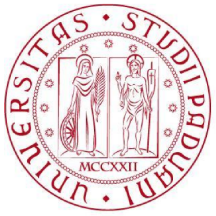

UNIVERSITÀ DEGLI STUDI DI PADOVA

Sede amministrativa: **Università degli Studi di Padova**
Dipartimento di Scienze Chimiche

SCUOLA DI DOTTORATO: Scienze Molecolari

CURRICOLO: Scienze Chimiche

CICLO: XXIX°

Coordinatore: *Ch.mo* Prof. Antonino Polimeno

Sede partner: **École Normale Supérieure - Paris**
Département de Chimie

ÉCOLE DOCTORALE: ED 388 - Chimie Physique et Chimie Analytique de Paris Centre.

Modeling motions in proteins at the molecular level

Theoretical and methodological approaches in an experimental perspective

Supervisore: *Ch.mo* Prof. Antonino Polimeno
Supervisore: *Ch.mo* Dr. Daniel Abergel

Dottorando: Marco Gerolin

Tesi redatta con il contributo finanziario della Fondazione Cariparo

Abstract

Dynamical aspects have a central role in the biological function of proteins. Each motional process has a characteristic time scale, amplitude and energy range. Proteins in particular display a broad range of characteristic motions, from very fast and localized, such as atomic fluctuations, to motions that occours on the scale of the whole molecule, such as folding transitions. Futhermore, dynamical events occourring at long time scales are actually coupled to faster motions, and they can be seen as rare events, the origin of which is found on the microdynamics of rotation around chemical bonds. Electron and nuclear magnetic spectroscopies are sensible to molecular motions at different time scales, making them very powerful tools in the study of molecular dynamics. Nevertheless the dynamical information contained in the experimental data is hidden and theorethical methodologies of interpretation are needed. In this work, we present different theoretical approaches which allow to better describe the stochastic descriptions of flexible macromolecules in solution. State-of-the-art approaches to dynamic models are first reviewed, aimed at the interpretation of magnetic resonance relaxation experiments. Next, the main focus is on *i)* the comparison between information content of the experiment and prediction capability of the model, using a Bayesian Markov-Chain Monte Carlo approach *ii)* defining a way to identify subgroups of atoms, the dynamics of which can be treated independently from the others *iii)* a new model of description of flexible macromolecules via projection operators, which allows to tune the description of the system to the preferred level of description in relation to the spectroscopic observable of interest.

Abstract

(italiano)

Aspetti dinamici hanno un ruolo centrale nella funzione biologica delle proteine. Ad ogni processo dinamico, sono associati un tempo, ampiezza ed energia caratteristici del moto. In particolare, le proteine presentano una vasta distribuzione dinamica, da moti rapidi e localizzati, come fluttuazioni atomiche, a processi che coinvolgono l'intera molecola, come le transizioni di folding. Inoltre, i processi che avvengono a scale temporali lente, sono accoppiati a moti piú rapidi, i primi possono essere quindi visti come eventi rari, la cui origine risiede nella microdinamica di rotazione attorno ai legami chimici. Le risonanze magnetiche, sia elettroniche che nucleari, sono estremamente sensibili ai moti molecolari a diverse scale temporali, rendendo di fatto tali tecniche, strumenti fondamentali per lo studio di aspetti dinamici. Tuttavia, l'informazione dinamica contenuta negli esperimenti è nascosta, per cui sono necessari modelli teorici interpretativi. In questo lavoro, vengono presentati diversi approcci teorici con lo scopo di migliorare la descrizione stocastica della dinamica di macromolecole flessibili in soluzione. Nella prima parte sono quindi descritte ed applicate metodologie teoriche avanzate ai fini dell'interpretazione di esperimenti di risonanza magnetica. Successivamente, l'attenzione viene posta *i)* sul confronto tra la quantità di informazione contenuta in un esperimento e le capacità predittive di un determinato modello teorico interpretativo, basato sulla combinazione di simulazioni Monte Carlo di catene Markoviane ed il teorema di Bayes, *ii)* sulla definizione di una metodica identificativa di sottogruppi di atomi la cui dinamica puó essere trattata in maniera indipendente rispetto agli altri, basata su metodi di *clustering* dinamico ottenuti da simulazioni di dinamica molecolare, *iii)* sulla descrizione di una nuova modellistica descrittiva di macromolecole flessibili attraverso tecniche di operatori proiettivi; tale metodologia permette di adattare la descrizione del sistema

al livello di dettaglio oppertuno, in funzione dell'osservabile spettroscopica d'interesse.

Abstract

(Français)

La dynamique des protéines ocupe un rôle central dans la réalisation de leurs fonctions biologiques. Ces mouvements sont très divers, aussi bien sur le plan de l'échelle de temps concernée que de son amplitude ou des énergies mises en jeu. Dans le cas des protéines, les mouvements moléculaires peuvent être aussi bien localisés et rapides, par exemple dans le cas de rotations ou de vibrations des liaisons chimiques, que diffus, à l'échelle spatiale de toute la protéine et concernant des échelles de temps longues, comme dans le cas du repliement des protéines. Ces mouvements moléculaires se produisant sur des échelles de temps longues sont en réalité couplés à la mobilité rapide, les premiers pouvant être considérés comme des évènements rares dont l'origine se trouve dans la microdynamique de rotation autour des liaisons chimiques. Les spectroscopies de résonance paramagnétique électronique (RPE) et de résonance magnétique nucléaires (RMN) sont sensibles à la mobilité se produisant à différentes échelles de temps et en font ainsi des outils puissants pour l'étude de la dynamique moléculaire. Cependant, le contenu dynamique des informations obtenues expérimentalement nécessite le recours à des modèles théoriques et des méthodologies d'analyse des données permettant d'interpréter les observations. Dans ce travail, nous présenterons diverses approches théoriques permettant d'améliorer la description de la dynamique stochastique de macromolécules flexibles en solution. Après avoir rappelé quelques méthodes et modèles usuels d'interprétation des expériences de relaxation de spins nucléaires, nous aborderons les thèmes principaux de ce travail, qui se répartissent sur trois axes. Il s'agit d'une part de revisiter la question de l'interprétation des mesures de vitesses de relaxation RMN afin d'extraire des paramètres dynamiques, et de préciser les attentes et les limites d'une telle approche. D'autre part, nous présenterons une nouvelle méthode, reposant sur une analyse de

simulations de dynamique moléculaire et consistant à identifier des sous-groupes d'atomes d'une protéine dont la dynamique peut être traitée indépendamment. Enfin, une approche nouvelle permettant de décrire la dynamique de macromolécules flexibles et utilisant des opérateurs de projection sera proposée. Cette dernière permet en particulier d'adapter la description de la molécule étudiée au niveau de description souhaité selon la technique spectroscopique utilisée et l'observable mesurée.

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

FIGURE **Page**

CHAPTER 1

INTRODUCTION

Internal and overall dynamics of proteins and macromolecules in general, are critically involved in the determination and the regulation of their physical and chemical properties, biological functions and spectroscopic signatures. Examples of dynamic-controlled classes of processes are the allosteric effects in enzyme catalysis, the formation of non-specific transient encounter complexes in the protein-protein association[2, 3] and the regulation of molecular recognition.

Monitoring and describing proteins dynamics is therefore a fundamental area of investigation in modern physical chemistry. Experimental study of protein dynamics is divided into two main strands. The former, which we shall call “ensemble techniques” is related to experiments done on samples containing order of the number of Avogadro of molecules. The latter regards single-molecule experiments [4] that are conducted on a very reduced number of molecules (ideally only one). A short overview of these methods is given in the following.

Ensemble techniques

Nuclear magnetic resonance

Nuclear magnetic resonance (NMR) is a widely employed technique in studying both structure and dynamics of proteins [5–13]. Information on dynamics is obtained interpreting relaxation

and cross-relaxation measurements (as T_1, T_2 and Nuclear Overhauser Effect, NOE) of spin labeled nuclei, in particular ^{15}N, ^{13}C and ^{2}H, by means of a proper model. The two most employed methhos are the Model Free (MF) approach [14, 15] and the Slowly Relaxing Local Structure (SRLS) model [16–18] which is described in detail in Section 2.1.2. Both the models associate the relevant dynamics of the protein to that of two diffusive stochastic rotors, one describing the rotational motion of the protein and the second collecting the local motions in which the probe (the aminoacid) is involved. SRLS is based on the full description of the diffusive model, taking into account rigorously the coupling of the two rotors due to a potential of mean force that emulates the constraints that chemical bonds and non-local interactions impose to the motion of the residue wiht respect to the rest of the protein. MF, instead, is based on the statistical decoupling of the rotators and represents an approximation of SRLS, which is valid within certain limits ("rigid" residues in globular proteins). The kind of information that can be extracted from NMR using these models is both structural and dynamical and regards local properties, i.e. residue-specific knowledge is gained from the analysis. In particular, from the structural point of view NMR analysis gives information on the local potential of mean force acting on the probe, thus having access to the amplitude of motion of each residue and the kind of possible movements. Secondly, on the dynamical point of view, information on local diffusive properties and correlation time scales is obteined. This kind of knowledge is important to understand how locally residues can induce / permit conformational changes, make the protein adapt to binding substrates, adapt to external stimuli, etc.

Recently high interest arose on a different kind of observable measured with NMR: the Residual Dipolar Coupling (RDC). The protein is constrained to not equally sample all the rotations, e.g. by use of diluite liquid crystals [19]. In this way, the anisotropic magnetic interactions, specially the dipolar-dipolar interaction, are not averaged. On one hand, this leads to a very precise definition of the orientation of internuclear bonds relative to a molecular-fixed frame, making RDC powerful for structure determination of proteins. On the other hand, RDC measurements report on averages over relatively long time-scales (up to millisecond range), opening access to dynamic information complementary to motions detected from NMR spin relaxation studies [20].

Electron paramagnetic resonance

Electron paramagnetic resonance (EPR) spectroscopy is widely employed to obtain information about molecular dynamics of system into which a stable free-radical probe group has been introduced. The most frequently used probes are nitroxides, which are often covalently attached to a particular residue of interest in proteins (site-directed spin labeling, SDLS-EPR).[21] The EPR specturm of a nitroxide is sensitive to molecular reorentation because the magnetic interactions of the unpaired electron with the applied magnetic field, as well as those with the nuclei on the probe, are inherently anisotropic. Depending on the EPR frequency, molecular motion is broadly classified in three regimes that depend on the relative magnitudes of the characteristic time scale of the motion, τ_c, and the inverse of the frequency width, $\Delta\omega$, of the spectrum. In the fast motion regime, $\tau_c\Delta\omega \ll 1$, the anisotropic interactions are averaged out, leading to simple Lorentzian line shapes are observed, and only estimates of molecular parameters (e.g., diffusion tensor values) are obtained. At very long correlation times, a static distribution of all the possible orientations, the "rigid limit" spectrum, is observed. When $\tau_c\Delta\omega \simeq 1$, the motion is said to fall in the slow motion regime for the given EPR frequency. The interpretation of slow motional spectra requires an analysis based upon sophisticated theory, combining the world of quantum mechanics, as far as the parameters of the spin Hamiltonian are concerned, and the world of molecular dynamics and statistical thermodynamics for the spectral lineshapes. Interpretation of EPR spectra will be discussed in Section 2.2.

Fluorescence anisotropy decay

The triptophan (Trp) aminoacid is a useful chromophore for UV studies of proteins. In particular time resolved fluorescence can give access to proteins dynamics [22]. In particular what is measured is the anisotropy decay of Trp probe fluorescence which, through appropriate moteling, can be associated to local motional properties of the probe and its surroundings. Thus the information

obtained from this kind of measurements is similar to that obtained from NMR and require some specific modeling of protein motions, specially when the technique is used to probe mobile regions as loops [23].

The fact that the information on dynamics that can be extracted is localized around the probe makes this technique not completely exhaustive because the probes are not distributed all over the protein structure as it happens in NMR, where all the residues contain the probe (a part proline in N-H relaxation measurements). Site specific labeling by mutating residues to Trp aminoacids can be used to improve the investigation of protein dynamics using fluorescence anisotropy decay. Anyway, care must be paid when exptrapolating information collected from mutants to the wild-type protein, specially in the case of very flexible systems.

Time-resolved X-Ray

A recently developed technique in studying protein dynamics is time-resolved X-ray scattering, which has been succesfully applied to the study, in crystal, of kinetics of CO migration from myoglobin to the water layer surrounding the protein.[24] Because X-ray crystallography gives direct access to the electron density of the molecule no model is required, a part the atomic model, to follow the molecular dynamics. So, it is possible to directly "watch" proteins moving while carrying out their function. However, a couple of issues are still to be solved.

The former is related to time resolution of the technique. At present it is around 100 ps. This was sufficient to carry out the cited study on myoglobin, but is certainly too large for detailed study of reactions, where bonds are broken and formed at very shorter time scales. A possible breaktrough seems to be given by a new technique for producing very bright X-ray beams which is called X-ray free electron laser [25]. It seems that this technique can produce 0.1 fs pulses, which are compatible to atomic motions.

The second issue concerns the fact that not all proteins can be crystallized and many processes cannot occur in the crystal. Moreover, doubts can be arisen about the extrapolation to solution of mechanisms and kinetics studied in the crystal state. To overcome this problem, time-resolved

X-ray scattering has been ported also to solution techniques. In particular small angle X-ray scattering (SAXS) and wide angle X-ray scattering (WAXS) techniques have been used to study the same kinetics of CO migration from myoglobin in water.[26] It was shown that the kinetic scheme of the reaction can be recognized and followed using the combined information from SAXS and WAXS, looking at the system for 10 ms. Unfortunately this approach cannot give a direct information on 3D structure of the molecule and it is not possible to directly access the molecular dynamics, as is possible in the crystal. To solve this problem a search of the correct structure giving simulated SAXS/WAXS spectra corresponding to experimental ones need to be used.

Single-molecule techinques

Förster fluorescence resonance energy transfer

A spectroscopic technique for the single-molecule study of molecular dynamics is Förster fluorescence resonance energy transfer (FRET).[27] It requires the presence, in the same molecule of both a donor fluorophore (DF) and an acceptor fluorophore (AF). What is meaused is the non-radiative transfer of energy from the excited DF to AF *via* a strongly distance-dependent dipole-dipole coupling. The measured FRET efficiency provides information on long-range molecular distances in the range of 20 - 100 Å. When performed at the single-molecule level, FRET studies can yield information about heterogeneities in terms of conformations and conformational dynamics that are unavailable from ensemble measurements. The particular ability of this technique of detecting different conformations has been exploited mainly in the investigation of folding/unfolding pathways and kinetics. Recent applications are on the determination of an upper bound for transition path times in folding of labeled small protein GB1 for which a simple two-state folding path exists. [28] FRET has also given new insights in the effect of molecular chaperones in affecting protein folding. In particular of rhodanese protein, by analyzing FRET trajectories outside and inside a chaperonin cage.[29] Another example of application of FRET is on the determination of rates of the fast-folding protein α_3D.[30]

In all the cited cases a two-state model was used to describe the folding kinetics. However, FRET, which is sensitive to different conformations, can be potentially used to detect a larger number of

subpopulations and the kinetics of transitions among them.

Opical tweezers

Optical tweezers are ideally suited to perform force microscopy experiments which isolate a single biomolecule, which then provides multiple binding sites for ligands [31]. At the extremes of the biomolecule are attached two beads. One is kept fixed, while the second one is trapped by a laser beam focused on the bead. Moving the laser makes the bead move, stretching the biomolecule. This technique as been used mainly in studies regarding DNA mechanical properties and binding to proteins. [31, 32] Also biological motors have been studied with optical tweezers [32]. The methodology is suited even for the direct study of folding kinetics [33]. The protein is unfolded by pulling away the two terminal residues and then system is released to let the protein re-fold. The profiles of the unfolding and folding forces versus extension contain important information on kinetics and thermodynamics of the process.
Coupling experimental observations to molecular dynamics simulations makes optical tweezres an important tool in the understanding the dynamic behaviour of individual protein molecules at the single-molecule level.

Atomic force microscopy

The atomic force microscopy (AFM) technique is a quite versatile tool for the *in singulo* study of biomolecules. In particual two kinds of experiments can be performed. The first is AFM imaging, which can lead to both structural and dynamical information. AFM imaging is performed by absorbing the biomolecule(s) on a solid surface (usually mica) and scan the surface with a tip. What is registered is the force required to keep the tip at a certain distance from what's below it and a sort of topographic image is obtained. Recent examples of studies of biomolecules with AFM imagaing regard the incorporation of membrane proteins in single lipid bilayers [34] and properties of single-stranded DNA-binding protein - DNA complexes. [35] Furthermore, with

time-resolved AFM imaging it is possible to study dynamic processes and their kinetics at single-molecule level. An example of application of this method is the kinetics of association/dissociation of the complex between the *E. Coli* chaperonin GroEL and its co-chaperonin GroES [36].
AFM can also be used in a way similar to optical tweezezers or hydrodynamic manipulation by attaching one ending of the molecule to the tip. With this setup it is possible to get information on mechanical properties of biomolecules [37, 38] as well as binding and folding forces [39]. As for the previously mentioned "dragging" experiments, coupling with molecular dynamics simulations is very important in order to interpret at atomistic level the experimental observations. An exmaple of this kind of approach can be found in the literature applied to the study of mechanical extension of a 26-residue long polyalanine chain. [40] Application of the idea to biomolecules require high performance calculation resources (both hardware and software) and *ad hoc* interpretative tools in order to understand the complex information that molecular dynamics simulations provide.

As previously stated, the acquired experimental observations need in most cases to be rationalized in order to acquire a meaningful description of (some of) the many complex relaxation processes, therefore of the underlying motions. These include global reorientation and small or large amplitude motions of entire domains, as well as limited local readjustments and restricted single-residue motions. In general, different spectroscopic techniques probe different physical observables which, in addition, provide information on motions taking place at different time windows. It seems therefore particularly important to introduce relevant sets of coordinates, the definition of which depends on the observable involved in the experimental method. This consideration is especially important in the case of magnetic resonance spectroscopies for which modeling methods for internal relaxation processes have been developed early on. Once, however, a precise model for the internal and global dynamics of the macromolecule has been defined, interpretation (and in a few cases, again mostly beloging to the subset of relaxation magnetic resonance techniques) of complex experimental results is fully feasible.

Naturally, the apparent simplicity of the above statement *"a precise model for the internal and global dynamics of the macromolecule has been defined"* is deceptive. Unless relatively small systems (such as oligopeptides) or well-defined, in terms of structural and internal dynamical

properties, are considered, the actual predictive power of most theoretical approaches is based on a trial-and-error philosophy: *i)* define a suitable set of markovian internal variables, describing the relevant motional properties of the macromolecule, *ii)* attempt to interpret any available experimental datum, *iii)* refine choice of point *i)* until a satisfactory interpretation is reached. It is tempting to bypass this modelistic approach by using brute force approaches (e.g. extensive molecular dynamics simulations, MD), but in practice the complexity of the macromolecule properties often defies this shortcut.

In this thesis we shall propose a number of investigations, carried out partly at the University of Padova and partly at the École Normale Supérieure in Paris, having the general purpose 1) of understanding some of the major difficulties implied by a modelistic approach to protein motions, or rather their limited description for the interpretation of EPR and NMR relaxation experiments, and 2) of proposing some possible novel developments to partially overcome such difficulties. Basically, one can identify several steps, requiring a careful definition of appropriate theoretical methods

1. The description of the dynamics of a large object, such as a protein, requires a careful definition of molecular frames to which the motions can be referred. Therefore, in general, some geometrical considerations are needed in order to actually relate a dynamic model with a set of observables amenable to experimental measurement, and care should be taken to accurately account for the tensorial nature of the magnetic interactions, by defining proper frames of reference; after this is done, integrated computational approaches based on a combination of stochastic models and effective Liouville dynamics can be computationally solved to describe most systems. *Chapter 2 presents a self-contained example of advanced interpretation of an actual experiment.*

2. The technical solution of the resulting theoretical apparatus (in the form an augmented stochastic equation) can be carried out, nowadays, with relative easiness; however one has first to question if the model employed is well-defined, i.e. if the information contained in the experiment is sufficient to justify a complex theoretical apparatus. *Chapter 3 discusses the conditions under which information on complex dynamics can be reliably obtained from*

NMR relaxation data.

3. Most existing modelistic approaches for intepreting relaxation experiments rely on a phenomenological definition of variables describing the molecular internal dynamics; a self-consistent procedure to identify internal degrees of freedom relaxing in different time-scales would be needed, as the basis for a rational model. *Chapter 4 is devoted to build a protein motion analysis on suitably chosen cross-correlated functions of the atomic coordinates, which can be used to perform a cluster reduction of the protein.*

4. Finally, once a number of effective coordinates has been selected, in order to apply the full machinery of stochastic methods, the problem remains of defining their reduced dynamics. *Chapter 5 introduces the theoretical framework for the derivation of stochastic descriptor of flexible macromolecules in solution from atomistic models, via projection operators. An example of an approximate approach to the interpretation of magnetic resonance relaxation experiments is presented.*

CHAPTER 2

INTEGRATED COMPUTATIONAL APPROACHES TO THE INTERPRETATION OF MAGNETIC RESONANCE RELAXATION

Interpretation of structural, and dynamical behaviour of molecules is of fundamental importance to understand their reactivity, biological function and activity. In general, one has to deal with complex systems in which a wide range of time scales are present, from localized fluctuations involving selected chemical groups (ps and fs) to global dynamics (μs and slower). Physico-chemical properties of molecules and macromolecules in solution depend on the action of different motions at several time scales. Information on multiscale dynamics can be gained, in principle, by a variety of spectroscopic techniques. In this work we focus our attention on magnetic spectroscopies, both electron paramagnetic resonance (EPR) and nuclear magnetic resonance (NMR).

2.1 Interpretation of NMR relaxation experiments

2.1.1 Introduction

Nuclear magnetic resonance spectroscopy showed to be a powerful method to elucidate protein dynamics because of the possibility to interpret nuclear spin relaxation properties in terms of

microdynamic parameters. Magnetic relaxation rates R_1, R_2 and nuclear Overhouser enhancement (NOE, η_{NH}) of ^{15}N, ^{13}C and ^{2}H nuclei, depend on dipolar (^{15}N and ^{13}C) and quadrupolar (^{2}H) interactions, chemical shift anisotropy (CSA) and cross-correlation effects. When the spin probe is the X-^{1}H bond, following standard Bloch theory [41], it is possible to express the NMR relaxation rates as functions of the spectral densities $J(\omega)$ evaluated at the Larmor frequencies 0, ω_H, ω_X, and $\omega_{\mathrm{H}\pm\mathrm{X}} = \omega_{\mathrm{H}} \pm \omega_{\mathrm{X}}$:

$$
\begin{aligned}
\eta_{\mathrm{NH}} &= 1 + \frac{\gamma_H}{\gamma_X}\frac{d^2}{R_1}\left(6J(\omega_{\mathrm{H+X}}) - J(\omega_{\mathrm{H\text{-}X}})\right) \\
R_1 &= d^2\left(3J(\omega_{\mathrm{X}}) + J(\omega_{\mathrm{H\text{-}X}}) + 6J(\omega_{\mathrm{H+X}})\right) + 2c^2 J(\omega_{\mathrm{X}}) \\
R_2 &= d^2\Big[2J(0) + \frac{3}{2}J(\omega_{\mathrm{X}}) + \frac{1}{2}J(\omega_{\mathrm{H\text{-}X}}) \\
&+ 3J(\omega_{\mathrm{H}}) + 3J(\omega_{\mathrm{H+X}})\Big] + c^2\Big(\frac{4}{3}J(0) + J(\omega_{\mathrm{X}})\Big)
\end{aligned}
\tag{2.1}
$$

where $d = \mu_0\hbar\gamma_{\mathrm{H}}\gamma_{\mathrm{X}}/4\sqrt{10}\pi\langle r^3_{\mathrm{XH}}\rangle$, $c = \gamma_{\mathrm{X}}B_0\Delta\sigma_{\mathrm{X}}/\sqrt{15}$, and r_{XH} is the XH distance. The parameters γ_H and γ_X are the gyromagnetic ratios of X and ^{1}H atoms, respectively, μ_0 is the vacuum magnetic susceptibility, $\hbar$ is the reduced Planck constant, and $\Delta\sigma_{\mathrm{X}}$ is the ^{15}N chemical shift anisotropy. Usage of stochastic modeling in simulating/interpreting NMR relaxation of complex systems dates back to 1982, when Lipari and Szabo [42, 43] introduced the so-called “Model-Free” (MF) approach, and applied it to globular proteins. In the MF framework, an analytic expression for the correlation function (from which the spectral density is obtained) is introduced. The whole system is then described by the combination of two uncoupled motions: the global tumbling of the whole molecule and the effective local motion of the ^{15}N-^{1}H (or ^{13}C-^{1}H) bond. The correlation function takes the form of a bi-exponential, which in many cases, is sufficient to fit NMR data. However, parameters entering the bi-exponential have a physical interpretation only if timescale separation between global and local dynamics is valid, limiting the cases of application.[44].

In parallel, a model that takes into account those cases excluded by MF analysis and allows a more complete local description of the protein has been introduced in the early 90s, the “slowly relaxing local structure” (SRSL) model. [45, 46] SRLS consists on a two-body Smoluchowski equation that describes the time evolution of the density probability of two relaxation processes (at different time scales) coupled by an interaction potential. To the description of protein dynamics, the two relaxing processes are interpreted as the slow global tumbling of the whole protein and the

relatively fast local motion of the spin probe, the local motion (*e.g.* the ^{15}N-^{1}H bond). Both the processes are described as rigid rotators the motion of which is coupled by a potential correlating their relaxation and that is interpreted as the local ordering that the molecule imposes to the probe. In the following section we give a summary overview on how the SRLS model is applied to the interpretation of NMR data.

2.1.2 The SRLS model

SRSL model is a two-body (protein and probe) coupled rotator model. To give a full definition of the model it is necessary to introduce a number of reference frames. In what follows, take Fig. 2.1 as reference:

- LF is the fixed inertial laboratory frame;
- M_1F is the protein fixed frame where the diffusion tensor of the protein, $\boldsymbol{D}_1$, is diagonal;
- M_2F is the protein fixed frame where the diffusion tensor of the probe, $\boldsymbol{D}_2$, is diagonal;
- VF is the protein fixed frame having the z-axis aligned with the director of the orienting potential;
- OF is the probe fixed frame the z-axis of which tends to be aligned to the director of the potential;
- DF is the probe fixed frame where the dipolar interaction is diagonal;
- CF is the probe fixed frame where the chemical shift tensor is diagonal.

To complete the picture, we have to define the set of Euler angles that transform among the different frames:

- $\boldsymbol{\Omega}_L$ transform from LF to VF, while $\boldsymbol{\Omega}_{LO}$ transform from LF to OF;
- $\boldsymbol{\Omega}$ transform from VF to OF;
- $\boldsymbol{\Omega}_V$ transform from M_1F to VF;

- $\boldsymbol{\Omega}_O$ transform from M_2F to OF;

- $\boldsymbol{\Omega}_D$ transform from OF to DF, while $\boldsymbol{\Omega}_{OC}$ transform from OF to CF;

- $\boldsymbol{\Omega}_C$ transform from CF to DF.

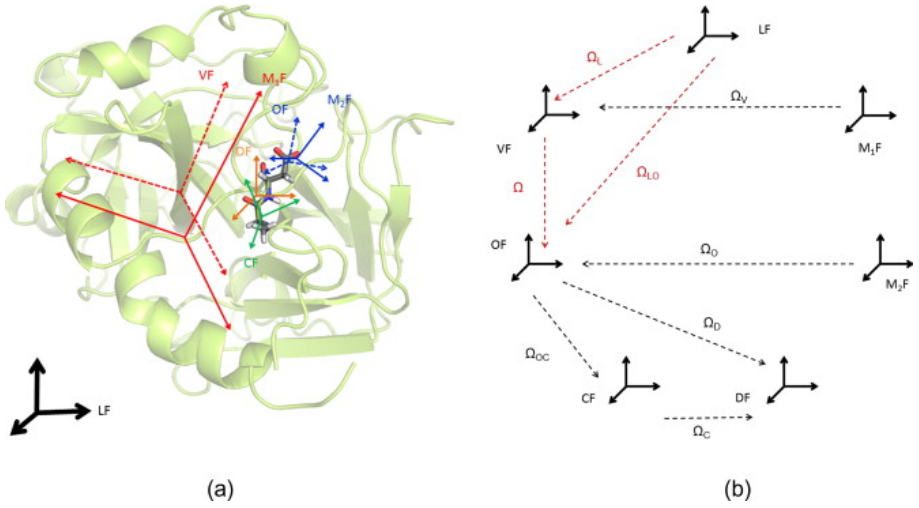

Figure 2.1: Definition of frames and Euler angles in the SRLS model applied to NMR.

The system is fully described with two set of stochastic Euler angles and in particular our choice is on the set of Euler angles $\boldsymbol{\Omega}_L$, giving the orientation of the protein respectively to the laboratory frame, and $\boldsymbol{\Omega}$, which represent the relative orientation of the probe and the protein. Using this set of stochastic variables, $\boldsymbol{X} = (\boldsymbol{\Omega}, \boldsymbol{\Omega}_L)$, the diffusion operator describing the time evolution of the density probability of the system is

$$\begin{aligned} \hat{\Gamma}(\boldsymbol{X}) \quad = \quad & {}^{O}\hat{\boldsymbol{J}}^{\dagger}(\boldsymbol{\Omega})\,{}^{O}\boldsymbol{D}_2 P_{eq}(\boldsymbol{X})\,{}^{O}\hat{\boldsymbol{J}}(\boldsymbol{\Omega}) P_{eq}^{-1}(\boldsymbol{X}) + \\ & + \left[{}^{V}\hat{\boldsymbol{J}}(\boldsymbol{\Omega}) - {}^{V}\hat{\boldsymbol{J}}(\boldsymbol{\Omega}_L)\right]^{\dagger} {}^{V}\boldsymbol{D}_1 P_{eq}(\boldsymbol{X}) \left[{}^{V}\hat{\boldsymbol{J}}(\boldsymbol{\Omega}) - {}^{V}\hat{\boldsymbol{J}}(\boldsymbol{\Omega}_L)\right] P_{eq}^{-1}(\boldsymbol{X}) \end{aligned} \tag{2.2}$$

where ${}^{O}\boldsymbol{D}_2$ is the diffusion tensor of the probe in OF, ${}^{V}\boldsymbol{D}_1$ is the diffusion tensor of the protein in VF and the equilibrium distribution, $P_{eq}(\boldsymbol{X})$ is given by

$$P_{eq}(\boldsymbol{X}) = \mathcal{N} \exp[-V(\boldsymbol{\Omega}, \boldsymbol{\Omega}_L)/kT] \tag{2.3}$$

with k the Boltzmann constant and T the absolute temperature.

We will assume that the protein is immersed in an isotropic medium, so the equilibrium distribution is independent on $\boldsymbol{\Omega}_L$ and the total potential is only the interaction potential between the

two processes for which we take the following expansion over Wigner matrices:

$$\begin{aligned} -V(\mathbf{\Omega})/kT &= c_0^2\mathscr{D}_{00}^2(\mathbf{\Omega}) + c_2^2\left[\mathscr{D}_{0-2}^2(\mathbf{\Omega}) + \mathscr{D}_{02}^2(\mathbf{\Omega})\right] + c_0^4\mathscr{D}_{00}^4(\mathbf{\Omega}) + \\ &+ c_2^4\left[\mathscr{D}_{0-2}^4(\mathbf{\Omega}) + \mathscr{D}_{02}^4(\mathbf{\Omega})\right] + c_4^4\left[\mathscr{D}_{0-4}^4(\mathbf{\Omega}) + \mathscr{D}_{04}^4(\mathbf{\Omega})\right] \end{aligned} \tag{2.4}$$

Due to the fact that this is a pure rotational problem, observables are expressed as spectral densities, i.e. Fourier - Laplace transforms of correlation functions of Wigner functions of the absolute probe Euler angles, $\mathbf{\Omega}_{LO} = \mathbf{\Omega} + \mathbf{\Omega}_L$

$$j_{k,k'}(\omega) = \langle \mathscr{D}^j_{m\,k}(\mathbf{\Omega}_{LO}) P_{eq}(\mathbf{\Omega}_{LO}) | \left(i\omega - \hat{\Gamma}\right)^{-1} | \mathscr{D}^{j'}_{m'\,k'}(\mathbf{\Omega}_{LO}) P_{eq}(\mathbf{\Omega}_{LO}) \rangle \tag{2.5}$$

Considering the symmetry of the magnetic interactions (dipolar and chemical shift anisotropy) contributing to the spin Hamiltonian of the system for ^{15}N-^{1}H probe, only physical observables with $j = j' = 2$ and $m = m' = 0$ have to be considered.

From these spectral densities it is possible to calculate the spectral densities for every magnetic interaction, μ (dipolar, CSA), as

$$J^{\mu}(\omega) = \sum_{k,k'=-2}^{2} \left[\mathscr{D}^{2\,*}_{k\,0}\left(\mathbf{\Omega}_{\mu}\right)\mathscr{D}^{2}_{k'\,0}\left(\mathbf{\Omega}_{\mu}\right)\right] j_{k,k'}(\omega) \tag{2.6}$$

being $\mathbf{\Omega}_{\mu}$ the set of Euler of angles transforming from OF to the frame where the μ-th magnetic tensor is diagonal.

Calculation of spectral densities $j_{k,k'}(\omega)$ is achieved by spanning the diffusive operator over a proper basis set. In such a way one moves the problem of calculating integrals in eq. (2.5) to a classical linear algebra problem. The basis onto which the operator is spanned is given by the direct product $|\Lambda\rangle = |\lambda_1\rangle \otimes |\lambda_2\rangle = |L_1 M_1 K_1\rangle \otimes |L_2 M_2 K_2\rangle$, where

$$|L_1 M_1 K_1\rangle = \sqrt{\frac{(2L_1+1)}{8\pi^2}}\,\mathscr{D}^{L_1}_{M_1 K_1}(\mathbf{\Omega}_L) \tag{2.7}$$

$$|L_2 M_2 K_2\rangle = \sqrt{\frac{(2L_2+1)}{8\pi^2}}\,\mathscr{D}^{L_2}_{M_2 K_2}(\mathbf{\Omega}) \tag{2.8}$$

This basis is infinite and to practically solve the problem the expansion have to be truncated at a certain value of the principal numbers L_1 and L_2. For what concerns the basis expansion for the protein ($\{\lambda_1\}$) the truncation is fixed by the symmetry of the physical observables to $L_1 = 2$ and

$M_1 = 0$. So only one truncation parameter remains, i.e. L_2. Given a maximum value, $L_{2,MAX}$, the dimension of the basis (in absence of other symmetries) will be

$$N = 5 \sum_{i=0}^{L_{2,MAX}} (2i+1)^2 = \frac{5}{3}\left(L_{2,MAX}+1\right)\left(2L_{2,MAX}+1\right)\left(2L_{2,MAX}+3\right) \tag{2.9}$$

It is simpler to work with auto-correlation functions so instead of calculating directly spectral densities in eq. (2.5) we define the function $2C_{k,k'} = \mathscr{D}^2_{0k} + \mathscr{D}^2_{0k'}$ and calculate the symmetrized spectral densities

$$j^S_{k,k'}(\omega) = \langle C_{k,k'}(\mathbf{\Omega}_{LO}) P_{eq}(\mathbf{\Omega}_{LO}) | \left(i\omega - \hat{\Gamma}\right)^{-1} | C_{k,k'}(\mathbf{\Omega}_{LO}) P_{eq}(\mathbf{\Omega}_{LO}) \rangle \tag{2.10}$$

and then obtain the $j_{k,k'}(\omega)$ functions as linear combinations of the symmetrized spectral densities:

$$j_{k,k'}(\omega) = \left[2\left(1+\delta_{k,k'}\right) j^S_{k,k'}(\omega) - j^S_{k,k}(\omega) - j^S_{k',k'}(\omega)\right]/10 \tag{2.11}$$

The absence of any assumptions on timescale separation between global and local motion and the intrinsically tensorial description, makes the SRLS model extremely versatile and it has been extensively applied to the interpretation of NMR relaxation in proteins, and more recently also to polysaccarides and soft coated metal nanoparticles. [47] In the following we report an example of interpretation of NMR relaxation measure on *E. Coli* Ribonuclease HI, taken from Ref. [48].

2.1.3 SRLS analysis of ^{15}N NMR spin relaxation from *E. Coli* Ribonuclease HI

Ribonuclease HI (RNase H) is an endonuclease that hydrolizes the RNA strand in RNA-DNA hybrid molecules [49, 50]. *E. Coli* RNase is a polypeptide chain composed of 155 amino acid residues. The three-dimensional structure has been resolved both from X-ray crystallography [51] and NMR [52]. The RNase H is composed by five α-helices (residues 43 to 58, 71 to 80, 81 to 88, 100 to 112, and 127 to 142), and five β-strands (residues 4 to 13, 18 to 27, 32 to 42, 64 to 69 and 115 to 120).The backbone structure of RNase H is shown in Figure 2.2. Results from NMR spectroscopy suggest that the loops between β_1 and β_2, α_C and α_D, and β_5 and α_E partecipate in substrate binding. [53, 54] Several analyses of RNase H NMR relaxation data based on model-free can be found in the literature [51, 55–57]. Nevertheless, several significant inconsistencies

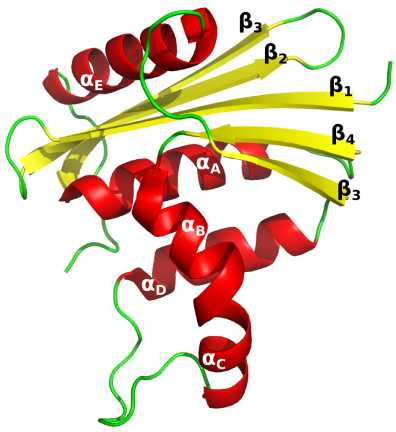

Figure 2.2: Structure of *E. Coli* RNase H. α-helices denoted α_A to α_E and β-strands denoted β_1 to β_5 (see text). The figure was drawn with the software PyMOL and using the PDB coordinate file 1RNH. [1]

emerged. For example, ^{15}N relaxation parameters of RNase H were also acquired at 14.1 and 18.8 T [58] to investigate the variability of the ^{15}N chemical shift anisotropy (CSA) tensor. The methodology used, in addition to good statistics, required that the N-H bonds employed in the analysis be free of slow local motional effects and conformational excange contributions, R_{ex}. these requirements were fulfilled by the single-field data. However, it has been found in combined data acquired at two or three magnetic fields yield different results. [59], R_{ex} contributions, often associated with different residues at different magnetic fields, emerged for approximately 50% of the relevant N-H bonds. clearly, many of these R_{ex} terms are artificial. Few of them can be eliminated accounting for an axial global diffusion tensor $\boldsymbol{D_1}$ with $D_{1,\parallel}/D_{1,\perp} = 1.23$ using standard MF methods. [56] However, it has been shown that such small deviations of $\boldsymbol{D_1}$ from spherical simmetry can be ignored. [60] Hence, one has to search for effects other than $\boldsymbol{D_1}$ axiality having been absorbed by the artificial R_{ex} terms.

The tensorial perspective offered by SRLS model can provide quantitative information on the strength of the local ordering and the magnitudes of the local motional rates at every N-H site in the protein. Figure 2.3 shows the results of the SRLS analisys on relaxation data measured at 14.1 and 18.8 T.

The profile of the local order parameter, S_0^2 (Figure 2.3a, left) suggests that in general, the

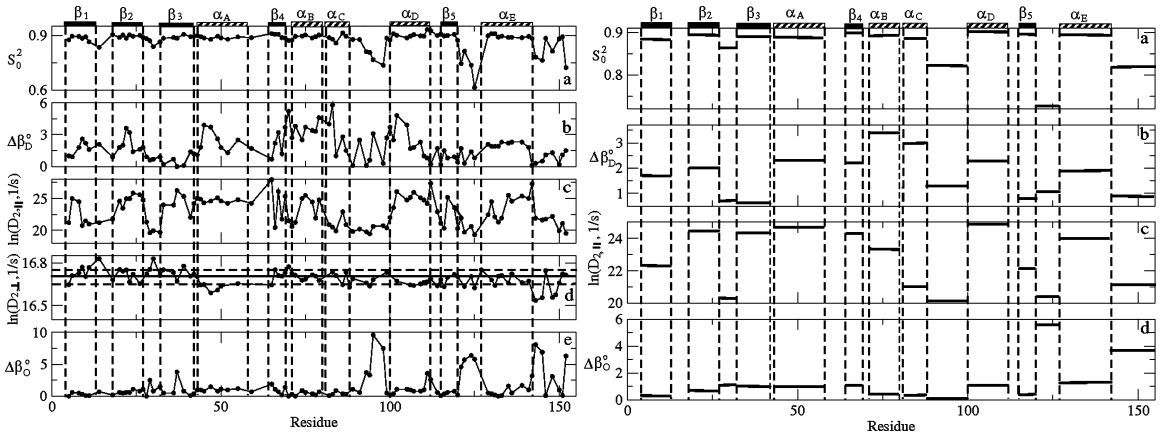

Figure 2.3: Left, best-fit value of S_0^2 ($S_0^2 = \langle D_{0,0}^2(\Omega_{VF-OF})\rangle$) (a), $\Delta\beta_D$ (b), ln($D_{2,\parallel}$, 1/s) (c), ln($D_{2,\perp}$, 1/s) (d) and $\Delta\beta_O$ (e), the tilt angle of the principal axis of the local ordering tensor, $\boldsymbol{S}$, from the N-H bond. Results were obtained by allowing c_2^0, $D_{2,\parallel}$, $D_{2,\perp}$, β_D, β_O to vary during the fitting routine. The errors are estimated at 2% for S_2^0, $\Delta\beta_D$, ln($D_{2,\parallel}$) and $\Delta\beta_O$ and 5% for ln($D_{2,\perp}$). For the calculations, ^{15}N CSA value of -172 ppm, a bond length r_{NH} = 1.02 Å, and a -17◦ tilt angle between the ^{15}N-^{1}H dipolar and ^{15}N CSA tensor frames. Right, average values for the various secondary structure elements and loops calculated from the results on the left.

loops exhibit lower local ordering than the secondary structure elements. In particular, the loops α_C/α_D and β_5/α_E exhibit significantly lower local ordering.

The deviation of the principal axis of the local ordering tensor from the $C_{i-1}^{\alpha} - C_i^{\alpha}$ axis, $\Delta\beta_D$ (Figure 2.3b, left) does not exceed 6°. Within the seconday structure elements the N-H bond vector fluctuations, described by $D_{2,\parallel}$ are fast, (described as $ln(D_{2,\parallel}, 1/s)$ in Figure 2.3c). The correlation time, $\tau_{\parallel}$ = 1/(6$D_{2,\parallel}$), extends from 3 to 125 ps. Slower N-H bond vector fluctuations are detected for the loops. The average values for the α_C/α_D and β_5/α_E loops are respectively 344 and 208 ps (Figure 2.3c, right). The perpendicular component of the local diffusion tensor $D_{2,\perp}$ is shown as $ln(D_{2,\perp}, 1/s)$ in Figure 2.3d. It can be seen that in most cases $ln(D_{2,\perp}, 1/s)$ is virtually the same as $ln(D_1, 1/s)$. This is consistent with the equilibrium orientation of the N-H bond being fixed in the protein backbone, with the only local motional mode experienced being N-H fluctuations. Loops β_1/β_2 and β_2/β_3 exhibit values of $D_{2,\perp}$ higher than D_1. This is consistent with localized ns motion of the protein backbone, taking place in addition to the collective rocking motions occurring on the 100 ns time scale, detected with solid-state NMR. [61] These two loops, together with loop β_5/α_E have been suggested to participate in substrate binding.

2.1.3.1 Conclusions

SRLS describes the local restrictions in terms of second-rank ordering tensor and the local restrictions in term of a second-rank diffusion tensor. When these tensors are taken axial but their principal axes are tilted, they describe an asymmetric setting, giving an insightful tensorial description of the N-H structural dynamics. The local ordering is found to be stronger within secondary structure elements ($\langle S_0^2 \rangle = 0.89$) in respect to the loops ($\langle S_0^2 \rangle = 0.84$). The principal local ordering axis is nearly parallel to the $C_{i-1}^{\alpha} - C_i^{\alpha}$. The parallel component of the local diffusion tensor is nearly the same of the global tumbling. The parallel component of the local diffusion tensor represents N-H bond vector fluctuations centered around the equilibrium N-H orientation. Fluctuations occour with average correlation times of 3-125 ps for secondary structure elements, 125-344 ps for loops, and 125 ps for the C-terminal part.

2.2 Interpretation of EPR relaxation experiments: characterization of a set of rigid 3_{10}-helical peptides with TOAC nitroxide spin labels

2.2.1 Introduction

Electron paramagnetic resonance (EPR) of spin-labeled compounds has become a powerful technique in biological structure determination. Most commonly, this latter issue relies on measuring distances between two paramagnetic centers, often spin labels, covalently linked to well defined positions in the biomacromolecule of interest. The methodologies to assess such distances by EPR are limited because: *i)* they work well for frozen solutions at low temperatures and *ii)* distance ranges between 0.8 and 1.5 nm are difficult to address. [62] Physiological conditions such as liquid solutions at room temperature pose additional challenges. The dipolar interactions between spins, so far the most reliable indicator for distance, can be partially averaged in liquid solution. Therefore, the isotropic exchange interaction , being of the short-distance (several tenths

of nanometer) type, is difficult to interpret in terms of separation between spins. Moreover, in liquid solution, the spin-spin interaction is extracted from lineshape. More specifically, since the differences in the spectra of the system of interest are evaluated in the absence/presence of the spin-spin interaction, weak spin-spin interactions and long separations are challenging to measure. In previous papers, [63–65] the authors synthesized and experimentally investigated by EPR a complete series of four 3_{10}-helical peptides, based on α-aminoisobutyric acid (Aib) , with pairs of TOAC (4-amino-1-oxyl-2,2,6,6-tetramethylpiperidine-4-carboxylic acid) nitroxide spin labels separated by three, four, five and six residues (see Table 2.1 for the exact amino acid sequences and the number of covalent bonds between the two TOAC labels). The nitroxide-containing TOAC is a residue as strongly helicogenic [66] as the well-known Aib,[67] in that they are both members of the same class of C_α-tetrasubstituted α-amino acids. Moreover, the TOAC side chain is rigidly connected to the peptide main chain so that the overall TOAC flexibility is reduced to a minimum. As reference compounds for our EPR analysis, we also investigated three size-matched mono-TOAC-labeled peptides. The bis-labeled peptides were classified according to the magnitude of the exchange interactions:i *i)* class I ($\simeq$800 MHz) with $HEPTA_{3,6}$ (three intervening residues) and $HEXA_{1,5}$ (four intervening residues) which show a large exchange interaction and five-line EPR spectra, and *ii)* class II ($<$ 9 MHz) with $OCTA_{2,7}$ (five intervening residues) and $NONA_{2,8}$ (six intervening residues) which exhibit a small exchange interaction and three-line EPR spectra. In this work, a full computational study is presented of the mentioned mono- and bis-labeled peptides employing an established integrated computational approach [68, 69] based on the definition and solution of a proper stochastic Liouville equation (SLE) for the system under study.[70] Such an approach has been applied with success in the interpretation of EPR spectroscopy of similar peptides, allowing the determination of molecular properties such as the main secondary structure in different solvents, proving the power of the interplay between EPR experiments and proper theoretical/computational modeling.[63, 64, 71] The coupling constant J is also a parameter of the SLE of the biradicals that need to be determined. One possibility, which would be in line with the philosophy of the ICA, is the calculation of this parameter. Quantum mechanical methods to access ab inito the value of J in biradicals are based on the difference in energy of the singlet and triplet states. Approaches based on density

functional theory [72–74], configuration interaction [75], and asymptotic methods [76, 77] have been applied to the calculation of J in biradicals for which the interaction is weak, were with "weak" it is intended that the exchange integral is on the order of $1 - 0.1$ cm^{-1}. This corresponds to energy differences of $10^{-3} - 10^{-4}$ kcal/mol. Such an accuracy is not reached in routine quantum mechanical calculations, for which it is usually of the order of 10^{-1} kcal/mol [78]. Higher accuracy can be obtained in configuration interaction-based methods, but those cannot be used to make calculations in reasonable times on medium-large molecules, like the poly-peptides studied in this work. Moreover, a coupling constant much smaller than the limits above mentioned is expected for the octa-, and nona-peptides, by inspection of their experimental EPR spectra. As will be shown in Section 2.2.3, the entity of the coupling has been found of the order of 0.1 Gauss, *i.e.* 10^{-3} cm^{-1} . This, in turn, means that if J had to be accessed by quantum mechanical calculations, energies more accurate of 10^{-6} kcal/mol would have been required: a still prohibitive limit. A second route, the one we opted for, is to obtain the coupling constant from a fitting procedure of experimental data. A point of strength of the SLE-based approach is that it allows to exactly account for inhomogeneous line broadening. Sensitivity on such a feature of the cw-EPR spectrum is particularly important in the present study since, as found in a previous work over bis-labeled fullerene moieties, [79] the sign of the constant affects differently (and in a specular way, if sign is changed) the left and right parts of the spectrum, with respect to the central, electron Larmor, frequency. As will be discussed in the Results, the sensitivity is sufficiently high such that it is possible to distinguish the sign even for a small, with respect to the isotropic hyperfine coupling constant, value of the exchange integral. This makes the SLE-based approach in a preferential position in the computational approaches to determine the constant from experimental measurements. On the other hand, the integrated computational approach allows at present the calculation of most of the parameters entering the SLE at a sufficient quality level such that the difficulties of complex multidimensional fitting procedures are avoided. As will be shown below, a very limited set of three fitting parameters will be employed, namely a correction to the isotropic hyperfine interaction of the unpaired electron with the ^{14}N nucleus, the constant and a homogeneous broadening accounting for secondary effects on spectral lines coming from details neglected in the model. Table 2.1 reports the chemical formulas and acronyms for the systems

studies and the distance between the two nitroxide moieties.

	compound	acronym	n_{CB}	radical state
(i)	Z-$(Aib)_5$-TOAC-Aib-OMe [a]	$HEPTA_6$	-	*mono*
(ii)	Z-$(Aib)_6$-TOAC-Aib-OMe	$OCTA_7$	-	*mono*
(iii)	Fmoc-Aib-TOAC-$(Aib)_7$-OMe [b]	$NONA_2$	-	*mono*
(iv)	Fmoc-$(Aib)_2$-TOAC-$(Aib)_2$-TOAC-Aib-OMe	$HEPTA_{3,6}$	15	*bis*
(v)	Fmoc-TOAC-$(Aib)_3$-TOAC-Aib-OMe	$HEXA_{1,5}$	18	*bis*
(vi)	Fmoc-Aib-TOAC-$(Aib)_4$-TOAC-Aib-OMe	$OCTA_{2,7}$	21	*bis*
(vii)	Fmoc-Aib-TOAC-$(Aib)_5$-TOAC-Aib-OMe	$OCTA_{2,7}$	24	*bis*

[a] Z, benzyloxycarbonyl; OMe, methyloxy [b] Fmoc, fluorenyl-9-methyloxycarbonyl

Table 2.1: Chemical formulas and acronyms for the peptides investigated

2.2.2 Modeling

2.2.2.1 The Stocastic Liouville Equation

As it was shown in previous works on similar systems,[63, 64, 71] Aib-based short peptides can be treated as rigid bodies from the point of view of the cw-EPR spectroscopy in solution. The relevant (slow) coordinates of the molecules are simply the three Euler angles, $\mathbf{\Omega}$, that describe the overall orientation of a molecular-fixed reference frame (MF) with respect to a laboratory-fixed (LF) frame. The remaining degrees of freedom, *i.e.* peptide internal dynamics and solvent, are treated at the level of a thermal bath, providing only fluctuation-dissipation to the angular momentum of the molecule. Within this level of description, the time behavior of the coordinate $\mathbf{\Omega}$. To describe its time evolution, the quantity $\rho(\mathbf{\Omega}, t|\mathbf{\Omega}_0, 0)$ is introduced, *i.e.* the conditional probability density of finding the molecule with an orientation $\mathbf{\Omega}$ at a time t, if it was in $\mathbf{\Omega}_0$ at some reference time. In this case the high friction approximation regime is used, under which the angular momentum is thought to relax in a time scale much faster with respect to the Euler angles, thus it can be projected out. Under this assumption, the time evolution of $\rho(\mathbf{\Omega}, t) = \rho(\mathbf{\Omega}, t|\mathbf{\Omega}_0, 0)$ is

$$\frac{\partial}{\partial t} = \rho(\mathbf{\Omega}, t) = -\hat{\boldsymbol{J}}^{tr}(\mathbf{\Omega})\boldsymbol{D}\hat{\boldsymbol{J}}(\mathbf{\Omega})\rho(\mathbf{\Omega}, t) = -\hat{\Gamma}(\mathbf{\Omega})\rho(\mathbf{\Omega}, t) \tag{2.12}$$

which is valid in an isotropic medium. In Eq. 2.12, $\hat{\boldsymbol{J}}(\mathbf{\Omega})$ is the angular momentum operator, describing the infinitesimal rotation of the molecule, and $\boldsymbol{D}$ is the rotational diffusion tensor. It *i)* MF is chosen as the frame that diagonalizes $\boldsymbol{D}$ and *ii)* and assuming a nearly axially simmetric

rotational diffusion tensor, then the diffusive operator reads

$$\hat{\Gamma} \simeq D_{\perp}\hat{J}^2 - \left(D_{\parallel} - D_{\perp}\right)\hat{J}_Z^2 \tag{2.13}$$

with $D_{\parallel} = D_{ZZ}$ the principal value of the rotational diffusion tensor about the direction nearly parallel to the axis of the 3_{10}-helix and $D_{\perp} = (D_{XX} + D_{YY})/2$ the average of the other two principal values for the rotation about two perpendicular axes, both nearly perpendicular to the helix axis. $\hat{J}^2$ and $\hat{J}_Z^2$ are, respectively, the square of the total angular momentum and its projection over the Z-axis of MF. Since the relaxation time scales characteristic for $\mathbf{\Omega}$ is likely to be comparable with spin relaxation rates, the quantum mechanical evolution of spin pseudo variables σ, and the classical motion need to be treated in a coupled way. The Stochastic Liouville equation [70] provides the correct framework to describe in complete and exact way the full set of relaxations in the system.

$$\begin{aligned} \frac{d}{dt}\hat{\rho}(\sigma,\mathbf{\Omega},t) &= -i\left[\hat{H}(\mathbf{\Omega}),\hat{\rho}(\sigma,\mathbf{\Omega},t)\right] - \hat{\Gamma}(\mathbf{\Omega})\hat{\rho}(\sigma,\mathbf{\Omega},t) \\ &= -\left(i\hat{H}^{\times}(\mathbf{\Omega}) + \hat{\Gamma}(\mathbf{\Omega})\right)\hat{\rho}(\sigma,\mathbf{\Omega},t) \\ &= -\hat{L}\hat{\rho}(\sigma,\mathbf{\Omega},t) \end{aligned} \tag{2.14}$$

where now the probability density is an operator (density matrix), $\hat{H}$ is the spin Hamiltonian, $\hat{H}^{\times}$ a super-operator that returns the commutator of $\hat{H}$ and $\hat{\rho}(\sigma,\mathbf{\Omega},t)$, and $\hat{L}$ the stochastic Liouvillean. Since in this work, we deal with both *mono*- and *bis*-labelled peptides, each spin label bearing an unpaired electron coupled with one nitrogen nucleus, the general shape of the spin hamiltonian (in units of frequency) is

$$\hat{H} = \frac{\beta_e}{\hbar}\sum_{i=1}^{n_{probes}} \boldsymbol{B}_0 \cdot \boldsymbol{g}_i \cdot \hat{\boldsymbol{S}}_i + \sum_{i=1}^{n_{probes}} \hat{I}_i \cdot \boldsymbol{A}_i \cdot \hat{\boldsymbol{S}}_i - 2\gamma_e J \hat{S}_1 \cdot \hat{S}_2 + \hat{S}_1 \cdot \boldsymbol{T} \cdot \hat{S}_1 \tag{2.15}$$

where β_e is the Bohr magneton and $\hbar$ the Plank constant. The first term is the Zeeman interaction of each electron spin with the magnetic field $\boldsymbol{B}_0$, depending of the $\boldsymbol{g}_i$ tensor; the second term is the hyperfine interaction of each ^{14}N/unpaired electron, defined with respect to the hyperfine tensor $\boldsymbol{A}_i$; the third and fourth terms are the electron exchange and spin-spin dipolar interactions, respectively. J is the exchange constant, while the tensor $\boldsymbol{T}$ is modeled here in the point

approximation

$$\boldsymbol{T} = \frac{\mu_0}{4\pi}\frac{g_e^2\beta_e^2}{hr^3}\left[\mathbf{1}_3 + \frac{3}{r^2}\boldsymbol{r}\otimes\boldsymbol{r}\right] \tag{2.16}$$

where μ_0 is the vacuum magnetic permeability, $\boldsymbol{r}$ the distance vector between the position of the two electrons, r its modulus, and $\otimes$ stands for the dyadic product. While in principle to evaluate the tensor $\boldsymbol{T}$, the distributions of the upaired electrons in their orbitals should be taken into account, the two N-O moieties in the *bis*-labeled radicals of this study are sufficiently far away (> 7 Å) so that is possible to consider electrons as point charges. [63] In the calculations, the electrons are placed in the center of the N-O bond. In Eq. 2.15, tensors $\boldsymbol{g}_i$ and $\boldsymbol{A}_i$ are taken diagonal in their local frames μ_iF (μ = g, A) rigidly fixed on the i-th nitroxide, and the set $\boldsymbol{\Omega}_{\mu_i}$ is introduced, as the time- indipendent set of Euler angles that transforms from MF to μ_iF. Operators $\hat{\boldsymbol{S}}_i$ and $\hat{\boldsymbol{I}}_i$ are defined in LF. For monoradicals n_{probes} = 1 and the third and fourth term of the Hamiltonian are not present, while for biradicals n_{probes} = 2 and the full Eq. 2.15 must be considered. Finally, the dependence of the spin Hamiltonian on $\boldsymbol{\Omega}$ is implicit due to the fact that Zeeman, hyperfine and dipolar interactions are modulated by tensorial quantities, which are constant in MF, but change in LF, which is the reference where the spin operators are defined. [70, 80]

The EPR spectrum is obtained as Fourier-Laplace transform of the correlation function for the X-componente of the magnetization, defined as

$$|v\rangle = (2I+1)^{-n_{probes}/2}\sum_{j=1}^{n_{probes}}|\hat{S}_{X,j}\rangle$$

where I is the nuclear spin. following standard definitions,[70] the spectral lineshape is obtained as

$$G(\omega-\omega_0) = \frac{1}{\pi}\mathscr{R}\left\{\left\langle v\left|\left[i(\omega-\omega_0)+\left(i\hat{H}^{\times}+\hat{\Gamma}\right)\right]^{-1}\right|vP_{eq}\right\rangle\right\} \tag{2.17}$$

where $P_{eq} = 1/8\pi^2$ is the (isotropic) distribution in the $\boldsymbol{\Omega}$ space. Here, ω is the sweep frequency, $\omega_0 = g_0\beta_e B_0/h = \gamma_e B_0$ and g_0 is the trace of the $\boldsymbol{g}_i$ tensor divided by three. The starting vector $|v\rangle$ of Eq. 2.17 is related to the allowed EPR transitions and it is actually an operator acting on the spin degrees of freedom.[70]

To summarize, the peptide is described as a diffusive rotor, and the TOAC probes are rigidly fixed.

Parameters are the principal values of the diffusion tensor D_{XX}, D_{YY}, D_{ZZ}; the principal values of $\boldsymbol{g}$ and $\boldsymbol{A}$ tensors, and the Euler angles $\boldsymbol{\Omega}_\mu$ describing the orientation of the magnetic local tensors with respect to MF; in the case of biradicals, the exchange interaction J and the dipolar tensor $\boldsymbol{T}$ must be added to the set.

2.2.2.2 Structure and magnetic tensors

The geometrical optimization of all the peptides has been carried out using the Gaussian 03 software package [81] at DFT level of theory in acetonitrile solvent, which is modeled at the level of the polarizable continuum model, PCM. [82] The hybrid counterpart PBE0 of the conventional functional PBE with the standard 6-31G(d) basis set was employed. Based on previous studies on Aib-based, TOAC-labeled peptides have been prepared in the 3_{10}-helix secondary structure and assumed a twist geometry for the TOAC rings.

Hyperfine and Zeeman tensors have been computed by the same functional and using the N06 basis set.[83] No vibrational averaging correction has been applied to the isotropic hyperfine term, $A_{iso} = tr\{\boldsymbol{A}\}/3$. Rather, the A_{iso} term has been kept as an adjustable parameter comparing the calculated spectra with the experimental ones. In biradicals, as described in Subsection 2.2.2.1 the spin-spin dipolar interaction tensor has been calculated within the point approximation in Eq. 2.16, taking the vector connecting the centers of the two N-O bonds as a measure of the distance between the electrons. Concerning the exchange interactions, calculation of J is still a difficult task, especially in cases when its absolute value is of few Gauss, or smaller. On the other hand, due to the inhomogeneous broadening of spectral lines, the spectrum is very sensitive not only to the absolute value of the electrons exchange constant, but also on its sign. [79]. Thus, J has been kept as a free parameter of the calculation, to be fitted over the experimental data.

2.2.2.3 Dissipative properties

The evaluation of the diffusion properties of the peptides was based on a hydrodynamic approach.[84] The melecule is described as a set of N rigid fragments made of atoms or groups of atoms immersed in a homogeneous isotropic fluid of known viscosity. the tensor $\boldsymbol{D}$ can be conveniently partitioned into translational, rotational, internal and mixed blocks. It is thus obtained as the

inverse of the friction tensor Ξ using Einstein's relation[85, 86]

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{TT} & \boldsymbol{D}_{TR} & \boldsymbol{D}_{TI} \\ \boldsymbol{D}_{TR}^{tr} & \boldsymbol{D}_{RR} & \boldsymbol{D}_{RI} \\ \boldsymbol{D}_{TI}^{tr} & \boldsymbol{D}_{RI}^{tr} & \boldsymbol{D}_{II} \end{bmatrix} = k_B T \Xi^{-1} \tag{2.18}$$

where k_B is the Boltzmann constant and T the absolute temperature. The friction tensor for the constrained system of spheres (the real molecule), Ξ, is calculated from the friction tensor of non-constrained extended atoms, as described in Ref. [84].

The complete diffusion tensor is represented by a 6×6 matrix. Due to the translational invariance of the magnetic tensors, one may project out the translational part and reduce the diffusion tensor to a 3×3 matrix made up only of the rotational tensor, $\boldsymbol{D} = \boldsymbol{D}_{RR}$. For all the peptides, the diffusion tensors have been calculated with this set of parameters: viscosity 0.343 cP,[87] temperature 293 K, an effective radius of 2 Å on all the non-hydrogen atoms and stick boundary conditions.

2.2.3 Results

The calculation of the cw-EPR spectra has been carried out with the E-SpiRes software package. [69] Relevant parameters are reported in Tables 2.3 and 2.4, respectively for the three *mono*-labeled and for the four *bis*-labeled peptides. since there was no g-calibration in the experimental spectra, a fixed correction g_{corr}, was applied in order to center the theoretical spectra with the experimental ones. A limited set of parameters have been adjusted via a non-linear least squares procedure, that is: the isotropic part of the hyperfine tensors (needed because librational effects were not accounted for in QM calculations), the value of the exchange integral in biradicals, and an intrinsic linewidth which provides an homogeneous broadening to the spectral lines. The latter parameter is added in order to simply take into account secondary effects of structure/dynamics on the spectrum neglected by the stochastic model. The values obtained for J are reported in Table 2.2, together with the geometric distance between the two nitroxide moieties, and the distance expressed in terms of the number of covalent bonds (n_{CB}).

Peptide	r / Å	n_{CB}	J / Gauss
$HEXA_{1,5}$	11.9	15	250
$HEPTA_{3,6}$	7.0	18	>300
$OCTA_{2,7}$	15.0	21	-0.38
$NONA_{2,8}$	12.9	24	0.31

Table 2.2: Summary of J values obtained by fitting of experimental spectra. Geometric distance (r) and distance in term of number of covalent bonds (n_{CB}) between the two TOAC labels are also reported.

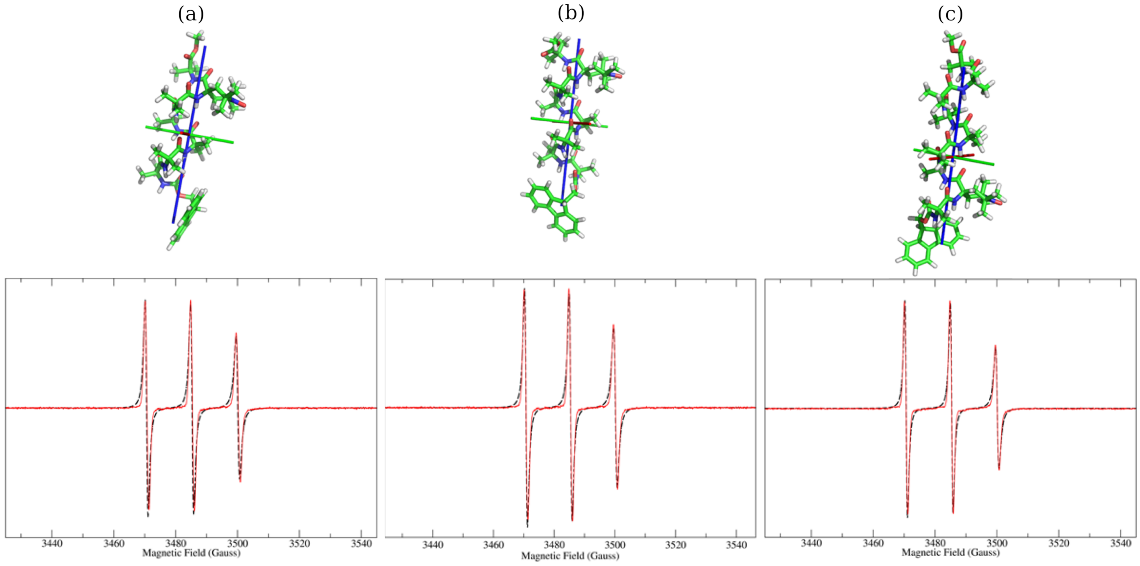

Figure 2.4: Experimental (red, solid line) and calculated (black, dashed line) cw-EPR spectra of the three monoradicals, and their QM-minimized structures. a) $HEPTA_6$, b) $OCTA_7$, c) $NONA_2$. The principal axes of rotational diffusion are also shown (coloring scheme: X red, Y green, Z blue).

Figure 2.4 reports the comparison between experimental and calculated spectra for the mono-labeled peptides, while Figures 2.5 and 2.6 show the comparison for the four biradicals. The good agreement of theoretical lineshapes with the experimental data obtained using a very limited set of parameters underlines the goodness of the stochastic model employed, *i.e.* despite its simplicity it is able to catch the relevant dynamics, with respect to EPR, of the molecules. While for the mono-radicals, all the the parameters, but the isotropic part of the hyperfine coupling, are all computed ab initio, for biradicals the most important physical parameter to discuss is the exchange integral.

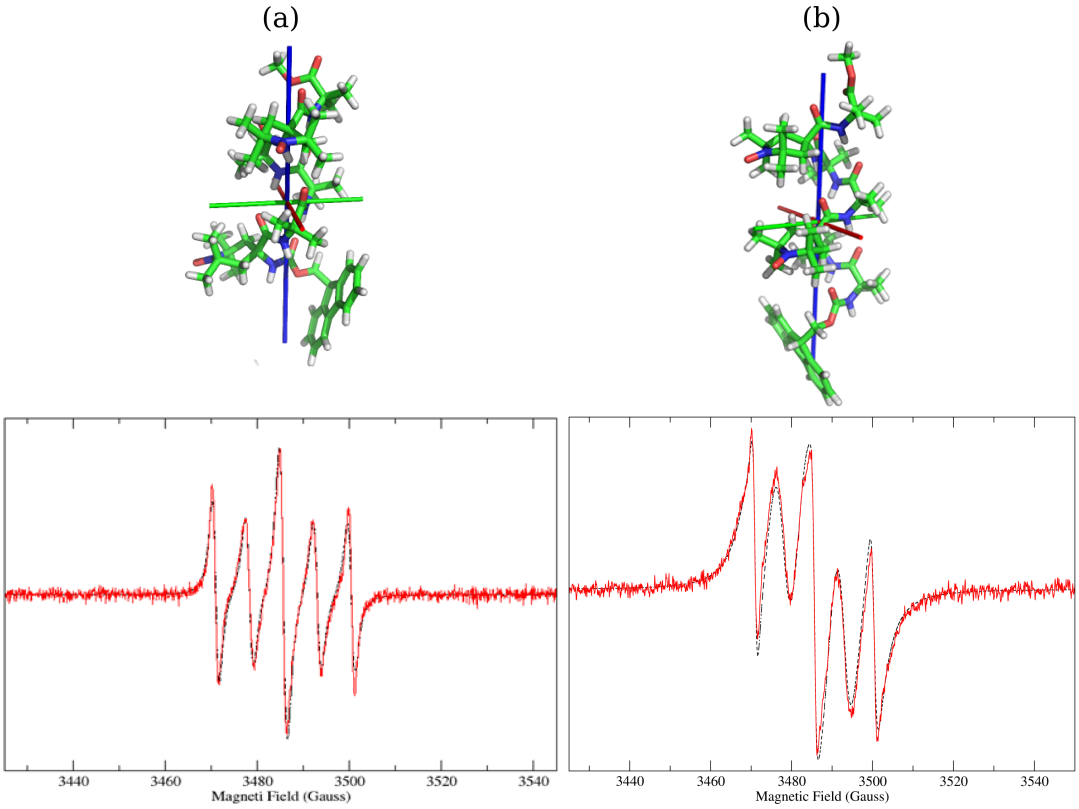

Figure 2.5: Experimental (red, solid line) and calculated (black, dashed line) cw- EPR spectra of the biradicals a) $HEXA_{1,5}$ and $HEPTA_{3,6}$, and their QM-minimized structures. The principal axes of rotational diffusion are also shown (coloring scheme: X red, Y green, Z blue).

2.2.3.1 The effect of the value of J on the spectrum

As underlined in the Introduction, in this study, the molecules showed to lay in two limiting cases. On one hand, spectra of $HEXA_{1,5}$ and $HEPTA_{3,6}$ are constituted of five lines, with the two extra lines with respect to the normal monoradical pattern placed exactly at $\pm A_{iso}/2$ and with high intensity (see Figure 2.5. Following Luckurst [88] this indicates that $J/A_{iso} >> 1$. In fact, for $HEXA_{1,5}$ the fit returned 250 Gauss, while for the $HEPTA_{3,6}$ peptide, the only possible estimation was that $J \geq 300$ Gauss, since beyond this value, the calculated spectrum started to become insensitive on the value of the exchange integral. To obtain a good agreement with the experimental spectra the assumption that a certain percentage of mono-radical peptide was present in the sample has been made (*e.g.* due to partial degradation of the sample [79]. In particular, a component 20% and 4% of monoradical for the $HEXA_{1,5}$ and $HEPTA_{3,6}$ has been

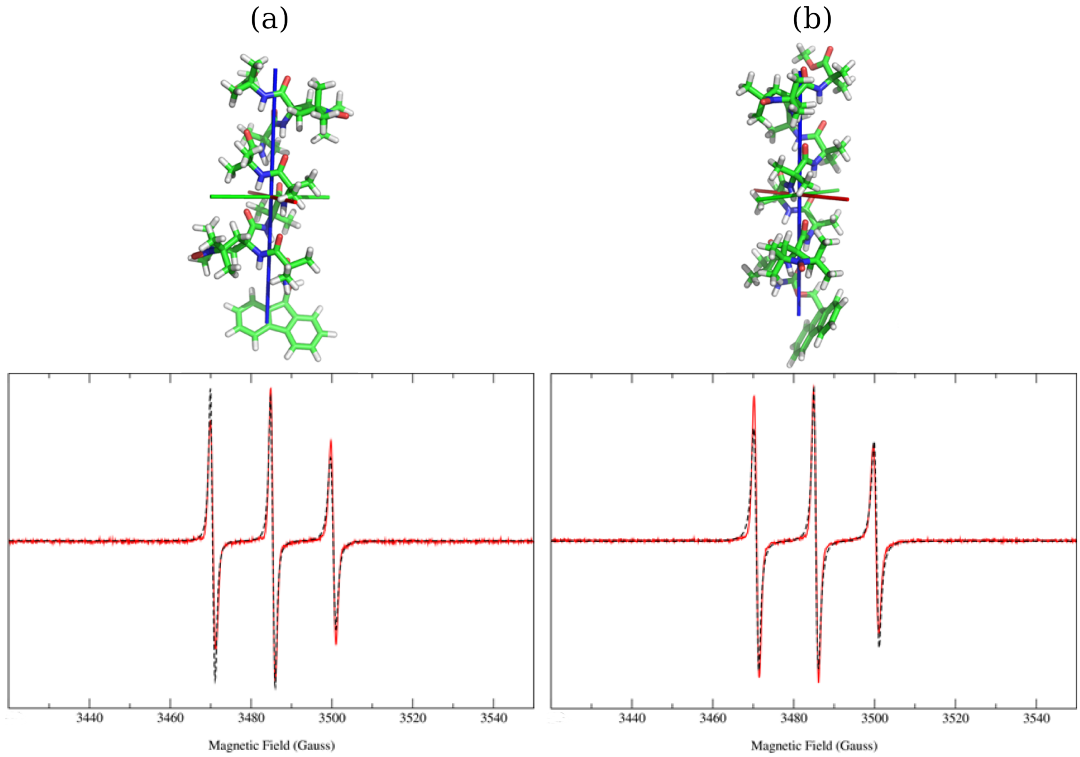

Figure 2.6: Experimental (red, solid line) and calculated (black, dashed line) cw- EPR spectra of the biradicals a) $OCTA_{2,7}$ and b) $NONA_{2,8}$, and their QM-minimized structures. The principal axes of rotational diffusion are also shown (coloring scheme: X red, Y green, Z blue).

used.

On the other hand, experimental spectra of $OCTA_{2,7}$ and $NONA_{2,8}$ show only three peaks, suggesting a very weak exchange interaction among the two unpaired electrons. also, due to the large distance of the electrons (15 and 13 Å, respectively), the dipolar interaction is not able to contribute to the inhomogeneous broadening of the peaks in a sensible way. Thus, it was not possible to estimate, if present, the quantity of *mono*-raical with accuracy, for which, the calculations were run with the *bis*-radical contribution only, neglecting any possible contamination by the *mono*-radical.

To find the value of J, two fits were run starting from either a positive or a negative value of the coupling constant. Values reported in Table 2.2 are those that gave the best chi-squared value. For the sake of completeness, the spectra calculated with both positive and negative values of J (together with intrinsic linewidth, γ, and the chi-squared, χ^2) are reported in Figure 2.7, for the two peptides. the spectra show that a small but decisive difference is noticeable between the two

calculations with the inverse sign of the exchange integral. Thus, not only the SLE approach is sensitive to very small, in absolute value J, but it is able to catch also the sign of the exchange integral. While the first information is in some way "hidden" in the spectral pattern, the sign is intrinsically related to inhomogeneous broadening, which is exactly taken into account in our approach, within the limits of the precision of the chosen model for the dynamics.

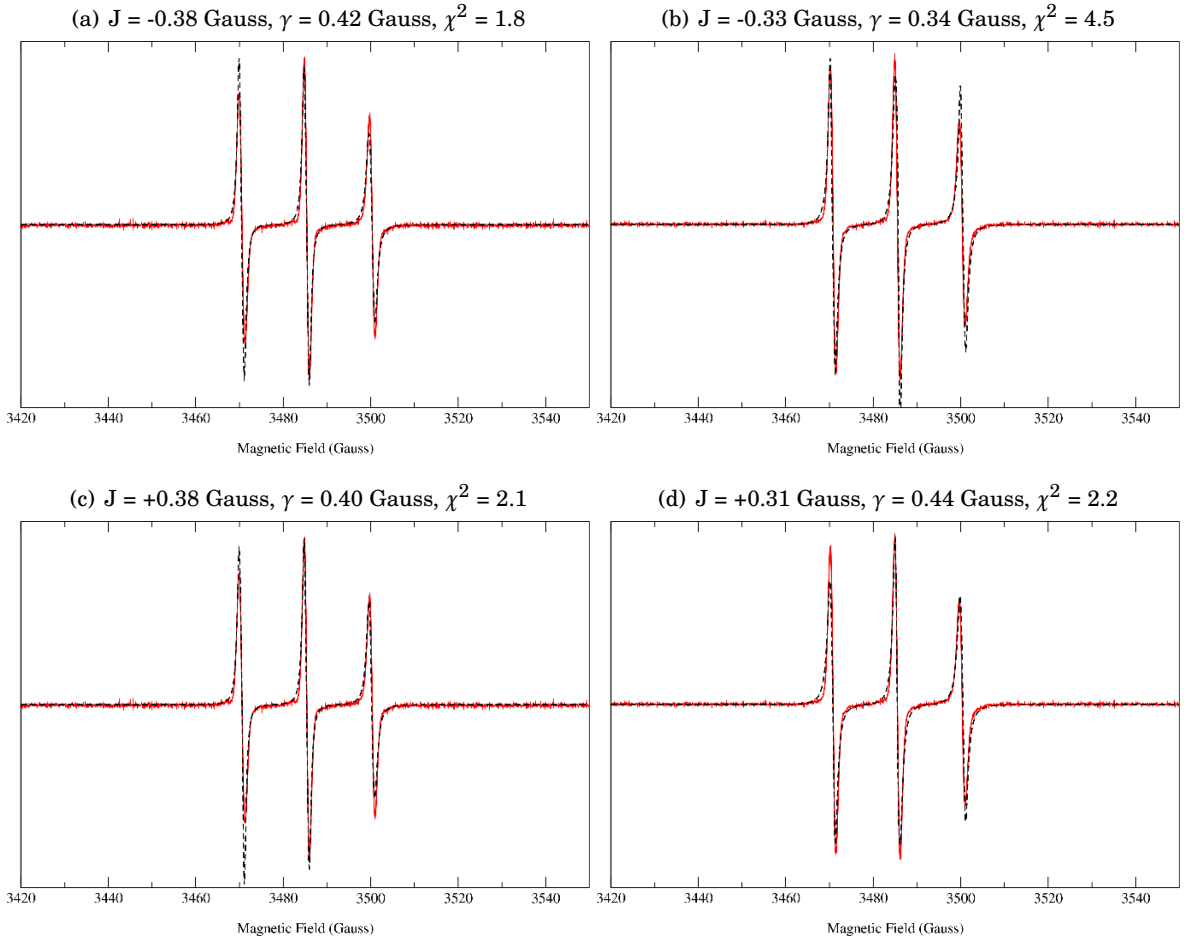

Figure 2.7: Comparison among experimental (red, solid line) and theoretical (black, dashed line) cw- EPR spectra of (a, c) $OCTA_{2,7}$ and (b, d) $NONA_{2,8}$ bis-radicals fitted using a negative or a positive initial guess for J (intrinsic linewidth was also fitted). χ^2 for the fittings are reported.

2.2.4 Conclusions

The combination of different computational methods, from MD to quantum mechanical calculations and stochastic modeling, has been proved to be a winning strategy to interpret cw-EPR

	$HEPTA_6$	$OCTA_7$	$NONA_2$
$\boldsymbol{D}$ / 10^9 Hz	1.03, 1.08, 2.75	1.36, 1.38, 3.78	1.08, 1.11, 3.35
$\boldsymbol{g} - g_e$ / 10^{-3}	6.41, 3.66, -0.29	6.48, 3.71, -0.22	6.97, 4.14, 0.16
$\boldsymbol{\Omega}_g$ / deg	-2.0, 89.9, -1.6	-107.5, 12.5, 170.2	153.0, 182.1, 249.4
$\boldsymbol{A}$ / Gauss	5.50, 5.64, 33.08	5.52, 5.66, 33.07	5.64, 5.80, 32.75
$\boldsymbol{\Omega}_A$ / deg	77.7, 172.1, -70.6	151.5, 10.4, 162.7	146.0, 174.5, -102.7
ω / 10^9 Hz	9.784351	9.784351	9.786595
g_{corr} / Gauss	+11.5	+11.5	+11.5
γ / Gauss	0.73	0.70	0.58

Table 2.3: Dissipative, geometric and magnetic parameters employed in the calculation of cw-EPR spectra of mono-labeled peptides. Principal values of the tensors are given, together with their transformation angles with respect to MF. Reported are also the spectrometer frequency (ω), the g shift correction (g_{corr}) and the intrinsic linewidth (γ).

	$HEXA_{1,5}$	$HEPTA_{3,6}$	$OCTA_{2,7}$	$NONA_{2,8}$
$\boldsymbol{D}$ / 10^9 Hz	0.89, 0.92, 1.69	0.83, 0.87, 2.04	1.01, 1.03, 1.93	0.51, 0.53, 1.48
$\boldsymbol{g}_1 - g_e$ / 10^{-3}	6.90, 4.20, 0.02	6.83, 4.08, 0.13	6.97, 4.14, 0.16	6.70, 3.90, -0.10
$\boldsymbol{\Omega}_{g_1}$/ deg	-87.4, 116.4, -147.9	171.5, 7.6, -28.0	154.0, 182.1, 249.4	153.0, 182.1, 249.4
$\boldsymbol{A}_1$ / Gauss	5.64, 5.80, 32.75	5.50, 5.64, 33.08	5.64, 5.80, 32.75	5.64, 5.80, 32.75
$\boldsymbol{\Omega}_{A_1}$ / deg	-157.2, 45.6, -139.3	171.5, 24.5, -98.3	146.0, 174.5, -102.7	146.0, 174.5, -102.7
$\boldsymbol{g}_2 - g_e$ / 10^{-3}	6.41, 3.66, -0.29	6.97, 4.14, 0.16	6.88, 4.10, 0.18	6.48, 3.71, -0.22
$\boldsymbol{\Omega}_{g_2}$ / deg	37.4, 87.2, 133.3	-40.3, 137.6, 169.3	-107.5, 12.5, 170.2	-107.5, 12.5, 170.2
$\boldsymbol{A}_2$ / Gauss	5.50, 5.64, 33.08	5.64, 5.80, 32.75	5.52, 5.66, 33.07	5.52, 5.66, 33.07
$\boldsymbol{\Omega}_{A_2}$ / deg	37.4, 87.2, 133.3	-40.3, 137.6, 169.3	151.5, 10.4, 162.7	6.3, 90.0, -180.0
r / Å	11.9	7.0	15.0	12.9
ω / 10^9 Hz	9.787400	9.787091	9.786611	9.785979
g_{corr} / Gauss	+11.5	+11.5	+11.5	+11.5
γ / Gauss	1.16	0.92	0.42	0.44

Table 2.4: Dissipative, geometric and magnetic parameters employed in the calculation of cw-EPR spectra of bis-labeled peptides. Principal values of the tensors are given, together with their transformation angles with respect to MF. Reported are also the spectrometer frequency (ω), the g shift correction (g_{corr}) and the intrinsic linewidth (γ).

spectra in bis-labelled polypeptides. [63, 64] It is thus possible to combine covergent and complementary computational techniques to obtain geometrical and dynamical information. The DFT geometry optimization procedure led to a 3_{10} helical structure for all the polypeptides studied in this work, this is consistent with previous theoretical and experimental studies.[63, 64]. For the systems $OCTA_{2,7}$ and $OCTA_{2,8}$ we found that the principal values of the hyperfine tensors ($\boldsymbol{A}_1$ and $\boldsymbol{A}_2$) are equal. In general, we found Aib/TOAC peptides having a stronger coupling

than that in the corresponding Ala/TOAC peptides investigated some time ago. [89] This result is not surprising in view of the known, much less strong helix-supporting properties of Ala versus Aib.[67] In the studies reported in ref. [89], the authors also determined the relaxation parameters by power-saturation experiments. They demonstrated that the relaxation parameters for all four bis-radical peptides differ significantly from those for the mono-radicals, showing that a different relaxation mechanism is operative in the bis-radicals. The different relaxation is attributed to the spin-spin interaction occurring in the latter compounds and proposed that it can be used as a tool for distance determination. In this work, we showed that for specific nitroxide spin labels in the distance regime of 0.8 to 1.5 nm, electron spin-spin relaxation could be used as an indicator for distances , thus expanding the tools available so far to EPR spectroscopists even further towards biologically significant conditions.

CHAPTER 3

BAYESIAN ANALYSIS OF DISPERSION RELAXATION EXPERIMENTS

3.1 Introduction

Nuclear magnetic relaxation has long been known for its unique capability to investigate molecular motions, and it has been used extensively to probe internal dynamics in nonrigid macromolecules. Moreover, given the development of two- and higher dimensional NMR experiments, NMR has become a unique tool to extract dynamical information with atomic resolution.

One of the characteristic features of the technique is the indirect connection between spin relaxation and molecular motions. Indeed, these latter appear in spin relaxation rates through correlation functions of the spatial part of fluctuating hamiltonians.[41] Apart from experimental issues, there are several fundamental problems when relating spin relaxation measurements to molecular dynamics. The connection between relaxation rates and the spectral density function $J(\omega)$, (i.e. the Fourier transform of the correlation function) is lacunary, due to the restricted number of frequencies entering the expression of the relaxation rates. Therefore the analysis of relaxation data requires models of spectral density functions. These can be built on various grounds, including physical models of dynamics, or empirical formulations of the correlation functions. In any case, the practical problem at hand is that of fitting experimental data to a model. Although the choice of a given model is of fundamental importance, by providing a physical

interpretation of the data, this question is not the subject of this work. Rather, the point is to address the question of the capacity, given a model for the dynamics, or the correlation functions, of NMR relaxation experiments to provide reliable estimates of model parameters. Difficulties arise from different factors. One is the relatively small number of independent relaxation rates that can be measured, which limits the number of adjustable parameters and therefore the complexity of the models to be fitted. Moreover, in addition to what was stated above, most of the values entering the expressions of the relaxation rates sample the higher frequency region of the spectrum, leaving largely undeterminate the sampling of the spectral density profile. Therefore, some of the time constants, such as fast internal time scales, may not be correctly probed by relaxation experiments. Another issue often overlooked is the correlation between model parameters during fitting. This is particularly true for overall correlation time(s) and internal dynamical parameters. This can be ascribed to the fact that contributions from the part of the correlation function that account for the overall diffusion of the molecule, and for the usually faster internal dynamics, may not be delineated from NMR measurements. This typically gives rise to multiple minima of the target function χ^2. When an analytical expression of $J(\omega)$ is available, this can be related to the existence of asymptotic expressions of the spectral density function that are independent of some of the model parameters, therefore leading to some undeterminacy.

Some of these aspects have been investigated in a recent work,[90] where the authors proposed a Markov Chain-Monte Carlo (MCMC) approach to analyze NMR relaxation data.[91] There, it was shown that such a strategy can be used for the determination of all, or a subset only of parameters of a dynamical model, thereby avoiding the model selection process. Moreover, the statistical correlation between internal and overall dynamic parameters was investigated through MCMC approach, which pinpointed the limitations of the use of only high-field relaxation studies. Indeed, simultaneous fitting of overall diffusion parameters, that contribute most to the spectral density function, and parameters pertaining to internal motions, contributing a small fraction of $J(\omega)$, leads to the similar kind of under determined minimization problem.

Therefore, the authors investigated the possible advantages of using relaxation measurements performed at different magnetic fields to extract these parameters simultaneously. Relaxometry,

i.e. the measurement of relaxation rates over a wide range of magnetic fields, dates back to the early days of NMR, and has been recently developed by several authors to match the standards of high-resolution NMR.[92–94] It is thus possible to perform relaxation experiments at magnetic fields as low as 0.5 T on a high-resolution spectrometer. The use of such a broad spectral measurement is likely to provide a better sampling of the spectral density functions.

The main purpose of this chapter is to determine the conditions under which both overall and internal dynamics can be reliably obtained from relaxation data, using a MCMC strategy. This work has been carried out during a prolonged stay at the École Normale Supérieure in Paris under the guidance of Dr Daniel Abergel.

3.2 Theory

3.2.1 Spin relaxation and dynamics

According to the Abragam-Redfield theory, [41, 95, 96] spin relaxation is determined by the nature of the fluctuating spin interactions and the correlation function of the spatial part of the hamiltonian, which contains information about the "lattice" fluctuations, i.e., the molecular motions. In the case of relaxation of a spin 1/2 X (X=^{15}N,^{13}C) bonded to a ^{1}H the stochastic motions of the internucleus X -^{1}H vector induce fluctuations of the dipole-dipole interaction with the directly attached proton and of the X chemical shift anisotropy (CSA) tensor $\boldsymbol{\sigma}$. For a molecule in solution, NMR relaxation rates are determined by time correlation functions $C(t) = \langle P_2(\boldsymbol{\mu}(t) \cdot \boldsymbol{\mu}(0)) \rangle$, where $\boldsymbol{\mu}(t)$ is a unit vector pointing along the X ^{1}H bond and $P_2(.)$ is the second order Legendre polynomial $P_2(\theta) = 1/2(\cos(\theta)^2 - 1)$. The spectral density function $J(\omega)$ is defined as the Fourier transform of the correlation function $C(t)$ as:

$$J(\omega) = \int_0^\infty C(t) \cos \omega t \, dt \tag{3.1}$$

Longitudinal and transverse ^{15}N relaxation rates (R_1, R_2), and ^{15}N{^{1}H} heteronuclear Overhauser enhancement (η_{NH}), as reported in Section 2.1.1 are expressed in terms of the spectral density function $J(\omega)$ evaluated at particular values of the Larmor frequencies. For convenience,

we report here the expressions of Equation 2.1:

$$
\begin{aligned}
\eta_{\mathrm{XH}} &= 1+\frac{\gamma_H}{\gamma_X}\frac{d^2}{R_1}(6J(\omega_{\mathrm{H+X}})-J(\omega_{\mathrm{H\text{-}X}})) \\
R_1 &= d^2(3J(\omega_{\mathrm{X}})+J(\omega_{\mathrm{H\text{-}X}})+6J(\omega_{\mathrm{H+X}}))+2c^2J(\omega_{\mathrm{X}}) \\
R_2 &= d^2\Big[2J(0)+\frac{3}{2}J(\omega_{\mathrm{X}})+\frac{1}{2}J(\omega_{\mathrm{H\text{-}X}}) \\
&+ 3J(\omega_{\mathrm{H}})+3J(\omega_{\mathrm{H+X}})\Big]+c^2\Big(\frac{4}{3}J(0)+J(\omega_{\mathrm{X}})\Big)
\end{aligned}
\tag{3.2}
$$

where $d=\mu_0\hbar\gamma_{\mathrm{H}}\gamma_{\mathrm{X}}/4\sqrt{10}\pi\langle r^3_{\mathrm{XH}}\rangle$, $c=\gamma_{\mathrm{X}}B_0\Delta\sigma_{\mathrm{X}}/\sqrt{15}$, and r_{XH} is the XH distance. The parameters γ_H and γ_X are the gyromagnetic ratios of X and ^{1}H atoms, respectively, μ_0 is the vacuum magnetic susceptibility, $\hbar$ is the reduced Planck constant, and $\Delta\sigma_{\mathrm{X}}$ is the ^{15}N chemical shift anisotropy. The presence of additional mobility on the $\mu s-ms$ time scale appears as a contribution R_{ex} to the observed transverse relaxation rate: $R_2^{app}=R_2+R_{ex}$.[97–99] In order to further the analysis of relaxation measurements, one therefore must assume a model of the correlation functions, based on physical considerations that depend on the kind of motions that are expected to take place. In this work, the fractional brownian dynamics (FBD) model has been used. The practical goal is then to extract the relevant model parameters from the experiments.

3.2.2 Fractional brownian dynamics

A model of fractional diffusion was recently introduced to account for the presence of long time tail decays of internal correlation functions appearing in NMR spectroscopy. [100–102] In the proposed model, overall tumbling and internal motions are statistically decorrelated, which enables the factorization of the correlation function as:

$$
C(t)=C_{\mathrm{I}}(t)\times C_{\mathrm{O}}(t)=e^{-t/\tau_o}\times C_{\mathrm{I}}(t), \tag{3.3}
$$

where overall diffusion is isotropic with correlation time τ_o. Internal rotational correlation functions $C_{\mathrm{I}}(t)$ were thus modeled as: [100, 103]

$$
C_{\mathrm{I}}(t)=S^2+(c_{el}-S^2)E_\alpha(-[t/\tau]^\alpha), \tag{3.4}
$$

$E_\alpha(z)$ is the Mittag-Leffler (ML) function, an entire function in the domain of complex numbers [104], and defined as:

$$
E_\alpha(z)=\sum_{k=0}^{\infty}\frac{z^k}{\Gamma(1+\alpha k)},\ \alpha\in\mathbb{C} \tag{3.5}
$$

The function $E_\alpha(-[t/\tau]^\alpha)$ in Equation (3.4) can be viewed as a generalization of a stretched exponential function with α and τ being the shape and scale parameters, respectively. For $\alpha = 1$ the stretched ML function reduces to the exponential, whereas for $0 < \alpha < 1$ it exhibits a power law decay at large times, $E_\alpha(-[t/\tau]^\alpha) \propto (t/\tau)^{-\alpha}$, and an infinitely steep decay at $t = 0$. For $0 < \alpha \leq 1$, the Mittag-Leffler function can be expressed as the continuous superposition of exponential relaxation functions $\exp(-\lambda t)$, with the relaxation rate distribution function $p_{\alpha,\tau}(\lambda)$:

$$E_\alpha(-[t/\tau]^\alpha) = \int_0^\infty d\lambda\, p_{\alpha,\tau}(\lambda)\exp(-\lambda t). \tag{3.6}$$

The spectrum of relaxation rates is positive and has the form: [105, 106]

$$p_{\alpha,\tau}(\lambda) = \frac{\tau}{\pi}\frac{(\tau\lambda)^{\alpha-1}\sin(\pi\alpha)}{(\tau\lambda)^{2\alpha} + 2(\tau\lambda)^\alpha\cos(\pi\alpha) + 1} \tag{3.7}$$

In Eq.3.7, $p_{\alpha,\tau}(\lambda)$ satisfies the normalization condition $\int_0^\infty p_{\alpha,\tau}(\lambda)d\lambda = 1$ and reduces to a Dirac distribution centered at the value τ^{-1} for $\alpha = 1$. Moreover, the inverse of the scaling parameter τ gives the median of the distribution $p_{\alpha,\tau}(\lambda)$, so that $\lambda_{1/2} = \tau^{-1}$ [103]. The stretched Mittag-Leffler function is the solution of a fractional differential equation [105, 107]. The spectral density function associated with Eqs (3.3-3.4) and (3.6) is given by:[100]

$$J(\omega) = \frac{S^2\tau_0}{1+(\omega\tau_0)^2} + (c_{el} - S^2)\frac{1}{\gamma}\frac{(\gamma\tau)^\alpha\cos\beta + \cos[\beta(1-\alpha)]}{(\gamma\tau)^\alpha + (\gamma\tau)^{-\alpha} + 2\cos\beta\alpha}, \tag{3.8}$$

where $\cos\beta = (\tau_0\gamma)^{-1}$, $\sin\beta = \omega/\gamma$, $\gamma = (\tau_0^{-2} + \omega^2)^{1/2}$.

As previously noted in Ref. [100], the initial decay of the correlation function, occurring at time lags typically shorter than $\approx 1\,ps$, is due to the presence of very fast processes that give rise to rapidly damped oscillations of the correlation function. These phenomena are not described by this diffusion process, and are empirically taken into account by introducing the parameter $c_{el} < 1$ in Eq. 3.4.[100] This parameter is just the value of the correlation function at the minimum time lag where the theory is assumed valid, what happens at shorter times remaining beyond the scope of the model. Thus, the use of Mittag-Leffler functions represents a way to account for the presence of multiple time scale internal dynamical processes, whilst keeping at the same time the number of model parameters as small as possible.[100]

3.2.3 Bayesian analysis of relaxation data using MCMC analysis

The usual parameter estimation strategy of spin relaxation measurements involve fitting to nested models of increasing complexity that are selected based on statistical F-tests.[108, 109] However, it has been shown that a Bayesian approach may have several advantages in this context.[90, 91]

Indeed, the latter provides a time effective parameter estimation strategy, because efficient algorithms that do not use time consuming minimizations allow to perform a search in the parameter space, rather than in the data space. In addition, a Bayesian approach is probabilistic, so that the whole statistical information content is retained, which therefore provides marginal probability distributions of the different model parameters. Moreover, any additional insight or knowledge of the system at hand can be used as input through the prior probability, and finally, it avoids the need for model selection strategies, and a may provide estimates of a subset of the model parameters. This may reveal of particular interest when measurements are insensitive to certain motional time scales, therefore, model parameters.[90]

In contrast to conventional treatments of experimental data, a Bayesian approach can accomodate asymmetrical or even multimodal probability distribution functions (PDFs) of model parameters that cannot be satisfactorily characterized by a few numbers representing their averages and spreads. Thus, MCMC provides a way to access the posterior probability $P(\boldsymbol{\theta}|\boldsymbol{R},I)$, of the parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{R}$. In this expression, I denotes variables accounting for any additional information about the system. For our purpose, $\boldsymbol{R} = \{R_1, R_2, NOE, \ldots\}$, the set of measured relaxation rates, whereas the model parameters $\boldsymbol{\theta}$ depends on the particular dynamical model used for the analysis. The posterior probability $P(\boldsymbol{\theta}|\boldsymbol{R},I)$ is related to the likelihood $P(\boldsymbol{R}|\boldsymbol{\theta},I)$ through the Bayes theorem of conditional probabilities:

$$P(\boldsymbol{\theta}|\boldsymbol{R},I) \propto P(\boldsymbol{R}|\boldsymbol{\theta},I) \times P(\boldsymbol{\theta},I) \tag{3.9}$$

Markov chain Monte Carlo (MCMC) methods represent very effective strategies for the simulation of samples of known probability distributions. A detailed account is beyond the scope of this work and the interested reader is referred to general references for detailed descriptions (see for instance Ref. [110]). In brief, the goal is to generate a sequence of random numbers that has

the posterior probability $P(\boldsymbol{\theta}|\boldsymbol{R},I)$ as its stationary distribution. However, one actually wants to sample from the *unknown* distribution $P(\boldsymbol{\theta}|\boldsymbol{R},I)$. But the problem is circumvented by the use of Eq.3.9 which allows to replace the posterior probability $P(\boldsymbol{\theta}|\boldsymbol{R},I)$ with the likelihood $P(\boldsymbol{R}|\boldsymbol{\theta},I)$. Indeed, the latter can be calculated from the model by assuming that the measurement error obeys a Gaussian white noise distribution centered on the experimental values $\boldsymbol{R}$ of the relaxation rates, with the standard deviation σ:

$$P(\boldsymbol{R}|\boldsymbol{\theta},I) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{[f(\boldsymbol{\theta}) - R]^2}{2\sigma^2} \tag{3.10}$$

where $f(\boldsymbol{\theta})$ is the theoretical relaxation rate. In the calculations, the value of σ was chosen so as to match a typical 3% relative error made on the relaxation rates in NMR experiments.

3.2.4 Implementation

In this work, we used a standard Metropolis-Hastings algorithm [111, 112] using a random-walk proposition law for the MCMC simulations, according to the following conventional algorithm ($\mathscr{U}(0,1)$ is the uniform distribution):

- for $i = 1,\ldots,N$ generate u from $\mathscr{U}(0,1)$ and θ^* from $\mathscr{N}(\boldsymbol{\theta}^{i-1},\sigma)$
  if $u \leq \dfrac{P(\boldsymbol{R}|\boldsymbol{\theta}^*,I)}{P(\boldsymbol{R}|\boldsymbol{\theta}^{i-1},I)}$ then
  $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$
  else
  $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$

Here, the components θ_η^{i-1} of $\boldsymbol{\theta}^{i-1}$ are sampled from independent normal laws $\mathscr{N}(\theta_\eta^{i-1},\sigma_\eta)$. MCMC simulations were used to provide estimates of the marginal probabilities of the various components of the parameter vector $\boldsymbol{\theta}$:

$$P(\theta_\eta|\boldsymbol{R},I) = \int d\theta_{\zeta\neq\eta} P(\boldsymbol{\theta}|\boldsymbol{R},I) \tag{3.11}$$

The estimated marginal probability distributions of the model parameters θ_η were computed from the histograms of values generated by the Markov chains. In order to do so, independent and

identically distributed (iid) samples must be selected, which is typically achieved by retaining one every N points in the chain. The value of N was determined based on the correlation function. Thus, points were considered statistically independent for a decay to less than 10% of the correlation function. A value of N was determined independently for each of the parameters, and the retained data points were then used to compute histograms of the marginal probability distributions of the model parameters.

3.2.5 The prior $P(\theta, I)$

Bayesian statistics provides a natural way to include a priori knowledge about the problem at hand through the prior probability $P(\boldsymbol{\theta}, I)$, which orients the search of the parameter space towards regions of greater probability. In the framework of the FBD , the prior, $P(\boldsymbol{\theta}, I) = P(S^2, c_{el}, \alpha, \tau, \tau_0, I)$ included constraints to ensure that only physically acceptable parameter values were retained. Thus, the following constraints $0 \leq S^2 \leq 1$, $S^2 < c_{el}$, and $0 \leq \alpha \leq 1$ were imposed, so that $P(S^2, I) = 0$ for $S^2 \notin [0,1]$, $P(\alpha, I) = 0$ for $\alpha \notin [0,1]$ and $P(S^2 \geq c_{el}, I) = 0$.

3.3 MCMC analysis of NMR relaxation rates

The behaviour of the MCMC method was tested on synthetic relaxation rates generated from the models presented in the previous section. R_1, R_2 and NOE relaxation rates were calculated for various selections of magnetic B_0 fields in the set given in Table 3.1.

B_0(T)	0.329	0.49	0.72	1	1.4	1.99	3	5	7	14.09	18.79	21.14	23.5
ν_0 (MHz)	14	21	31	43	60	85	128	213	298	600	800	900	1000

Table 3.1: Set of chosen values of B_0 fields

For the FBD model, two cases with different overall correlation time have been studied respectively with $\tau_0 = 15.1$ ns and $\tau_0 = 4.0$ ns. Each system is composed by six residues, with $S^2 = 0.8$, $\alpha = 0.7$, $c_{el} = 0.97$ and characterized by a different internal correlation time: 10, 50, 100, 200, 500 and 1000 ps. MCMC simulations were typically performed on samples of sizes on the order of 1.50^6 "time" points. The initial burn-in period, during which the points are not distributed from the

posterior distribution, was determined by inspection and discarded from the analysis. Following standard procedure, independent and identically distributed (iid) samples were extracted from the Markov chains, based on the correlation function of the sequence.[110] Thus, only one every N points were selected to build the probability distributions of the parameters, where N corresponds to the "time" lag at which the correlation function has decayed to zero. In this study, a threshold of 10% was used, so that points of the MC for which the correlation function has decayed by more than $\sim 90\%$ were considered independent samples of the equilibrium distribution. MCMC simulations were implemented in the Scilab software.[113]

3.4 Results

Relaxation rates were calculated for different internal mobility parameters. These rates were then fitted through a Bayesian MCMC approach described in the previous section, from three subsets of fields were compared: fields larger than 14.09 T (600 MHz), fields larger than 5 T (213 MHz), and all fields.

Simulations were performed with the FBD model of Eq. 3.8, for the chosen sets of parameters. These are meant to represent typical situations of a more or less fast fluctuating spin bearing backbone amide N-H moieties, associated with internal time scales τ comprised between 10 ps and 1 ns, in a macromolecule with overall tumbling time τ_0 =4 ns or 15.1 ns.

The present study aims at evaluating a simultaneous estimation strategy of both internal and overall dynamical parameters. However, in order to reduce the dimension of the parameter space and make the parameter search easier, c_{el} in Eq. 3.8 was assumed known and kept fixed to its exact value in all the simulations.
The investigation of FBD involved internal characteristic time (see Eq. 3.8) τ =10, 50, 100, 200, 500, and 1000 ps, whilst motion restriction S^2 and the fractionarity α of the FBD, were fixed to the respective values $S^2 = 0.8$ and $\alpha = 0.7$.

3.4.1 Residues with $\tau_0 = 15.1$ ns

Markov chain simulations were performed using synthetic relaxation rates obtained for the three different combinations of field strengths described above. In each case, nine simulation runs with $1.5{\cdot}10^6$ points were performed. Typical results for residues with internal correlation times ranging from 10 ps to 1 ns are shown in Figures 3.4 and 3.1 and 3.5. The results are grouped in these three different figures to emphasize the different behaviours of the Markov chains. For values of the characteristic time τ = 50 ps to 200 ps, the convergence of the MCMC simulations was overall very satisfactory. As a matter of fact, the Markov chains reached stationarity during each run for each of the model parameters (see Fig. 3.1). Interestingly, we found similar behaviours for all subsets of magnetic fields, although somewhat "noisier" the case where only high fields ($B_0 \geq 14.1$ Tesla $\rightarrow$ 23.5 Tesla) were used. Observation of the MC does not indicate improvement through the use of additional lower fields (respectively $B_0 \geq 5$ Tesla and $B_0 \geq 0.33$ Tesla).

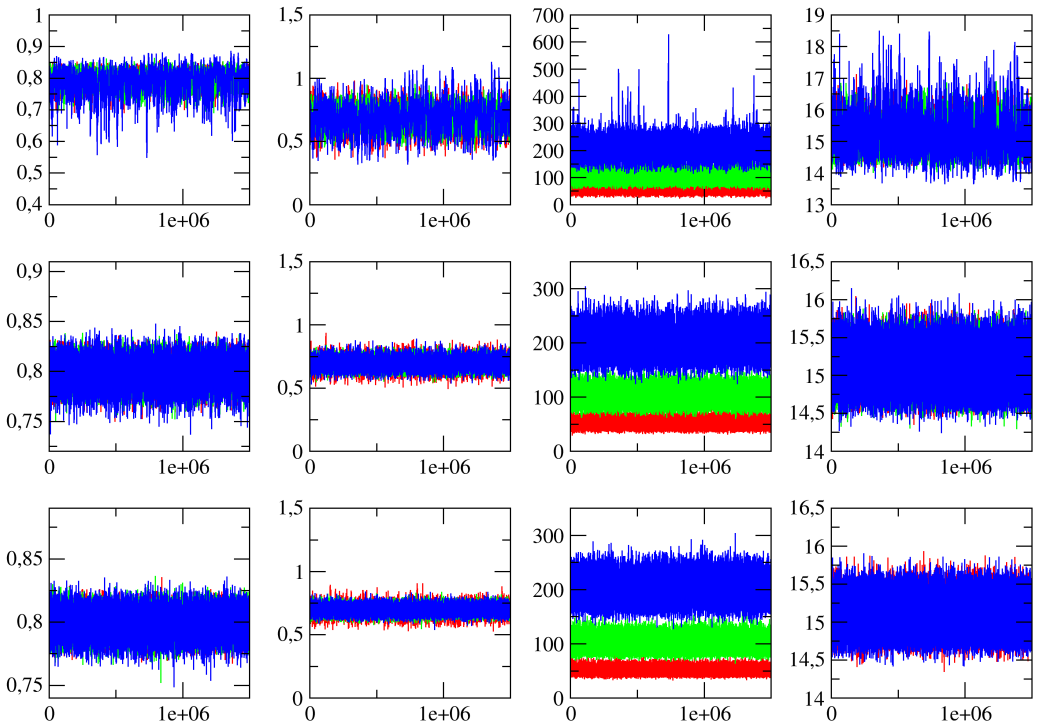

Figure 3.1: MCMC trajectories of the model parameters for relaxation rates measured for three sets of fields and $\tau_0 = 15.1$ ns. The following cases of internal characteristic times are depicted: $\tau = 50$ ps (red), $\tau = 100$ ps (green), $\tau = 200$ ps (blue). From left, S^2, α, τ, τ_0.

In order to go beyond these qualitative observations, an analysis based on the computation of marginal probability distribution functions (PDFs) for each model parameter was performed. To this aim, correlation functions of the Markov chains were calculated to estimate their correlation lengths and extract approximately iid samples from the trajectories to construct the relevant PDFs.[114]

For "intermediate" values of the internal characteristic time $\tau = 50$, 100, 200 ps, the marginal PDFs for each of the parameter θ_i exhibited well-defined, single-peaked shapes when only fields larger than $B_0 = 14$ T were used. This is shown in Figure 3.2. In this case, these distributions were well fitted by Gaussian distributions $\mathcal{N}(\mu_i, \sigma_i)$ with averages and standard deviations μ_i, σ_i that were computed. Results are indicated in Table 3.2. Importantly, averages of the order parameter ($\overline{S}^2$) and of the overall correlation time ($\overline{\tau}_0$) were accurate estimates of the true value. Moreover, the associated standard deviations $\overline{\sigma}_{S^2}$ and $\overline{\tau}_0$ had values on the order of the pseudo-experimental error used in the simulations of the relaxation rates. Interestingly, none of the average values obtained for S^2, τ_0, α and τ was significantly improved when additional B_0 fields were used, although the standard deviations $\overline{\sigma}_\theta$ were slightly lower for the "medium-to-high", with respect to the "high" field set of fields. This therefore suggested only a modest precision improvement of the model parameters. Moreover, no additional improvement was noticed in the “all field” case, which indicates that, overall, fields lower than ~ 3 T do not provide significant improvement in this case (compare with Figure 3.3 and Table 3.2). In the case of very short internal characteristic time $\tau = 10$ ps in Eq. 3.8, the situation is less favorable (see Fig. 3.4), as the Markov chains explore different regions of the parameter space, all of which are compatible with the simulated relaxation rates, and therefore correspond to possible solutions of the spectral density model $J(\omega)$, although only one of them corresponds to the true solution. Moreover, correlations between the model parameters ($S^2, \alpha, \tau, \tau_0$) are clearly visible in these trajectories. Sampling of multiple regions of the parameter space by the Markov chains therefore involves significantly longer correlations, which reduces the number of points available for the computation of the marginal PDFs. When only high magnetic fields are used, model parameters remain largely underdetermined. In particular, the marginal PDFs obtained for S^2 and τ_0 show bimodal distributions, indicating two sets (S^2, τ_0) of maximum probability, one of them corresponding to the true theoretical values.

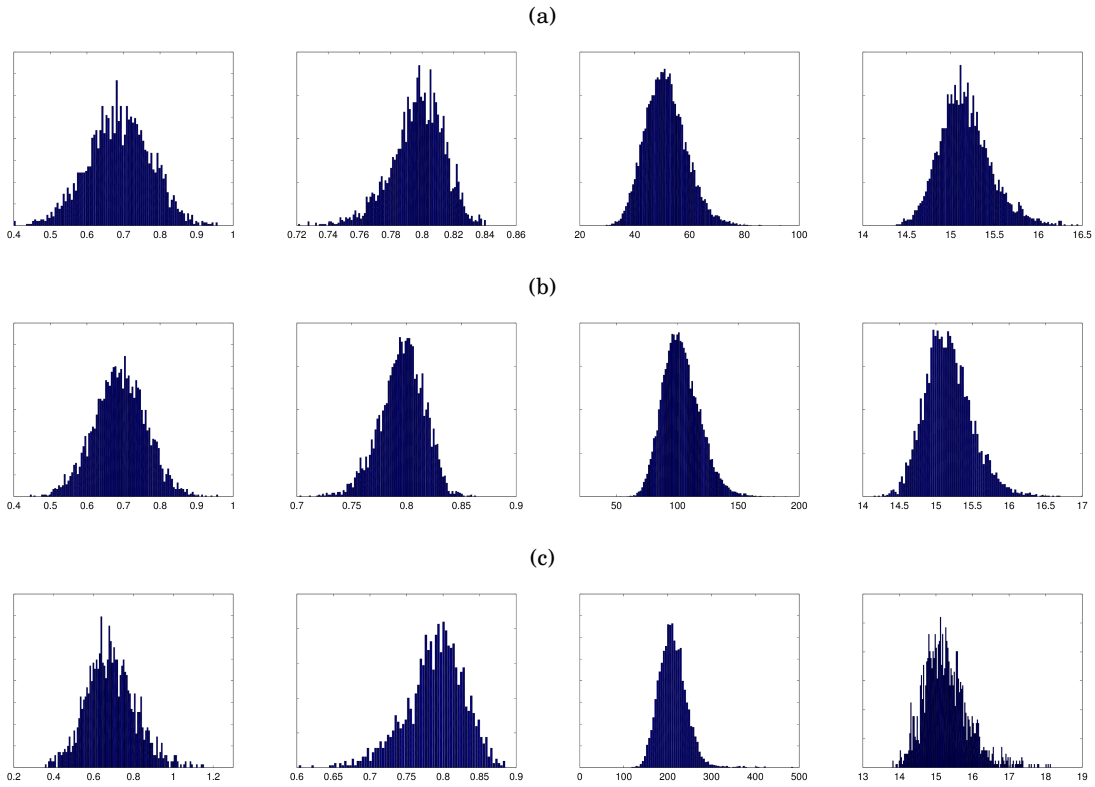

Figure 3.2: From left, PDF distributions of S^2, α, τ = a) 50, b) 100, c) 200 ps, τ_0 =15.1 ns, τ_0 obtained from MCMC simulations using synthetic measurements at high fields ($\omega_0(^1H) = 600, 800, 900, 1000$ MHz).

This feature was previously ascribed [90] to the existence of too short τ, which prevents adequate sampling of such fast internal motions, even using fields as high as 23.5 T. Interestingly, this ambiguity can be removed by using all field measurements. In this case, unimodal marginal PDFs, and centered about the true values of these parameters, can be obtained. However, although the decay of the correlation function of the Markov chains become faster as more fields are used, this is a particularly demanding situation in terms of simulation length, as only few points of the simulation remain in order to yield an iid sample of the parameters' PDFs. Nevertheless, the distributions of S^2 and τ_0 could be determined. This is also somewhat surprising, as it shows that the combination of lower frequency sampling of the spectral density functions may improve the determination of timescale parameters that are relevant for very fast motions.

Parameters α and τ showed significantly large standard deviations, which gives an indication of their relative undeterminacy. Results are indicated in Table 3.2. This therefore indicates the

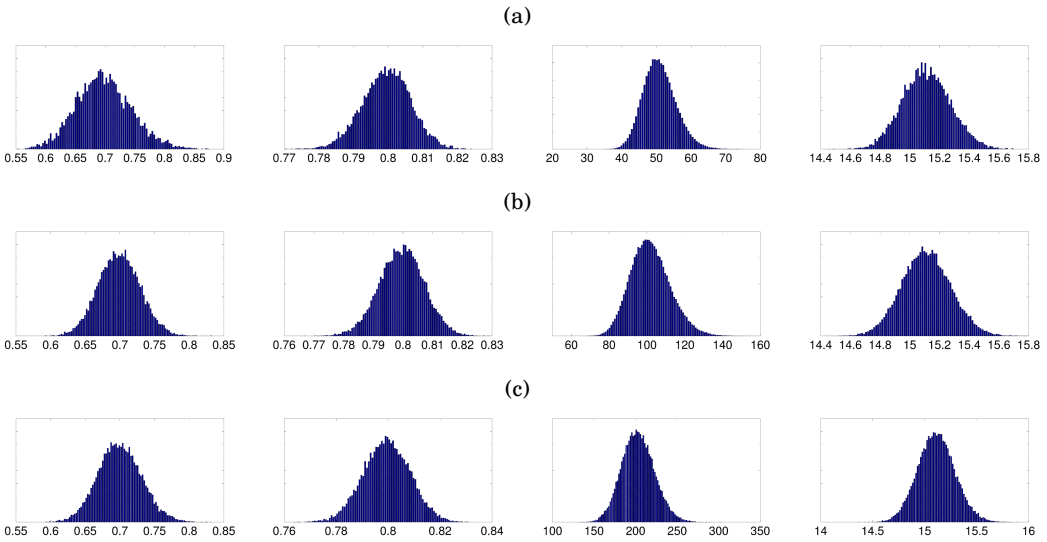

Figure 3.3: From Left, PDF distributions of the model parameters α, S^2, τ, τ_0 (from left to right) obtained from MCMC trajectories using *all fields*. Synthetic relaxation rates are obtained with $\tau_0 = 15.1$ ns, $S^2 = 0.8$, $\alpha = 0.7$, a) $\tau = 50$ ps, b) $\tau = 100$ ps, c) $\tau = 200$ ps

relative advantage of using multiple field measurements in this case. In the case where $\tau = 500$ ps, well constrained Markov chains require "medium-to-high" field rates (see Fig.3.5). Well defined marginal PDFs for all the parameters of the model can be obtained using the complete set of fields (see Figure 3.6) . Finally, when $\tau = 1000$ ps, the determination of accurate PDFs remained challenging, for the same reasons above. Overall, Figure 3.6 illustrates the fact that the use of a wide range of magnetic fields improves the parameter search in cases where $\tau = 10$ ps and $\tau = 1$ ns, although with much lower efficiency than for the other τ values.

3.4.2 Residues with $\tau_0 = 4.0$ ns

This was intended to mimick the case of a faster tumbling molecule, with identical internal parameters as in the τ_0 =15 ns case ($S^2 = 0.8$, $\tau = 10$ ps to 1 ns $\alpha = 0.7$). The Bayesian MCMC strategy was applied and probability distribution functions of the model parameters were computed. Results are summarized in Figs. 3.7-3.8.

Results are in sharp contrast with the previous simulations. Indeed, in this case, it appears that the Markov chains explore a significantly larger range of parameter values, irrespective of the value of the characteristic internal time τ. As expected, this behaviour is associated with slow

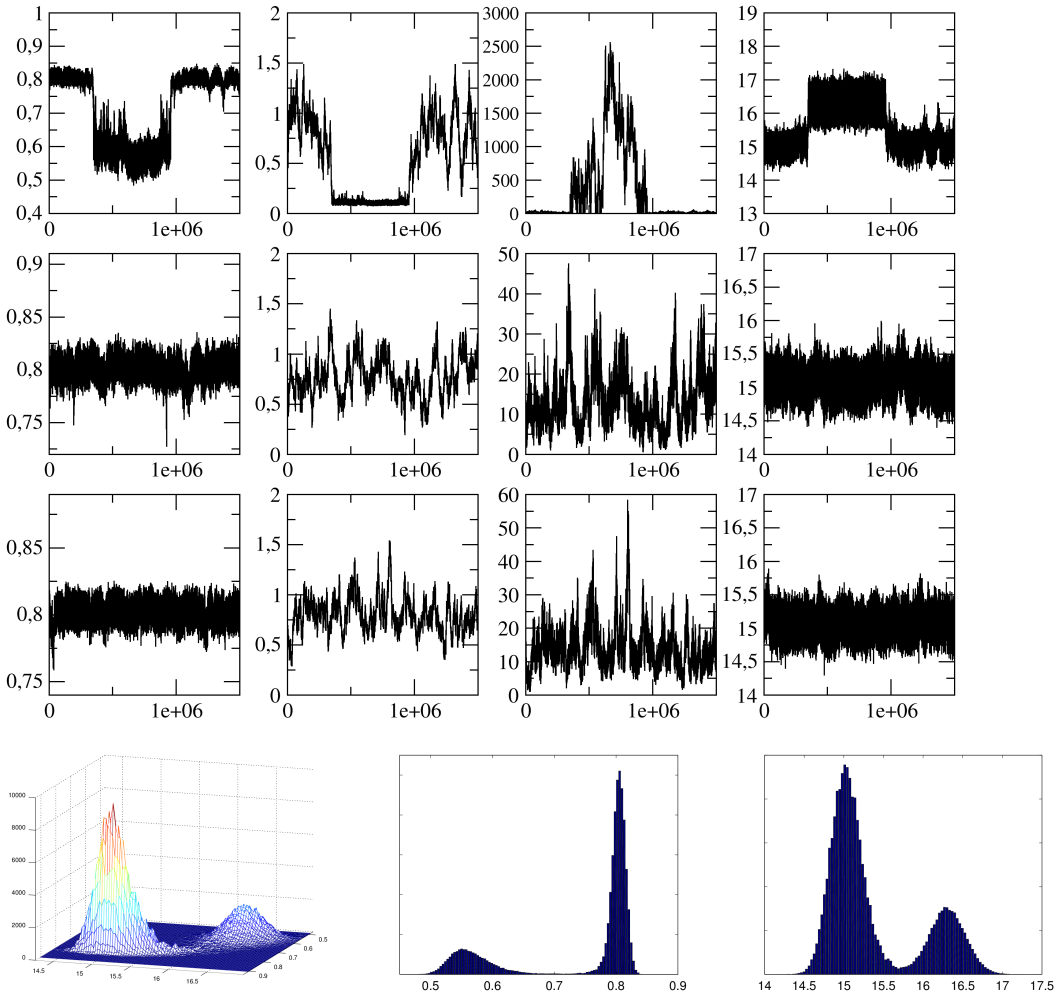

Figure 3.4: MCMC trajectories of the model parameters for relaxation rates measured for three sets of fields and $\tau_0 = 15.1$ ns. The case of short internal characteristic times $\tau = 10$ ps is depicted). From left, S^2, α, τ, τ_0.

decays of the parameters' correlation functions. Therefore, only few points could be selected to achieve iid samples from which the marginal PDFs were computed. Typically, one point every 10^4 (or more) was selected, so that more than 35 simulations were needed to achieve the PDF histograms displayed in Figs. 3.7-3.8. The undeterminacy of the parameters is clearly attested by the presence of asymmetric, or bimodal distributions (exemplified in Figs. 3.7-3.8 for $P(S^2)$ in the case where $\tau = 10$ ps). Nevertheless, averages and standard deviations associated with the distributions of these parameters were estimated and are summarized in Table 3.3

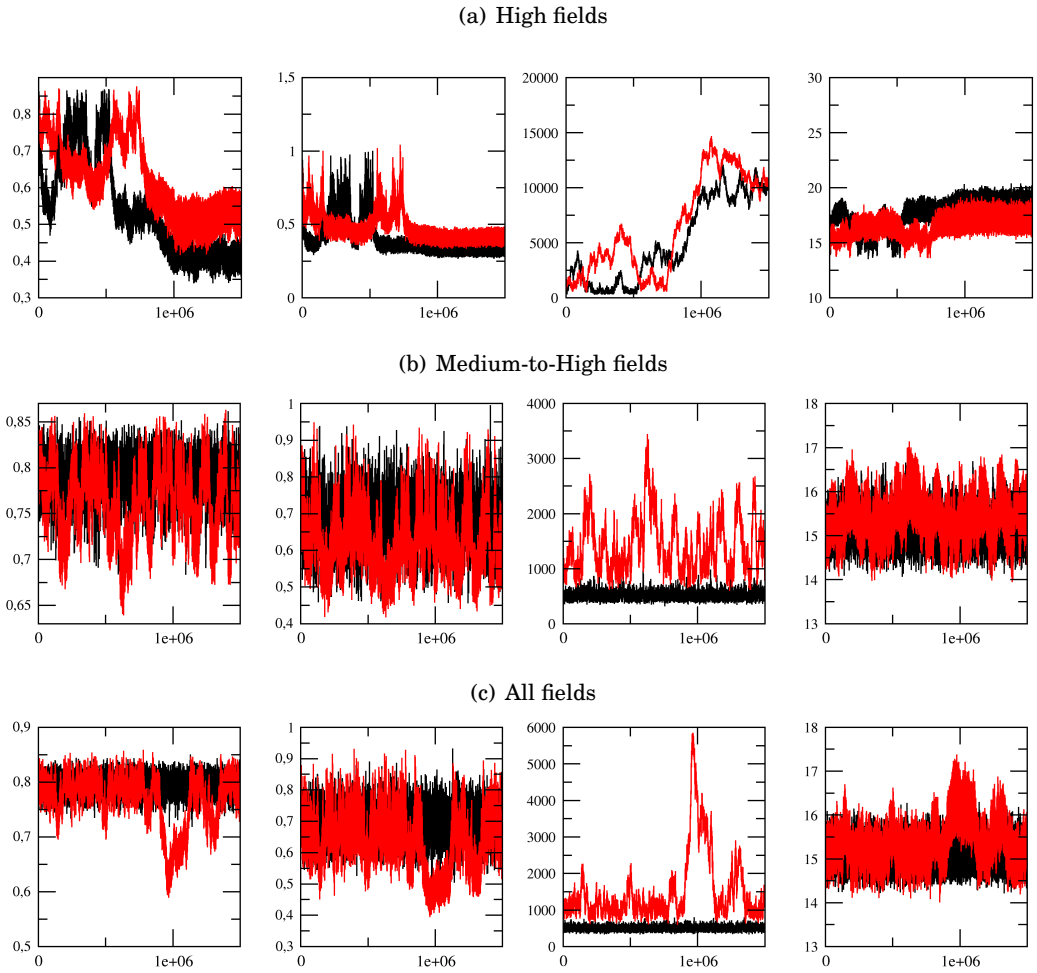

Figure 3.5: MCMC simulations of the model parameters (S^2, α, τ, τ_0, from left to right) for relaxation rates measured for three sets of fields and τ_0 = 15.1 ns. The case of long internal characteristic time τ is shown: τ = 500 ps (black), τ = 1000 ps (red).

3.4.3 Use of the probability distribution function of τ_0 as prior information

The above discussion shows that the use of a relaxometry strategy may improve in some cases the determination of internal dynamics parameters. However, possible advantage depends on several factors, one of them being the magnitude of the overall diffusion. But in any event, the larger the number of simultaneously free model parameters the lower the estimation efficiency. This search inside a larger parameter space seems to be in this context one of the main obstacles to satisfactory.

This state of affairs can nevertheless be improved through a MCMC Bayesian approach, as

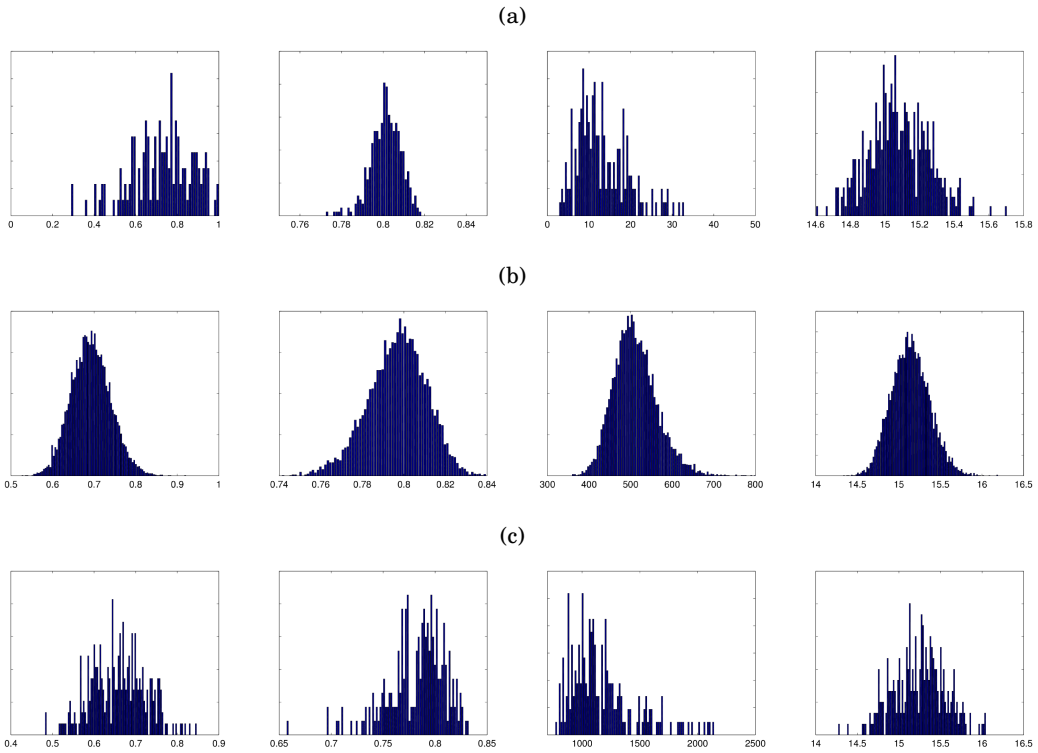

Figure 3.6: From left, PDF distributions of the model parameters S^2, α, τ, τ_0 obtained from the simulations of Fig. 3.4 and 3.1 using *all fields*. Synthetic relaxation rates are obtained with $\tau_0 = 15.1$ ns, $S^2 = 0.8$, $\alpha = 0.7$, a) $\tau = 10$ ps, b) $\tau = 500$ ps, c) $\tau_0 = 1000$ ps

additional knowledge on the model can be included through the use of the prior probability $P(\boldsymbol{\theta}, I)$ (see section 3.2.3).

Indeed, in actual situations, measurements are performed on an ensemble of residues that belong to the same protein and therefore share the same overall diffusion properties.

Some of these residues can be expected to be amenable to the above MCMC strategy of relaxation rate analysis. Typically, in cases of less restricted motions, with lower internal characteristic time scales, probability distribution functions of the model parameters can be more easily extracted. This would be the case, for instance when S^2 = 0.8, τ = 50 ps, α = 0.7, and τ_O =15.1 ns). Therefore, since all residues in the molecule are assumed to share the same overall diffusion, an estimate of the marginal $P(\tau_0)$ can be obtained from such residues. Assuming N such residues with relaxation rates R_i, $i = 1, \ldots, N$, one has $P(\tau_0) \propto \Pi_{1,\ldots,N} P_i(\tau_0)$. [91] $P(\tau_0)$ can then be used as a prior in the MCMC search for the internal parameters of other residues of the protein. The

τ		high fields	medium-to-high fields	all fields
10 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	0.80 ± 0.01	0.8 ± 0.01
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a \pm n.a$	0.77 ± 0.20	0.79 ± 0.20
	$\overline{\tau} \pm \sigma_\tau$	$n.a \pm n.a$	14.19 ± 16.28	13.34 ± 6.61
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a \pm n.a$	14.08 ± 0.17	15.10 ± 0.17
50 ps	$\overline{S^2} \pm \sigma_{S^2}$	0.80 ± 0.01	0.80 ± 0.01	0.80 ± 0.01
	$\overline{\alpha} \pm \sigma_\alpha$	0.69 ± 0.09	0.70 ± 0.05	0.70 ± 0.04
	$\overline{\tau} \pm \sigma_\tau$	51.35 ± 7.46	51.07 ± 5.63	50.93 ± 4.77
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	15.16 ± 0.29	15.11 ± 0.17	15.11 ± 0.16
100 ps	$\overline{S^2} \pm \sigma_{S^2}$	0.79 ± 0.02	0.80 ± 0.01	0.80 ± 0.01
	$\overline{\alpha} \pm \sigma_\alpha$	0.70 ± 0.07	0.70 ± 0.03	0.70 ± 0.03
	$\overline{\tau} \pm \sigma_\tau$	103.29 ± 15.17	101.95 ± 11.78	101.89 ± 10.49
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	15.17 ± 0.32	15.11 ± 0.17	15.11 ± 0.16
200 ps	$\overline{S^2} \pm \sigma_{S^2}$	0.78 ± 0.04	0.80 ± 0.01	0.80 ± 0.01
	$\overline{\alpha} \pm \sigma_\alpha$	0.68 ± 0.12	0.70 ± 0.04	0.70 ± 0.03
	$\overline{\tau} \pm \sigma_\tau$	212.34 ± 35.50	203.11 ± 21.24	203.03 ± 16.63
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	15.28 ± 0.62	15.11 ± 0.20	15.11 ± 0.18
500 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	0.79 ± 0.02	0.80 ± 0.01
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	0.68 ± 0.07	0.69 ± 0.05
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	529.13 ± 73.39	511.00 ± 52.77
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	15.21 ± 0.31	$15.14 \pm 0.0.23$
1000 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	0.75 ± 0.05	0.78 ± 0.03
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	0.60 ± 0.10	0.66 ± 0.07
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	$n.a \pm n.a$	1201.30 ± 423.74
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	15.52 ± 0.53	15.25 ± 0.31

Table 3.2: Average and standard deviation of the PDFs for the different sets of magnetic fields. The overall correlation time is $\tau_0 = 15.1$ ns (n.a.: not applicable)

result of such a strategy is illustrated in Figs. 3.10-3.11 and Figs. 3.12-3.13, where the marginal PDFs of the parameters obtained for $\tau_0 = 15.1$ ns and $\tau_0 = 4$ ns are depicted.

In the example with $\tau_0 = 15.1$ ns, this prior information makes it possible to extract the other model parameters using only high fields ($B_0 \geq 14.1$) T. Alternatively, in the case $\tau_0 = 4$ ns, using the prior in the form of a Gaussian with known average and standard deviation, significant improvement can be achieved in the definition of the marginal parameters of the system model. However, in contrast with the previous example, the use of measurements at all fields were still necessary to obtain significant improvement.

The fact that introducing additional information improves the estimation of the parameters' probability distribution functions is actually not surprising, as it amounts to constrain the parameter search to a reduced region of the parameter space. Thus, the narrower the prior PDF of the overall τ_0, the easier the search. Of course, in the limit of an infinitely narrow distribution

τ		high fields	medium-to-high -high fields	all fields
10 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.80 ± 0.03
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.49 ± 0.36
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	178.78 ± 353.33
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	4.02 ± 0.04
50 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.79 ± 0.02
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.61 ± 0.16
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	73.65 ± 62.55
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	4.01 ± 0.04
100 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.79 ± 0.02
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.65 ± 0.10
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	131.03 ± 59.84
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	4.01 ± 0.04
200 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.79 ± 0.02
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.67 ± 0.07
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	245.97 ± 82.76
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	$4.01 \pm 0.0.04$
500 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.79 ± 0.03
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.68 ± 0.07
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	661.10 ± 283.61
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	4.01 ± 0.04
1000 ps	$\overline{S^2} \pm \sigma_{S^2}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.80 ± 0.03
	$\overline{\alpha} \pm \sigma_\alpha$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	0.71 ± 0.08
	$\overline{\tau} \pm \sigma_\tau$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	1124.4 ± 452.92
	$\overline{\tau_0} \pm \sigma_{\tau_0}$	$n.a. \pm n.a.$	$n.a. \pm n.a.$	4.01 ± 0.04

Table 3.3: Average and standard deviation of the PDFs for the differents sets of magnetic fields. The overall correlation time is $\tau_0 = 4$ ns. (n.a.: not applicable)

of τ_0, the search for the $n = 4$ model parameters is reduced to a region of the parameter space with a fixed value τ_0 ($P(\tau_0) = \delta(\theta - \tau_0)$), which is equivalent to a space with $n - 1$ dimensions.

The significant difference observed between the MCMC simulations performed with the two overall correlation times $\tau_0 = 4$ ns and $\tau_0 = 15.1$ ns was at first surprising, and the fact that using a large set of magnetic fields does not necessarily improve the "resolving power" of relaxometry experiments a bit disconcerting. A qualitative explanation can be found if one considers the contributions to the relaxation rates R_1 of the various dynamical parameters. In the case of fractional brownian dynamics, It is seen from Eq. 3.8 that R_1 is the sum of two terms. The first one, which we denote $R_1(S^2, \tau_0)$ depends on S^2 and τ_0 only, whereas the second one depends on all model parameters. For each value of the overall correlation time τ_0, and for all τs, the relative contribution $R_1(S^2, \tau_0)/R_1$ were computed. It was assumed that the respective contributions of both these terms can explain part of the behaviour of the relaxation rate analysis. Indeed, when

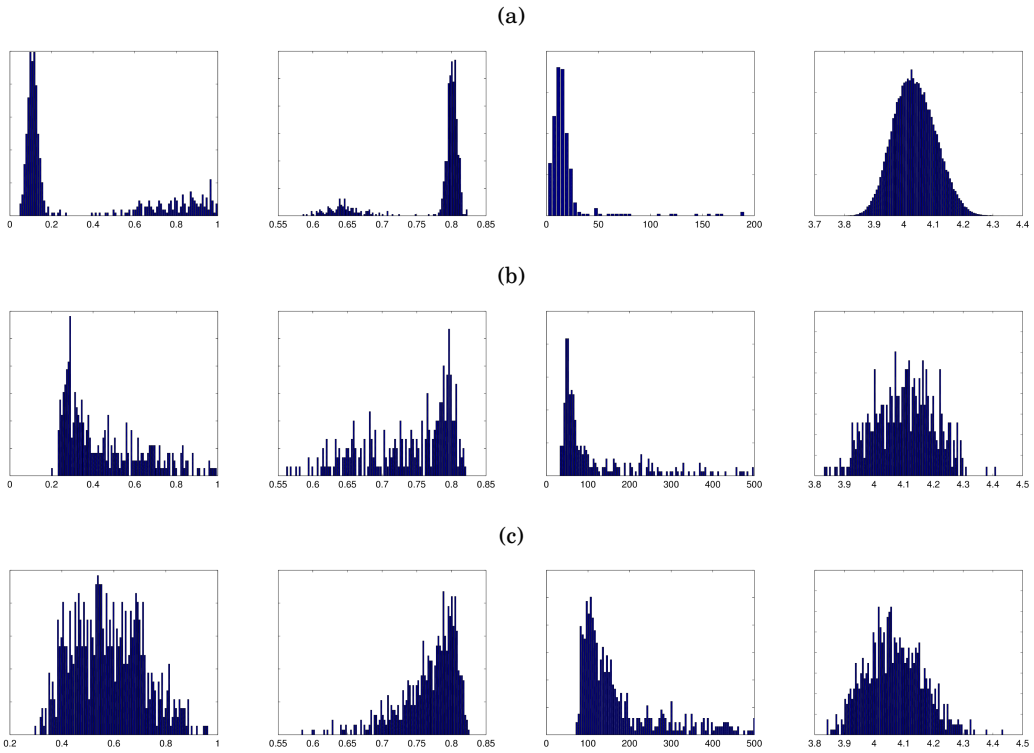

Figure 3.7: PDF distributions of the model parameters S^2, α, τ, τ_0 (from left) for the case of τ_0=4.0 ns extracted from relaxation rates using all fields. Parameter values are $S^2 = 0.8$, $\alpha = 0.7$. a) 10, b) 50, c) 100 ps of internal correlation time.

$R_1(S^2,\tau_0)/R_1 \sim 1$, R_1 is completely defined by $R_1(S^2,\tau_0)$. Consequently, the remaining α and τ parameters become irrelevant in the fitting process and the Markov chains are expected to slowly reach stationarity. Results are shown in Figures 3.14(a) and 3.14(b), where the horizontal line indicates a 90% $R_1(S^2,\tau_0)/R_1$ ratio, above which only $R_1(S^2,\tau_0)$ predominantly contributes. Calculations show that, for $\tau_0 = 15.1$ ns, $R_1(S^2,\tau_0)$ contributes always less than 90% of R_1 for $B_0 \geq 14$ T, except in the case $\tau = 10$ ps. This means that measurements at high fields tend to provide sufficient constraint for the determination of the remaining α and τ parameters. Alternatively, at lower fields, the contributions of the latter to the relaxation rates are moot, so that these mainly reflect the values of S^2 and τ_0. Note that when $\tau = 10$ ps, even high-field relaxation rates fail to correctly sample the associated fast motions, so that parameter extraction remains difficult to achieve.

This kind of argument can be also applied to the $\tau_0 = 4$ ns case, where, in contrast to the above,

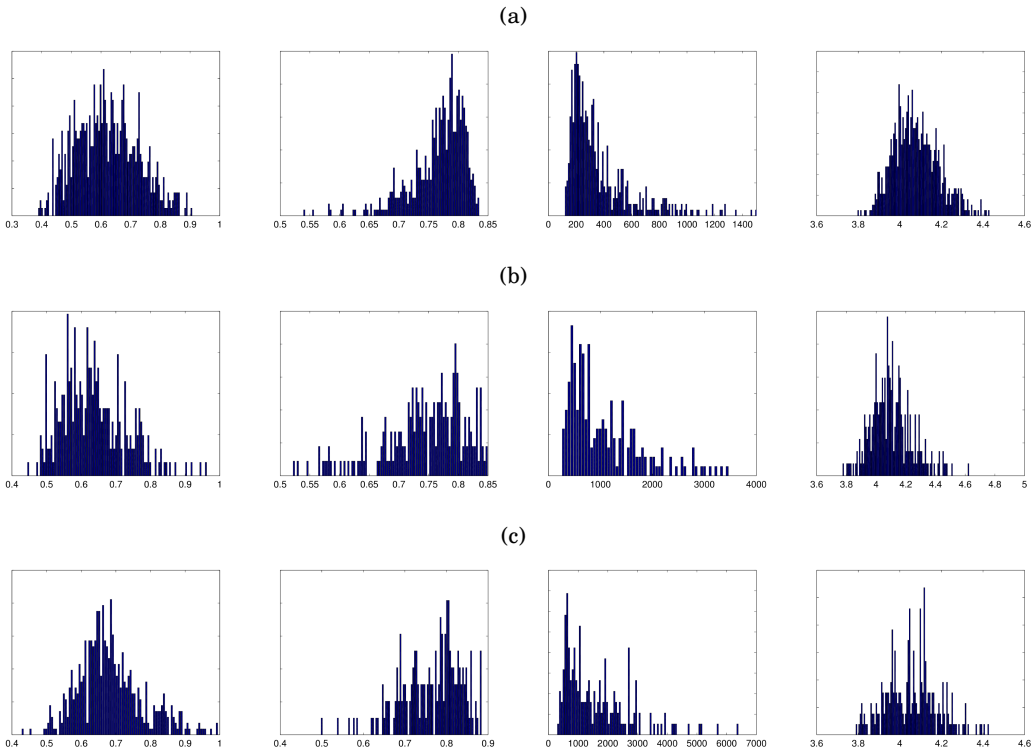

Figure 3.8: PDF distributions of the model parameters S^2, α, τ, τ_0 (from left) for the case of τ_0=4.0 ns extracted from relaxation rates using all field combinations. Parameter values are $S^2 = 0.8$, $\alpha = 0.7$. a) 200, b) 500, c)1000 ps of internal correlation time.

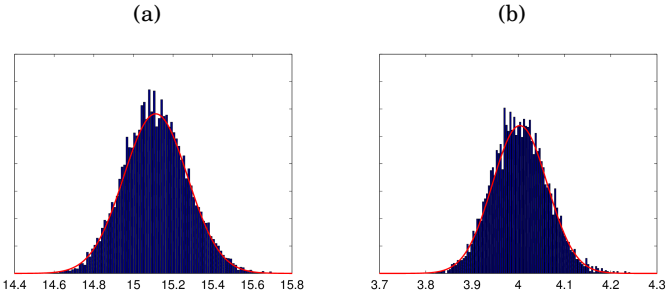

Figure 3.9: Gaussian distributions extracted for the parameter τ_0, from a residue with a) $S^2 = 0.8$, $\alpha = 0.7$, $\tau = 50$ ps, $\tau_0 = 15.1$ ns, b) $S^2 = 0.6$, $\alpha = 0.6$, $\tau = 50$ ps, and $\tau_0 = 4.0$ ns.

$R_1(S^2,\tau_0)/R_1$ is larger than 90%, for most values of τ, so that parameter estimation remains ambiguous and clearly, combining high and low fields does not significantly improve matters.

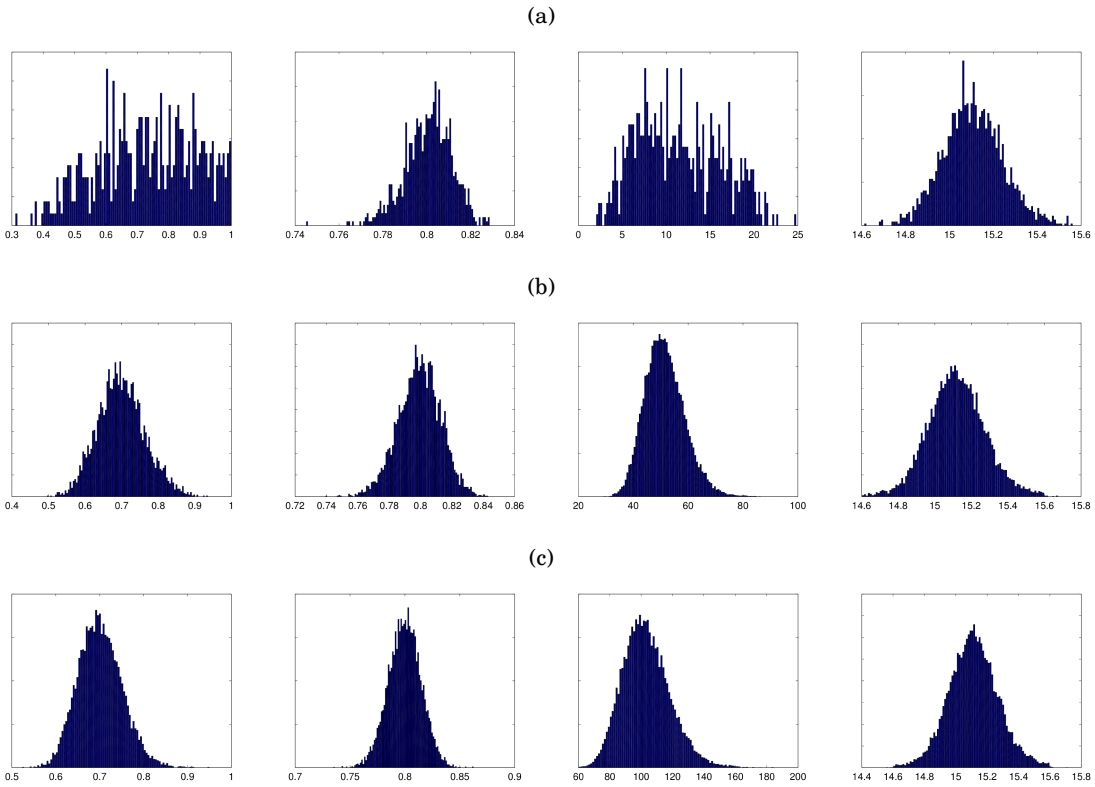

Figure 3.10: PDFs extracted using *high fields only* and prior information on τ_0 =15.1 ns. From left α, S^2, τ, τ_0. (a) 10 ps, (b) 50 ps, (c) 100 ps. Compare with Figs 3.2 and 3.3

3.5 Conclusions

In this chapter, we investigated the potential use of a Markov Chain-Monte Carlo (MCMC) approach to the analysis of NMR relaxometry measurements. The primary goal was to investigate the possibility to perform simultaneous fitting of all dynamical parameters, i.e., relative to both internal and overall motions, thereby avoiding . We have shown the potentially beneficial use of an analysis of NMR relaxation rates along these lines on synthetic data. With respect to the conventional approach of data fitting, this method is much more time effective since it avoids time consuming minimization routines. In addition, our use of the MCMC approach allows one to avoid commonly used model selection strategies. In contrast to conventional fitting, insufficient characterization of some of the parameters does not imply the rejection of the model in favor of a simpler one, the method indicates what information may be extracted from the data, but also what cannot. The Bayesian approach naturally leads to estimates of the total and the

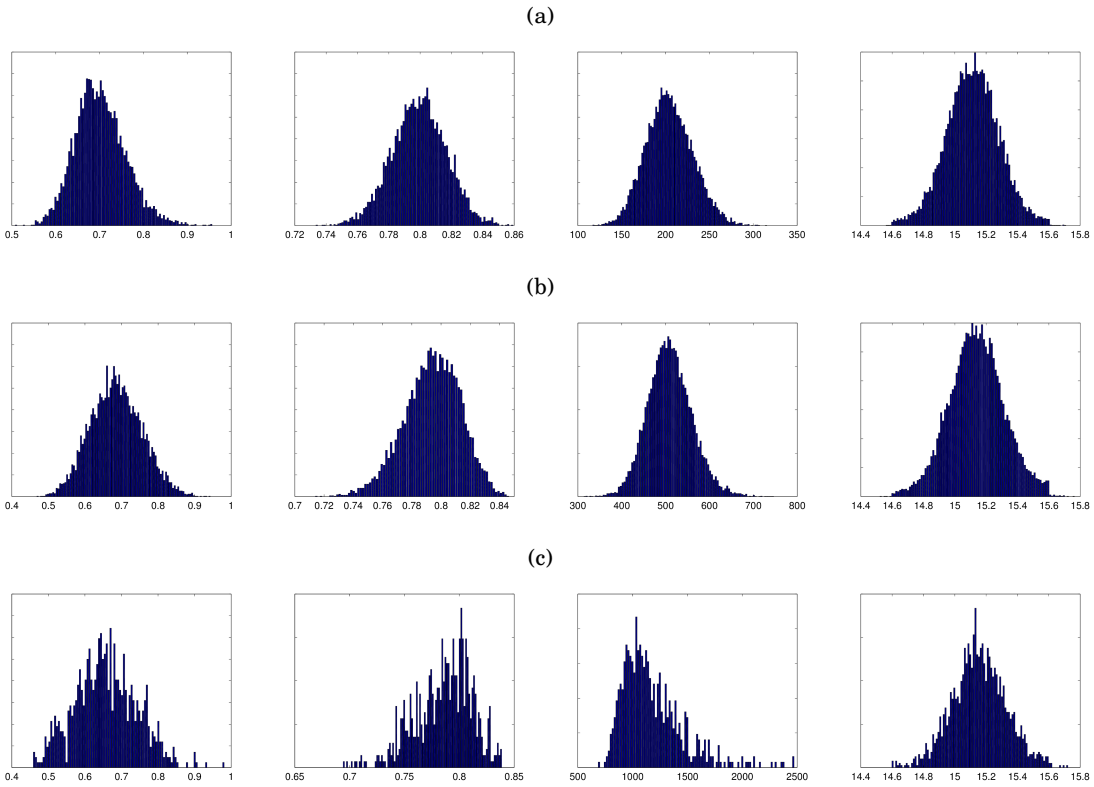

Figure 3.11: PDFs extracted using *high fields only* and prior information on τ_0 =15.1 ns. From left α, S^2, τ, τ_0.(a) 200 ps, (b) 500 ps, (c) 1000 ps. Compare with Figs 3.2 and 3.3

various marginal probability distribution functions of the model parameters. It has been shown that adding measurements at lower fields helps to better describe the spectral density function, leading to improvements in the search of the correct parameter distribution. This effect is more evident in the case of slow overall motion while for faster global tumbling such improvement is not always present. The introduction of additional knowledge that retain only physically meaningful solutions through the prior probability allows to better constrain the search within smaller regions of the parameter space, thereby increasing the search efficiency and PDF determination. However, it appeared that the use of (synthetic) relaxation rates at several fields improves the determination of the parameters under certain conditions only. These depend on the relative values of the internal and overall correlation times, for instance. It was qualitatively argued that favorable cases correspond to situations where each of the various dynamical parameters significantly affect the relaxation rates, or the spectral density function in distinct regions of

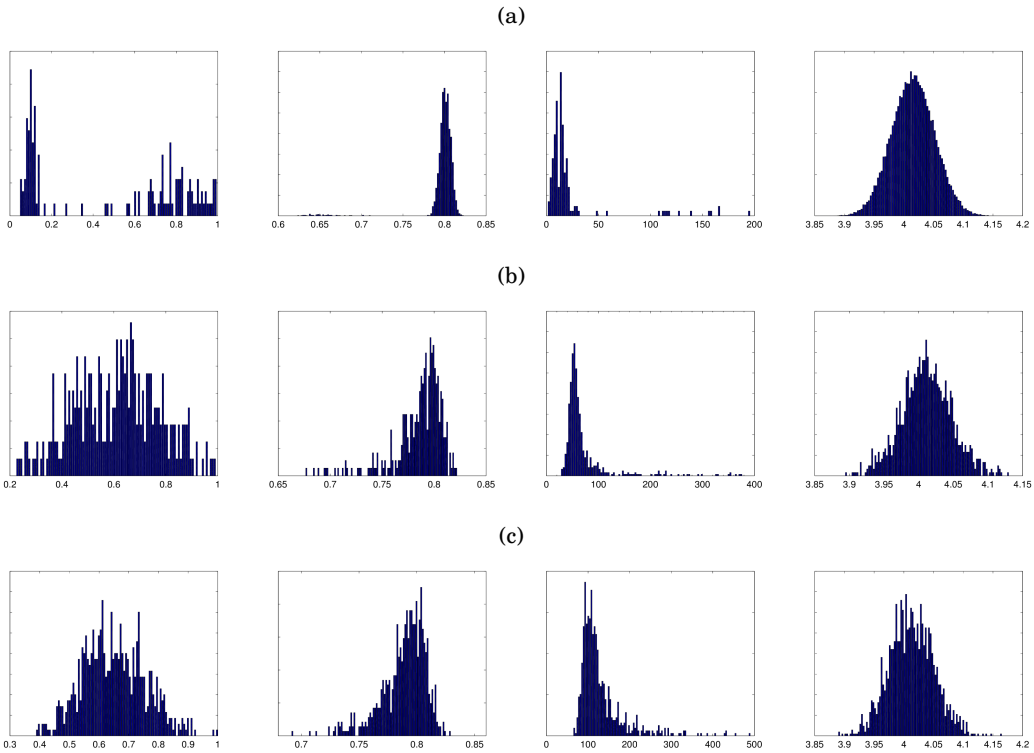

Figure 3.12: PDF distributions of the model parameters, from left, α, S^2, τ, τ_0 extracted from relaxation rates using *all fields and prior information on* τ_0. Parameter values are $S^2 = 0.8$, $\alpha = 0.7$, $\tau_0 = 4$ ns, and a) $\tau = 10$ ps, b) $\tau = 50$ ps, c) $\tau = 100$ ps (compare Fig. 3.7). $P(\tau_0)$ was estimated from the parameter set $S^2 = 0.6$, $\alpha = 0.9$, $\tau_0 = 4$ ns, and $\tau = 50$ ps.

the frequency spectrum. These temporary conclusions relied on a study based on a particular physical model. In order to extend our study and test these conclusions to other situations, the same MCMC analysis will be tested on different models of dynamics, such as the SRLS model, for instance.

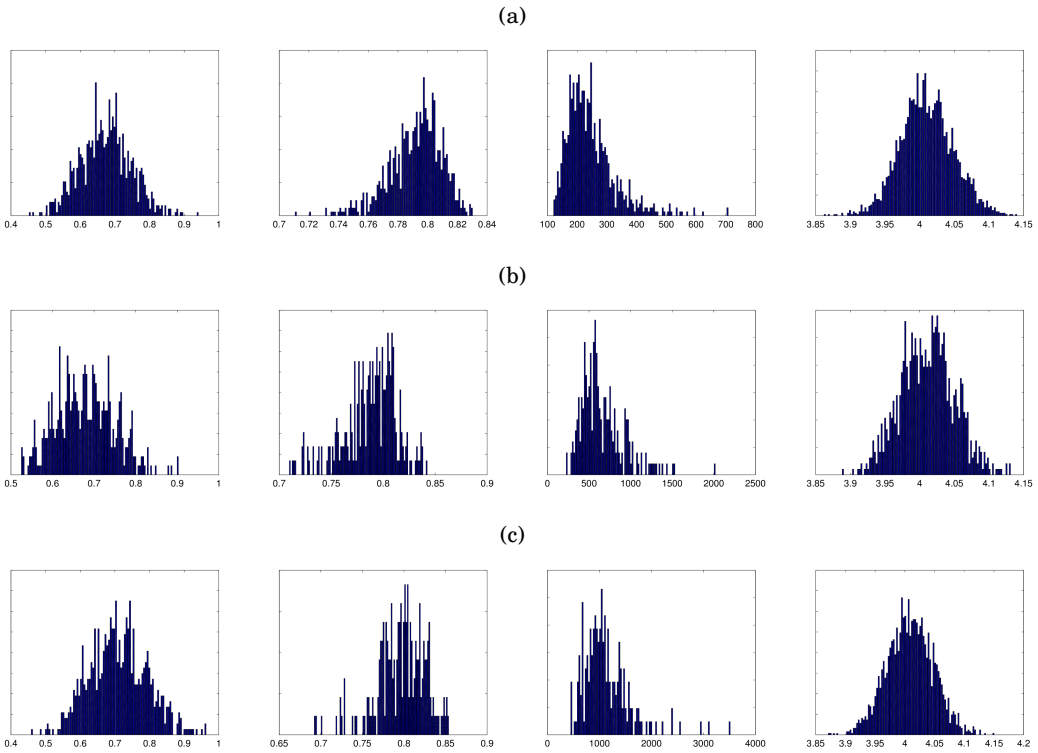

Figure 3.13: PDF distributions of the model parameters, from left, α, S^2, τ, τ_0 extracted from relaxation rates using *all fields and prior information on* τ_0. Parameter values are $S^2 = 0.8$, $\alpha = 0.7$, $\tau_0 = 4$ ns, and a) $\tau = 200$ ps, b) $\tau = 500$ ps, c) $\tau = 1000$ ps (compare Fig. 3.8). $P(\tau_0)$ was estimated from the parameter set $S^2 = 0.6$, $\alpha = 0.9$, $\tau_0 = 4$ ns, $\tau = 50$ ps.

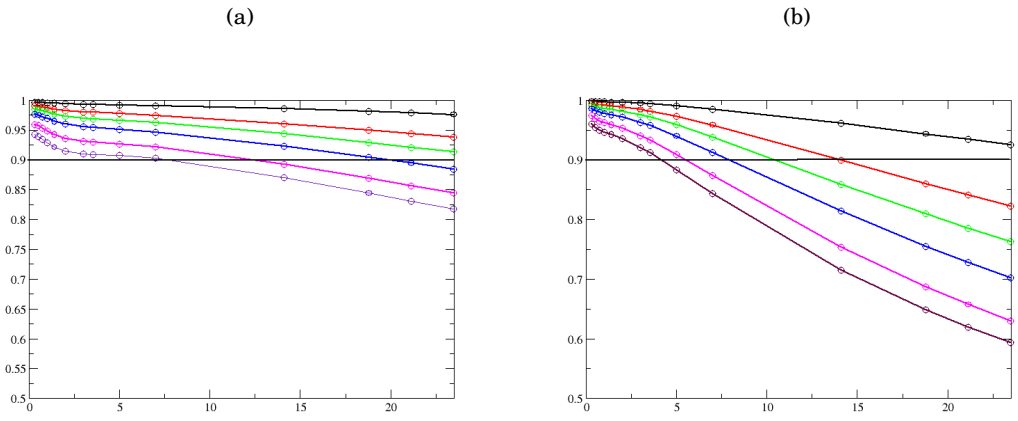

Figure 3.14: Contribution of the $R_1^{(o)}$ part to the relaxation rate for a) $\tau_0 = 4$ ns and b) $\tau_0 = 15.1$ ns. From top, 10, 50, 100, 200, 500, 1000 ps of τ. In abscissa, field in Tesla.

CHAPTER 4

DECOMPOSITION OF PROTEINS INTO DYNAMIC UNITS FROM ATOMIC CROSS-CORRELATION FUNCTIONS

4.1 Introduction

Internal mobility of a protein is recognized as a basic factor affecting its mechanism of action at the molecular level, and therefore its function. This has motivated the development of various techiques to study protein dynamics, amongst which NMR has the unique capability to provide dynamical information covering time scales ranging from the pico-second to the microsecond and beyond, together with localization at atomic resolution. However, an accurate characterization of internal motions in the protein that goes beyond the mere identification of characteristic time scales and includes possible correlations across the protein, remains an open issue. Molecular dynamics (MD) simulations represent a most valuable tool to contribute to shed some light on these questions.

When analyzing experimental observables of protein dynamics, one of the main issues is to characterize the underlying key motions in relation with some particular aspect of the dynamics probed by the experiments. Several methods, such as principal components analysis (PCA) [115–117] or essential dynamics (ED), [118–120] are based on analyses of the covariance matrix through the projection of the atomic position fluctuations in an all-atom MD trajectory onto the

eigenspace of coordinate fluctuations. This allows one to distinguish possibly different domains and to analyze their motions separately. Alternatively, the dynamic behaviour of a protein can be described in a coarse-grained manner by clustering together subsets of atoms in the molecule to form elementary rigid units. As far as coarse-graining is concerned, it is usually performed on structural (spatial proximity) or energetics (free-energy profiles) bases. [121–124] Since these approaches are more or less explicitly related to the protein structure and the amplitudes of the fluctuations thereof, through the covariance matrix, so are the derived methods to identify motional domains, giving rise to *structure-based motional domains*. Such approaches do not explicitly take into account the time scales at which motions occur. And due to the wide range of time scales usually present, the "correlation maps" obtained by such analyses may significantly depend on the length of the simulation, which can make it difficult to reach stationary correlation matrix through MD simulations.[125] In fact, relative motions of such ensembles occur on an extremely wide range of time scales, including both long times such as those compatible with allostery, and short ones, in the THz range, where collective motions also seem to exist.[126]
In this chapter, a new clustering approach of the ensemble of atoms representing the protein into subsets is introduced. This new methodology is based exclusively on the characteristic times of their position cross-correlation functions.

Atomic correlation is solely defined on the basis of the characteristic times of their correlation functions, without reference to the amplitudes of the correlated motions. To this aim, we build our protein motion analysis on suitably chosen cross-correlated functions of the atomic coordinates, and we show that the characteristic times of these correlation functions, defined as the area under the curve of that part of the correlation function that decays to zero, can be used to perform a cluster reduction of the protein.

A metric in the space of correlation times of the protein is introduced, which is used to define groups of dynamically nearby atoms. This time-windowed clustering analysis was performed on MD simulation trajectories through implementation of the Affinity Propagation algorithm[127]. This allows one to identify subsets of atoms in the protein belonging to common "motional units" that are defined without any direct reference to the protein structure and are unrelated to its structural domains. Our approach is therefore in contrast with widely used covariance

matrix methods relying on the analysis of the coordinate fluctuation amplitudes[125, 128, 129] to determine groups of 'linked' atoms in a protein.

This work has been carried out under the guidance of Dr. Paolo Calligari at the University of Padova.

4.2 Methodology

In this section, the protocol developed for the determination of time-dependent similarities between pairs of atoms, in terms of effective correlation times, is described. Such a goal requires the definition of the atoms that represent the protein (all atoms, heavy atoms, etc), as well as the type of correlation functions to be used for the analysis (atomic positions, distances, relative orientations, etc).

4.2.1 Reference atoms

Since there is no unique choice of the atoms representing a protein, this should be tailored to the problem at hand. In order to reduce the computational burden of the protocol, a few representative atoms only were selected for each residue to probe the dynamical properties of the latter. This choice may also depend on the experimental observables (e.g. X-ray diffraction, NMR spectroscopy). Moreover, if the investigation is focused, for example, on slow backbone motions, a rather natural set of *representative atoms* could be the backbone C_α or amide N atoms. In the present work, backbone motions of different proteins are based on C_α as reference atoms.

4.2.2 Correlation functions

To describe correlated motions several, potentially complementary, observables can be envisaged. The correlation functions of atomic coordinates represent the most straightforward and natural tool to analyze internal motions in proteins. However, coordinate cross-correlation functions do not necessarily decay to zero nor have a constant sign, and are therefore more difficult to handle from the computational point of view.

To takle this problem, an alternative useful and computationally cheap set of observables is introduced, which is provided by normalized distance correlation functions:

$$D_{ij}(t) := \frac{\langle u_{ij}(0)|u_{ij}(t)\rangle}{\langle \mathbf{u}_{ij}^2\rangle} = \frac{1}{\langle \mathbf{u}_{ij}^2\rangle T_{MD}} \int_0^{T_{MD}} u_{ij}(\tau)u_{ij}(t-\tau)d\tau, \tag{4.1}$$

where $u_{ij} = r_{ij} - \langle r_{ij}\rangle$ and $r_{ij} = |\mathbf{r_{ij}}| = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between atoms i and j, of coordinates $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$. The integral in the above equation is normalized with respect to $\langle \mathbf{u}_{ij}^2\rangle = D_{ij}(0)$. These quantities suffer much less from non-ideal sampling than coordinate correlation functions, because they only refer to correlated motions along the direction given by the distance vector $\mathbf{r_{ij}}$. Additionally, they decay monotonously to zero and are easily integrated (see next Section), and at the same time, still account for cross-correlated motions of pairs of atoms. Finally their calculation do not require the global motions of the protein to be removed beforehand. This is particularly interesting, as it prevents the introduction of the additional assumption that global and local motions are statistically uncorrelated. This condition may not be satisfied for small molecules in which overall tumbling happens in the same time scale of internal motions. For these reasons, such distance correlation functions are less prone to statistical and numerical problems and are therefore good candidates to probe internal motions.

4.2.3 Convergence of correlation functions

An automated and reliable procedure for the assessment of convergence of the correlation functions is therefore designed, and performed as follows. The numerical correlation functions (CF) presented in Section 4.2.2 approximate the corresponding ideal correlation functions ($T_{\mathrm{MD}} = \infty$). The analyses of numerical CF are thus restricted to time lags shorter than a maximum value t_{max} with T_{MD}/t_{max} large enough to limit the effects of the finite value of T_{MD} and thus to allow the correlation functions to be computed with good enough statistics.[130] In the definition of correlation function Eq. (4.1) has converged if its long-time tail reaches a plateau value around zero without significant large fluctuations. To automatically detect these properties we evaluated the average value (α), the standard deviation (σ) and the linear slope ($\rho = |D_{ij}(t_{max}) - D_{ij}(t_{max} - t_{plateau})|/t_{plateau}$) of each CF's long-time tail. These quantities are calculated over the time range

$t_{plateau}$, with $t_{plateau} \sim 0.1 t_{max}$.

The zero thresholds of α, σ and ρ are given by the set of control parameters $\{\epsilon_\alpha, \epsilon_\sigma, \epsilon_\rho\}$.

The procedure applied to each CF is described below and illustrated in the flow-chart in Fig. 4.1.

1 If $\alpha < \epsilon_\alpha$, $\sigma < \epsilon_\sigma$ and $\rho < \epsilon_\rho$ then a close-to-zero plateau has been reached and convergence is assumed if also step 5 is verified.

2 If $\alpha \geq \epsilon_\alpha$ or $\sigma \geq \epsilon_{tail}$, then the CF does not reach convergence.

3 Else if $\alpha < \epsilon_\alpha$ and $\rho \geq \epsilon_\rho$, then protocol proceeds to step 4.

4 Compute the integrals Γ_{CF} and Γ_{tail} of the CF in the ranges $[0 : t_{max}]$ and $[t_{max} - t_{tail} : t_{max}]$, respectively (Here and in the following, integrals are numerically calculated as the area under the curve simply using Simpson's rule. Then, if $\Gamma_{CF} \leq 4\Gamma_{tail}$, the integral Γ_{CF} is assumed to be dominated by noise and CF has not reached convergence. Else, if $\Gamma_{CF} \geq 4\Gamma_{tail}$, a supplementary verification at the next and final step.

5 The integral of CF is computed at N_{check} different time lags in the interval $[0.5 t_{max} : t_{max}]$ and convergence is considered to be reached only if $\Delta_{k+1} \leq \Delta_k$ for all $k = 1...N_{check}$, where $\Delta_k = \Gamma_{CF}[k+1] - \Gamma_{CF}[k]$.

The automatic detection of CF convergence is compared with the decisions made by visual inspection on a subset of one hundred CFs randomly selected from the original set. The optimal value for the set of control parameters $\{\epsilon_\alpha, \epsilon_\sigma, \epsilon_\rho\}$ is taken as the one that maximizes the correlation between automatic detection and visual inspection. In an iterative procedure, the values $\{\epsilon_\alpha = 0.2, \epsilon_\sigma = 5. \cdot 10^{-5}, \epsilon_\rho = 1. \cdot 10^{-6}\}$ are found to give the highest consensus.

4.2.4 Effective correlation times

As mentioned in the Introduction, the usual approach to the study of internal motion correlations in protein makes use of *time-independent* quantities, such as the covariance matrix, for instance. However, true time independence cannot always be ascertained, due to the inherent time limitations of MD simulations. In this respect, it has been pointed out by many authors

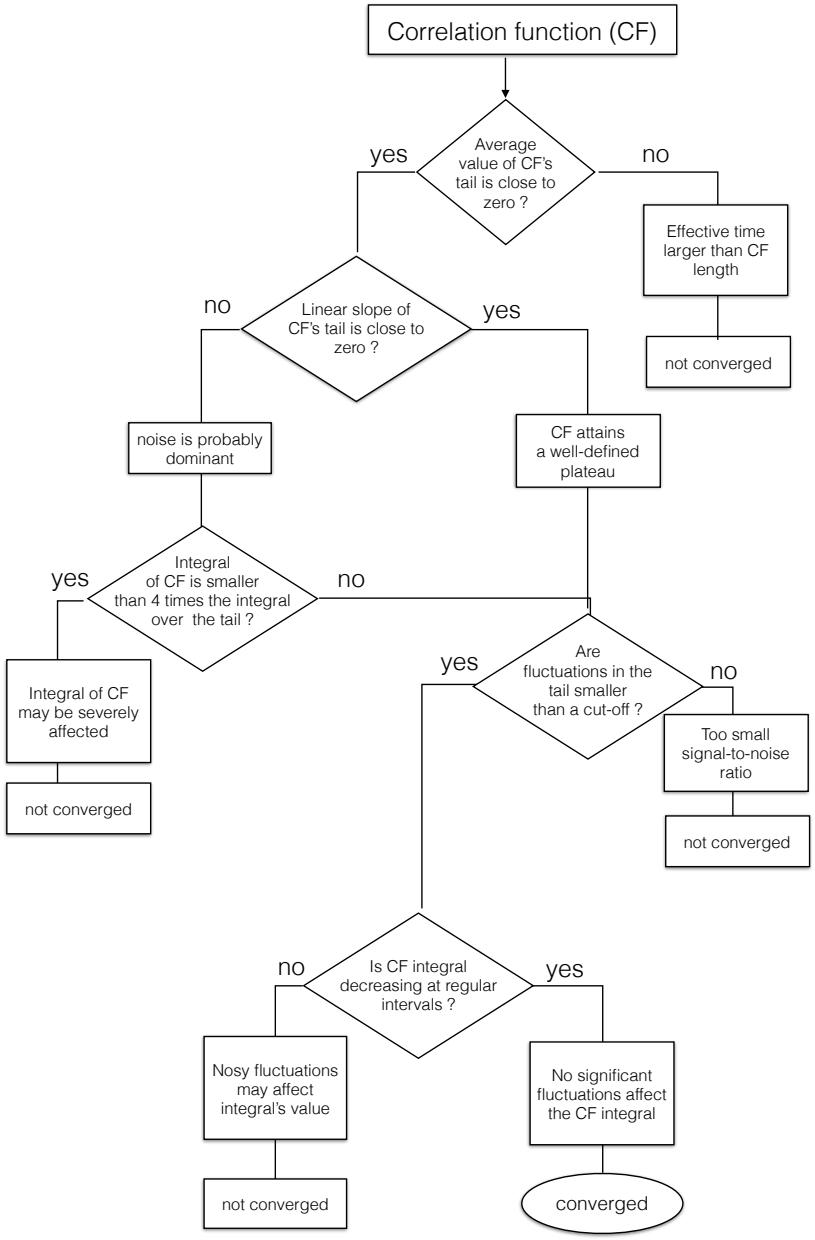

Figure 4.1: Flow Chart of the protocol used to automatically detect the convergence of correlation function calculated from molecular dynamics simulations.

that covariance matrix computed on simulations of different durations yield different correlation patterns.[125, 128, 129] This severely impair the interpretation of MD simulations in terms of protein dynamics.[125]

Here, a new strategy is proposed, that is, analyze motion correlation in proteins through atom pair distance correlation functions. The latter provides a characterization of pairwise atomic motions in the molecule through an effective correlation time τ_{ij}, which defines the characteristic time over which correlated dynamics takes place between a pair of atoms i and j:

$$\tau_{ij} := \int_0^{+\infty} (D_{ij}(t) - D_{ij}(\infty))\, dt \tag{4.2}$$

Note that for an exponentially decay of $D_{ij}(t) - D_{ij}(\infty) = e^{-t/\tau_{ij}}$, the characteristic time defined by Eq. 4.2 is exactly the exponential decay rate. Thus, τ_{ij} is computed from the atom pair distance correlation functions obtained from the MD simulations, leading to a *cross-correlation time map* (CCTM). In the following, the analysis is restricted to backbone C_α carbon atoms, which are thus connected in a pairwise manner through their effective correlation times τ_{ij}.

Measures of (Dis)similarity One can reconstruct ensembles of atoms with mutually correlated motions, together with the associated time scales, by using the complete set of effective cross-correlation times extracted from an MD simulation. In this new framework, the i^{th} atom is represented by its set Λ_i of cross-correlation times with all other atoms $j = 1,\ldots,N$, $j \neq i$: $\Lambda_i = \{\tau_{i1}, \tau_{i2}, ..., \tau_{ij}, \tau_{iN}\}$. Thus, each Λ_i defines a point in the space of the cross-correlation times of the protein, in which a metric is defined. A distance between points Λ_m, Λ_n, $m,n = 1,\ldots,N$ in this correlation-time space therefore allows to define proximities between atoms according to their sets of cross-correlated times.

One of the most robust and efficient method of comparing point sets of arbitrary dimension is the Hausdorff distance,[131, 132] which allows to measure the distance of two set of points by taking into account also the similarity of their shapes.[131] The Hausdorff distance between atoms n and m is defined as follows,

$$d_H(m,n) = max\left\{ \sup_{\tau_{mj}\in\Lambda_m} d(\tau_{mj}, \Lambda_n), \sup_{\tau_{nj}\in\Lambda_n} d(\tau_{nj}, \Lambda_m) \right\}, \tag{4.3}$$

where

$$d(\tau_{mj}, \Lambda_n) = \inf_{\tau_{nk} \in \Lambda_n} (\tau_{mj} - \tau_{nk})^2. \tag{4.4}$$

From these pairwise Hausdorff distances, a *distance*, or *similarity, matrix*, $\mathbf{S} = [s_{ij}]_{NxN} = [d_H(m,n)]_{NxN}$ is then constructed for the protein from MD simulations. The similarity matrix is obviously symmetric and with only zeroes on the diagonal.

Clustering of residues The partitioning of a protein structure into dynamically independent domains can be essentially treated as a problem of graph clustering. Indeed, one may in this context represent a protein structure as a graph $G(\mathcal{N},\mathcal{E})$, comprising $\mathcal{N}$ nodes connected by $\mathcal{E}$ edges. Two nodes of the graph are connected by an edge when a certain *degree of similarity* can be defined between them. In the problem at hand, the nodes $\mathcal{N}$ represent selected atoms or residues that represent the structure of the protein.

The clustering of nodes (residues) is performed here by using the Affinity Propagation algorithm (AP).[127] The latter is recently applied to several kinds of problems, and is shown to be faster and more accurate than other common clustering algorithms. In the authors' words, "AP detects the most representative nodes by exchanging real-valued messages among all nodes in the graph".[127] Nodes are then grouped with their most representative *exemplar*, i.e., around which nodes will cluster. The main principles of its implementation are briefly presented in the following.

Firstly, each node is labelled by a *preference value* P according to which this node should, or should not be chosen as an exemplar by the affinity propagation algorithm. If no prior hypothesis can be made as to which nodes should be favored as exemplars, all nodes are initially assigned the same P value. The magnitude of P can be used to control the granularity of clusters, *e.g.* the extent to which the algorithm can describe the graph/network in terms of discrete components. The preliminary search of the optimal value of P is described in the Appendix.

Secondly, AP performs an iterative search of the so-called "responsibility" $r(i,k)$ and "availability" $a(i,k)$ parameters, for each pair of nodes i and k in the graph $G(\mathcal{N},\mathcal{E})$. The responsibility $r(i,k)$ is a measure of how well suited node k is as an exemplar for node i; and the "availability" $a(i,k)$ reflects the level of evidence that i should choose k as an exemplar. In the AP search, these

quantities are iterated according to the following algorithm:

$$r(i,k) \leftarrow s(i,k) - \max_{k':k' \neq k} \{a(i,k') + s(i,k')\} \tag{4.5}$$

$$a(i,k) \leftarrow \min \left\{ 0, r(k,k) + \sum_{i':i' \notin \{i,k\}} \max\{0, r(i',k)\} \right\} \tag{4.6}$$

In Eq. (4.5) $s(i,k) = -s_{ik}$, where s_{ik} is the element of the similarity matrix $\mathbf{S}$ for the two nodes i and k, and the diagonal element $s(i,i)$ contains the preference for node i. The node k that maximizes $a(i,k) + r(i,k)$ is the exemplar for node i or is itself the exemplar, if $k = i$. It is worth noting here that negative values for $s(i,k)$ are used to enhance the quality of clustering, as prescribed in the original work by Frey *et al.*[127] Equations (4.5) and (4.6) are iterated for a fixed number of iteration or until the local assignments remain constant for a given number of iterations.

4.2.4.1 Assessment of clustering robustness

In order to assess the quality of the clustering protocol we calculated the *silhouette*, $\mathscr{S}(i)$, of each node (residue) which is defined as follows:[133]

$$\mathscr{S}(i) = \frac{\mathscr{B}(i) - \mathscr{A}(i)}{\max(\mathscr{A}(i), \mathscr{B}(i))}. \tag{4.7}$$

Here $\mathscr{A}(i)$ is the "within' dissimilarity", *e.g* the average distance (in the Hausdorff metric) between residue i and all other residues belonging to the same cluster as i; $\mathscr{B}(i)$ is the "between dissimilarity", *i.e.* the smallest average distance between residue i and all other residues belonging to other clusters. This definition implies that $-1 \leq \mathscr{S}(i) \leq 1$, and it is seen that $\mathscr{S}(i)$ represents a practical and efficient way to classify clusters according to their extent and the definiteness of their boundaries: $\mathscr{S}(i) \to 1$ implies that $\mathscr{A}(i)$ is much smaller than $\mathscr{B}(i)$ and therefore means that residue i is ‚Äòwell-clustered‚Äô; on the contrary, if $\mathscr{S}(i) \to -1$, then $\mathscr{A}(i)$ is much larger than $\mathscr{B}(i)$. In this case, residue i is probably mis-classified, likely because it lies at a boundary between clusters. In addition, we also use in the following the *silhouette overall score* $\overline{\mathscr{S}} = \frac{1}{N}\sum_N \mathscr{S}(i)$, which is just the average silhouette over the residues, to synthetically describe the overall quality of the clustering.

4.2.4.2 Through-Space Proximity

The usual segmentation of protein structures in domains on the basis of knowledge of structural or dynamic properties raises the question as to whether the connectedness of domains should be assumed "a priori" or should emerge from the clustering algorithm [134]. From the viewpoint adopted here, there is no fundamental reason why residues sharing the same dynamical properties should also be contiguous in space or neighbours in the sequence. However, if decomposition is performed to provide a coarse-grained model, the space contiguity of residues that belong to the same domain may be taken into account to increase the efficiency of theoretical models. The introduction of a penalty function assigns larger weights to short range interactions and therefore emphasizes the effects of local dissimilarities in the resulting clustering procedure. Here, the effects induced by introducing a penalty function are tested:

$$f(n,m) = 1 + \epsilon \left[1 - \frac{1}{2}(1 + \tanh(R_c - r_{nm})) \right] \tag{4.8}$$

into equation (4.3). Here ϵ is strength factor, whose value is typically of the same order of magnitude than the largest dissimilarity in $\mathbf{S}$ and R_c is the cut-off distance between atoms n and m.

4.3 Results

The new methodology has been applied to the analysis of the C-Terminal Headpiece subdomain of Human Villin (HP35), for which a $\sim 1\,\mu$s molecular dynamics simulation has been run. The internal dynamics of this protein has been extensively studied both experimentally[135–139] and by computational methods.[140, 141] The choice of these rather small proteins was motivated by the need to explore a wide range of motions while keeping the size of the data to be analysed to a reasonable value. Moreover the relatively small size of the protein makes the existence of well defined structural domains *a priori* unlikely. Nevertheless, as shown in the following the existence of clearly identified *dynamical units* or *dynamical domains* can be determined solely by the effective times of the associated inter-atomic distance correlation functions. In this section, the method is presented in its most straightforward implementation. Moreover, for sake of comparison with alternative, more conventional, approaches based on structural information,

additional assumptions involving local interactions and space connectedness are also investigated. And a comparison with a protein clustering technique using a rigid-block decomposition will be discussed. The 35-residue polypeptide HP35 is composed of three short α-helices that has been often used as a model system for computational and theoretical methods since the pioneering work by Duan and Kollman.[142] It has been also extensively investigated experimentally, as it is a good candidate for folding studies, which in this case occurs in both fast (< 100 ns) and slower (< $1\mu s$) regimes.[141]

In this work, HP35 is used as a model system to assess if, and to what extent, macromolecules can be decomposed into structural fragments solely on the basis of, possibly lacunary, internal dynamics information.

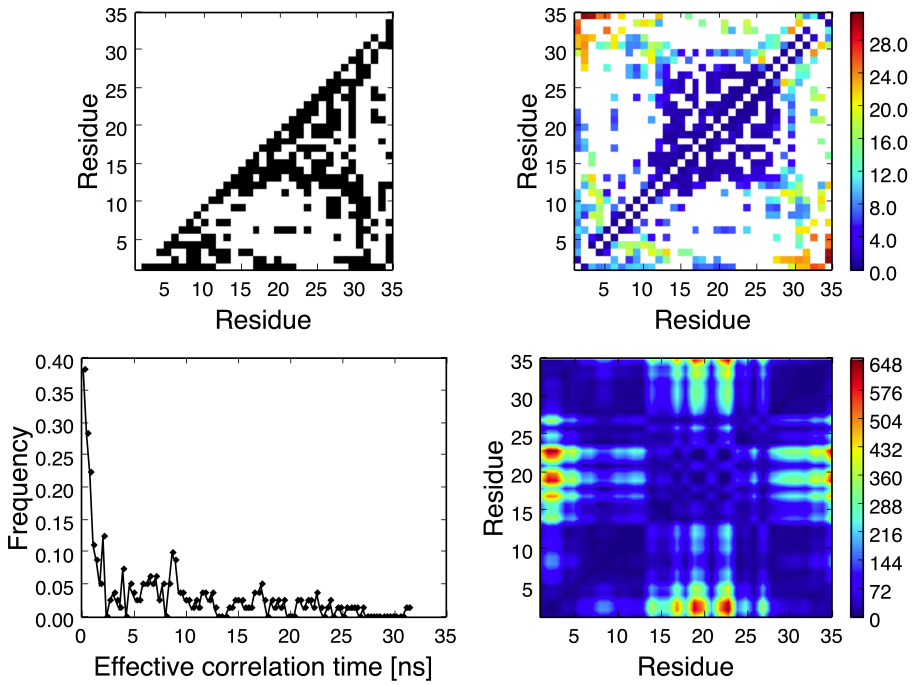

Figure 4.2: **Distance cross-correlation map of the protein HP35.** Top left: the binary map of cross-correlation times shows in black the existence of a well-converged correlation function, as determined according to the criteria discussed in Section 4.2; top right: the time-correlation map of interatomic distances is color-coded as indicated on the scale (in units of ps); bottom left: histogram of cross-correlation effective times; bottom right : the similarity matrix as defined by the distance in Eq. (4.3) in the text.

Among the 595 correlation functions computed from the MD simulation, 43% (257) were

shown to have reached convergence to a well-defined plateau value. This illustrates that the $1.2\,\mu s$ trajectory captures much of the diversity and heterogeneity of the protein internal motions. Nevertheless, the fact that $\approx 57\%$ of the cross-correlation functions failed to converge, and were therefore discarded from the analysis, indicates the presence of slower dynamical processes in the protein that could not be taken into account in the analysis of the MD simulation. The binary map on the top left corner of Fig. 4.2 gives a synthetic overview of the converged (black) and unconverged (white) cross-correlation functions extracted from our MD computations. The distribution of effective characteristic times (Eq. 4.2) over a wide range of values is represented by the histogram of all τ_{ij} (bottom left graph of Fig. (4.2)). The latter exhibits correlations times ranging from the ~ps to several tens of ns. The mapping of this distribution of time scales onto the protein topology is shown on the top right graph of Fig. 4.2. This plot shows that nearby atoms in the sequence exhibit very fast cross-correlation times (dark blue points). Alternatively, groups of atoms regions of the protein, rather that individual atoms, are correlated with long characteristic correlation times, above 5 ns (green to red). Remarkably, only the region between residues 12-27, which encompasses two α-helices, clearly displays times far below 5ns also for non-contiguous atoms. It is also worth noting that the same region has very long effective cross-correlation times with residues 3 to 8. This behaviour correlates well with the large amplitude and slow motions performed by the HP35 N-terminus

Before proceeding with the analysis of these correlation times, the relatively low ratio of converged correlation functions deserves some comment. Indeed, considering the long duration of the MD simulation (1.2 μs), it may be surprising that no more than 50% of the distance correlation functions have converged. In the present context, convergence means both satisfactory statistical averaging, and sufficient decay towards a plateau value. Several papers show that the computation of reliable correlation functions from MD simulations as well as the extraction of correlation times are extremely demanding[130, 143], and require large T_{MD}/τ ratios (T_{MD} is the length the MD simulation and τ the characteristic decay time of the correlation function). Statistical calculations suggested that numerical correlation functions can be sampled with good enough statistics only for time lags shorter than approximately $0.1 \times T_{\mathrm{MD}}$, [130] whilst more recent analyses[143] indicate that, on a model of rotational diffusion, a ratio $T_{\mathrm{MD}}/\tau < 50$ would

make a correct estimate of τ difficult. Therefore, in this work, the convergence criterion, which ensures that a correlation function has reached a plateau value, implies that this plateau is actually reached for time lags smaller than $\sim 0.1 \times T_{\mathrm{MD}}$. And of course, the computation of the effective correlation time requires such a convergence. For a decaying exponential with time constant τ, a plateau is reached at a time lag of approximately 5τ, which is verified for $\tau \approx 30$ ns in our simulations. Thus, the surprisingly small ratio of correlation functions admissible for our clustering analysis is nevertheless consistent with accepted statistical quality criteria.

The similarity (Hausdorff distance) matrix obtained from these correlation functions is shown in Fig. 4.2. It is seen that the motions of the residues located in the center of the primary sequence of HP35 (residues 12-27) form a uniform block of nodes of the protein graph that are characterized by relative motions taking place on similar time scales. This is indicated by the region in blue of the similarity matrix $\mathbf{S}$. Alternatively, residues that belong to the N-terminus are (in the correlation time space of the protein) significantly distant from the latter, as the color scale (green to red) indicates. Interestingly, these observations are consistent the above analysis of the cross-correlation time map. Thus, the metric in the space of correlation times of the protein, in the form of the Hausdorff distance between residues, captures the information conveyed by the cross-correlation times, and therefore provides a sound mathematical tool for their analyses.
.However, at this point, the similarity matrix (Eq.4.3) between points (atoms/residues), shown in Fig. 4.2, only provides a qualitative view of HP35 in terms of units of dynamics defined on the basis of similar time scale properties. The precise extent of these dynamical units must still be determined. This was achieved through application of the *Affinity Propagation* (AP) clustering algorithm presented in section 4.2 to the HP35 similarity matrix.

The result of the clustering protocol on HP35 obtained with these settings are shown in Figure 4.3, and analysed as follows. HP35 is decomposed into two clusters. One of them (cluster A) encompasses the central part of the protein, while the second one (cluster B) is essentially localized near the N- and C- termini. The protein partition into this set of two clusters obtained here agrees with the fuzzy picture suggested by the correlation-time correlation maps. But beyond a simple qualitative representation, the approach developed in this work provides a sound

methodology to perform such an analysis. The relevance and quality of the obtained dynamics cluster decomposition were assessed by the computation of the silhouette $\mathscr{S}(i)$ for each residue of the protein during the AP procedure. Interestingly, all the $\mathscr{S}(i)$ values are positive, and most of

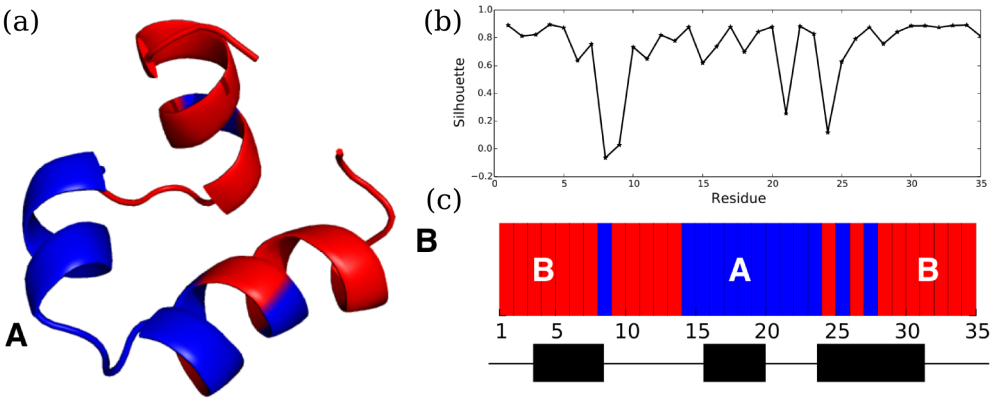

Figure 4.3: Domain decomposition of HP35. *Panel A:* Pictorial representation of the two clusters A (blu) and B (red) onto the HP35 structure. *Panel B:* Silhouette values of each residues representing the quality of the clusters detected by the AP algorithm. *Panel C:* Linear representation of the clusters along the primary sequence of the protein. Secondary structure of HP35 is pictorially sketched on the bottom of the figure.

them are larger than 0.6, which indicates that residues are clearly *well-clustered*. Only a small number of relatively lower values ($\mathscr{S}(i) < 0.4$) were observed. It is also worth noting that in two regions (Gln8-Ala9 and Lys24-Asn27) lower silhouette values correspond to discontinuities of both clusters. This may indeed occur and should not be surprising, as no additional continuity constraint with respect to the residue sequence was imposed on the clusters. However, for residues Gln8-Ala9 the uncertainty of cluster assignment, as indicated by values of $\mathscr{S}(i)$ that are very close to zero or slightly negative, which is not trivial to interpret, but may be at least in part ascribed to the lack of data. In fact, as seen in Fig. 4.2, only few cross-correlated functions could be found for these residues. Alternatively, in region Lys24-Asn27 clustering results exhibit a strong correlation with the local secondary structure of the protein. Indeed, in the α-helix encompassed by this region, the pairs of residues Pro21-Gln25 and Trp23-Asn27 are assigned to cluster A, while Gln26-Lys30 and Lys24-Leu28 are associated to cluster B. The resulting pattern is reminiscent of the α-helix $\{i, i+4\}$-periodicity for the backbone hydrogen bonding network. This indicates that in Lys24-Asn27 local non-covalent interactions have stronger effect than

long-range interactions in determining the assignment to a specific domain. These observations are also consistent with root-mean-squared fluctuations found via MD simulation.
Overall, the results discussed so far suggest that similarities in terms of motional time scales are strongly correlated to the molecular environment around each pair of residues.

4.3.1 Distribution of cross-correlated times

In order to further understand the origin of the clustering provided by the analysis of the internal dynamics of the protein, the distributions of the effective cross-correlation times were computed (Fig. 4.4). The histograms of cross-correlated times relative to each clusters were calculated. These include all the effective correlation times that were obtained for each clusters node (atom or residue). On the other hand, histograms where the correlation times involving nodes of different clusters were excluded; and the histogram of correlation times involving nodes of different clusters only, were computed. These histograms of correlation times *within* and *between* clusters A and B, are depicted in Figure 4.4.

Inspection of Fig. 4.4(a) shows that the distributions of correlation times scales τ_i in clusters A and B are rather different. Cluster B shows a largest contribution at faster (≤ 5 ns) time scales, whereas cluster A presents a more scattered distribution in the $10-20$ ns range. In addition, only cluster A has very long correlation times, extending over ~ 30 ns. Alternatively, it is seen from Fig.4.4(b) that, when *intra-* and *inter-cluster* correlation times are distinguished, the histograms of both clusters become similar on faster time-scales. More interesting is the fact that the short time scales are common to *inter-cluster* and to the *intra-cluster* A, suggesting that the fast time scale contribution to correlation functions between clusters mainly originates from residues in cluster A.
This is of particular interest, as it shows that the connectivity of the atoms defined by time dependence properties, *i.e.*, based on the effective correlation times of the cross-correlation functions, and provided by the AP clustering algorithm, may rely on some statistical decorrelation of the internal molecular motions, without actual time scale separation as a basis for subdividing the molecule into independent parts. Therefore, in the perspective of a coarse-graining description of the internal protein dynamics, this implies that such dynamical properties as domain motions,

for instance, could be studied in terms of the respective dynamics of the clusters as distinct protein units, and without reference to the structure of the protein. These results therefore suggest that it is possible to empirically characterize distinct motional units in proteins without invoking a priori assumptions on the motions' statistics.

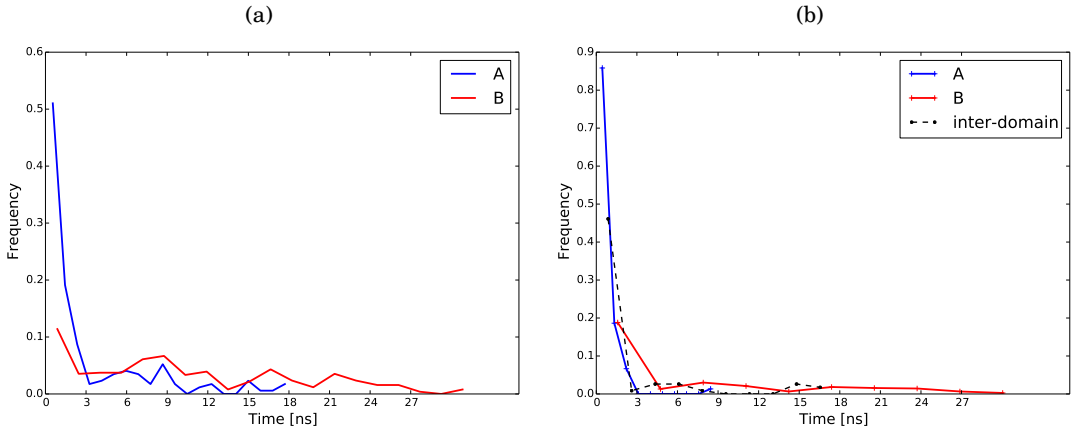

Figure 4.4: Histograms for the distributions of cross-correlation times in HP35. (a) Overall distributions for cluster A and B (b) The same distribution as in *panel a)* but with *intra-* and *inter-domain* correlation times plotted separatelly.

4.3.2 Complementarity with other approaches

As shown in the previous Section, this new method is based on information derived from the time decays of correlation functions, therefore on clearly different grounds than the various *structure-based* analyses customarily used to detect correlated motions and quasi-rigid domains in proteins.[129, 134, 144–146] It may be therefore instructive to compare results from both kinds of approaches to assess their potential complementarity. A very illustrative example is given in Figure 4.5, where the normalized covariance map of HP35 (or Dynamic Cross-Correlation Map - DCCM [147]), is shown along with the cross-correlation time map (CCTM). Both quantities were cmputed from the same 1.2μs MD simulation of HP35

DCCM is a correlation map, which therefore allows one to identify positively or negatively correlated motions between pairs of atoms (red/blue regions in the upper map of Figure 4.5. At a first look, the variability in DCCM seems to have a counterpart in CCTM. For instance,

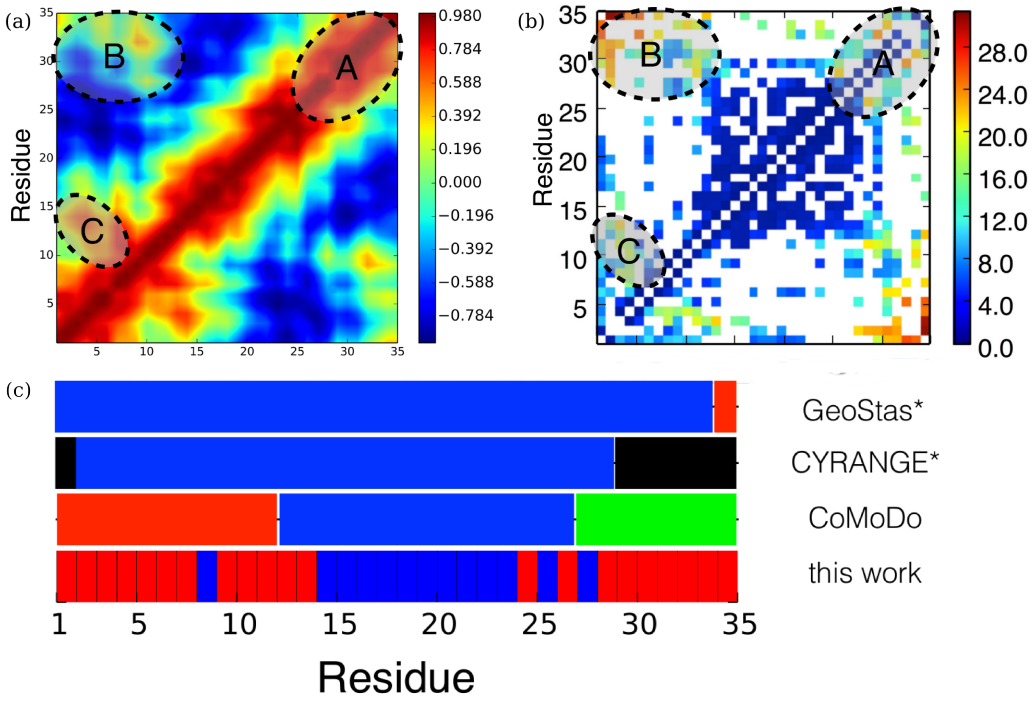

Figure 4.5: Comparison with other approaches: Dynamic Cross-Correlation Map (a) and Cross-correlation Time Map (b). The highlighted regions "A", "B" and "C" evidence the complementarity of the descriptions given by the two maps. (c) Direct comparison of different methods for protein domain decomposition. The asterisks indicate results obtained from configurations sampled every 2ns (instead of 1ps).

in off-diagonal regions (see for example the highlighted region B in Figure 4.5), positive and negative correlations may be related to short and long time scales, respectively. Alternatively, in other regions (see highlighted regions A and C in the upper and central panels of Figure 4.5 large areas of positively correlated pairs of atoms are linked to areas of CCTM where insetead characteristic times may vary from ~ 100 picoseconds to tens of nanoseconds. Interestingly, this simple visual inspection shows that while DCCM is effective in detecting groups of atoms with correlated motions, CCTM completes this information allowing to distinguish the different time-scales at which such motions occur. Moreover, Figure 4.5 also suggests a strong relation between the blank areas in CCTM and the negatively correlated areas in DCCM. This may indicate that the anti-correlated motions in these regions of the protein have larger time-scale of those spanned by our MD simulation and for this they cannot be correctly sampled and yield well-converged cross-correlation functions.

As already mentioned, most decomposition methods of proteins in terms of quasi-rigid domains are based on the analysis of the covariance matrix or on similarity matrices derived from it.[129, 144, 145] Although results from these methods may show local significant differences, a general consensus among them can often be outlined. The results obtained for HP35 from three such methods are shown in the lower panel of Figure 4.5. The comparison with our method shows a clear consensus between the two approaches: HP35 can be decomposed into two domains, a fast-relaxing one localised in the central part of the protein and a slow-relaxing domain which encompasses the two termini of the molecule. Overall, these observations illustrate the complementarity of the *structure-based* and the *dynamic-based* approaches.

In these *structure-based* methods, the simulation length may be a limiting factor to obtain stationary covariance matrices, hence impairing detailed comparison with the new method.[125] In such cases, it may be useful to investigate alternative structure-based approaches. As an example, a rigid-block clustering of HP35 through the PiSQRD method[134, 148] was performed (Figure 4.6 panel c)). The latter is a most widespread method for the detection of rigid domains based on structural information and Gaussian network models. Network models rely on (harmonic) pair potentials that are defined through the three-dimensional structure of the protein. Therefore, such an approach explicitly favors the clustering of the protein in terms of regions of the molecule that are contiguous in space, through the use of a penalty function that truncates the range of interactions. The quality of the clustering obtained by this method is usually assessed by comparing the root mean squared fluctuation (rmsf) of the coarse-grained protein model to the one obtained from the complete network of atoms. The optimal decomposition corresponds to rmsf value that is 80% of that of the complete network.[134, 148] The analysis is firstly performed on HP35 using PiSQRD with this usual criterion, which led to a decomposition into four clusters. If instead one choses to retain five clusters, the result accounts for 86% of the protein mobility. This is illustrated in Fig.4.6

PiSQRD belongs to a class of clustering methods based on the structure of the protein.[149] In order to elucidate possible connections between such commonly used approach and the method introduced in this work, the effect of structure-based information has been investigated. Thus, the penalty function defined in equation (4.8) has been used for the calculation of the similarity

matrix **S**, where the Hausdorff distances between atoms m and n were multiplied by the weight $f(n,m)$. The strength parameter ϵ was set to a value that was approximately one order of magnitude larger than the maximal dissimilarity found in **S**. Besides, the cut-off distance was assigned different values. Results of the effect of changing R_c on the cluster decomposition of the protein are illustrated in Figure 4.6 (panel A).

When the penalty function is introduced into the distance calculation, the number of clusters is consistently increased with respect to results shown in Figure 4.3. For values lower than $R_c = 0.6$ nm, the quality of the clusters deteriorated, as attested by a low $\overline{\mathscr{S}}$. However, when $R_c = 0.7$ nm, the silhouette score increases and the dispersion over the $\mathscr{S}(i)$ is slightly reduced. As R_c increases beyond 0.7 nm, the number of clusters slowly decreases, whilst $\overline{\mathscr{S}}$ does not significantly vary.
It is useful to note that $R_c = 0.7$ nm is the typical distance between neighbouring C_α in proteins, and for this reason has been used in various structure-based, coarse-grained analyses of protein dynamics.[121, 123, 124, 150–153] The new correlation time clustering method has been then performed using this kind of spatial constraint, by introducing the penalty function of Eq. 4.8 with the value $R_c = 0.7$ nm. In this case $\mathscr{S}(i)$ values are quite large ($\mathscr{S}(i) \geq 0.5$) except for four residues (Thr13, Pro14, Leu20, Leu28), for which $\mathscr{S}(i) \sim 0.2$ and three residues (Thr7,Gln8,Ala9) for which $\mathscr{S}(i) \sim -0.4$. Figure 4.6 (panel B) shows that with this additional ingredient, both new and the PiSQRD methods lead to similar results.

Here, the two original correlation time distributions are split into five: each of the former clusters A and B in Figure 4.4 is decomposed in smaller clusters in Figure 4.7 according to the scheme A → A,B,E and B → C,D. Figure 4.7 shows that clusters detected by introducing the spatial connectivity may display very similar distributions of correlation times. In particular, the distinctive contents of *within* and *between* cross-correlation times has disappeared from the histograms. This represents a clear information loss as far as mutual atom dynamics is concerned. These results show that using a "generic" structural information, such as the cut-off radius $R_c = 0.7$ nm, which is related to the typical number of atom neighbours in a protein, drives the predictions of this new approach towards those of structure-based models. However,

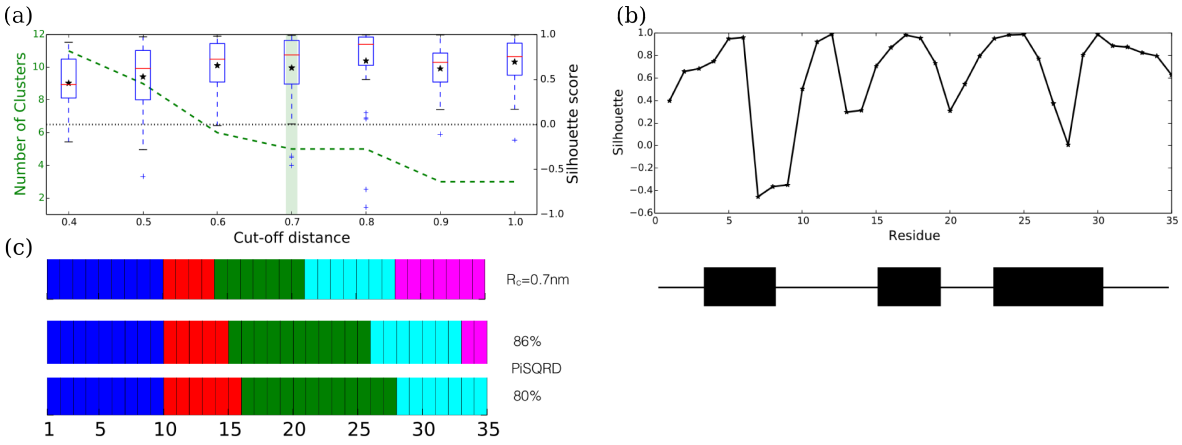

Figure 4.6: Effects of the through-space proximity in the domain decomposition of HP35. (a) Number of clusters and Silhouette score as a function of the cut-off value used in the penalty function in Eq. 4.8. (b) Silhouette per residue obtained by using a cut-off radius $R_c = 7$nm. (c) Linear representation of the domain decomposition obtained with $R_c = 7$nm. Results obtained by the rigid-domain decomposition method PiSQRD are shown with a similar pictorial representation for comparison. See text for details.

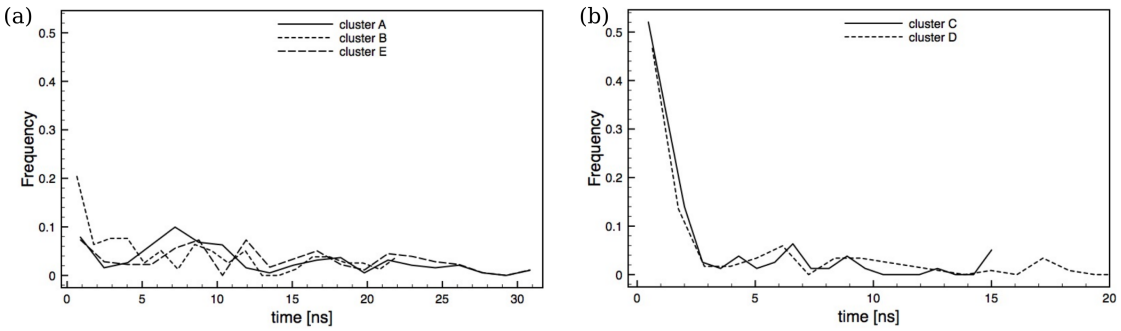

Figure 4.7: Effects of the through-space proximity: Global distributions of correlation time in each of the four cluster found in HP35.

the persistent differences between new and PiSQRD model highlights the fact that the three-dimensional structure of the protein is an important, not the only element of its dynamics. Also, it shows that mixing these two different and independent kinds of information seems to impair the specificity of the perspective of our correlation-time based, and "structure-free" model of dynamic clustering. Therefore, these findings suggest that the introduction of a penalty function to favour through-space proximity of atoms that belongs to the same domain/cluster should be avoided to obtain a domain decomposition that is based on relaxation dynamics.

4.4 Conclusions

In this chapter a new computational approach has been proposed, based on time domain properties of interatomic correlation functions, in contrast with usual methods based on configurational characteristics. To our knowledge, this is the first computational approach that performs atom clustering of proteins in terms of cross-correlation time scales only, without any structural information. Interatomic distance correlation functions calculated from state-of-the-art molecular dynamics simulations were used to estimate the effective cross-correlation times between pairs of atoms. Despite the relatively sparse data obtained from these calculations, our clustering approach performed well, and possibly misassigned atoms were identified through a low silouhette score. The proposed method of atom clustering in proteins on the basis of the time scales of the motions should provide the basis for an adaptive strategy to achieve coarse-graining of proteins where the identified atom clusters are considered as subunits of the protein, the dynamics of which are independent of one another. Such clustering could therefore be used in the derivation of coarse-grained stochastic models for flexible macromolecules, and this approach may serve as a basis for the development of a unified framework for the derivation of dynamic models allowing to extend the range of time scales accessed by MD simulations. Thus, in the perspective of protein dynamics studies, the introduction of time-scale dependent domain decomposition of proteins seems advantageous as compared with methods based on rigid block approximations.

CHAPTER 5

STOCHASTIC MODELING OF FLEXIBLE SYSTEMS IN SOLUTION

5.1 Introduction

The description of the dynamics of a large object, such as a protein, requires a careful definition of molecular frames to which the motions can be referred. Therefore, in general, some geometrical considerations are needed in order to actually relate a dynamic model with a set of observables amenable to experimental measurement, and care should be taken to accurately account for the tensorial nature of the magnetic interactions, by defining proper frames of reference. To this aim, we define the following frames of reference, as shown in Fig. 5.1: i) a laboratory frame (LF), i.e. a fixed external frame; ii) a molecular frame (MF), i.e. a frame fixed on the molecule, where the exact way of defining the MF is actually model-dependent and will be left temporarily undefined; iii) an interaction frame (μF), i.e. a local frame linked to the MF where some specific second rank tensor spectroscopic property μ is well-represented. This could be for instance frame where the ^{15}N-^{1}H dipolar tensor is diagonal or a ^{15}N chemical shift (CSA) tensor is diagonal.[41, 154] For sake of simplicity we shall assume that experimental observables are identified with the normalized correlation function of 2-nd rank Wigner matrix elements for each interaction frame, or its associated spectral density (more precisely, the real part of the spectral density):

$$J_{\mu\nu}(\omega) = \frac{5}{8\pi^2}\int_0^\infty dt e^{-i\omega t}\langle \mathscr{D}^2_{0,0}\left[\Omega_\mu(t)\right]^* \mathscr{D}^2_{0,0}\left[\Omega_\nu(0)\right]\rangle \tag{5.1}$$

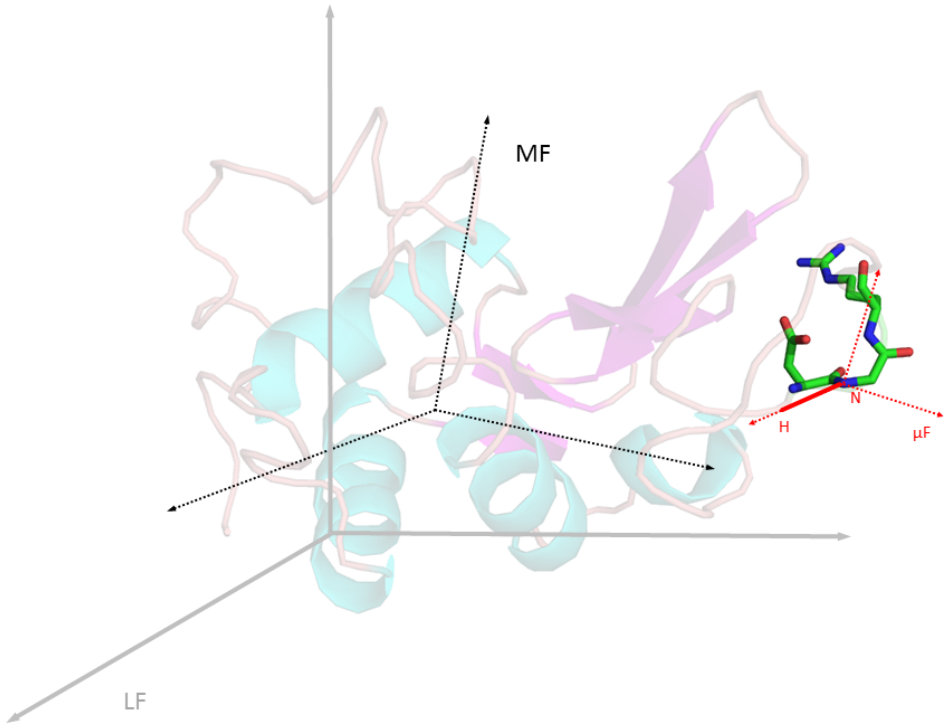

Figure 5.1: Scheme of reference frames, structure and an example of local frame (see text) for Lysozyme; the RF is centered on GLY 67, located between residues ASP 66 and ARG 68.

where μ, ν are indices referring to two tensorial properties and $\Omega_{\mu(\nu)}$ are the sets of Euler angles defining the orientations of the two interaction frames with respect to the LF. The factor $5/8\pi^2$ accounts for normalization in a isotropic medium. It is useful to separate contributions to the spectral density function that originate from molecular motions, on the one hand, from those that come from the geometrical features of the interactions involved, on the other hand. This is achieved using standard algebra[155] of Wigner matrices. Denoting Ω the Euler angles describing the orientation of the MF with respect to the LF, and $\overline{\Omega}_{\mu(\nu)}$ the tilt of $\mu(\nu)$F with respect to the MF, the measurable spectral density for auto-correlated relaxation is given by (cfr. Fig. 5.2):

$$J_{\mu\nu}(\omega) = 5 \sum_{k,k'=-2}^{2} \mathscr{D}^2_{k,0}\left(\overline{\Omega}_\mu\right)^* \mathscr{D}^2_{k',0}\left(\overline{\Omega}_\nu\right) \int_0^\infty dt e^{-i\omega t} G^2_{kk'}(t) \tag{5.2}$$

where

$$G^2_{kk'}(t) = \langle \mathscr{D}^2_{0,k}[\Omega(t)]^* \mathscr{D}^2_{0,k'}[\Omega(0)] \rangle \tag{5.3}$$

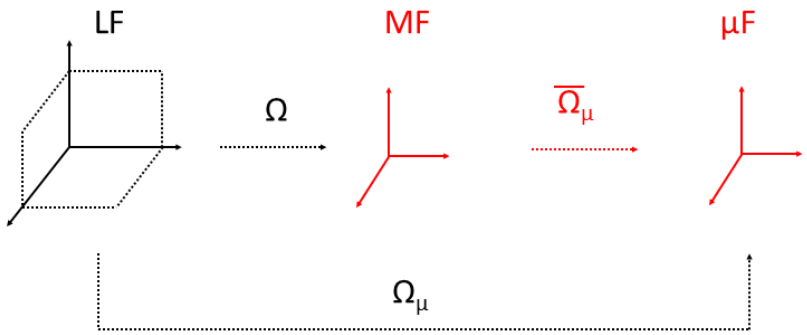

Figure 5.2: Reference frames and Euler angles

A fast and accurate method to evaluate Eq. 5.3 is the main objective of a dynamic modeling approach. For the study of dynamics of such large and presumably non-rigid objects as proteins, this requires a rigorous and adapted kinematic and dynamical description of the motions, as well as consistent stochastic models. Current dynamical models are devised to provide a description of the dynamics of observables, the time dependence of which is due to the presence of local motions, and possibly coupled to global protein tumbling or limited domains motions. Within certain additional approximations, one can give up a detailed parametrization of the local motions (MF approach). Finally, one can make the simplifying assumption that local motions are due, at least for semi-rigid systems, to a network of dynamically coupled neighbors (Network Model) [156] with specific statistical characteristics (diffusive or Brownian dynamics, fractional Brownian dynamics [100] etc.); or caused by partial local reorientation (SRLS) [157]. With the aim of a general strategy that possibly combines the respective advantages of the above methods, we find it crucial to propose a detailed and systematic description of the system geometry, based on its

molecular structure, as well as of the associated dynamical features.
In essence, we aim at obtaining a well-defined time evolution in the form of a stochastic differential or master equation, based on a relevant set of degrees of freedom,and that defines, at least approximately, a Markovian process. In order to do so, one can proceed in a straightforward manner, by setting up the Liouville equation of motion [158] for a generic flexible body defined as a set of material points (atoms or extended atoms), in terms of roto-translational and natural internal coordinates. The general description of a macromolecule in solution is then carried out in terms of a collection of flexible bodies, to which a standard Nakajima-Zwanzig [159, 160] projection method is applied in order to eliminate the "irrelevant" i.e. not directly observed, degrees of freedom. In the next Section we apply this method to the case of a partially rigid Brownian system, i.e. with only fast relaxing internal modes.

5.2 Semi-flexible Brownian body

5.2.1 Semi-rigid Brownian body: hard internal coordinates

Numerically solving the general case of a flexible body model, is a formidable task, although manageable in some specific circumstances.

Here, we address to the case of a semi-rigid folded protein, assuming the absence of internal motions of large amplitude. Single residues are supposed here to undergo "restricted" motions, without the possibility of large rearrangements. Notice that this implies *i)* neglecting activated torsional kinetic and/or *ii)* crankshaft motions [161, 162]. Instead, we concentrate on the description of internal motions adopting the common view of a harmonic or boson bath, retaining full coupling with external motion and including dissipative/stochastic effects. We want to show that a very viable and cost-effective treatment is possible in this limited, but interesting, regime.

Since the total energy is a quadratic form, the model is quite manageable. In particular, as we shall see in the remaining of this section, neither the neglect of inertial effects due to the presence of fast relaxing momenta nor the internal-rotational uncoupling hypotheses need to be invoked in order to calculate observable equivalent to those defined by Eq. 5.3. Let $\boldsymbol{y} = (\boldsymbol{q}, \boldsymbol{L}, \boldsymbol{p})$, where $\boldsymbol{q}$, $\boldsymbol{L}$, $\boldsymbol{p}$ are respectively internal coordinates, angular momenta and momenta

associated to internal coordinates of the system. The total energy is $H = (\boldsymbol{y} - \boldsymbol{y}^{(0)}) \cdot \boldsymbol{k} \cdot (\boldsymbol{y} - \boldsymbol{y}^{(0)})/2$, where $\boldsymbol{y}^{(0)} = (\boldsymbol{q}^{(0)}, \boldsymbol{0}, \boldsymbol{0})$. Defining $\mathcal{Q} = (\Omega, \boldsymbol{y})$, where Ω are the three Euler angles describing the orientation of the MF with respect to the LF; the time evolution operator is defined by

$$\hat{\Gamma}_{FP} = \hat{\mathrm{P}}r - \hat{\nabla}_{\mathcal{Q}} \cdot \boldsymbol{J}_{FP} \overline{\rho}(\mathcal{Q}) \cdot \hat{\nabla}_{\mathcal{Q}} \overline{\rho}(\mathcal{Q})^{-1} \tag{5.4}$$

and it contains $\rho(\mathcal{Q})$, Boltzmann distribution defined on H, the precessional term $\hat{\mathrm{P}}r$ is defined

$$\hat{\mathrm{P}}r = \boldsymbol{L} \cdot \left(\hat{\nabla}_{\boldsymbol{L}} H \times \hat{\nabla}_{\boldsymbol{L}} \right) \tag{5.5}$$

and $\boldsymbol{J}_{FP}$

$$\boldsymbol{J}_{FP} = k_B T \begin{pmatrix} \boldsymbol{0} & -\boldsymbol{1} \\ \boldsymbol{1} & \boldsymbol{\xi} \end{pmatrix} \tag{5.6}$$

In $\boldsymbol{J}_{FP}$, the friction tensor $\boldsymbol{\xi}$ is supposed constant. The matrix $\boldsymbol{k}$ has the form

$$\boldsymbol{k} = \begin{pmatrix} \boldsymbol{K} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}^{-1} + \boldsymbol{A}^{\mathrm{tr}} \boldsymbol{g} \boldsymbol{A} & -\boldsymbol{A}^{\mathrm{tr}} \boldsymbol{g} \\ \boldsymbol{0} & -\boldsymbol{g}\boldsymbol{A} & \boldsymbol{g} \end{pmatrix} \tag{5.7}$$

Here the inertia tensor $\boldsymbol{I} = \sum_\alpha M_\alpha (c_\alpha^2 \boldsymbol{1}_3 - \boldsymbol{c}_\alpha \boldsymbol{c}_\alpha^{tr})$ and the gauge potential $\boldsymbol{A}_\mu = \boldsymbol{I}^{-1} (\sum_\alpha M_\alpha \boldsymbol{c}_\alpha \times \partial \boldsymbol{c}_\alpha / \partial q^\mu)$ are defined, where $\boldsymbol{c}_\alpha(\boldsymbol{q})$ are the relative position vectors of the particles with respect to the center of mass ; $\boldsymbol{g}$ is the (contra)variant metric tensor, defined in terms of the internal coordinates values.[158, 163, 164] All tensor can be evaluated from geometrical considerations, explicitly. It is convenient to introduce scaled, rotated variables: let us define $\boldsymbol{x} = (k_B T)^{-1/2} \boldsymbol{S} (\boldsymbol{y} - \boldsymbol{y}^{(0)})$, where $\boldsymbol{S} = \boldsymbol{\Lambda} \boldsymbol{U}^{\mathrm{tr}}$. $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$ are defined as $\boldsymbol{U}^{tr} \boldsymbol{k} \boldsymbol{T} = \boldsymbol{\Lambda}^2$, where $\boldsymbol{\Lambda}$ diagonal. In terms of the new set of coordinates $\mathcal{Q} = (\Omega, \boldsymbol{x})$, we finally get

$$\frac{\partial \rho(\mathcal{Q}, t)}{\partial t} = -\hat{\Gamma} \rho(\mathcal{Q}, t) \tag{5.8}$$

$$\hat{\Gamma} = \hat{\mathrm{P}}r - \hat{\nabla}_{\mathcal{Q}} \cdot \boldsymbol{\omega} \rho(\mathcal{Q}) \cdot \hat{\nabla}_{\mathcal{Q}} \rho(\mathcal{Q})^{-1} \tag{5.9}$$

where $\hat{\nabla}_{\mathcal{Q}} = (\hat{\boldsymbol{M}}, \hat{\nabla}_{\boldsymbol{x}})$, $H = -x^2/2$, $\rho(\mathcal{Q}) = \rho(\boldsymbol{x}) = p(\boldsymbol{x})/8\pi^2 = \exp(-x^2/2)/(2\pi)^{(6n-9)/2} 8\pi^2$

$$\boldsymbol{\omega} = \begin{pmatrix} 0 & -\boldsymbol{\omega}^{int} \\ (\boldsymbol{\omega}^{int})^{\mathrm{tr}} & \boldsymbol{\omega} \end{pmatrix} \tag{5.10}$$

$\boldsymbol{\omega}$ and $\boldsymbol{\omega}^{int}$ are matrices (with $\boldsymbol{\omega}^{int}$ non-symmetric) $(6n-9)\times(6n-9)$ and $3\times(6n-9)$ respectively, with dimensions of frequencies: $\boldsymbol{\omega}$ is related to dissipative properties (friction tensors) and it has the form:

$$\boldsymbol{\omega} = \boldsymbol{S}\begin{pmatrix} \mathbf{0} & \mathbf{0} & -\mathbf{1} \\ \mathbf{0} & \boldsymbol{\xi}_{\mathrm{RR}} & \boldsymbol{\xi}_{\mathrm{RS}} \\ \mathbf{1} & \boldsymbol{\xi}_{\mathrm{SR}} & \boldsymbol{\xi}_{\mathrm{SS}} \end{pmatrix}\boldsymbol{S}^{\mathrm{tr}} \tag{5.11}$$

while $\boldsymbol{\omega}^{int}$ is

$$\boldsymbol{\Omega}^{int} = (k_B T)^{1/2}\boldsymbol{e}\boldsymbol{S}^{\mathrm{tr}} \tag{5.12}$$

where $\boldsymbol{e} = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} \end{pmatrix}$ The precessional operator can be written in the general form $\hat{\mathrm{P}}r = \omega^P_{ijk} x_i x_j \frac{\partial}{\partial x_k}$, where coefficients ω^P_{ijk} can be found straightforwardly (from now on we neglect tensorial notation), where frequencies ω^P_{ijk} are obtained as

$$\omega^P_{ijk} = (k_B T)^{1/2}\sum_{\alpha\beta\gamma}\epsilon_{\alpha\beta\gamma}(\boldsymbol{e}\boldsymbol{S}^{-1})_{\gamma i}(\boldsymbol{e}\boldsymbol{S}^{\mathrm{tr}})_{\alpha j}(\boldsymbol{e}\boldsymbol{S}^{\mathrm{tr}})_{\beta k} \tag{5.13}$$

5.2.2 Semi-rigid Brownian body: approximate solution

The model described in the previous section, is the simplest description, recoverable for an initially atomistic model for the Brownian probe, [165] retaining fully inertial effects on rotation/shape coupling: it described the semi-rigid macromolecule as a rotator, coupled to $6n-9$ (i.e. $3n-6$ internal coordinates, $3n-6$ internal momenta and 3 components of the $\boldsymbol{L}$ vector) harmonic degrees of freedom, in a fashion quite similar to standard spin-boson quantum mechanical approaches. Indeed, the similarity can be exploited and quite manageable expression can be found for the evaluation of orientational correlation functions, at least in specific dynamic regimes. To clarify, let us consider the different contributions to the time evolution operator, as weighted by $\boldsymbol{\omega}$, $\boldsymbol{\omega}^{int}$, $\boldsymbol{\omega}^P$; $\boldsymbol{\omega}$ elements are of the order $O(\Lambda\xi)$, while $\boldsymbol{\omega}^{int}$, $\boldsymbol{\omega}^P$ are of the order $O(k_B T)^{1/2}\Lambda$. We shall limit our analysis to the case of relatively high friction, i.e. significantly large ξ. This allows to neglect at least as a first approximation the direct contribution of precession terms. We shall therefore write the approximate SRB operator in Eq. 5.9 in the form:

$$\hat{\Gamma} = \hat{\Gamma}_0 + \hat{\Gamma}_{int} = -\sum_{ij}\omega_{ij}\frac{\partial}{\partial x_i}p(x)\frac{\partial}{\partial x_j}p(x)^{-1} + \sum_{i\alpha}\omega^{int}_{i\alpha}x_i\hat{M}_\alpha \tag{5.14}$$

where $i = 1,\ldots,6n-9 = n_T$, $\alpha = 1,2,3$. Our purpose is is to estimate spectral densities associated to correlation functions as in eq. 5.3; for a property of rank j we need to evaluate

$$J^{j}_{m,kk'}(s) = (2j+1)\langle \mathscr{D}^{j}_{m,k}(\Omega)^{*} | (s+\hat{\Gamma})^{-1} | \mathscr{D}^{m}_{j,k'}(\Omega)\rho\rangle \tag{5.15}$$

and $j = 2$, $m = 0$ are for instance employed in the case of nuclear magnetic resonance relaxation. Following a standard expansion of the kernel $(s+\hat{\Gamma})^{-1}$,[166–169] with respect to $\hat{\Gamma}_{int}$, we obtain

$$J^{j}_{m,kk'}(s) = \frac{2j+1}{8\pi^2}\langle \mathscr{D}^{j}_{m,k}(\Omega)^{*} | \left[s+\hat{R}(s)\right]^{-1} | \mathscr{D}^{m}_{j,k'}(\Omega)\rho\rangle_{\Omega} \tag{5.16}$$

where

$$\begin{aligned} \hat{R}(s) &= \sum_{l=1}^{\infty} \hat{R}_l(s) \\ \hat{R}_1(s) &= \overline{\Gamma} \\ \hat{R}_l(s) &= (-1)^{l-1}\langle \hat{\Gamma}_{int}(s+\hat{\Gamma}_0)^{-1}(\hat{Q}\hat{\Gamma}_{int})^{l-2}(s+\hat{\Gamma}_0)^{2-l}\hat{\Gamma}_{int}p(\boldsymbol{x})\rangle_{\boldsymbol{x}} \quad l \geq 2 \end{aligned} \tag{5.17}$$

each term is of l order in $\hat{\Gamma}_{int}$. The terms $\hat{R}_l(s)$, not surprisingly, become rapidly very complex. One can attempt, following analogous treatments in the literature, [166–169] to evaluate the first few terms and then employ a resummation technique. usually, this approach is employed on quantum master equations with memory functions which are approximated via Padé or continuous-fraction expansions. Here, we shall limit our analysis to a restricted regime of relatively high friction, therefore arresting the expansion to the second order term. For simplicity, we prefer to work directly on the kernel *operator*, evaluating only at the end the *memory matrix* after projecting on a suitable subspace of rotational basis functions (see below). We sketch here the basic points. First of all, we define the symmetrized time evolution operator $\tilde{\Gamma} = p(\boldsymbol{x})^{-1/2}\hat{\Gamma}p(\boldsymbol{x})^{-1/2}$, which is written as

$$\tilde{\Gamma} = \tilde{\Gamma}_0 + \tilde{\Gamma}_{int} = \sum_{ij}\omega_{ij}\hat{S}^{+}_i\hat{S}^{-}_j + \sum_{i\alpha}\omega^{int}_{i\alpha}(\hat{S}^{+}_i + \hat{S}^{-}_i)\hat{M}_{\alpha} \tag{5.18}$$

where $\hat{S}^{\pm}_i = \mp e^{(\mp x_i^2/4)}\frac{\partial}{\partial x_i}e^{(\pm x_i^2/4)}$ are raising and lowering operators acting on the orthonormal set in x_i $|s_i\rangle_i$, with $\hat{S}^{+}_i|s_i\rangle_i = \sqrt{s_i+1}|s_i+1\rangle_i$, $\hat{S}^{-}_i|s_i\rangle_i = \sqrt{s_i}|s_i-1\rangle_i$, $\hat{S}^{+}_i\hat{S}^{-}_i|s_i\rangle_i = s_i|s_i\rangle_i$, $\langle s_i|s'_i\rangle = \delta_{s_i,s'_i}$. A state for the n_T bath coordinates is indicated in the following with $|\boldsymbol{s}\rangle_{\boldsymbol{x}} = |s_1,\ldots,s_{n_T}\rangle_{\boldsymbol{x}}$, and $\langle \boldsymbol{s}|\boldsymbol{s}'\rangle_{\boldsymbol{x}} = \delta_{\boldsymbol{s},\boldsymbol{s}'}$. In particular, we indicate with $|\boldsymbol{0}\rangle_{\boldsymbol{x}}$ the symmetrized equilibrium distribution

$p(\boldsymbol{x})^{1/2}$, i.e. the eigenstate with null eigenvector $\tilde{\Gamma}_0|\mathbf{0}\rangle_{\boldsymbol{x}} = 0$, and with $|\mathbf{1}_i\rangle_{\boldsymbol{x}}$ the first excited state with respect to the i-th coordinate, i.e. $|\mathbf{1}_i\rangle_{\boldsymbol{x}} = |0,\ldots,1,\ldots,0\rangle_{\boldsymbol{x}}$. Eqns. (5.17) are now

$$\begin{aligned} \hat{R}(s) &= \sum_{l=1}^{\infty} \hat{R}_l(s) \\ \hat{R}_1(s) &= \langle \mathbf{0}|\tilde{\Gamma}|\mathbf{0}\rangle_{\boldsymbol{x}} \\ \hat{R}_l(s) &= (-1)^{l-1}\langle \mathbf{0}|\tilde{\Gamma}_{int}(s+\tilde{\Gamma}_0)^{-1}(\tilde{Q}\tilde{\Gamma}_{int})^{l-2}(s+\tilde{\Gamma}_0)^{2-l}\tilde{\Gamma}_{int}|\mathbf{0}\rangle_{\boldsymbol{x}} \quad l \geq 2 \end{aligned} \tag{5.19}$$

where $\tilde{Q} = 1 - |\mathbf{0}\rangle_{\boldsymbol{x}}\langle\mathbf{0}|$. Using the properties of $\tilde{\Gamma}_0$, $\tilde{\Gamma}_{int}$, one shows that all odd terms are identically zero, while the first non-zero contributions is:

$$\hat{R}_2(s) = -\sum_{\alpha\beta} D^{(2)}_{\alpha\beta}(s)\hat{M}_\alpha \hat{M}_\beta \tag{5.20}$$

where

$$D^{(2)}_{\alpha\beta}(s) = \sum_{ij} \omega^{int}_{i\alpha}[\boldsymbol{\omega}(s)^{-1}]_{ij}\omega^{int}_{j\beta} \tag{5.21}$$

here $\omega_{ij}(s) = \langle 1_i|s+\tilde{\Gamma}_0|1_j\rangle_{\boldsymbol{x}} = s\mathbf{1}+\boldsymbol{\omega}$ is the matrix representation of $s+\tilde{\Gamma}_0$ on the set of first excited states $|\mathbf{1}_i\rangle_{\boldsymbol{x}}$. In the limit of high friction we set $\hat{R}(s) = \hat{R}_2(s)$. A numerical estimate of Eq. (5.16) is now possible. We consider the $2j+1$ subspace of normalized Wigner matrix functions $|k\rangle_\Omega = \sqrt{(2j+1)/8\pi^2}\mathscr{D}^j_{m,k}(\Omega)$. Eq. (5.16) takes the form

$$\boldsymbol{J}^j_{m,kk'}(s) = \left\{[s\mathbf{1}+\boldsymbol{R}(s)]^{-1}\right\}_{kk'} \tag{5.22}$$

i.e. is the kk' element of the $(2j+1)\times(2j+1)$ matrix $[s\mathbf{1}+\boldsymbol{R}(s)]^{-1}$; the generic element of the matrix representation of the resolvent $\hat{R}(s)$ is $R_{kk'}(s) = \langle k|\hat{R}(s)|k'\rangle_\Omega$. $\boldsymbol{R}(s) \approx \boldsymbol{R}_2(s)$ can be obtained as follows

$$\boldsymbol{R}(s) = -\sum_{\alpha\beta} D^{(2)}_{\alpha\beta}(s)\boldsymbol{M}_\alpha \boldsymbol{M}_\beta \tag{5.23}$$

i.e. in terms of (complex) matrix elements of the Cartesian components of the rotational operator, $M_{\alpha,kk'} = \langle k|\hat{M}_\alpha|k'\rangle_\Omega$, which are found immediately from well-known properties of the Wigner matrix functions:[155] $M_{1,kk'} = -i(c^-_{j,k'}\delta_{kk'-1} + c^+_{j,k'}\delta_{kk'+1})$, $M_{2,kk'} = -(c^-_{j,k'}\delta_{kk'-1} - c^+_{j,k'}\delta_{kk'+1})$, $M_{3,kk'} = -ik'\delta_{kk'}$, where $c^\pm_{j,k} = \sqrt{j(j+1)-k(k\pm1)}$. For the case of second rank properties, for instance, only 5×5 matrices are involved.

5.3 Case studies: polyalanine peptides

The semi-rigid model as been applied in an explorative study of the dynamics of a set of short polyalanine peptides. In particular, nine peptides with 2 to 10 alanine units have been analyzed. The solution of the Fokker-Plank equation followed the route of the expansion of the resolvent, as provided in equation 5.17 and stopping to the second order, $\hat{R}_2(s)$. While for large molecules with a large number of internal coordinates and a large range of characteristic frequencies of the dynamics (ranging from the fast bond librations to the slow global tumbling) such a truncation may result insufficient, we expect that the approximation is good enough for the small peptides presented in this preliminary application of the model.

The complete series of calculations has been carried out using the Hessian of the internal energy (properly converted in internal coordinates) as $\boldsymbol{K}$. To this purpose, the geometry of each of the polyalanine peptides has been minimized using a molecular mechanics approach. The MMTK software [170] has been emplyed for such a minimization procedure. The Amber99ff force field has been used to describe interactions among nuclei in the peptide.

To perform calculations, the C_α and C atoms of the first residues and the N atom of the second residue have been used to build the AF reference frame, from which the molecular frame, MF, is obtained by means of a roto-translation (translation to the center of mass and rotation to align to the principal axes of inertia). Also, the hydrodynamic parameters that have been chosen are: effective radius of 0.5 Å, stick boundary conditions, temperature 298.15 K, and viscosity $8.9 \cdot 10^{-4}$ Pa s (water). Rank 2 Wigner functions spectral densities have been computed.

As comparison, the same spectral densities have been calculated in the diffusive rigid body limit. In this case, the diffusion tensor has been calculated with DiTe [84] using the same set of hydrodynamic parameters and in absence of hydrodynamic interactions (as was implemented for the semi-rigid body). For all the polyalanines the diffusion tensor is nearly axial symmetric. By approximation of the tensor to an axial symmetric one, the exact solution of the rigid body diffusive equation is analytical.

In Figs 5.3-5.4 are shown the real part of the spectral density and the spectral density Cole-

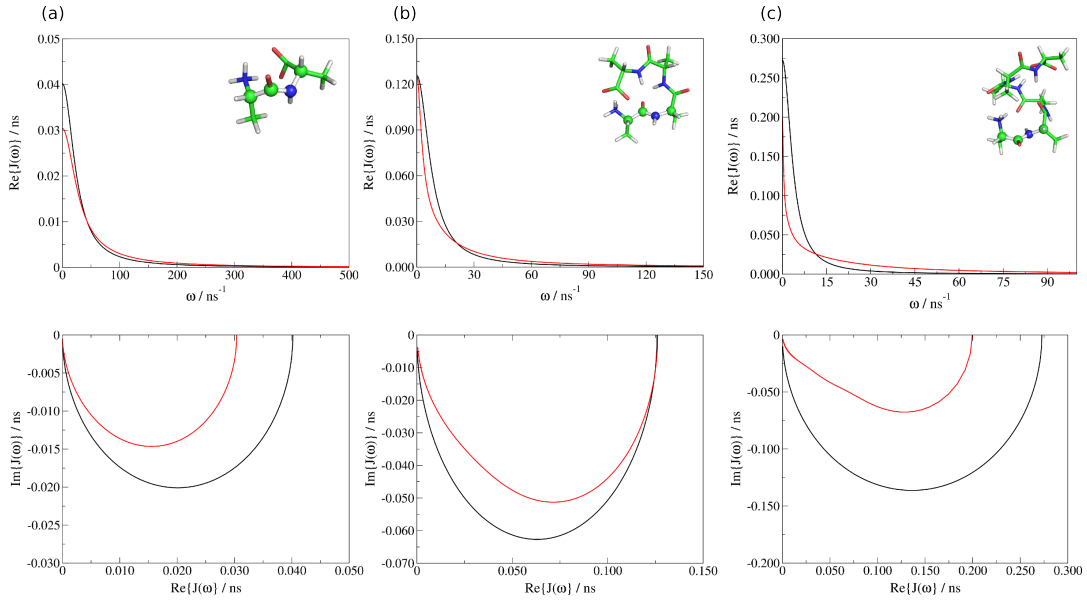

Figure 5.3: Comparison between diffusive rigid body (black) and intertial semi-rigid body (red) description of the dynamics of (a) dialanine, (b) tetra-alanine and (c) esa-alanine.
Top: real part of the spectral density of $D^2_{0,0}(\mathbf{\Omega})$, inset: energy-minimized structures, atoms used to build the AF reference system are showed as spheres.
Bottom: Cole-Cole plot of the same spectral density.

Cole plot specifically for the $D^2_{0,0}$ observable, for selected peptides with 2-, 6-, 8-, and 10-alanine pepdite. The spectral density of $D^2_{0,0}$, in case of the nearly-axial ridig body, is described by a single relaxation process with characteristic frequency equal to $6D_\perp = 3(D_{XX} + D_{YY})$. What is observed is that, on one hand, internal motions affect the total correlation time (value of the zero-frequency spectral density) by lowering it, i.e. a faster relaxation with respect to the rigid body case. Such an effect becomes more evident while increasing molecular size due to an increasing separation of global and local time scales. A second effect, even if less evident, is on the tail of the spectral density, which is higher in the semi-rigid body model with respect to the rigid body. This effect can be rationalized as follows: the high frequency region of the spectral density is related to the short-time relaxation. What happens is that in the initial moments the fast coordinates relax while the global motion is still ”frozen”. Thus, global motion does not contribute to high frequencies (i.e., it is a slow motion).

The Cole-Cole plots show that going from the dialanine peptide to the decaalanine peptide an increasing number of different time scales is affecting relaxation. In the dialanine peptide the

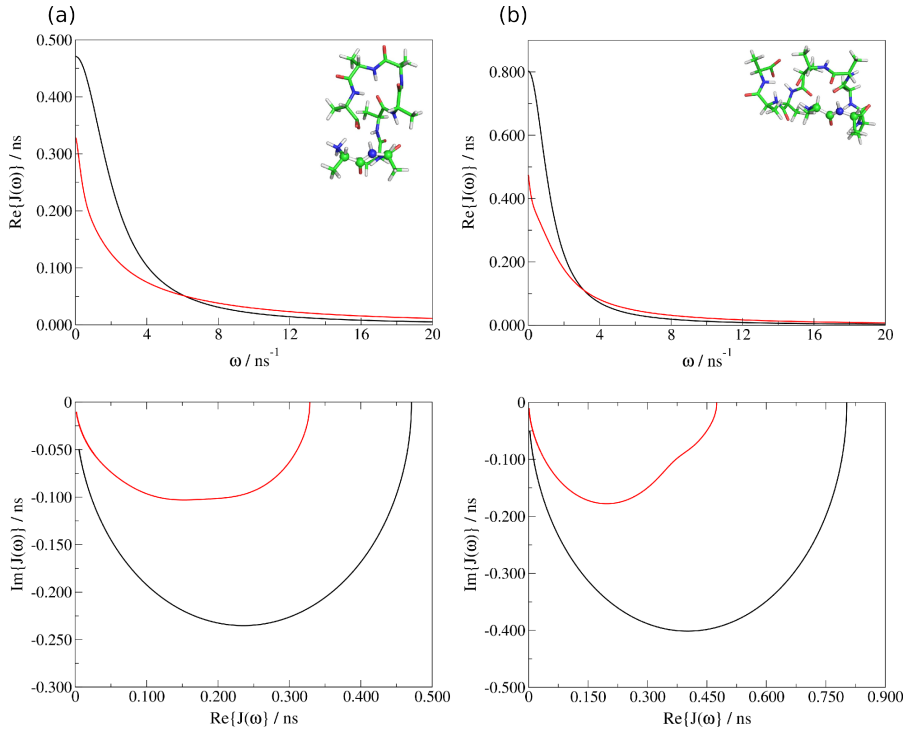

Figure 5.4: Comparison between diffusive rigid body (black) and intertial semi-rigid body (red) description of the dynamics of (a) octa-alanine and (b) deca-alanine.
Top: real part of the spectral density of $D^2_{0,0}(\mathbf{\Omega})$, inset: energy-minimized structures, atoms used to build the AF reference system are showed as spheres.
Bottom: Cole-Cole plot of the same spectral density.(a) octa-alanine, (b) deca-alanine.

global and internal motions relax in similar time scales, thus the Cole-Cole plot is very similar to the single-relaxation time behavior (i.e., that of the rigid body). Concerning the deca-alanine peptide, the separation of time scales is seen in the shape of the Cole-Cole plot, where a good separation can be seen between high frequencies (real part of the spectral density tending to 0) and low frequencies (real part of the spectral density approaching the total correlation time). For the dialanine only, the same calculation has been conducted by parametrizing the internal energy from a molecular dynamics trajectory. The simulation has been carried out with the NAMD software package. The peptide, parametrized with the CHRMM22 force field, has been solvated with 458 TIP3P water molecules in a cubic box of 24 Åside length. After energy minimization, the system has been heated to 298.15 K. A 2 ns run of equilibration has been carried out, followed by 5 ns of production. Calculations have been done using periodic boundary conditions, particle mesh Ewald for electrostatics with a cutoff at 12 Å, NpT ensemble. Then, water was removed

from the trajectory and the RMS mass weighted superposition of the peptide on the first snapshot has been carried out. Here, no clustering has been performed assuming that, because of the short trajectory, the peptide was only fluctuating about one important local energy minimum.
From the aligned trajectory, the covariance matrix in Cartesian coordinates has been calculated and then converted in internal coordinates as described above.
Figure 5.5 compares the $D^2_{0,0}$ spectral densities for the rigid body model and the semi-rigid model in the two cases where internal energy has been obtained from the Hessian of the internal energy or from the covariance matrix. What is observed is that in this second case, the molecule is more "floppy", i.e. the force constants are smaller with respect to those obtained from the Hessian of the internal energy in vacuum.

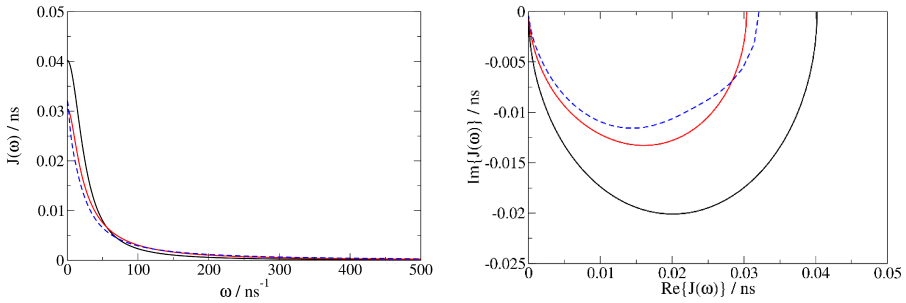

Figure 5.5: Comparison among diffusive rigid body (black), inertial semi-rigid body with internal energy from Hessian (red), and inertial semi-rigid body with internal energy from covariance matrix description of the dynamics of dialanine. Left: real part of the spectral density of $D^2_{0,0}(\mathbf{\Omega})$. Right: Cole-Cole plot of the same spectral density.

5.4 Discussion

In this Section we have presented a systematic approach that is able to describe the dynamics of a non rigid body, based on elaborations from fundamental classical and statistical mechanics. This is particularly relevant for the description of large molecular objects, such as proteins, which represent the main domain of application in the our perspective. But of course, it is a versatile methodology that enables one to tackle virtually any classical non rigid dynamics. The model therefore provides a physically sound framework, which in addition is well suited for computational developments. One of the particularly interesting features of our approach, from

the spectroscopist's viewpoint, is that it allows one to to build fine- to coarse-grained model of dynamics of large molecular objects, depending on additional *a priori* extraneous information that may guide possible approximation strategies. For instance, as described in the typical situations envisaged in this Section, the singling out of hard/soft shape variables allows one to subdivide regions of a molecule into smaller entities (bodies) that are treated in a simplified manner from the dynamical viewpoint. However, in so doing, the exact same theoreticalframework is used. Only the relevance or the validity of the approximations should be discussed. In addition, this versatility remains beyond the sole mechanical description of the molecule. Indeed, the Zwanzig projection technique introduces a memory kernel, the details of which should also be discussed. Adequate approximations and assumptions are likely to depend on the kind of coarse graining itself. Investigation of these aspects are delayed to future work.

BIBLIOGRAPHY

[1] Yang., W.; Hendrickson, W. A.; Crouch, R. J. *Science* **1990**, *249*, 1398.

[2] Kim, Y.; Tang, C.; Clore, G.; Hummerg, G. *P. Natl. Acad. Sci. Usa* **2008**, *105*, 12855–12860.

[3] Mittag, T.; Kay, L.; Forman-Kay, J. *J. Mol. Recognit.* **2009**, *23*, 105–116.

[4] Bustamante, C.; Macosko, J. C.; Wuite, G. J. L. *Nat. Rev. Mol. Cell Biol.* **2000**, *1*, 130.

[5] Palmer, A. G. *Chem. Rev.* **2004**, *104*, 3623–3640.

[6] Mittermaier, A.; Kay, L. E. *Science* **2006**, *312*, 224–227.

[7] Igumenova, T. I.; Frederick, K. K.; Wand, A. *J. Chem. Rev.* **2006**, *106*, 1672–1699.

[8] Jarymowycz, V. A.; Stone, M. J. *J. Chem. Rev.* **2006**, *106*, 1624–1671.

[9] Hall, J. B.; Fushman, D. *J. Am. Chem. Soc.* **2006**, *128*, 7855–7870.

[10] Loth, K.; Pelupessy, P.; Bodenhausen, G. *J. Am. Chem. Soc.* **2006**, *127*, 6062–6068.

[11] Wang, T.; Weaver, D. S.; Cai, S.; Zuiderweg, E. R. P. *J. Biomol. NMR* **2006**, *36*, 79–102.

[12] Sheppard, D.; Li, D.-W.; Godoy-Ruiz, E.; Bruschweiler, R.; Tugarinov, V. *J. Am. Chem. Soc.* **2010**, *132*, 7709–7719.

[13] Sheppard, D.; Spranger, R.; Tugarinov, V. *NMR Spettrosc.* **2010**, *56*, 1–45.

[14] Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.

[15] Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559–4570.

[16] Polimeno, A.; Freed, J. H. *J. Phys. Chem.* **1995**, *99*, 10995–11006.

[17] Meirovitch, E.; Polimeno, A.; Freed, J. H. *J. Phys. Chem. B* **2006**, *110*, 20615.

[18] Zerbetto, M.; Buck, M.; Meirovitch, E.; Polimeno, A. *J. Phys. Chem. B* **2011**, *115*, 376–388.

[19] Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 9279.

[20] Bouvignies, G.; Bernadó, P.; Blackledge, M. *J. Magn. Res.* **2005**, *173*, 328.

[21] Berliner, L. *Biological magnetic resonance: Spin labeling, the next millenium*, 14th ed.; Springer US, 2002.

[22] Bucci, E.; Steiner, R. F. *Biophys. Chem.* **1988**, *30*, 199.

[23] Vergani, B.; Kintrup, M.; Hillen, W.; Lami, H.; Piémont, E.; Bombarda, E.; Alberti, P.; Doglia, S. M.; Chabbert, M. *Biochem.* **2000**, *39*, 2759.

[24] Schotte, F.; Lim, M.; Jackson, T. A.; Smirnov, A. V.; Soman, J.; Olson, J. S.; Jr., G. N. P.; Wulff, M.; Anfinrud, P. A. *Science* **2003**, *300*, 1944.

[25] Cho, A. *Science* **2002**, *296*, 1008.

[26] Cho, H. S.; Dashdorj, N.; Schotte, F.; Graber, T.; Henning, R.; Anfinrud, P. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 7281.

[27] Widengren, J.; Kudryavtsev, V.; Antonik, M.; Berger, S.; Gerken, M.; Seidel, C. A. M. *Anal. Chem.* **2006**, *78*, 2039.

[28] Chung, H. S.; Louis, J. M.; Eaton, W. A. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11837.

[29] Hofmann, H.; Hillger, F.; Pfeil, S. H.; Hoffmann, A.; Streich, D.; Haenni, D.; Nettels, D.; Lipman, E. A.; Schuler, B. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 11793.

[30] Chung, H. S.; Gopich, I. V.; McHale, K.; Cellmer, T.; Louis, J. M.; Eaton, W. A. *J. Phys. Chem. A* DOI: 10.1021/jp1009669.

[31] McCauley, M. J.; Williams, M. C. *Biopolymers* **2008**, *91*, 265.

[32] Kimura, Y.; Bianco, P. R. *Analyst* **2006**, *131*, 868.

[33] Bechtluft, P.; van Leeuwen, R. G. H.; Tyreman, M.; Tomkiewicz, D.; Nouwen, N.; Tepper, H. L.; Driessen, A. J. . M.; Tans, S. J. *Science* **2007**, *318*, 1458.

[34] Milhiet, P.-E.; Gubellini, F.; Berquand, A.; Dosset, P.; Rigaud, J.-L.; Grimellec, C. L.; Lévy, D. *Biophys. J.* **2006**, *91*, 3268.

[35] Hamon, L.; Pastré, D.; Dupaigne, P.; Breton, C. L.; Cam, E. L.; Piétrement, O. *Nucl. Acids Res.* **2007**, *35*, e58.

[36] Viani, M. B.; Pietrasanta, L. I.; Thompson, J. B.; Chand, A.; Gebeshuber, I. C.; Kindt, J. H.; Richter, M.; Hansma, H. G.; Hansma, P. K. *Nat. Struct. Biol.* **2000**, *7*, 644.

[37] Fisher, T. E.; Oberhauser, A. F.; Carrion-Vazquez, M.; Marszalek, P. E.; Fernandez, J. M. *TIBS* **1999**, *24*, 379.

[38] Kim, D.-H.; Park, J.; Kim, M. K.; Hong, K.-S. *J. Mech. Sci. Tech.* **2008**, *22*, 2203.

[39] McAllister, C.; Karymov, M. A.; Kawano, Y.; Lushnikov, A. Y.; Mikheikin, A.; Uversky, V. N.; Lyubchenko, Y. L. *J. Mol. Biol.* **2005**, *354*, 1028.

[40] Masugata, K.; Ikai, A.; Okazaki, S. *Appl. Surf. Sci.* **2002**, *188*, 372.

[41] Abragam, A. *Principles of Nuclear Magnetism*; Clarendon Press, Oxford, 1961.

[42] Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546.

[43] Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559.

[44] Zerbetto, M.; Polimeno, A.; Meitovitch, E. *J. Phys. Chem. B* **2009**, *113*, 13613.

[45] Polimeno, A.; Freed, J. *Adv. Chem. Phys.* **1993**, *83*, 89–163.

[46] Polimeno, A.; Freed, J. *J. Phys. Chem.* **1995**, *99*, 10995–11012.

[47] Zerbetto, M.; Polimeno, A. *Int. J. Quantum Chem.* **2016**, *116*, 1706–1722.

[48] Meirovitch, E.; Shapiro, Y. E.; Zerbetto, M.; Polimeno, A. *J. Phys. Chem. B* **2012**, *116*, 886–894.

[49] Stein, H.; Hausen, P. *Science* **1969**, *166*, 393.

[50] Crouch, R. *J. New Biol.* **1990**, *2*, 771.

[51] Mandel, A. M.; Akke, M.; Palmer, I., A. G. *J. Mol. Biol.* **1995**, *246*, 144.

[52] Yamazaki, T.; Yoshida, M.; Nagayama, K. *Biochemistry* **1993**, *32*, 5656.

[53] Nakamura, H.; Oda, Y.; Iwai, S.; Inoue, H.; Ohtsuka, E.; Kanaya, S.; Kimura, S.; Katsuda, C.; Katayanagi, K.; Morikawa, K.; Miyashiro, H.; Ikehara, M. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 11535.

[54] Oda, Y.; Iwai, S.; Ohtsuka, E.; Ishikawa, M.; Ikehara, M.; Nakamura, H. *Nucleic Acids Res.* **1993**, *21*, 4690.

[55] Mandel, A. M.; Akke, M.; Palmer, I., A. G. *Biochemistry* **1996**, *35*, 16009.

[56] Butterwick, J. A.; Loria, J. P.; Astrof, N. S.; Kroenke, C. D.; Cole, R.; Rance, M.; Palmer, I., A. G. *J. Mol. Biol.* **2004**, *339*, 855.

[57] Butterwick, J. A.; Palmer, A. G. *Protein Sci.* **2006**, *15*, 2697.

[58] Kroenke, C. D.; Loria, J. P.; Lee, L. K.; Rance., M.; Palmer, A. G. *J. Am. Chem. Soc.* **1998**, *120*, 7905.

[59] Baber, J. L.; Szabo, A.; Tjandra, N. *J. Am. Chem. Soc.* **2001**, *123*, 3953.

[60] Meirovitch, E.; Shapiro, Y. E.; Zerbetto, M.; Polimeno, A. *J. Phys. Chem. B* **2012**, *116*, 886–894.

[61] Lewandowski, J. R.; Sein, J.; Blackledge, M.; Emsley, L. *J. Am. Chem. Soc.* **2010**, *132*, 1246.

[62] Jeschke, G.; Koch, A.; Jonas, U.; Godt, A. *J. Magn. Reson.* **2002**, *155*, 72–82.

[63] Zerbetto, M.; Carlotto, S.; Polimeno, A.; Corvaja, C.; Franco, L.; Toniolo, C.; Formaggio, F.; Barone, V.; Cimino, P. *J. Phys. Chem. B* **2007**, *111*, 2668–2674.

[64] Carlotto, S.; Zerbetto, M.; Shabestari, M.; Moretto, A.; Formaggio, F.; Crisma, M.; Toniolo, C.; Huber, M.; Polimeno, A. *J. Phys. chem. B.* **2001**, *115*, 13026–13036.

[65] Shabestari, M.; van Son, M.; Moretto, A.; Crisma, M.; Toniolo, C.; Huber, M. *Biopolymers(Pept. Sci.)* **2014**, *102*, 244–251.

[66] Toniolo, C.; Crisma, M.; Formaggio, F. *Biopolymers(Pept. Sci.)* **1998**, *47*, 153–158.

[67] Toniolo, C.; Crisma, M.; Formaggio, F.; Peggion, C. *Biopolymers(Pept. Sci.)* **2001**, *60*, 396–419.

[68] Barone, V.; Polimeno, A. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4609–4629.

[69] Zerbetto, M.; Polimeno, A.; Barone, V. *Comput. Phys. Comm.* **2009**, *180*, 2680–2697.

[70] Schneider, D.; Freed, J. *Adv. chem. Phys.* **1989**, *73*, 387–528.

[71] Carlotto, S.; Cimino, P.; Zerbetto, M.; Franco, L.; Corvaja, C.; Crisma, M.; Formaggio, F.; Toniolo, C.; Polimeno, A.; Barone, V. *J. Am. Chem. Soc.* **2007**, *129*, 11248–11258.

[72] Khafizov, N.; Madzhidov, T.; Kadkin, O.; Tamura, R.; Antipin, I. *Int. J. Quantum. Chem.* **2016**, *116*, 1064–1070.

[73] Coulaud, E.; Hagebaum-Reignier, D.; Siri, D.; Tordo, P.; Ferré, N. *Phys. Chem. Chem. Phys.* **2012**, *14*, 5504–5511.

[74] Illas, F.; Moreira, I.; de Graaf, C.; Barone, V. *Theor. chem. Acc.* **2000**, *104*, 265–272.

[75] Miralles, J.; Castell, O.; Caballol, R.; Malrieu, J. *Chem. Phys.* **1993**, *172*, 33–43.

[76] Umanskii, S.; Golubeva, E.; Plakhutin, B. *Russ. Chem. Bulletin* **2013**, *7*, 1511–1518.

[77] Umanskii, S. *Russ. J. Phys. Chem. B* **2015**, *9*, 1–8.

[78] Curtiss, L.; Redfern, P.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 084108.

[79] Polimeno, A.; Zerbetto, M.; Franco, L.; Maggini, M.; Corvaja, C. *J. Am. Chem. Soc.* **2006**, *128*, 4734–4741.

[80] Meirovitch, E.; Igner, D.; Igner, E.; Moro, G.; Freed, J. *J. Chem. Phys.* **1982**, *77*, 3915–3938.

[81] Frisch, M. J. et al. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[82] Scalmani, G.; Rega, N.; Cossi, M.; Barone, V. *J. Comp. Meth. Sci. Eng.* **2002**, *2*, 469–474.

[83] Improta, V., R.and Barone *Chem. Rev.* **2004**, *104*, 1231–1253.

[84] Barone, V.; Zerbetto, M.; Polimeno, A. *J. Comput. Chem.* **2008**, *30*, 2–13.

[85] Moro, G. *Chem. Phys.* **1987**, *118*, 167–180.

[86] Moro, G. *Chem. Phys.* **1987**, *118*, 181–197.

[87] Haynes, W. *CRC Handbook of Chemistry and Physics, 93rd Edition*; CRC Handbook of Chemistry and Physics; Taylor & Francis, 2012.

[88] Berliner, L., Reuben, J., Eds. *Spin Labeling, Theory and Applications*; Academic Press: New York, 1976; pp 133–181.

[89] Hanson, P.; Millhauser, G.; Formaggio, F.; Crisma, M.; Toniolo, C. *J. Am. chem. Soc.* **1996**, *118*, 7618–7625.

[90] Abergel, D.; Volpato, A.; Coutant, E. P.; Polimeno, A. *Journal of Magnetic Resonance* **2014**, *246*, 94–103.

[91] Andrec, M.; Montelione, G. T.; Levy, R. M. *Journal of Magnetic Resonance* **1999**, *139*, 408–421.

[92] Redfield, A. G. *Magn. Reson. Chem.* **2003**, *41*, 753–768.

[93] Redfield, A. G. *J Biomol NMR* **2012**, *52*, 159–177.

[94] Charlier, C.; Khan, S. N.; Marquardsen, T.; Philippe Pelupessy, V. R.; Sakellariou, D.; Bodenhausen, G.; Engelke, F.; Ferrage, F. *J. Am. Chem. Soc.* **2013**, 18665.

[95] Redfield, A. G. *IBM J. Res. & Dev.* **1957**, *1*.

[96] Redfield, A. G. *Adv. Magn. Reson.* **1965**, *1*, 1–32.

[97] Cavanagh, J.; Fairbrother, W. J.; Palmer, A. G.; Skelton, N. J. *Protein NMR Spectroscopy*; Academic Press, Inc., 1996.

[98] Korzhnev, D.; Billeter, M.; Arseniev, A.; Orekhov, V. *Progr. NMR Spectrosc.* **2001**, *38*, 197–266.

[99] Ernst, R. R.; Bodenhausen, G.; Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*; Oxford University Press, London, 1987.

[100] Calandrini, V.; Abergel, D.; Kneller, G. *J. Chem. Phys.* **2010**, *133*, 145101.

[101] Calligari, P.; Calandrini, V.; Kneller, G.; Abergel, D. *J. Phys. Chem. B* **2011**, *115*, 12370–12379.

[102] Calligari, P.; Abergel, D. *J. Phys. Chem. B* **2012**, *116*, 12955–65.

[103] Calandrini, V.; Abergel, D.; Kneller, G. *J .Chem. Phys.* **2008**, *128*, 145102.

[104] Erdélyi, A.; Magnus, W.; Oberhettinger, F.; Tricomi, F. *Higher Transcendental Functions*; McGraw Hill: New York, 1955.

[105] Glöckle, W.; Nonnenmacher, T. *Biophys. J.* **1995**, *68*, 46–53.

[106] Kneller, G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2641 – 2655.

[107] Kilbas, A.; Srivastava, H.; Trujillo, J. *Theory and applications of fractional differential equations*; North Holland Mathematics Studies; Elsevier, 2006; Vol. 204.

[108] Mandel, A.; M.Akke,; Palmer, A. *J. Mol. Biol.* **1995**, *246*, 144–163.

[109] d'Auvergne, E. J.; Gooley, P. R. *J. Biomol. NMR* **2008**, *40*, 121–133.

[110] Robert, C.; Casella, G. *Monte Carlo Statistical Methods*; Springer Texts in Statistics; Springer-Verlag, 2010.

[111] Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[112] Hastings, W. *Biometrika* **1970**, *57*, 97–109.

[113] *Scilab, a free scientific software package*; Copyright 1989-2005. INRIA ENPC,www.scilab.org.

[114] Robert, C.; Casella, G. *Introducing Monte Carlo Methods with R*; Springer-Verlag New York: New York, 2010.

[115] Das, A.; Mukhopadhyay, C. *J .Chem. Phys.* **2007**, *127*, 165103.

[116] Caves, L. S.; Evanseck, J. D.; Karplus, M. *Protein Sci* **1998**, *7*, 649–66.

[117] David, C. C.; Jacobs, D. J. *Methods Mol Biol* **2014**, *1084*, 193–226.

[118] Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Proteins: Struct., Funct., Bioinf.* **1993**, *17*, 412–25.

[119] Hayward, S.; de Groot, B. L. *Methods Mol Biol* **2008**, *443*, 89–106.

[120] de Groot, B. L.; van Aalten, D. M.; Amadei, A.; Berendsen, H. J. *Biophys. J.* **1996**, *71*, 1707–13.

[121] Hinsen, K.; Thomas, A.; Field, M. J. *Proteins: Struct., Funct., Bioinf.* **1999**, *34*, 369–382.

[122] Fuglebakk, E.; Reuter, N.; Hinsen, K. *J Chem Theor Comput.* **2013**, *9*, 5618–5628.

[123] Dhulesia, A.; Bodenhausen, G.; Abergel, D. *J Chem Phys* **2008**, *129*, 095107.

[124] Abergel, D.; Bodenhausen, G. *J Chem Phys* **2005**, *123*, 204901.

[125] Tiberti, M.; Invernizzi, G.; Papaleo, E. *J Chem. Theory Comput.* **2015**, *11*, 4404–14.

[126] He, Y.; Chen, J.-Y.; Knab, J.; Zheng, W.; Markelz, A. *Biophys. J.* **2011**, *100*, 1058–1065.

[127] Frey, B.; Dueck, D. *Science* **2007**, *315*, 972.

[128] Lange, O.; Grubmuller, H. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 1053–1061.

[129] Romanowska, J.; Nowiński, K. S.; Trylska, J. *J Chem Theor Comput.* **2012**, *8*, 2588–2599.

[130] Zwanzig, R. *Adv. Chem. Phys.* **1969**, *40*, 325–331.

[131] Hausdorff, F. *Set theory*; New York: Chelsea Publishing Co., 1957.

[132] Vlasblom, J.; Wodak, S. J. *BMC Bioinformatics* **2009**, *10*, 99.

[133] Rousseeuw, P. J. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.

[134] Potestio, R.; Pontiggia, F.; Micheletti, C. *Biophys J* **2009**, *96*, 4993–5002.

[135] Havlin, R. H.; Tycko, R. *Proc Natl Acad Sci U S A* **2005**, *102*, 3284–3289.

[136] Frank, B. S.; Vardar, D.; Buckley, D. A.; McKnight, C. J. *Protein Sci* **2002**, *11*, 680–687.

[137] Bunagan, M. R.; Gao, J.; Kelly, J. W.; Gai, F. *J Am Chem Soc* **2009**, *131*, 7470–7476.

[138] Brewer, S. H.; Vu, D. M.; Tang, Y.; Li, Y.; Franzen, S.; Raleigh, D. P.; Dyer, R. B. *Proc Natl Acad Sci U S A* **2005**, *102*, 16662–16667.

[139] Vermeulen, W.; Vanhaesebrouck, P.; Van Troys, M.; Verschueren, M.; Fant, F.; Goethals, M.; Ampe, C.; Martins, J. C.; Borremans, F. A. M. *Protein Sci* **2004**, *13*, 1276–1287.

[140] Saladino, G.; Marenchino, M.; Gervasio, F. L. *J Chem Theory Comput* **2011**, *7*, 2675–2680.

[141] Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Proc Natl Acad Sci U S A* **2012**, 17845.

[142] Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740–744.

[143] Lu, C.-Y.; Bout, D. V. *J. Chem. Phys* **2006**, *125*, 124701.

[144] Kirchner, D. K.; Güntert, P. *BMC bioinformatics* **2011**, *12*, 1.

[145] Wieninger, S. A.; Ullmann, G. M. *J Chem Theor Comput.* **2015**, *11*, 2841–2854.

[146] Ponzoni, L.; Polles, G.; Carnevale, V.; Micheletti, C. *Structure* **2015**, *23*, 1516–25.

[147] McCammon, J.; Harvey, S. *Dynamics of proteins and nucleic acids*; Cambridge Univ Pr, 1988.

[148] Aleksiev, T.; Potestio, R.; Pontiggia, F.; Cozzini, S.; Micheletti, C. *Bioinformatics* **2009**, *25*, 2743–2754.

[149] Dziubiński, M.; Daniluk, P.; Lesyng, B. *Bioinformatics* **2016**, *32*, 25–34.

[150] Brueschweiler, R.; Wright, P. E. *J Am Chem Soc* **1994**, *116*, 8426–8427.

[151] Tirion, *Phys Rev Lett* **1996**, *77*, 1905–1908.

[152] Hinsen, K. *Proteins: Struct., Funct., Bioinf.* **1998**, *33*, 417–29.

[153] Sanejouand, Y.-H. *Methods Mol Biol* **2013**, *924*, 601–16.

[154] Peng, J. W.; Wagner, G. In ; James, T. L., Oppenheimer, N. J., Eds.; Methods in Enzymology; Academic Press, 1994; pp 563–595.

[155] Zare, R. N. *Angular momentum: understanding spatial aspects in chemistry and physics*; Wiley, 1988.

[156] Abergel, D.; Bodenhausen, G. *J. Chem. Phys.* **2005**, *123*, 204901.

[157] Meirovitch, E.; Shapiro, Y. E.; Polimeno, A.; Freed, J. H. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2010**, *56*, 360–405.

[158] Goldstein, H. *Classical Mechanics*; Addison-Wesley, 1980.

[159] Zwanzig, R. *Physica* **1964**, *30*, 1109–1123.

[160] Zwanzig, R. *J. Stat. Phys.* **1973**, *9*, 215–220.

[161] Moro, G. *J. Phys. Chem.* **1996**, *100*, 16419–16422.

[162] Nigro, B.; Moro, G. *J. Phys. Chem. B* **2002**, *106*, 7365–7375.

[163] Shabana, A. *Dynamics of Multibody Systems*; Cambridge University Press, 2005.

[164] Littlejohn, R. G.; Reinsch, M. *Review of Modern Physics* **1997**, *69*, 213–275.

[165] Polimeno, A.; Zerbetto, M.; Calligari, P. in preparation.

[166] Sparpaglione, M.; Mukamel, S. *J. Chem. Phys.* **1988**, *88*, 3263–3280.

[167] Cho, M.; Silbey, R. *J. Chem. Phys.* **1997**, *106*, 2654–2661.

[168] Mavros, M.; Van Vooris, T. *J. Chem. Phys.* **2014**, *141*, 054112.

[169] Chen, H.; Berkelbach, T.; Reichman, D. *J. Chem. Phys.* **2016**, *144*, 154106.

[170] Hinsen, K. *J. Comput. Chem.* **2000**, *21*, 79–85.