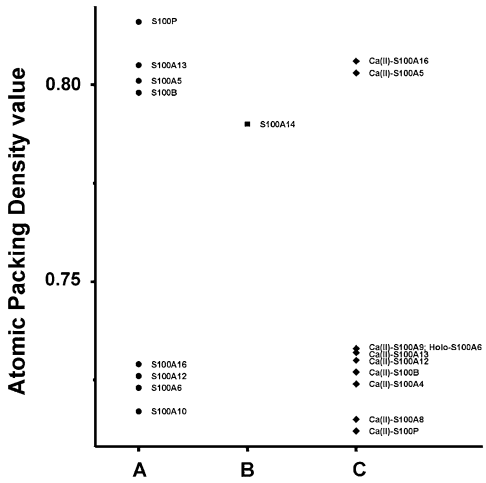


Fig. 9 Analysis of atomic-scale packing for human S100 proteins using the Voronoia toolkit with the PyMOL plug-in. The packing densities for S100 proteins in the "closed" and "open" conformations are reported in columns *A* and *C*, respectively. The packing density for S100A14 is reported in column *B*

In this respect, the analysis performed on S100A14 shows that in this protein the packing density is similar to the average observed for S100 proteins in the closed conformation. The high packing density can be explained by the arrangement of the hinge loop, where the N-terminal fragment is more ordered and buried in the protein.

Conclusion

We have presented the solution structure of human S100A14. The protein has a high degree of sequence homology with S100A13, although it exhibits a negligible affinity for calcium(II) ions and undergoes aggregation and precipitation in the presence of zinc(II) or copper(II) ions. The protein shows close structural similarities with the C-terminal domain of myosin light chains and folds in a semi-open conformation, with helices III and IV showing an offset X-shaped arrangement and helices II and III showing a high number of interactions creating a V-shaped arrangement. The most remarkable feature of this structural conformation involves the packing of the helices, which is reduced with respect to the closed structures of S100 proteins but is still sizably larger than for the corresponding open structures. At the same time, the analysis of the electrostatic potential surface suggests that the protein is permanently activated and it is not calcium(II)-regulated.

Acknowledgments We thank Leonardo Gonnelli for light-scattering measurements. This work was supported by MIUR-FIRB contracts RBLA032ZM7, RBRN07BMCT, and RBIP06LSS2, and by the European Commission, contracts Bio-NMR no. 261863, East-NMR no. 228461, STREP–SFMET no. 201640, WeNMR no. 261572, and INSTRUCT.

References

1. Marenholz I, Lovering RC, Heizmann CW (2006) Biochim Biophys Acta Mol Cell Res 1763:1282–1283
2. Donato R (1986) Cell Calcium 7:123–145
3. Kojetin DJ, Venters RA, Kordys DR, Thompson RJ, Kumar R, Cavanagh J (2006) Nat Struct Mol Biol 13:641–647
4. Marenholz I, Heizmann CW, Fritz G (2004) Biochem Biophys Res Commun 322:1111–1122
5. Heizmann CW, Fritz G (2004) Handbook on metalloproteins, 1st edn. Dekker, New York
6. Donato R (2001) Int J Biochem Cell Biol 33:637–668
7. Nelson MR, Chazin WJ (1998) Protein Sci 7:270–282

8. Bhattacharya S, Large E, Heizmann CW, Hemmings B, Chazin WJ (2003) Biochemistry 42:14416–14426
9. Okada M, Tokumitsu H, Kubota Y, Kobayashi R (2002) Biochem Biophys Res Commun 292:1023–1030
10. Agamennone M, Cesari L, Lalli D, Turlizzi E, Del Conte R, Turano P, Mangani S, Padova A (2010) ChemMedChem 5:428–435
11. Rety S, Sopkova J, Renouard M, Osterloh D, Gerke V, Tabaries S, Russo-Marie F, Lewit-Bentley A (1999) Nat Struct Biol 6:89–95
12. Babini E, Bertini I, Borsi V, Calderone V, Hu X, Luchinat C, Parigi G (2011) J Biol Inorg Chem 16:243–256
13. Babini E, Bertini I, Capozzi F, Luchinat C, Quattrone A, Turano M (2005) J Proteome Res 4:1961–1971
14. Capozzi F, Luchinat C, Micheletti C, Pontiggia F (2007) J Proteome Res 6:4245–4255
15. Moroz OV, Burkitt W, Wittkowski H, He W, Ianoul A, Novitskaya V, Xie JJ, Polyakova O, Lednev IK, Shekhtman A, Derrick PJ, Bjoerk P, Foell D, Bronstein IB (2009) BMC Biochem 10(11):1–18
16. Arnesano F, Banci L, Bertini I, Fantoni A, Tenori L, Viezzoli MS (2005) Angew Chem Int Ed 44:6341–6344
17. Heizmann CW, Cox JA (1998) Biometals 11:383–397
18. Pietas A, Schluns K, Marenholz I, Schafer BW, Heizmann CW, Petersen I (2002) Genomics 79:513–522
19. Theorell H, Ehrenberg A (1952) Arch Biochem Biophys 41:442–461
20. Jin Q, Chen H, Luo A, Ding F, Liu Z (2011) PLoS ONE 6:e19375
21. Leclerc E, Fritz G, Vetter SW, Heizmann CW (2009) Biochim Biophys Acta 1793:993–1007
22. Keller RLJ (2004) The computer aided resonance assignment tutorial. Cantina Verlag, Goldau
23. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) J Biomol NMR 44:213–223
24. Herrmann T, Guntert P, Wuthrich K (2002) J Mol Biol 319:209–227
25. Ponder JW, Case DA (2003) Adv Protein Chem 66:27–85
26. Bertini I, Case DA, Ferella L, Giachetti A, Rosato A (2011) Bioinformatics 27:2384–2390
27. Bhattacharya A, Tejero R, Montelione GT (2007) Proteins 66:778–795
28. Barbato G, Ikura M, Kay LE, Pastor RW, Bax A (1992) Biochemistry 31:5269–5278
29. Kay LE, Torchia DA, Bax A (1989) Biochemistry 28:8972–8979
30. Peng JW, Wagner G (1994) Methods Enzymol 239:563–596
31. Dosset P, Hus JC, Marion D, Blackledge M (2001) J Biomol NMR 20:223–231
32. Garcia de la Torre JG, Huertas ML, Carrasco B (2000) J Magn Reson 147:138–146
33. Zhou T, Zheng X, Yi D, Zhang Q (2009) Molecular modeling and structure analysis of S100 calcium binding protein A14: molecular modeling and structure analysis of S100A14. In: Proceedings of 2nd international conference on biomedical engineering and informatics, BMEI 2009, Tianjin, China, 17–19 October 2009, pp 1775–1778
34. Inman KG, Baldisseri DM, Miller KE, Weber DJ (2001) Biochemistry 40:3439–3448
35. Zhukov I, Ejchart A, Bierzynski A (2008) Biochemistry 47:640–650
36. Dutta K, Cox CJ, Basavappa R, Pascal SM (2008) Biochemistry 47:7637–7647
37. Bertini I, Dasgupta S, Hu X, Karavelas T, Luchinat C, Parigi G, Yuan J (2009) J Biol Inorg Chem 14:1097–1107
38. Bertini I, Fragai M, Luchinat C, Parigi G (2000) Magn Reson Chem 38:543–550
39. Maler L, Sastry M, Chazin WJ (2002) J Mol Biol 317:279–290
40. Malik S, Revington M, Smith SP, Shaw GS (2008) Proteins 73:28–40
41. Rustandi RR, Baldisseri DM, Inman KG, Nizner P, Hamilton SM, Landar A, Landar A, Zimmer DB, Weber DJ (2002) Biochemistry 41:788–796
42. Otterbein L, Kordowska J, Witte-Hoffmann C, Wang CL, Dominguez R (2002) Structure 10:557–567
43. Smith SP, Shaw GS (1998) Structure 6:211–222
44. Drohat AC, Baldisseri DM, Rustandi RR, Weber DJ (1998) Biochemistry 37:2729–2740
45. Imai FL, Nagata K, Yonezawa N, Nakano M, Tanokura M (2008) Acta Crystallogr Sect F Struct Biol Cryst Commun 64:70–76
46. Zhou Y, Yang W, Kirberger M, Lee HW, Ayalasomayajula G, Yang JJ (2006) Proteins 65:643–655
47. Vogel HJ, Brokx RD, Ouyang H (2002) In: Vogel HJ (ed) Calcium-binding protein protocols: volume 1: reviews and case studies. Springer, New York
48. Houdusse A, Cohen C (1996) Structure 4:21–32
49. Santamaria-Kisiel L, Rintala-Dempsey AC, Shaw GS (2006) Biochem J 396:201–214
50. Rustandi RR, Baldisseri DM, Weber DJ (2000) Nat Struct Biol 7:570–574
51. Rother K, Hildebrand PW, Goede A, Gruening B, Preissner R (2009) Nucleic Acids Res 37:D393–D395
52. Wright NT, Varney KM, Ellis KC, Markowitz J, Gitti RK, Zimmer DB, Weber DJ (2005) J Mol Biol 353:410–426
53. Koch M, Bhattacharya S, Kehl T, Gimona M, Vasak M, Chazin W, Heimann CW, Kroneck PMH, Fritz G (2007) Biochim Biophys Acta Mole Cell Res 1773:457–470
54. Fritz G, Mittl PRE, Vasak M, Grutter MG, Heizmann CW (2002) J Biol Chem 277:33092–33098
55. Garrett SC, Hodgson L, Rybin A, Toutchkine A, Hahn KM, Lawrence DS, Bresnick AR (2008) Biochemistry 47:986–996
56. Schafer BW, Fritschy JM, Murmann P, Troxler H, Durussel I, Heizmann CW, Cox JA (2000) J Biol Chem 275:30623–30630
57. Kordowska J, Stafford WF, Wang CLA (1998) Eur J Biochem 253:57–66
58. Allen BG, Durussel I, Walsh MP, Cox JA (1996) Biochem Cell Biol 74:687–694
59. Sivaraja V, Kumar TKS, Rajalingam D, Graziani I, Prudovsky I, Yu C (2006) Biophys J 91:1832–1843
60. Zimmer DB, Weber DJ (2010) Cardiovasc Psychiatry Neurol 2010:1–17
61. Becker T, Gerke V, Kube E, Weber K (1992) Eur J Biochem 207:541–547

Solution structure and dynamics of human S100A14*

Ivano Bertini[1,2,‡], Valentina Borsi[1], Linda Cerofolini[1], Soumyasri Das Gupta[1], Marco Fragai[1,2], Claudio Luchinat*[,1,2]

[1] Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy.

[2] Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy.

Address correspondence to: Claudio Luchinat via L. Sacconi 6, 50019 Sesto Fiorentino, Italy, Tel.: +390554574272. Fax: +390554574271. E-mail: luchinat@cerm.unifi.it

Supplementary Material

‡ Ivano Bertini passed away on July 7th, 2012

INDEX

T**able S1**. ^{13}C Resonance Assignments for S100A14 (pH 6.5, T=310K)

Chemical shifts (ppm)

Residue		C	Cα	Cβ	Cγ	Cδ	Cε
MET	1						
GLY	2						
GLN	3						
CYS	4						
ARG	5	176.35	56.02	30.68	26.89	43.22	
SER	6	174.12	57.99	63.69			
ALA	7	177.24	52.60	19.02			
ASN	8	174.94	52.86	38.80	177.04		
ALA	9	177.84	52.72	18.92			
GLU	10	176.43	56.88	29.63	36.16	183.83	
ASP	11	176.09	54.28	40.75	179.89		
ALA	12	177.78	52.51	18.78			
GLN	13	175.68	55.91	29.11	33.64	180.35	
GLU	14	175.78	55.80	29.95	36.11	183.65	
PHE	15	176.90	55.33	39.25			
SER	16	174.80	56.97	65.60			
ASP	17	179.69	57.97	39.45			
VAL	18	176.86	66.48	31.94	CG1 24.34, CG2 21.04		
GLU	19	178.62	59.99	28.66	39.29		
ARG	20	178.71	59.10	29.71	28.45	43.10	
ALA	21	179.34	55.22	17.93			
ILE	22	177.49	66.03	37.44	CG1 29.29, CG2 16.94	CD1 14.76	
GLU	23	178.02	60.30	29.51	36.47		
THR	24	177.03	67.09	68.22	CG2 21.31		
LEU	25	178.15	59.33	41.74	26.39	CD1 26.28, CD2 24.47	
ILE	26	177.95	64.09	37.26	CG1 28.33, CG2 17.50	CD1 12.89	
LYS	27	179.64	60.06	32.05	25.69	29.68	41.81
ASN	28	175.96	57.70	40.34			
PHE	29	176.16	61.21	38.61			
HIS	30	177.40	58.54	29.61			
GLN	31	176.58	57.76	28.37	33.52		
TYR	32	174.84	59.73	38.70			
SER	33	175.03	57.24	63.24			
VAL	34	177.79	64.36	32.11	CG1 21.39, CG2 21.39		
GLU	35	177.76	57.49	29.78	36.44		
GLY	36	175.05	45.26				
GLY	37	174.09	45.79				
LYS	38	176.02	56.14	32.72	24.57	28.86	41.97
GLU	39	175.61	56.81	30.21	36.39		
THR	40	173.38	59.32	72.14	CG2 21.79		
LEU	41	177.02	53.34	45.30	26.66	CD1 26.46, CD2 23.12	
THR	42	174.13	59.84	68.06	CG2 21.70		
PRO	43	178.59	66.21	31.71	28.21	50.01	
SER	44	176.17	61.51	62.17			
GLU	45	178.97	59.05	29.67	36.06		
LEU	46	177.64	57.27	40.85	26.61	CD1 26.41, CD2 23.81	
ARG	47	178.85	59.98	29.04	27.09	43.09	

ASP	48	176.71	57.54	39.43	179.09		
LEU	49	179.11	59.04	41.26	27.60	CD1 25.63, CD2 24.56	
VAL	50	178.76	66.35	31.46	CG1 22.44, QG2 22.17		
THR	51	175.99	65.76	68.86	CG2 21.31		
GLN	52	178.29	57.82	28.54	34.07	181.05	
GLN	53	176.94	54.96	28.43	32.07		
LEU	54	176.88	53.94	42.13	26.78	CD1 24.73, CD2 23.33	
PRO	55	178.29	65.33	30.75	27.01	49.79	
HIS	56	176.35	57.76	28.25			
LEU	57	176.39	55.44	42.55	26.61	CD1 25.10, CD2 22.36	
MET	58	171.61	53.51	32.66	28.86		17.11
PRO	59	177.44	63.11	31.89	27.25	49.91	
SER	60	174.71	59.13	63.23			
ASN	61	175.52	52.99	37.90	177.33		
CYS	62	175.34	59.23	27.36			
GLY	63	175.61	46.25				
LEU	64	177.85	57.57	41.98	26.66	CD1 24.43, CD2 23.66	
GLU	65	179.07	60.05	28.39	36.18	183.28	
GLU	66	178.18	59.06	29.14	36.31	182.06	
LYS	67	179.39	58.86	32.11	24.99	28.92	41.87
ILE	68	177.81	64.54	37.34	CG1 29.11, CG2 17.24	CD1 12.90	
ALA	69	179.62	54.44	17.94			
ASN	70	176.48	54.30	38.66			
LEU	71	178.33	56.30	42.34	26.03	CD1 25.03, CD2 22.86	
GLY	72	174.44	45.73				
SER	73	175.03	58.88	63.57			
CYS	74	174.51	58.66	27.72			
ASN	75	174.75	53.80	38.90			
ASP	76	175.53	54.42	40.55	180.59		
SER	77	173.31	59.04	63.14			
LYS	78	175.26	55.58	33.35	24.29	28.94	42.00
LEU	79	176.07	54.25	41.96	27.07	CD1 25.40, CD2 24.22	
GLU	80	177.08	55.16	31.64	36.65		
PHE	81	177.01	62.33	38.50			
ARG	82	178.03	59.19	29.18	26.54	43.21	
SER	83	175.81	61.69	62.71			
PHE	84	175.05	60.01	39.22			
TRP	85	179.08	59.75	29.16			
GLU	86	179.06	59.29	29.49	36.20		
LEU	87	178.90	57.95	41.35	26.62	CD1 26.59, CD2 24.63	
ILE	88	177.35	62.97	35.55	CG1 26.29, CG2 18.20	CD1 11.46	
GLY	89	175.69	45.38				
GLU	90	179.13	58.18	28.81	35.75		
ALA	91	178.70	54.02	17.97			
ALA	92	177.93	53.91	17.89			
LYS	93	177.04	57.45	31.83	24.58	28.92	41.82
SER	94	174.54	58.81	63.93			
VAL	95	175.42	63.16	31.49	CG1 21.03, CG2 21.03		
LYS	96	175.99	55.48	32.46	24.38	28.66	42.05
LEU	97	176.72	54.44	42.25	26.53	CD1 25.30, CD2 23.25	
GLU	98	175.87	56.21	30.10	36.04		
ARG	99	173.86	53.60	29.92	26.76	43.14	
PRO	100	176.70	62.80	31.73	27.10	50.38	

VAL	101	176.00	62.13	32.24	CG1 20.81, CG2 20.81		
ARG	102	176.40	55.88	30.74	26.71	43.09	
GLY	103	172.87	45.04				
HIS	104	179.02	56.88	30.29			

Table S2. All proton assigned resonances of S100A14 (pH 6.5, T=310K)

Chemical shift (ppm)

Residue		N	HN	Hα	Hβ	Hγ	Hδ	Hε	Others
MET	1								
GLY	2								
GLN	3								
CYS	4								
ARG	5	124.10	8.57	4.05	(1.90)	(1.71)	(3.29)		
SER	6	117.46	8.37	4.51	3.96, 3.92				
ALA	7	125.90	8.41	4.40	(1.46)				
ASN	8	117.45	8.36	4.77	2.91, 2.81				HD21 7.57, HD22 6.87
ALA	9	124.38	8.27	4.35	(1.47)				
GLU	10	119.31	8.47	4.31	2.04, 1.91	2.32			
ASP	11	120.61	8.20	4.61	(2.70)				
ALA	12	123.80	8.07	4.09	(1.37)				
GLN	13	118.55	8.23	4.27	2.02, 1.95	2.37, 2.46			
GLU	14	120.60	8.10	4.30	2.05, 1.91	2.37, 2.22			
PHE	15	121.08	8.03	4.88	3.46, 3.09		(6.92)	(7.14)	HZ 7.35
SER	16	119.43	9.53	4.85	(4.28)				
ASP	17	121.31	9.27	4.55	2.92, 2.80				
VAL	18	120.64	8.23	3.68	2.28	(1.15)			QG2 (0.93)
GLU	19	121.01	8.18	3.74	1.98, 1.52	3.20			
ARG	20	118.07	8.92	4.30	(2.12)	2.06, 1.90	3.47, 3.34		
ALA	21	123.83	8.06	4.42	(1.88)				
ILE	22	118.18	8.01	3.72	1.96	(0.98)	(0.74)		QG2 (0.78)
GLU	23	120.11	8.45	3.96	2.55	2.39, 2.20			
THR	24	117.32	8.72	4.01	4.01				QG2 (1.25)
LEU	25	123.44	8.15	4.28	(2.67)	2.23	(1.09)		QD2 (1.16)
ILE	26	119.53	8.51	4.00	2.01	1.60, 1.20	(0.68)		QG2 (0.98)
LYS	27	120.51	9.00	4.15	(2.05)	1.75, 1.55	(1.77)	(3.04)	
ASN	28	117.26	7.99	4.56	2.96, 2.48				HD21 7.62, HD22 7.25
PHE	29	119.13	7.99	3.91	3.33, 2.67		(6.49)	(7.10)	HZ 7.27
HIS	30	114.68	8.18	4.32	3.27				HD2 7.21, HE1 7.09
GLN	31	118.53	8.28	4.08	2.15	2.29			HE21 7.45, HE22 6.80
TYR	32	115.68	7.41	4.21	3.11, 2.70		(7.52)	(6.84)	
SER	33	113.97	7.26	4.53	(3.76)				
VAL	34	126.63	8.49	4.01	2.18	(1.04)			QG2 (0.99)
GLU	35	122.08	8.98	4.25	2.11, 2.05	2.38			
GLY	36	108.46	8.49	4.01					
GLY	37	108.71	8.14	4.00					
LYS	38	118.05	8.05	4.47	(1.96)	1.52, 1.45	1.74, 1.68	(3.05)	
GLU	39	118.14	8.63	4.56	2.38, 2.19	2.42, 2.33			
THR	40	108.10	7.49	5.16	4.15				QG2 (1.21)
LEU	41	120.75	8.66	5.04	1.73, 1.50	1.24	(0.58)		QD2 (0.68)
THR	42	113.47	9.13	5.00	4.85				QG2 (1.40)
PRO	43			4.15	2.41, 2.08	2.37, 2.24	4.04, 3.96		
SER	44	111.60	8.09	4.17	(3.95)				
GLU	45	123.94	7.70	4.26	2.16	2.48			
LEU	46	118.92	8.40	4.08	2.00, 1.47	1.60	(0.97)		QD2 (1.08)
ARG	47	119.26	8.26	3.77	2.16, 1.95	(1.58)	3.26, 3.07		
ASP	48	122.30	8.16	4.43	3.08, 2.81				

LEU	49	122.57	7.98	3.17	1.75, 1.09	1.24	(0.68)		QD2 (0.65)
VAL	50	118.23	8.39	3.55	2.18	(1.05)			QG2 (1.01)
THR	51	114.98	8.48	3.99	4.35				QG2 (1.32)
GLN	52	116.42	8.18	4.25	2.32, 2.22	2.84, 2.60			
GLN	53	112.15	8.49	4.85	2.28, 2.03	2.68, 2.59			HE21 7.20, HE22 6.77
LEU	54	119.81	7.92	5.26	1.96, 1.74	1.57	(0.86)		QD2 (0.82)
PRO	55			4.47	2.20, 1.58	1.93, 1.21	3.50, 3.38		
HIS	56	116.48	9.77	4.59	3.46, 3.01				HD2 7.47, HE1 7.13
LEU	57	116.64	7.83	4.33	1.99, 1.55	1.74	(0.77)		QD2 (0.75)
MET	58	117.98	7.97	4.54	2.42, 1.86			(1.96)	
PRO	59			4.45	2.37, 2.04	(2.00)	3.52, 3.44		
SER	60	114.90	8.36	4.34	4.06, 3.99				
ASN	61	120.75	7.99	4.81	(3.00)				
CYS	62	117.83	8.26	4.57	3.09, 3.03				
GLY	63	111.28	8.52	4.09					
LEU	64	122.42	7.91	4.05	1.77, 1.63	1.62	(0.92)		QD2 (0.83)
GLU	65	117.90	8.66	3.91	2.08, 1.97	2.36			
GLU	66	120.04	8.22	4.12	2.14, 2.08	2.34			
LYS	67	119.32	7.78	4.14	2.00, 1.87	1.50, 1.61	1.69, 1.62	(2.97)	
ILE	68	118.41	8.23	3.72	1.97	1.80, 1.63	(0.82)		QG2 (0.97)
ALA	69	122.85	8.07	4.22	(1.55)				
ASN	70	115.44	7.88	4.87	2.96, 2.90				HD21 7.62, HD22 6.97
LEU	71	120.60	7.82	4.33	1.98, 1.62	2.04	(0.96)		QD2 (0.95)
GLY	72	106.78	8.24	4.05					
SER	73	115.63	8.22	4.57	4.06, 4.02				
CYS	74	119.30	8.26	4.65	3.11				
ASN	75	120.02	8.43	4.84	3.04, 2.83				HD21 7.75, HD22 7.09
ASP	76	119.19	8.30	4.67	2.88, 2.73				
SER	77	112.82	8.22	4.26	4.04, 4.00				
LYS	78	120.52	7.93	4.56	(1.90)	1.52, 1.43	(1.78)	(3.08)	
LEU	79	123.72	9.02	4.78	2.05, 1.77	1.55	(0.95)		QD2 (0.92)
GLU	80	121.18	8.63	4.90	2.58, 2.25	2.49			
PHE	81	123.46	9.36	3.53	2.92, 2.89		(6.56)	(6.86)	HZ 6.70
ARG	82	116.45	8.94	4.04	2.08, 1.92	(1.77)	3.21, 3.28		
SER	83	115.73	7.61	4.51	4.14, 4.04				
PHE	84	122.19	7.49	4.22	(3.10)		(6.88)	(7.20)	HZ 7.33
TRP	85	120.70	8.77	3.76	3.12, 2.59				HD1 6.92, HE1 10.04, HE3 7.15, HH2 6.41, HZ2 5.87, HZ3 6.83
GLU	86	119.09	8.15	4.08	2.33, 2.17	2.53			
LEU	87	120.84	7.60	4.05	(1.78)	1.72	(0.82)		QD2 (0.85)
ILE	88	115.18	7.96	3.59	1.66	0.92, 0.82	(0.32)		QG2 (0.58)
GLY	89	108.14	7.62	2.85,2.16					
GLU	90	120.00	7.41	3.96	2.14, 2.05	2.28			
ALA	91	122.45	7.61	4.01	(1.37)				
ALA	92	117.96	7.74	3.67	(0.90)				
LYS	93	114.64	7.20	3.77	1.83, 1.74	1.40, 1.29	(1.61)	(2.92)	
SER	94	113.04	7.53	4.45	4.15, 4.00				
VAL	95	122.14	7.44	3.94	2.17	(0.99)			QG2 (0.96)
LYS	96	126.31	8.47	4.40	(1.90)	1.52, 1.44	(1.76)	(3.05)	
LEU	97	123.36	8.14	4.41	1.60, 1.51	1.57	(0.82)		QD2 (0.80)
GLU	98	121.36	8.32	4.34	2.07	2.33			
ARG	99	122.83	8.28		1.86, 1.75	(1.67)	3.26, 3.21		
PRO	100			4.52	2.33, 1.91	(2.03)	3.80, 3.65		
VAL	101	120.49	8.23	4.13	2.06	(0.96)			
ARG	102	124.58	8.39	4.42	(1.83)	(1.69)	(3.25)		
GLY	103	110.43	8.40	3.96					
HIS	104	123.38	7.89	4.50	3.25, 3.11				HD2 7.10, HE1 6.81

Table S3. ^{15}N R1, R2, NOE evaluated on the S100A14 at 298 K and 700 MHz.

		R1 (s-1)		***R2 (s-1)***		***NOE***
Res	**N°**	**Exp**	**Calc**	**Exp**	**Calc**	**Exp**
	1					
G	2	2.50		20.97		-1.12
Q	3	2.53		14.34		-0.85
C	4	2.12		20.61		-0.85
R	5	2.36		18.44		-0.23
S	6	2.36		9.56		-0.09
A	7	2.39		12.92		-0.07
N	8	2.29		17.33		-0.01
A	9	2.21		9.69		0.26
E	10	1.89		8.03		0.20
D	11	1.80		7.52		0.23
A	12	1.63		10.04		0.25
Q	13	1.70		12.67		0.40
E	14	1.72	0.91	13.02	18.76	0.35
F	15	1.30	0.86	17.51	19.84	0.43
S	16		0.92		18.48	0.75
D	17	1.69	0.86	24.98	19.72	0.88
V	18	1.51	0.82		20.58	0.45
E	19	1.27	0.89	22.60	19.12	0.90
R	20	0.96	0.90	25.35	18.94	0.69
A	21	0.92	0.81	26.35	20.83	0.75
I	22	0.89	0.82	16.34	20.58	0.99
E	23	0.66	0.83	18.82	20.37	0.72
T	24	0.45	0.82	17.73	20.79	0.67
L	25	0.80	0.82	25.54	20.66	
I	26	1.17	0.83	25.00	20.49	0.81
K	27	0.92	0.83	29.45	20.37	0.72
N	28	0.60	0.84		20.24	0.77
F	29	0.94	0.81		20.92	
H	30	1.05	0.81	21.76	20.88	0.73
Q	31	0.81	0.84	26.42	20.24	0.80
Y	32	0.57	0.85		19.96	0.83
S	33		0.87		19.57	0.80
V	34		0.96		17.86	
E	35	1.75	0.94	21.43	18.12	0.56
G	36		0.93	19.26	18.28	0.64
G	37	1.76	0.84	14.89	20.24	0.60
K	38	1.63	0.96	31.57	17.79	0.65
E	39		0.81		20.88	0.56
T	40		0.92		18.42	0.37
L	41		0.97		17.57	
T	42		0.97		17.64	0.72

P	43					
S	44	1.17	0.94	22.05	18.18	0.87
E	45	0.82	0.96	24.75	17.70	0.97
L	46	0.69	0.93		18.35	0.79
R	47	0.94	0.90	19.78	18.83	0.79
D	48	0.99	0.95	22.72	17.92	0.97
L	49	0.71	0.96		17.67	0.92
V	50	0.62	0.91		18.69	0.77
T	51	0.96	0.90	23.77	18.98	0.84
Q	52	0.74	0.96	21.98	17.83	0.81
Q	53	1.08	0.94	27.53	18.08	0.78
L	54		0.84	31.88	20.16	0.59
P	55					
H	56		0.95		18.02	0.71
L	57		0.96		17.79	0.75
M	58	1.14	0.96	35.78	17.73	0.76
P	59					
S	60	1.82	0.96	19.94	17.76	0.60
N	61		0.92	12.69	18.52	0.13
C	62		0.95		17.89	0.44
G	63		0.95	19.58	17.86	0.44
L	64	1.34	0.96	23.13	17.79	0.77
E	65	1.22	0.96	22.17	17.86	0.70
E	66	1.08	0.93	18.63	18.25	0.80
K	67	0.91	0.96	21.05	17.70	0.61
I	68	0.66	0.96		17.83	0.59
A	69	1.10	0.95	26.06	17.95	0.83
N	70	1.04	0.96	25.06	17.67	0.80
L	71	0.89	0.92	18.94	18.55	0.58
G	72	1.29	0.92	16.42	18.48	0.48
S	73	1.55	0.94	19.02	18.05	0.62
C	74	2.01	0.87	18.87	19.49	0.54
N	75		0.86	23.73	19.80	0.66
D	76	1.63	0.96	20.11	17.79	0.47
S	77		0.93		18.28	0.49
K	78	1.30	0.87	18.54	19.61	0.58
L	79		0.87		19.42	
E	80		0.95		17.89	
F	81		0.96		17.86	
R	82	0.92	0.95		17.95	0.70
S	83	0.76	0.93		18.25	0.64
F	84		0.96		17.83	
W	85		0.95		17.99	
E	86	1.31	0.91		18.69	0.89
L	87	1.13	0.94	35.53	18.08	0.50
I	88		0.96		17.86	
G	89		0.94		18.12	
E	90	1.26	0.92	31.21	18.48	0.69
A	91		0.95		18.05	0.73
A	92		0.95		17.99	

K	93	1.73	0.86	28.23	19.69	
S	94	1.36	0.94	22.31	18.08	0.55
	95		0.94		18.08	
	96		0.96		17.86	
L	97		0.95		18.02	0.36
E	98	1.54		16.54		0.44
R	99	1.68				0.43
P	100					
V	101			10.77		0.33
R	102	2.12		11.47		0.16
G	103	2.13		14.72		-0.19
H	104	1.58		10.34		-0.34

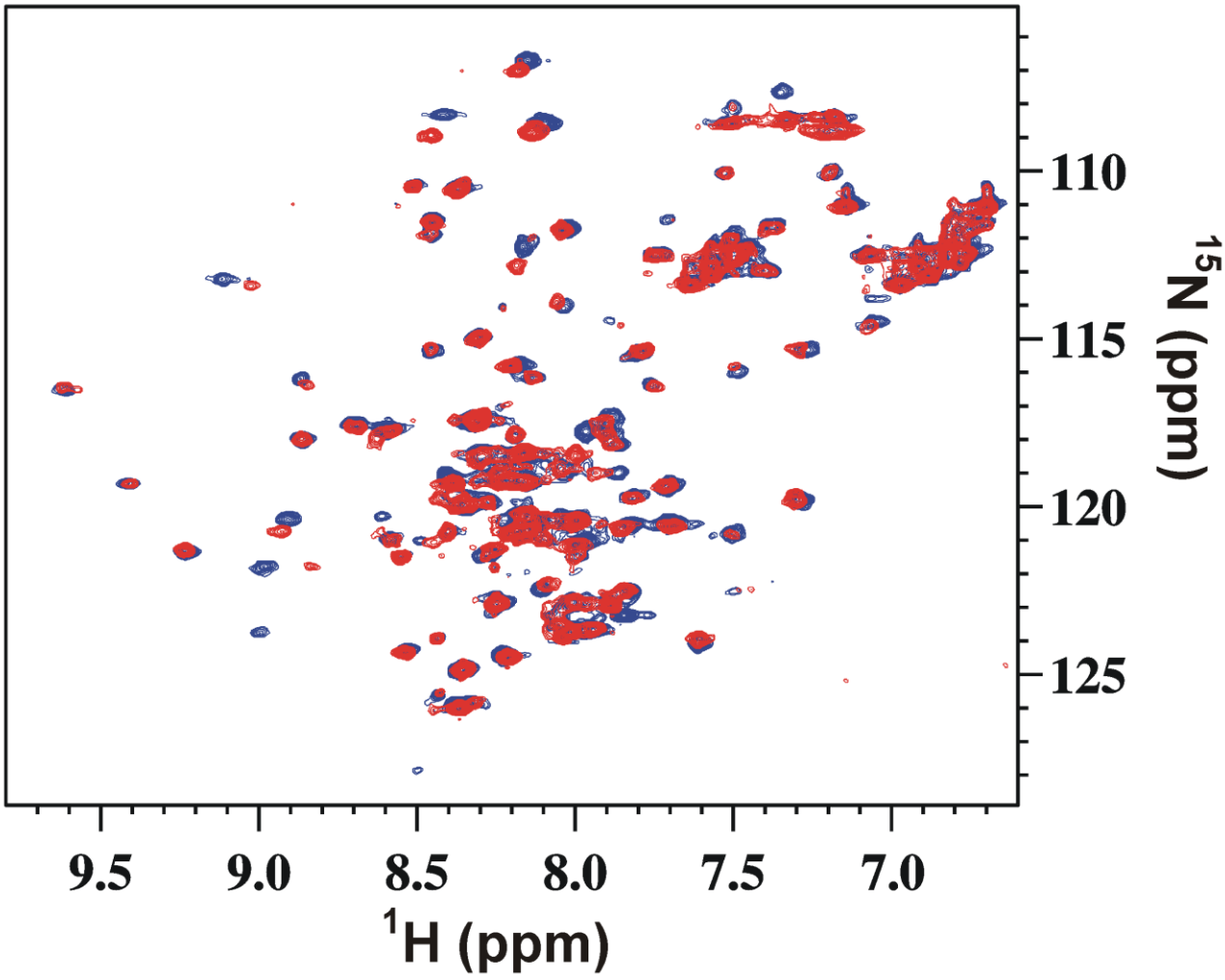


Figure S1. 2D ^{1}H-^{15}N-HSQC spectra of S100A14 in the absence (blue cross-peaks) and in the presence (red cross-peaks) of 80 mM calcium(II) ions.

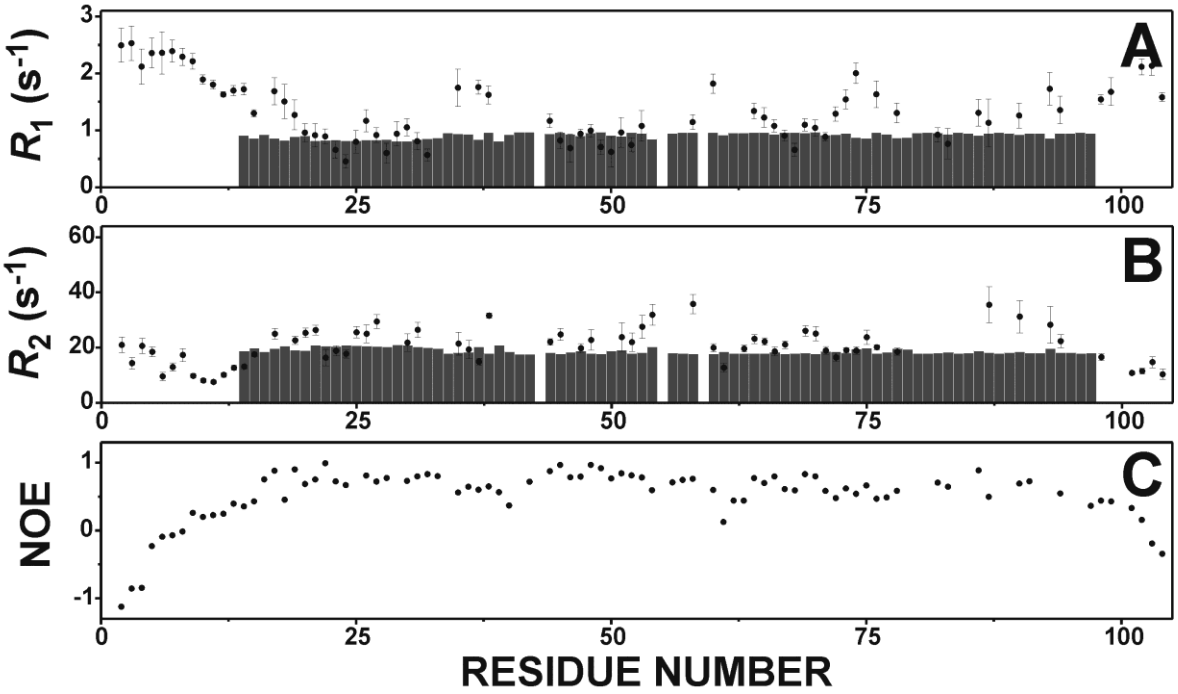


Figure S2. Relaxation data of S100A14 measured at 298K and pH 6.5. The experimental data are reported as black dots, instead the grey bars represent the predicted values by HydroNMR considering the homodimeric NMR solution structure of S100A14.

3.5 NMR characterization of the C-terminal tail of full-length RAGE in a membrane mimicking environment

Valentina Borsi [1], Linda Cerofolini [1], Marco Fragai [1,2], Claudio Luchinat [1,2]

[1] Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy.

[2] Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

*J Biomol NMR. (2012), **54**, 285-290*

J Biomol NMR (2012) 54:285–290
DOI 10.1007/s10858-012-9671-0

ARTICLE

NMR characterization of the C-terminal tail of full-length RAGE in a membrane mimicking environment

Valentina Borsi · Linda Cerofolini · Marco Fragai · Claudio Luchinat

Received: 13 June 2012 / Accepted: 2 September 2012 / Published online: 22 September 2012

Abstract Targeting the receptor for the advanced glycation endproducts (RAGE) signalling has a potential for the prevention and treatment of several pathologies. Extracellular activation of RAGE triggers the interactions of the RAGE cytoplasmic tail with intracellular protein partners. Here the cytoplasmic tail of RAGE has been investigated by NMR as part of the full-length protein, in the presence of a membrane-mimicking environment. The isolated cytoplasmic tail has also been studied for comparison. The NMR spectra of the whole receptor show that some but not all residues belonging to the C-terminal region of the cytoplasmic tail have a large flexibility, while the membrane proximal region seems to be rigidly connected to the trans-membrane domain and ectodomains. The analysis indicates that the behavior of the cytoplasmic tail is strongly affected by its being part of the whole receptor. These results provide new insight towards the understanding of signal transduction by RAGE.

Keywords RAGE · Signal transduction · Membrane proteins · NMR spectroscopy

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9671-0) contains supplementary material, which is available to authorized users.

V. Borsi · L. Cerofolini · M. Fragai · C. Luchinat (✉)
Magnetic Resonance Center (CERM), University of Florence, via L. Sacconi 6, 50019 Sesto Fiorentino, Italy
e-mail: luchinat@cerm.unifi.it

V. Borsi · M. Fragai · C. Luchinat
Department of Chemistry, University of Florence, via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Introduction

The receptor for advanced glycation endproducts (RAGE) is a multiligand receptor of the immunoglobulin superfamily present on the cell surface and involved in inflammation and immune responses (Schmidt et al. 2000; Neeper et al. 1992; Sparvero et al. 2009; Chavakis et al. 2004). RAGE consists of an N-terminal extracellular portion, comprising domains V, C1 and C2 (23–317), of a transmembrane helix (343–363), and of a short cytoplasmic tail (364–404). The latter plays a crucial role in the transduction of the RAGE-ligand interactions (Hudson et al. 2008; Rai et al. 2012; Sakaguchi et al. 2011), but the details of the molecular mechanism and the modulation of RAGE activation are still largely unrevealed. Sequence alignment shows that there is a high level of conservation between the cytoplasmic tails of the various species, and excludes any enzymatic activity for the C-terminal domain of RAGE. According to the amino acid composition and to the described functional roles, this short sequence has been divided into three distinct regions, one proximal to the cell membrane, rich of basic amino acids, a central fragment rich of acidic amino acids, and a low-conserved C-terminal region (Ishihara et al. 2003). Here, the cytoplasmic tail of RAGE has been investigated as part of the full-length receptor in the presence of a membrane-mimicking environment, and compared with the isolated peptide. The analysis indicates that the behavior of the cytoplasmic tail is strongly affected by its being tethered to the whole receptor.

Materials and methods

Preparation of the protein samples

The cytoplasmic domain of RAGE, including three residues belonging to the trans-membrane domain (I361-P404,

cytRAGE hereater), was amplified by PCR from full-length RAGE DNA (GenBank NM_001136) with primers containing 5′ Bam HI and 3′ Xho I restriction sites. Then DNA was subcloned into the pGEX 4T-1 vector in order to express the protein fused with glutatione S-transferase (GST) followed by a thrombin cleavage site. The vector was transformed in *E. coli* BL21(DE3) strain (Novagen) and cells were grown in LB media at 37 °C till an $OD_{600} \sim 0.7$ was reached. Then expression of the recombinant protein was induced by addition of 1 mM IPTG and growth was allowed for 5 h at 37 °C. Cells were harvested by centrifugation at 9,000*g* and lysed by sonication in ice in Tris 50 mM pH 8, PMSF 1 mM, DNase 0.02 mg/ml. Clarified lysate, obtained by centrifugation at 40,000*g*, was first purified on DEAE column (GE healthcare) equilibrated with lysis buffer and eluted with a 10 column volume linear gradient to Tris 50 mM, pH 8, NaCl 1 M. The GST-tag was removed by thrombin cleavage (1 unit for 1 mg of protein) incubated at room temperature for 5 h followed by separation over Superdex 16.60 75 (GE healthcare) with HEPES 10 mM, pH 7.2, NaCl 150 mM. Expression and purity of the protein sample were verified by SDS-PAGE in 17 % polyacrylamide stained with Coomassie brilliant blue R-250 against Protein marker. Samples of ^{15}N- and ^{15}N,^{13}C-enriched RAGE-cytoplasmic construct were produced as described above except for the use of M9 minimal media containing $^{15}(NH_4)_2SO_4$ and ^{13}C-glucose as the sole nitrogen and carbon source, respectively.

Full-length RAGE DNA (GenBank NM_001136) was subcloned into the pET15b vector in order to express the protein fused with a histidine-tag followed by a thrombin cleavage site. The vector was transformed in *E. coli* BL21(DE3)-C43 strain and cells were grown in LB media at 25 °C till an $OD_{600} \sim 0.7$ was reached. Then expression of the recombinant protein was induced by addition of 1 mM IPTG and growth was allowed overnight at 25 °C. Cells were harvested by centrifugation at 9000*g* and lysed by sonication in ice with Tris 20 mM NaCl 300 mM, PMSF 1 mM, DNase 0.02 mg/ml at pH 8. Then the protein, precipitated as inclusion bodies, was solubilized with a buffer containing Tris 20 mM NaCl 300 mM, Urea 8M, PMSF 1 mM, SDS 0.2 % at pH 8. The solubilized protein was first purified by size-exclusion chromatography on the HiLoad 26/60 Superdex 75 (GE healthcare) equilibrated and eluted with the same Tris buffer. Then the protein, bearing the His-tag, was loaded in a Ni Sepharose FF column and refolded by using a renaturation buffer containing Tris 20 mM, NaCl 300 mM, DPC 0.2 %, at pH 8 and eluted with imidazole 500 mM. The refolded protein was further purified on a DEAE column (GE healthcare) equilibrated with a buffer containing Tris 20 mM, NaCl 40 mM, DPC 0.25 at pH 8 and eluted with a 10 column volume linear gradient to Tris 20 mM, NaCl 500 mM, DPC 0.2 % at pH 8. Then the elution buffer was replaced with the final buffer containing Tris 20 mM, NaCl 300 mM, DPC 5.7 % at pH 8. Samples of ^{13}C- and ^{15}N-enriched full-length RAGE were produced as described above except for the use of M9 minimal media containing $^{15}(NH_4)_2SO_4$ and ^{13}C-glucose as the sole nitrogen and carbon source, respectively.

NMR measurements and protein assignment

The experiments for the sequence-specific assignment of the cytoplasmic domain of RAGE (0.2–0.3 mM in water buffer solution, 10 mM HEPES pH 7.2, 150 mM NaCl) were performed at the temperature of 298 K on Bruker AVANCE spectrometers operating at 900 and 700 MHz and equipped with triple resonance cryoprobes. The assignment of H^N, N, C', C^{α}, and C^{β} resonances was obtained by the analysis of a set of ^{1}H-detected (2D ^{1}H–^{15}N HSQC, 3D HNCA, 3D HNCO, 3D CBCA(CO)NH, 3D HNCACB) and ^{13}C-detected *protonless* (2D hCON, 2D hCACO, 2D hCBCACO) NMR experiments (Bermel et al. 2006). The aliphatic side-chain ^{1}H and ^{13}C resonances were assigned through the analysis of 3D (H)CCH-TOCSY spectrum at 700 MHz. All the assigned atoms are reported in Table S1 and Table S2. The experiments on the full-length RAGE (0.1 mM in water buffer solution 20 mM Tris pH 8, 300 mM NaCl, 20 mM $CaCl_2$), and on its isolated cytoplasmic domain in the presence of DPC micelles were acquired at 298 K on a Bruker AVANCE spectrometer operating at 900 MHz and equipped with a cryoprobe.

R_1, R_2 and NOE measurements

The experiments for measuring ^{15}N relaxation rates and ^{1}H-^{15}N NOE were acquired at 298 K on a 700 MHz Bruker Avance spectrometer on ^{15}N-enriched samples of the cytoplasmic domain and full-length RAGE at the concentration of 0.2 and 0.1 mM, respectively. The ^{15}N longitudinal relaxation rates (R_1) were measured by collecting a series of 2D ^{1}H–^{15}N HSQC spectra using a sequence modified to remove cross correlation effects during the relaxation delay and considering variable inversion recovery delays ranging between 20 and 1500 ms for the isolated cytoplasmic domain and between 20 ms and 2000 ms for the full-length RAGE, respectively (Kay et al. 1989; Barbato et al. 1992). The recycle delay was 4.0 s with an acquisition time of 81 ms for all the relaxation measurements. The ^{15}N transverse relaxation rates (R_2) were measured using a CPMG sequence (Kay et al. 1989; Peng and Wagner 1994) with a CPMG refocusing delay, τ_{CPMG}, of 450 μs and with the variable delays ranging from 16.96 to 750 ms for the isolated cytoplasmic domain and from 8.48 to 245.92 ms for the full-length receptor, respectively. Heteronuclear ^{1}H – ^{15}N NOEs were measured with and

without ^{1}H saturation. The relaxation data are reported in Table S3 and S4.

The effects of DPC on the isolated cytoplasmic domain were evaluated by adding minute amounts of a concentrated DPC (dodecylphosphocholine) solution (400 mg/ml) to a sample of the protein at the concentration of 0.33 mM in water buffer solution with 10 % of D_2O. 2D ^{1}H–^{15}N HSQC spectra were recorded after each addition and on the final sample, where the DPC concentration as monomer, was 240 mM.

Results and discussion

The isolated cytoplasmic tail of RAGE including three residues of the trans-membrane domain (I361-P404, cytRAGE) was expressed and purified as ^{15}N- and ^{15}N–^{13}C-enriched protein. The 2D ^{1}H–^{15}N HSQC spectra of the peptide show a poor spreading of the signals, as commonly observed for unfolded proteins (Fig. 1a). The NMR analysis of the protein was performed by a combination of ^{1}H-detected and ^{13}C-detected experiments that enabled the assignment of all but six (Gly-359, Ser-360, Gln-367, Arg-368, Arg-369 and Glu-380) NH resonances (Fig. 2a). The incompleteness of the assignment was caused by the low sequence diversity and the severe overlap involving these few residues. Secondary structure prediction was carried out with the program TALOS+ (Shen et al. 2009) using the H^N, N, C′, C^{α}, H^{α} and C^{β} chemical shift values as input data. The prediction suggests a random-coil conformation for the whole sequence with all the residues experiencing high mobility. Further and more detailed information on the dynamics of cytRAGE were provided by relaxation data. In particular, the characterization of fast motions occurring on picosecond-nanosecond time scales was performed by exploiting the spin relaxation properties of the amide ^{15}N nuclei through R_1, R_2, and NOE experiments (see Fig. 3a, c, e). Random-coil polypeptides and flexible protein regions are characterized by fast local motions with NOE values below 0.5 or negative. In our construct of cytRAGE, that includes also three aminoacids of the transmembrane region and two of the thrombin cleavage site, the heteronuclear NOEs for all amino acids are close to zero or even negative at the C-terminal region (Fig. 3e). These NOE values indicate that all the isolated cytRAGE is highly flexible and unfolded in solution. The absence of the α-turn structure observed by Rai and co-workers at the membrane-proximal region of the cytoplasmic tail (Rai et al. 2012) is possibly related to the slightly different experimental conditions and/or the three additional residues that are present in this construct. This is an indication that the presence or absence of secondary structure elements in cytRAGE can be critically dependent on its immediate environment.

Information on the cytoplasmic tail of RAGE in a more realistic environment can be achieved by analyzing the whole receptor in the presence of a membrane mimicking media. DPC micelles are frequently used not only as membrane mimetics for the structural characterization of peptides and proteins by solution NMR but also to solubilize and purify membrane proteins in their native conformation. Therefore, the full-length RAGE was cloned, expressed and characterized by NMR. In the 2D ^{1}H–^{15}N TROSY-HSQC spectra, performed on ^{15}N-enriched samples of the receptor in the presence of DPC micelles, only few poorly spread peaks are readily detectable (Fig. 1b).

The comparison of the spectra of the full-length receptor with those of the isolated cytoplasmic tail shows that most of the few observed cross-peaks belong to the cytRAGE. More in detail, most of the cross-peaks in the spectra of the full-length receptor correspond to signals of the isolated cytRAGE, with negligible chemical shift variations. The absence in the spectra of the signals relative to the amino acids of the other domains of the receptor (V, C1, C2 and the transmembrane domain) is interpreted as due to excessive broadening arising from the high molecular weight of the DPC-bound full-length RAGE. Very broad, unresolved additional peaks can be seen by lowering the intensity threshold with respect to that used in Fig. 1b (as shown in Figure S1, panel B). Self-assembly into oligomeric specie(s), suggested by native polyacrylamide gel (Figure S2) can also contribute to the broadening. Oligomerization phenomena have been reported also for short constructs of the receptor such as sRAGE at concentrations larger than 1 mg/ml (Sàrkàny et al. 2011).

Sixteen cross-peaks in the spectra of the full-length RAGE have been reassigned with high confidence to the residues A375, E377, E381, E382, E383, E384, A386, E387, L388, E393, E395, A396, G397, E398, G402 and G403, respectively (Fig. 1b), while for other six resonances (R373, Q379, R385, N389, Q390, E392) the assignment is less certain. Few other cross peaks, present in the spectra of the receptor, could not be assigned. At the same time, the cross peaks belonging to I361, L362, W363, R365, R366, G370, N378, S391, S399, S400, T401, respectively, are not present in the spectra of the full receptor. A detailed classification of the residues belonging to cytRAGE in the full-length receptor based on assignment certainty and disappearance of the cross-peaks from the spectra, is shown in Fig. 2b. It is immediately apparent from the distribution of the assigned residues along the sequence that the membrane proximal region of cytRAGE behaves very differently from the rest of the cytoplasmic tail. In particular, the assigned peaks belong to the amino acids downstream of the cytRAGE region, that has been recently reported to form an α-turn structure in the isolated peptide (Rai et al. 2012). The small or negative ^{1}H–^{15}N-NOE values, the large R_1 and the small R_2 values (see Fig. 3, panels b, d, f) of the assigned signals with respect to the isolated cytosolic tail in water buffered solution

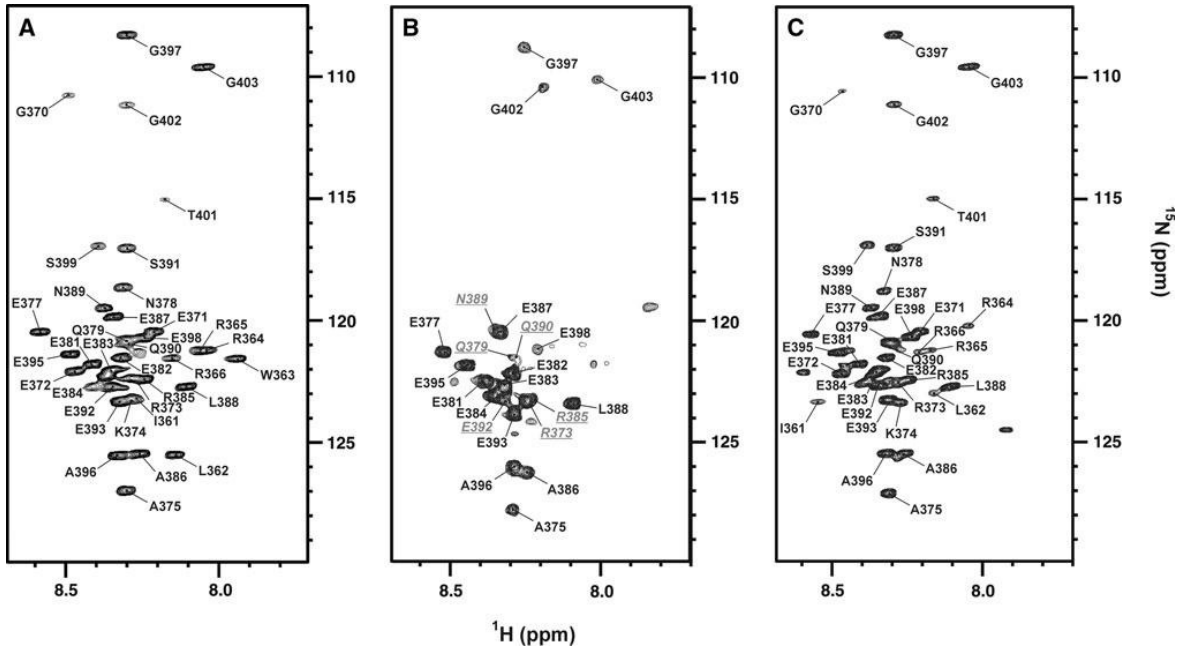


Fig. 1 **a** 2D ^{1}H–^{15}N HSQC spectra of cytRAGE in water buffer solution at 298 K. All resonances belonging to the isolated cytoplasmic tail but six have been assigned. **b** 2D ^{1}H–^{15}N TROSY-HSQC spectra of the full-length RAGE in the presence of DPC micelles at 298 K where sixteen signals belonging to the cytRAGE have been assigned with high confidence. For six additional aminoacids (*underlined* residues) a tentative assignment has been obtained. **c** 2D ^{1}H–^{15}N HSQC spectra of the isolated cytRAGE in the presence of DPC micelles at 298 K

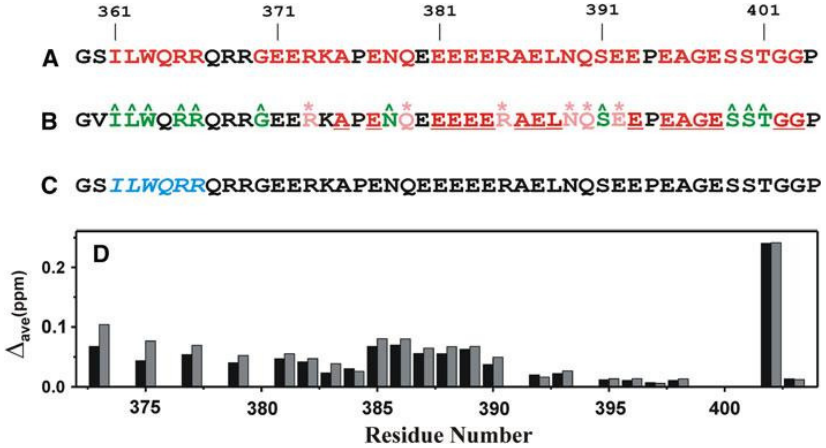


Fig. 2 Sequence of the expressed region of RAGE receptor (cytRAGE). The first two aminoacids (Gly–Ser in **a** and **c**) derive from the thrombin cleavage site, while the three contiguous residues (Ile–Leu–Trp) belong to the transmembrane domain. **a** In red, the residues of cytRAGE assigned in the 2D ^{1}H–^{15}N HSQC spectra when the polypeptide is dissolved in water buffer solution at 298 K; **b** in *red* (*underlined* characters), the residues of cytRAGE in the full-length receptor that are visible and reassigned with high confidence in the 2D ^{1}H–^{15}N HSQC spectra; in *pink* (indicated by *stars*), the visible residues for which the assignment is less certain; in green (indicated by *hats*), the residues whose peaks are not visible in the spectra of the full receptor and without any peak in the close proximity; in *black*, the residues that are not assigned even in the isolated cytRAGE, or those for which neither a tentative assignment, nor conclusive information on their disappearance can be provided. In the full-length receptor the serine, belonging to the thrombin cleavage site, is replaced by a valine; **c** in *cyan italics* fonts, the residues of the polypeptide undergoing chemical shift variations in the 2D ^{1}H–^{15}N HSQC spectra upon addition of DPC micelles; **d** plot of mean shift differences ($\Delta\delta = [(\Delta\delta_H)^2 + (\Delta\delta_N/5)^2]^{0.5}$) (Grzesiek et al. 1996) experienced by the residues of the cytoplasmic tail in the full-length receptor with respect to the isolated cytRAGE in presence (*black* columns) and in absence (*gray* columns) of DPC micelles

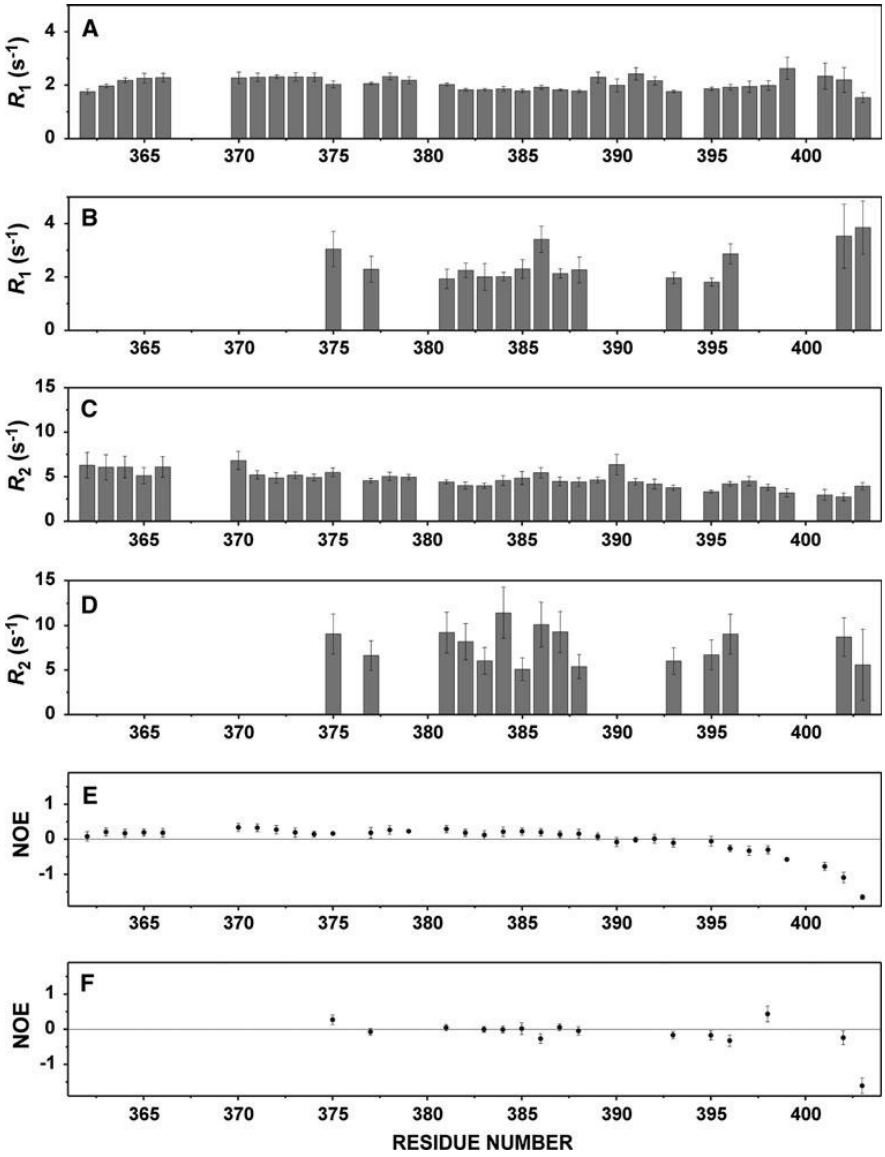


Fig. 3 Backbone ^{15}N relaxation data measured at 700 MHz and 298 K. R_1, R_2 and heteronuclear NOE values for the isolated cytRAGE are reported in panels **a**, **c** and **e**, respectively. R_1, R_2 and heteronuclear NOE values for the cytoplasmic tail in the full-length receptor are reported in panels **b**, **d** and **f**, respectively

indicates that the terminal region of cytRAGE in the full-length protein has a sizable degree of freedom. Conversely, the absence in the spectra of signals corresponding to the first six amino acids of the cytoplasmic tail suggests that the membrane proximal region of cytRAGE is probably rigidly held to the trans-membrane domain. This observation is consistent with the presence of an immobilized region at the N-terminus of the cytoplasmic tail that reduces the flexibility of this peptide-chain segment (Rai et al. 2012). The disappearance or shift observed for the other signals of cytRAGE (I361, L362, W363, Q364, R365, R366, G370, E371, E372, K374, N378, S391, S399, S400, T401) can be attributable to local interactions with the DPC micelles, when the cytoplasmic tail is tethered to the whole receptor. In particular, the increase of $^{1}H–^{15}N$-NOE values for the residues 398 and 402 and the relatively large change in the chemical shift for the latter residue in the full-length construct (see Fig. 2, panel d), seem to be consistent with the presence of transient interactions of the C-terminal region of the cytoplasmic tail with the membrane-mimicking micelles.

In an attempt to obtain additional structural information on the region of the cytoplasmic tail facing the membrane, the isolated cytRAGE was dissolved in the same membrane-mimicking environment used to characterize the full-

length receptor. The analysis of the 2D ^{1}H–^{15}N HSQC spectra (Fig. 1c) shows that the residues belonging to the trans-membrane domain, plus three residues at the membrane proximal region, experience sizable chemical shift variations upon increasing the DPC concentration (see Fig. 2c). The interaction of this cytRAGE construct with the membrane-mimicking environment is probably driven by the three hydrophobic residues, I361, L362, and W363 present in the N-terminal of cytRAGE and was not observed in a shorter construct (Rai et al. 2012). Conversely, the absent or negligible chemical shift variations observed for the residues at the C-terminal end indicates that this protein region does not interact appreciably with the micelles. Therefore, the addition of DPC micelles cannot reproduce in the isolated cytRAGE the effects observed in the full-length receptor, since all residues remain visible even in the presence of DPC micelles and none of the residues in the C-terminal region is shifted or disappears. Collectively, the effect of the membrane-mimicking media on the peptide resonances thus indicates that the isolated construct might not be a representative model of the cytoplasmic tail tethered to the whole receptor on the cell membrane.

Concluding remarks and biological implications

Although the molecular details of the signal transduction are still missing, it is known that the activation of RAGE signalling involves the interaction of the cytoplasmic domain with intracellular partners. cytRAGE is reported to bind different intracellular proteins such as mDia-1(Rai et al. 2012), ERK-1/2 (Ishihara et al. 2003) and after phosphorylation by PKCξ, TIRAP and MyD88 (Sakaguchi et al. 2011). In particular, it has been pointed out that the intracellular partner mDia-1 binds the α-turn at the membrane proximal region of cytRAGE, while the phosphorylation site Ser391 in the flexible tail is reported to be critical in the recognition of TIRAP and MyD88 proteins. It is reasonable to assume that the heterogeneous structural and dynamic properties of the cytoplasmic tail in the full-length receptor are functional to its broad binding capability toward multiple partners.

Acknowledgments This work has been supported by MIUR-FIRB contracts RBLA032ZM7, RBRN07BMCT and RBIP06LSS2, by Ente Cassa di Risparmio di Firenze, and by the European Commission, contracts Bio-NMR n. 261863, East-NMR n. 228461, STREP–SFMET n. 201640, SPINE2-COMPLEXES 031220, and We-NMR 261572.

References

Barbato G, Ikura M, Kay LE, Pastor RW, Bax A (1992) Backbone dynamics of calmodulin studied by 15N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. Biochemistry 31:5269–5278

Bermel W, Bertini I, Felli IC, Lee Y-M, Luchinat C, Pierattelli R (2006) Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. J Am Chem Soc 128:3918–3919

Chavakis T, Bierhaus A, Nawroth PP (2004) RAGE (receptor for advanced glycation end products): a central player in the inflammatory response. Microbes Infect 6:1219–1225

Grzesiek S, Bax A, Clore GM, Gronenborn AM, Hu JS, Kaufman J, Palmer I, Stahl SJ, Wingfield PT (1996) The solution structure of HIV-1 Nef reveals an unexpected fold and permits delineation of the binding surface for the SH3 domain of Hck tyrosine protein kinase. Nat Struct Biol 3:340–345

Hudson BI, Kalea AZ, Arriero MD, Harja E, Boulanger E, D'Agati V, Schmidt AM (2008) Interaction of the RAGE cytoplasmic domain with diaphanous-1 is required for ligand-stimulated cellular migration through activation of Rac1 and Cdc42. J Biol Chem 283:34457–34468

Ishihara K, Tsutsumi K, Kawane S, Nakajima M, Kasaoka T (2003) The receptor for advanced glycation end-products (RAGE) directly binds to ERK by a D-domain-like docking site. FEBS Lett 550:107–113

Kay LE, Torchia DA, Bax A (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to *Staphytococcal nuclease*. Biochemistry 28:8972–8979

Neeper M, Schmidt AM, Brett J, Yan SD, Wang F, Pan YC, Elliston K, Stern D, Shaw A (1992) Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins. J Biol Chem 267:14998–15004

Peng JW, Wagner G (1994) Investigation of protein motions via relaxation measurements. Methods Enzymol 239:563–596

Rai V, Maldonado AY, Burz DS, Reverdatto S, Schmidt AM, Shekhtman A (2012) Signal transduction in receptor for advanced glycation end products (RAGE). Solution structure of C-terminal rage (ctRAGE) and its binding to mDia1. J Biol Chem 287:5133–5144

Sakaguchi M, Murata H, Yamamoto K, Ono T, Sakaguchi Y, Motoyama A, Hibino T, Kataoka K, Huh N (2011) TIRAP, an adaptor protein for TLR2/4, transduces a signal from RAGE phosphorylated upon ligand binding. PLoS ONE 6(8):e23132

Sàrkàny Z, Ikonen TP, Ferreira-da-Silva F, Saraiva MJ, Svergun D, Damas AM (2011) Solution structure of the soluble receptor for advanced glycation end products (sRAGE). J Biol Chem 286:37525–37534

Schmidt AM, Yan SD, Yan SF, Stern DM (2000) The biology of the receptor for advanced glycation end products and its ligands. Biochim Biophys Acta Mol Cell Res 1498:99–111

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Sparvero LJ, Asafu-Adjei D, Kang R, Tang DL, Amin N, Im J, Rutledge R, Lin B, Amoscato AA, Zeh HJ, Lotze MT (2009) RAGE (receptor for advanced glycation endproducts), RAGE ligands, and their role in cancer and inflammation. J Trans Med 7:17

NMR CHARACTERIZATION OF THE C-TERMINAL TAIL OF FULL-LENGTH RAGE IN A MEMBRANE MIMICKING ENVIRONMENT

Valentina Borsi[1,2], Linda Cerofolini[1], Marco Fragai[1,2], Claudio Luchinat[1,2]*

[1] Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy.
[2] Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy.
Address correspondence to: Claudio Luchinat, via L. Sacconi 6, 50019 Sesto Fiorentino, Italy, Tel.: +390554574263. Fax: +390554574253. E-mail: luchinat@cerm.unifi.it

Supplementary Material

INDEX

Table S1. ^{13}C Resonance Assignments for cytRAGE polypeptide in water buffer solution (pH 7.2, 298K).

Residue		Chemical Shifts (ppm)					
		C	Cα	Cβ	Cγ	Cδ	Cε
GLY	359						
SER	360						
ILE	361	176.14	61.35	38.21	CG1 26.91, CG2 17.11	CD1 13.54	
LEU	362	177.09	55.39	41.74	30.04	CD1 26.42, CD2 25.32	
TRP	363	176.26	57.51	28.97			
GLN	364	175.74	56.07	29.23	33.60		
ARG	365	176.27	56.47	30.24			
ARG	366	176.72	56.20	30.47			
GLN	367						
ARG	368						
ARG	369						
GLY	370	174.08	45.24				
GLU	371	176.33	56.26	30.07	36.09		
GLU	372	176.16	56.55	29.96	36.12		
ARG	373	175.75	55.81	30.61	26.89	43.07	41.87
LYS	374	175.64	55.72	32.87	24.48	28.79	41.89
ALA	375	175.37	50.35	17.85			
PRO	376	176.93	63.14	31.87	27.16	50.29	
GLU	377	176.12	56.81	29.80	36.16		
ASN	378	174.78	53.01	38.72	176.90		
GLN	379	175.76	56.10	29.16	33.53		
GLU							
GLU	381	176.49	56.75	29.99	36.03		
GLU	382	176.44	56.56	30.13	35.99		
GLU	383	176.41	56.64	30.26	36.13		
GLU	384	176.33	56.58	30.00	35.95		
ARG	385	175.89	55.81	30.52	26.83	43.10	
ALA	386	177.67	52.54	19.00			
GLU	387	176.32	56.37	29.87	36.08		
LEU	388	176.91	54.99	42.17	26.70	CD1 24.73, CD2 23.16	
ASN	389	174.91	53.05	38.66	176.86		
GLN	390	175.71	55.73	29.15	33.63		
SER	391	174.10	58.26	63.68			
GLU	392	175.93	56.02	30.27	35.99		
GLU	393	174.31	54.05	29.46	35.74		
PRO	394	176.82	62.86	31.90	27.14	50.38	
GLU	395	176.23	56.36	29.85	36.04		
ALA	396	178.03	52.46	19.08			
GLY	397	174.01	44.94				
GLU	398	176.57	56.34	30.07	36.00		
SER	399	174.58	58.15	63.61			

SER	400	174.80	58.21	63.59			
THR	401	174.97	61.72	69.50	CG2 21.24		
GLY	402	174.01	45.05				
GLY	403	179.67	44.21				
PRO	404						

Table S2. ^{1}H and ^{15}N Chemical Shifts for cytRAGE polypeptide in water buffer solution (pH 7.2, 298K).

Residue		Chemical Shifts (ppm)							
		N	H^N	Hα	Hβ	Hγ	Hδ	Qε	Other
GLY	359								
SER	360								
ILE	361	122.93	8.29	4.11	1.88	1.36, 1.16	(0.84)		QG2 0.91
LEU	362	125.35	8.15	3.69	(2.96)	1.86	(1.67)		QD2 1.73
TRP	363	121.45	7.94	4.08	(3.19)				
GLN	364	121.11	8.05	4.44	(2.05)	(2.29)			
ARG	365	121.10	8.07		(1.95)				
ARG	366	121.46	8.17						
GLN	367								
ARG	368								
ARG	369								
GLY	370								
GLU	371	120.37	8.22	4.19	1.99, 1.82	(2.19)			
GLU	372	121.83	8.46	4.18	1.98, 1.83	(2.16)			
ARG	373	122.13	8.27	4.25	1.76, 1.67	(1.55)	(3.12)	(2.91)	
LYS	374	123.05	8.28		1.72, 1.64	(1.34)	(1.59)	(2.91)	
ALA	375	126.77	8.30	4.49	(1.29)				
PRO	376			4.32	2.20, 1.89	(1.93)	3.71, 3.59		
GLU	377	120.12	8.60	4.14	2.01, 1.82	(2.21)			
ASN	378	118.24	8.28	4.59	2.78, 2.66				
GLN	379	120.51	8.28	4.25	(1.89)	(2.27)			
GLU	380								
GLU	381	121.60	8.42	4.21	1.94, 1.84	(2.19)			
GLU	382	121.26	8.31	4.17	1.96, 1.84	(2.17)			
GLU	383	121.97	8.34	4.17	1.99, 1.86	(2.20)			
GLU	384	122.34	8.38	4.15	1.96, 1.85	(2.18)			
ARG	385	122.53	8.26	4.15	1.75, 1.66	(1.52)	(3.11)		
ALA	386	125.50	8.27	4.23	(1.31)				
GLU	387	119.78	8.36	4.19	1.97, 1.84	(2.19)			
LEU	388	122.70	8.12	4.18	(1.54)	1.53	(0.83)		QD2 0.78
ASN	389	119.50	8.38	4.24	2.77, 2.68				
GLN	390	120.94	8.31	4.28	2.03, 1.90	(2.27)			
SER	391	116.99	8.31	4.37	(3.79)				
GLU	392	122.60	8.36	4.36	(2.03)	(2.25)			
GLU	393	123.20	8.32	4.25	1.94, 1.78	(2.20)			
PRO	394			4.32	2.20, 1.82	(1.93)	3.71, 3.62		
GLU	395	121.24	8.48	4.33	1.95, 1.84	(2.20)			
ALA	396	125.46	8.33	4.14	(1.33)				
GLY	397	108.13	8.29	4.23, 3.90					
GLU	398	120.64	8.26	3.89	1.96, 1.85	(2.18)			
SER	399	116.87	8.40	4.42	(3.85)				
SER	400	118.08	8.41	4.48	(3.80)				
THR	401	114.99	8.17	4.33	4.23				QG2 1.13
GLY	402	111.02	8.30						
GLY	403	109.50	8.05						
PRO	404								

Table S3. ^{15}N relaxation values for cytRAGE in water buffer solution measured at 700 MHz ^{1}H frequency, and 298 K.

		R1 (s-1)		*R2(s-1)*		*NOE*
Residue		Exp.	Calc.	Exp.	Calc.	Exp.
GLY	359					
SER	360		1.79		9.12	
ILE	361	2.30	1.79	4.91	9.14	0.15
LEU	362	1.75	1.78	6.27	9.14	0.08
TRP	363	1.97	1.73	6.05	9.49	0.21
GLN	364	2.18	1.73	6.09	9.45	0.17
ARG	365	2.26	1.84	5.12	8.84	0.20
ARG	366	2.28	1.75	6.10	9.35	0.18
GLN	367		1.82		8.95	
ARG	368		1.83		8.90	
ARG	369		1.81		9.01	
GLY	370	2.27	1.68	6.81	9.75	0.34
GLU	371	2.29	1.83	5.20	8.87	0.33
GLU	372		1.83		8.88	
ARG	373	2.32	1.64	4.85	10.04	0.27
LYS	374	2.31	1.82	5.15	8.98	0.19
ALA	375	2.03	1.74	5.49	9.38	0.16
PRO	376					
GLU	377	2.05	1.77	4.56	9.23	0.18
ASN	378	2.33	1.69	5.04	9.72	0.26
GLN	379	2.18	1.83	4.94	8.90	0.23
GLU	380		1.79		9.10	
GLU	381	2.02	1.67	4.41	9.82	0.29
GLU	382	1.82	1.65	4.00	9.98	0.18
GLU	383	1.83	1.78	3.99	9.14	0.12
GLU	384	1.86	1.74	4.58	9.39	0.22
ARG	385	1.78	1.76	4.84	9.31	0.22
ALA	386	1.92	1.83	5.43	8.90	0.20
GLU	387	1.82	1.63	4.47	10.07	0.13
LEU	388	1.77	1.79	4.40	9.12	0.15
ASN	389	2.29	1.72	4.63	9.53	0.08
GLN	390	1.99	1.81	6.35	8.98	-0.08
SER	391	2.43	1.84	4.43	8.85	-0.02
GLU	392	2.16	1.67	4.18	9.81	0.02
GLU	393	1.76	1.83	3.76	8.87	-0.10
PRO	394					
GLU	395	1.86	1.69	3.31	9.68	-0.06
ALA	396	1.93	1.84	4.19	8.85	-0.26
GLY	397	1.94	1.78	4.51	9.15	-0.33
GLU	398	1.98	1.69	3.82	9.70	-0.30
SER	399	2.63	1.76	3.18	9.29	-0.57
SER	400		1.78		9.18	
THR	401	2.34	1.72	2.95	9.55	-0.78
GLY	402	2.19	1.73	2.73	9.45	-1.09

GLY PRO	403 404	1.53	1.82	3.94	8.94	-1.65

Table S4. Experimental ^{15}N relaxation values for RAGE full-length receptor in the presence of DPC micelles measured at 700 MHz ^{1}H frequency, and 298 K.

Residue		R1 (s-1)	R2(s-1)	NOE
GLY	359			
SER	360			
ILE	361			
LEU	362			
TRP	363			
GLN	364			
ARG	365			
ARG	366			
GLN	367			
ARG	368			
ARG	369			
GLY	370			
GLU	371			
GLU	372			
ARG	373			
LYS	374			
ALA	375	3.05	9.04	0.27
PRO	376			
GLU	377	2.29	6.62	-0.07
ASN	378			
GLN	379			
GLU	380			
GLU	381	1.93	9.20	0.05
GLU	382	2.25	8.18	
GLU	383	2.01	6.03	0.00
GLU	384	2.01	11.41	-0.01
ARG	385	2.30	5.08	0.02
ALA	386	3.41	10.11	-0.27
GLU	387	2.13	9.26	0.06
LEU	388	2.26	5.38	-0.05
ASN	389			
GLN	390			
SER	391			
GLU	392			
GLU	393	1.96	6.00	-0.17
PRO	394			
GLU	395	1.81	6.71	-0.17
ALA	396	2.86	9.03	-0.33
GLY	397			
GLU	398			0.43
SER	399			
SER	400			
THR	401			
GLY	402	3.53	8.70	-0.24
GLY	403	3.85	5.57	-1.61
PRO	404			

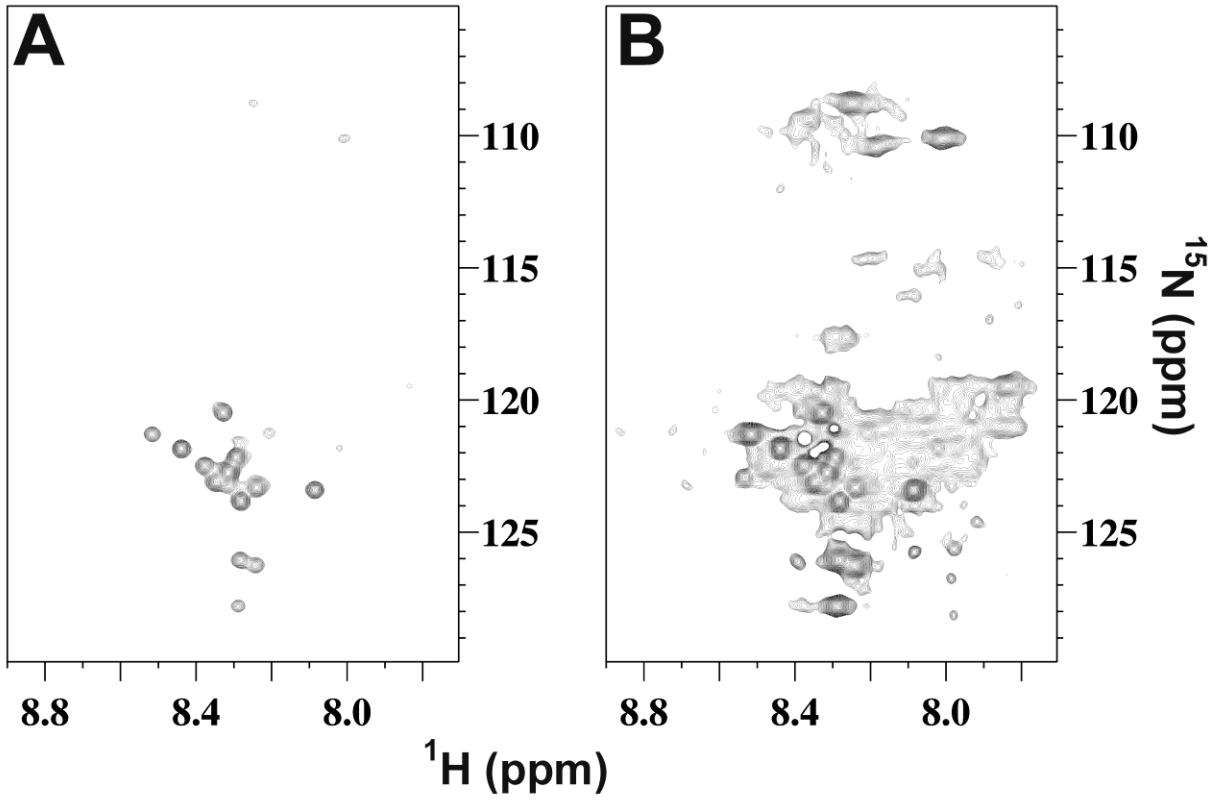


Figure S1. 2D ^{1}H-^{15}N TROSY-HSQC spectra of the full-length RAGE represented with two different intensity thresholds in order to highlight, besides the intense signals of the cytoplasmic tail (panel A), the additional broad signals present in the spectra (panel B).

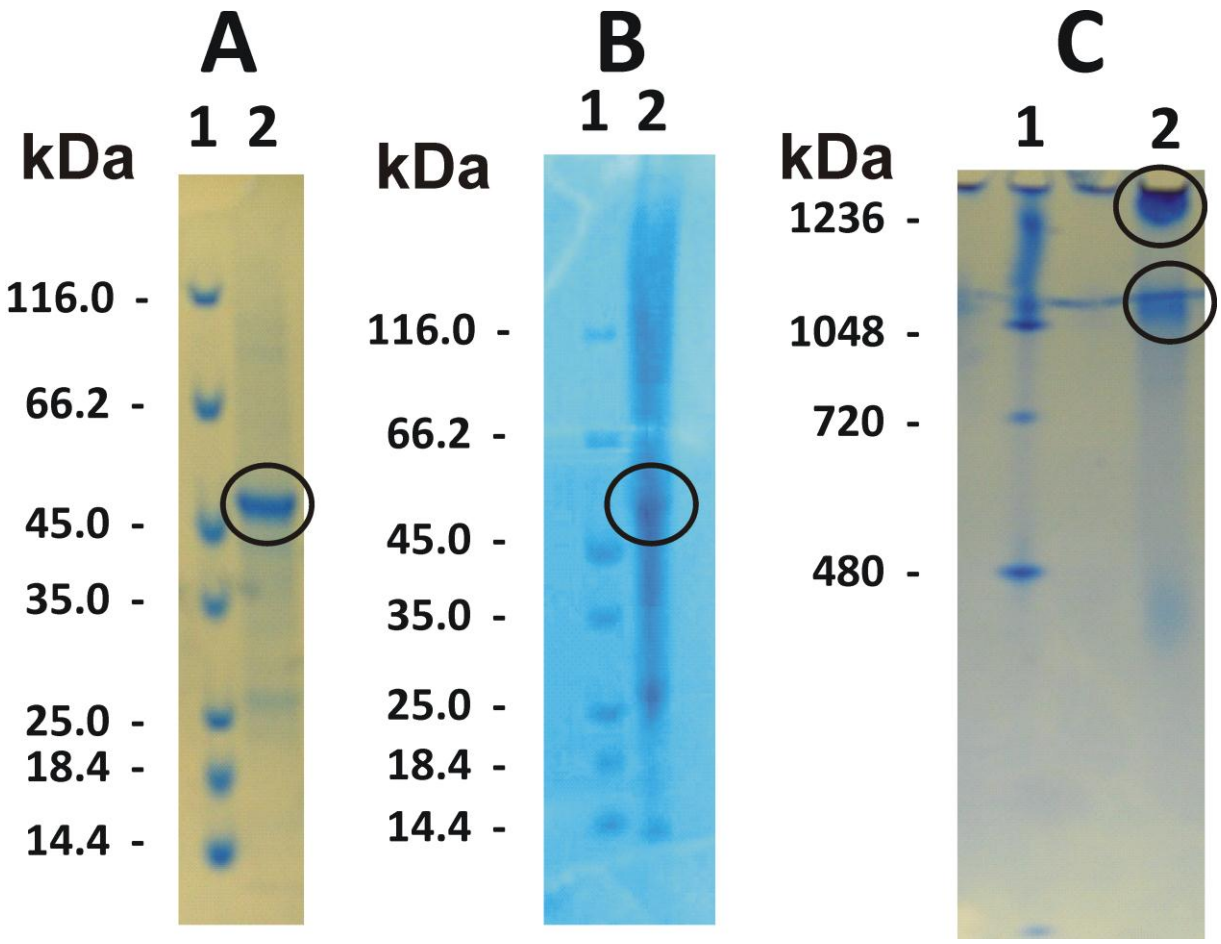


Figure S2. Reducing denaturing (**A**) and non-reducing denaturing (**B**) 4-12% Bis-Tris polyacrylamide gels. The spot corresponding to the monomeric full-length RAGE are marked with open black circles. Non-reducing native 10-20% Tris-Tricine polyacrylamide gel (**C**). Only two spots corresponding to species with very high molecular weight can be observed in the gel.

4. Conclusions and perspectives

A complete description of the activity of biological systems often requires the rigorous investigation of their conformational heterogeneity and the analysis of the energy landscape related to the transitions between the different conformations.

In the present research project, structural and dynamics information provided by NMR studies have been fruitfully used for the analysis of four different biomolecules, with the aim to answer specific questions about their biological function and the molecular mechanisms in which they are involved.

Starting from the analysis of the truncated G-quadruplex construct, the NMR analysis has demonstrated, for the first time, the existence of the computational predicted G-triplex structure. The existence of this new folding is significant because it can open new opportunities for the comprehension of the mechanisms that control the maintenance of genome stability.

In fact, this innovating finding supplies a step forwards for the understanding of the prologue of the process of chromosomal elongation due to the telomerase activity that is associated with aging and cancer. In particular, since this new conformation can be considered the possible intermediate of the folding/unfolding pathway of more complex G-quadruplex structures, it may also create new chance to block the telomerase activity, thus providing new tools for anticancer therapy.

Relevant are also the results obtained in the analysis of matrix metalloprotease-1 in solution. Actually, the *maximum occurrence* (MO) analysis performed on NMR and SAXS data has provided the disclosure of the conformations of matrix metalloprotease-1, that can exist for the maximum per cent of time in solution. This is a fundamental clue for the inspection of the still not clarified mechanism of collagenolysis.

In fact, this innovating strategy, scoring any wished conformation with a weight, allows to estimate also the probability of existence of the crystal structures. The latter, surprisingly, appear not to be the most probable ones, but to have only a maximum occurrence in solution of 20%. Therefore, the conformations possessing the lowest energy in crystallization conditions are not necessarily the most significantly represented in solution. This also explains the necessity of the refinement of crystal structures with NMR-based experimental constraints prior to use them as representative models of the protein in the

physiological conditions. Furthermore, the results we got, demonstrate that multidomain proteins should be analysed in terms of ensembles of conformations and not as single structures.

Collectively, our results suggest that collagenolysis by MMP-1 can be described considering the conformations with the highest MO values as the starting point of a multistep mechanism where interdomain flexibility, *conformational selection* and *induced fit* cooperate in the degradation of triple helical collagen.

Obviously, for the description of this very complex process both *conformational selection* and *induced fit* should be considered. A sequence of events, where the most probable (highest MO) conformations are the most easily *selected* and accessible, followed by the *induced fit*-related reorientation of the multidomain enzyme to proceed along the steps of collagenolysis, is plausible and possible.

In conclusion, our results provide a contribution to elucidate the not trivial mechanism of collagenolysis that is crucial not only to clarify the molecular basis of the pathologies where the uncontrolled collagen degradation play a critical role, but also to design selective inhibitor of MMP-1.

In the last few weeks new paramagnetic constraints have been acquired on MMP-1 in the presence of a collagen model. This new data will provide more details about the mechanism of collagenolysis.

Furthermore, the analysis of the conformational space sampled by other members of the MMPs' family in the absence and in the presence of collagen models, hopefully, will allow us to answer the questions concerning the different role of each MMP in vertebrates.

One major advantage of NMR over x-ray crystallography is the possibility to solve the three-dimensional structure of biomolecules in experimental conditions closer to what they experience in vivo. Moreover, many proteins do not crystallize at all, and NMR remain the unique possible experimental approach to get structural information. In the present research project, the NMR analysis performed at physiological temperature, has revealed the unusual 'semi-open' structure for the human EF-hand protein S100A14. This particular structural organization of the EF-hand domain together with the inability to bind calcium(II) are strictly related to the biological function of this protein, that seems to be constitutively activated and not to be modulated by the concentration of calcium(II) ion.

The analysis of the interaction of this peculiar member of the S100 family with the RAGE receptor could shed light on the function and localization of S100A14, providing also hints to target this biologically relevant interaction.

The major challenge in the modern biomolecular NMR is the characterization of membrane proteins and the analysis of intracellular signal transduction mediated by membrane receptors after their activation by extracellular ligands.

Cell signalling is, actually, an important mechanism to govern basic cellular activities and coordinates cell actions. Therefore, the elucidation of the cascade of events involved in signal transduction is important to design strategies to identify potential pharmacological targets.

In particular, RAGE is an important pharmacological target that is involved in many inflammation-related pathological states. The cytosolic domain of RAGE receptor plays a pivotal role in signal transduction but its function, structural features and dynamical properties are still largely unknown.

In the present research project, it has been highlighted that the C-terminal part of the cytosolic domain maintains its flexibility and structural heterogeneity also when tethered to the full-length receptor. In this way, the heterogeneous structural and dynamic properties of the cytoplasmic tail in the full-length receptor explain its broad binding capability toward multiple partners.

The structural and dynamical analysis of the cytosolic domain reported in this thesis is only the first step in the characterization of RAGE pathway. The use of more physiological membrane-like environments and the analysis of the interaction with its “adaptor proteins”, will allow us to move deep inside in the mechanism of signal transduction mediated by the RAGE receptor.

Collectively, the results obtained in the present research projects stress about the role of conformational heterogeneity as driving force for many biological processes, remarking the role of biomolecular NMR as a key tool in the study of systems biology.

Acknowledgment…. ovvero Ringraziamenti.

Sono passati ormai tre anni…

Al solo pensiero sembravano infiniti ed invece sono passati in un soffio… Ieri che ero l'ultima arrivata al CERM e oggi mi sento una delle colonne portanti…

Durante questo percorso ho incontrato tante persone nuove che sono entrate a far parte della mia vita in maniera più o meno diretta e che arrivata a questo punto mi sento di ringraziare…

Il mio primo Grazie va al Professor Ivano Bertini per avermi accolto al CERM, per avermi permesso di fare scienza e per tutti i progetti interessanti in cui sono stata coinvolta…

Un Grazie anche al professor Claudio Luchinat per la sua grande competenza in biologia strutturale e per le discussioni scientifiche che abbiamo affrontato. Il suo consiglio è sempre stato quello determinante per la riuscita di un progetto.

Un infinito grazie al Dottor Marco Fragai… Sempre presente per i piccoli problemi e per quelli grandi, con la battuta sempre pronta e con i suoi scherzi che spesso mi hanno fatto trasalire. Ma soprattutto per le sue idee geniali, per tutto quello che mi ha insegnato e senza il quale sarebbe stato impossibile questo lavoro di tesi!

Ringrazio anche il professor Giacomo Parigi per il suo indispensabile aiuto nel paramagnetic NMR.

Non posso non citare il dottor Maxime Melikian che al mio arrivo mi ha seguito passo-passo, mi ha introdotto alla biologia strutturale e al gruppo di ragazzi del CERM.

Ringrazio anche il mio compagno di "sventure" Joao Teixeira con il quale ho lavorato ai progetti più difficili senza mai darci per vinti, ottenendo alla fine tanta soddisfazione per i risultati ottenuti!

Un grazie particolare anche a Stefano Detito che ha lavorato con me al progetto sul DNA, per i momenti critici e per le risate. Ringrazio anche il dottor Antonio Randazzo per la sua simpatia e l'aiuto offerto.

Un grazie sincero alle dottoresse Valentina Borsi e a Soumyasri Das Gupta per la loro costante disponibilità.

Grazie anche ad Enrico Ravera per i suoi consigli relativi al *maximum occurrence* (vedi tesi).

Un sentito grazie anche alla dottoressa Malini Nagullapalli per la sua amicizia e per il confronto scientifico. Grazie anche al dottor Mauro Rinaldelli per l'aiuto informatico!

Ringrazio tutti i miei compagni di mensa, i vecchi (Maxime, Mirco, Olga, Malini, Mao, Gianluca, Jeff, Stefano…) e i nuovi (Magda, Tommaso, Mauro, Alessandro, Jeff, Tomas…) per i momenti di svago.

Un grazie anche a tutti gli altri colleghi e amici del CERM… Angelo, Daniela, Camilla, Vasantha, Lucio, Serena, Sara, Eduardo, Carmelo, Enrico, Riccardo, Letizia, Erica… e a tutti i tecnici per il loro supporto e gentilezza.

Ringrazio anche i miei amici di sempre che anche se sempre da più lontano mi hanno comunque seguito in questa mia nuova avventura: Marpia, Zi Chiara, Marco, Silvia, Siani, Sara, Giugi, Chellus, Ire, France…

Un grazie ai miei genitori che mi sono stati vicino anche nei momenti in cui ero di cattivo umore e hanno seguito questa mia scelta pur non sapendo esattamente di cosa si trattasse…

E infine, per ultimo ma non per importanza, un grazie speciale, il più grande a Samu per avere assistito a tutto il backstage di questo dottorato, per aver partecipato ai miei momenti felici e a quelli tristi, per avermi ascoltato sempre… anche quando non volevo parlare.

Ringrazio tutti quelli che hanno letto questa tesi e che sono arrivati infondo a leggere i ringraziamenti.. e anche quelli che sono andati subito alla fine per leggere direttamente i ringraziamenti…

Grazie anche a me perché, anche questa volta, nonostante sembrasse impossibile, sono arrivata alla fine…