Utrecht University

On the Use of EPR and sm-FRET Data to Guide the Modeling of Biomolecular Complexes

Master thesis

Iva Pritisanac
Student no. 3734188
Biomolecular Sciences
Supervisors:
Dr. Ezgi Karaca
Prof. dr. Alexandre Bonvin
Computational Structural Biology
NMR Spectroscopy Research Group

Abstract

At the moment, the macromolecular complexes having high molecular weights, containing intrinsically disordered components, interacting transiently or having membrane associated parts challenge the state-of-the-art high-resolution structural biology techniques. At the same time, the variety and coverage of the low-resolution experimental techniques, which can reveal some of the structural features of these challenging complexes, is increasing and thus boosting the development and the use of information-driven modeling approaches. Such approaches aim the translation of sparse low-resolution information into a structural model. One of the best examples of information-driven modeling approaches is the High Ambiguity DOCKing (HADDOCK). HADDOCK uses a wide range of experimental data and was so far successfully applied to the docking of various biomolecular partners. Recently, two low-resolution methods, EPR and sm-FRET were recognized by the structural biology community, since the long-range distance information derived from these methods, can be used to highlight the structural features of the challenging macromolecular complexes. Here we tested, for the first time, HADDOCK's ability to model complexes using the EPR and sm-FRET data, and showed that with the limited amount of distance restraints acceptable solutions could be generated. Moreover, we highlighted the critical aspects that need to be taken into account when using the sm-FRET and EPR restraints in HADDOCKing. The expansion of this study to a larger number of biomolecular complexes would pave the route for including EPR and sm-FRET distance restraints in the rich repertoire of data used by HADDOCK.

Outline

General Introduction

Genomics, proteomics and interactomics have revealed and addressed many aspects of cellular molecular biology. However, adding structural dimension into those "omics" is necessary, in order to complete our understanding of cellular processes. X-ray crystallography and Nuclear Magnetic Resonance Spectroscopy (NMR) are the most accurate experimental methods to give high-resolution information about the structure and the dynamics of the biomolecules and their complexes. X-ray crystallography can be used for high molecular weight structures and was successfully applied to gain the high resolution structural models of large macro-molecular complexes as ribosome[1,2,3,4], the transcription complex[5,6], RecBCD helicase[7] and plant photosystem II (PSII)[8]. However, membrane associated, flexible and transiently interacting complexes often cannot be crystallized or their data analysis suffers from the averaging effect, which makes the positions of flexible domains and unstructured regions undeterminable. Furthermore, constraining the biomolecules in a solid-state crystal might result in experimenting with a non-functional form of the molecule. On the other hand, high-resolution liquid NMR spectroscopy, can provide the structural and dynamical information of the macromolecules in solution, thus probing them in more "native-like conditions"[9,10,11]. Nevertheless, NMR requires relatively high sample concentration and suffers from the size limitation (~80kDa)[12,13]. Currently rapidly improving, high-resolution magic angle spinning (MAS) solid-state NMR (ss-NMR) techniques can overcome this size limitation and also can investigate the non-soluble biomolecular assemblies in their native settings (e.g. membrane bilayer, cellular organelles, bacterial cell wall) to address, for instance, ligand-protein, or protein-protein interactions[14]. Also, when combined with optimized protein production and isotopic labelling schemes ss-NMR can provide the information on structure and dynamics of the biomolecules in cellular environment[15]. Another promising method, cryo-electron microscopy (cryo-EM), is currently at the interface of low- and high-resolution (8-20Å)[16,17] and as a major advantage over the mentioned methods, has no requirements for the crystallization or large sample quantities[18-21].

Although the mentioned techniques helped immensely to generate accurate high-resolution data for many biomolecular complexes, many of the molecular machines, as well as flexible, intrinsically disordered, transiently interacting or membrane associated complexes still remain to be discovered. Currently, large efforts have been made in optimizing the high-resolution techniques to structurally address such complexes. However, in the meantime, different types of biochemical and biophysical experiments can already generate fast and accurate low-resolution information, even for those challenging biomolecular assemblies[16]. Therefore, for the moment, "an efficient approach" would be to focus on to the information coming from these methods and try to make best use of it to get structural insight on these challenging systems.

1. Low resolution methods- reliable source of information for structural biology?

The currently available low-resolution methods can provide information on the existence of the interaction, complex stoichiometry/composition, relative positions and/or orientations of the complex components, binding interface, proximity of the individual components in the complex, long range distances and complex's shape. The information content coming from these techniques differs in terms of spatial features, resolution, quality and quantity[22]. General overview for each method is given in *Table 1*.

Table 1: An overview of the techniques available for the detection and characterization of varies features of the biomolecular interactions

Technique	Affinity Purification	Tandem Affinity Purification	Mass spectrometry	Chemical cross-linking/MS	Native MS	Yeast two-hybrid assay	Protein Complentation Assay	Site Directed Mutagenesis	Quantitative Immunoblotting	Labeling with fusion proteins	Immuno-EM	Analytical ultra centrifugation	Size exclusion chromatography	Atomic force microscopy	Surface plasmon resonance	Flourescence Polarization/Anisotropy	Flourescence correlation spectroscopy	Dual polarization interferometry	Static Light Scattering	Dynamic Light Scattering	(H/D) exchange / MS	(H/D) exchange / NMR	Radical footprinting	Small angle X-ray scattering	Small angle neutron scattering	Cryo electron microscopy	Scanning transmission EM	Negative stain EM	RDCs (NMR)	Relaxation anisotropy (NMR)	CSPs (NMR)	Electron paramagnetic resonanccee	Sm-FRET
Interaction detection	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖			✖	✖	✖	✖	✖	✖	✖	✖	✖	✖									✖	✖	✖
Composition/stoichiometry	✖	✖			✖				✖		✖	✖								✖	✖													
Shape					✖								✖	✖											✖	✖	✖	✖	✖					
Size/mass			✖		✖							✖	✖					✖	✖	✖				✖	✖	✖	✖	✖						
Conformation/orientation					✖													✖								✖	✖	✖	✖	✖				
Inter-atomic distances				✖																												✖	✖
Interface		✖						✖						✖							✖	✖	✖								✖		
Energetics/kinetics														✖	✖	✖	✖	✖	✖	✖													

As shown in *Table 1*, a number of biochemical methods are available for the simple detection of the presence of, for instance, protein-protein interaction (e.g. yeast-two-hybrid assay, affinity purification, protein complementation assay, etc.). Even though such interaction identification is not directly structurally informative, in the case the stoichiometry of the complex is known, some of these methods can account for the low-resolution spatial information. For instance, a set of the independent affinity purifications or mass spectrometry fragmentation experiments can reveal the configurations and interaction profiles, or the architectures of the large complexes' subunits[22-26].

Apart from such indirect structural information, some methods can provide direct low-resolution information on the shape of a complex (cryo-electron microscopy (cryo-EM), Small Angle X-ray Scattering (SAXS), Scanning Transmission Electron Microscopy (STEM)); long range distances between its components (single-molecule Forster Energy Resonance Transfer (sm-FRET), Electron Paramagnetic Resonance (EPR), chemical cross-linking/Mass Spectrometry (MS)); positions (Native MS, Cryo EM) or relative orientations of its components (Residual dipolar couplings (RDCs), Relaxation anisotropy restraints (NMR)); and the binding interfaces among its components (site-directed mutagenesis, radical footprinting, H/D exchange/MS, H/D exchange/NMR, Chemical Shift Perturbations (CSPs)).

From the structural perspective, some of these techniques are quantitatively and/or qualitatively more informative than the others. For instance, electron-optical density distribution derived from the single particle cryo-EM can be used in combination with the high resolution structures of the individual complex constituents to result in the "quasi-atomic" models of the large assemblies (~250 kDa)[19,22]. On the other hand, shape information derived from SAXS experiments is considered to be less informative. Namely, SAXS data can be translated into a low-resolution shape envelope (with a resolution range of 10-30 Å) by rotational averaging, which makes it challenging to extract the spatial restraints[16,22]. Considerable efforts have been employed to optimize the usage of SAXS data in structural biology, as the technique is well suited for probing assemblies having a large mass-range of 50-250 kDa, which is not the case for cryo-EM or NMR spectroscopy[16,22,27,28].

As an alternative to those, Electron Paramagnetic Resonance (EPR) and single-molecule Förster Resonance Energy Transfer (sm-FRET) can provide long-range distance restraints without the limitations to the size of the assembly (*Section 3*)[29-31]. Moreover, transient or weak interactions of the multi- and/or flexible component complexes of high molecular weights can be probed with those methods, which is a great challenge for all of the currently available techniques. Besides, the measurements can be performed in solution, which allows the manipulation of the experimental conditions by adding, for instance, certain ligands or varying the concentrations of specific ions to mimic the biologically relevant environment [32](see *Sections 3,4*).

2. Computational modeling approaches-Translating low-resolution data into high-resolution structural information

Currently there are two approaches used in modeling of the biomolecular complexes, *ab initio* and information-driven. The major differences among the information-driven methods come in the choice of the information which is predominately used in the structure calculation[22]. Increase in the variety of the low-resolution experimental techniques revealing the structural features of biomolecular complexes (*Table 1*) boosted the development and the use of information-driven approaches. The information used can come from a single experimental source (e.g. chemical shift perturbations or NOE distance restraints from NMR, chemical cross-linking data from MS); or broader repertoire of the experimental and/or predicted data can be used simultaneously (e.g. epitope mapping, mutagenesis, and sequence conservation data, combined with the bioinformatics' interface predictions[109]). In this thesis, special attention is given to the information-driven computational docking, due to the fact that this approach is particularly well suited for the translation of varies (sparse) low-resolution information into a structural model[16].

Computational docking is the modeling of the three-dimensional structure of a biomolecular complex, on the basis of the structures of the individual molecules in their "free" unbound form[33]. When compared to template based methods its particular advantage is in its ability to predict the new binding modes[22]. Moreover, in recent Critical Assessment of PRedictions of the Interactions (CAPRI)[34,35], docking methods show considerable progress and are currently one of the most promising approaches in modeling of the biomolecular complexes[16].

During a classical docking run, as a first step, an extensive search of the conformational space is performed, starting from the high-resolution structures of the complex's individual components[16,22]. This sampling of the conformational space aims at the maximization of the geometrical and physicochemical complementarities of the complex constituents [16,36-40]. Challenges are encountered when the complex contains larger number and/or diverse components and/or experiences large conformational changes of the components upon complexation[16]. This is due to an increase in the number of the degrees of freedom in the system, which requires more extensive sampling of the conformational space[16,41,42]. The result of a docking run is usually an ensemble of possible conformations and the biologically relevant ones must be picked among numerous alternative solutions[16,43,44]. This step is called scoring that is performed by a scoring function encompassing terms to address thermo-dynamical and physicochemical properties of the interface[16]. Thus, the accuracy of the final solution strongly depends on the accuracy of the scoring function[16]. Yet, current, conventional scoring functions have failed to show the direct correlation with binding free energy and thus none of them can assure consistent correct ranking of the docking solutions[16,45]. A prominent way to overcome the challenges encountered during sampling and scoring is an information-driven docking. In contrast to the *ab initio* docking, during sampling, it is directed by the available experimental and/or evolutionary information, in order to focus to the "relevant part of the conformational space"[16,22,33,46,47]. Furthermore, the

probability of recognizing the correct solution in the scoring step is increased by filtering generated solutions based on their discrepancy with the present experimental data[16].

One of the best information-driven docking example is the High Ambiguity DOCKing (HADDOCK)[33] approach. HADDOCK's docking protocol consists of three stages: rigid body docking (*it0*), semi-flexible refinement (*it1*) and refinement in explicit solvent (*it2*). Before the start of the docking, molecules are separated and random rotations are applied, followed by *it0*, which docks the molecules as rigid bodies. Before the next stage, scoring of the solutions is performed and the energetically favourable structures submitted to (*it1*), where the interface is optimized via semi-flexible alignment. In the last stage (*it2*), structures are refined in explicit solvent, followed by the scoring and clustering of the generated solutions. Each successfully finished docking run has the docking statistics as a part of the output, allowing the user to efficiently and critically analyse the docked solutions[33,46,47]. HADDOCK protocol has been successfully applied to docking of proteins with various biomolecular partners, e.g. proteins, nucleic acids, oligosaccharides and small-molecules (*Figure 1*) and its latest released version (HADDOCK 2.1) supports nucleic acids, small molecules and can deal with a wide range of experimental data[33]. HADDOCK is readily available via web-server with user friendly interfaces allowing different degrees of control over various docking parameters depending on the user access level[16,33].

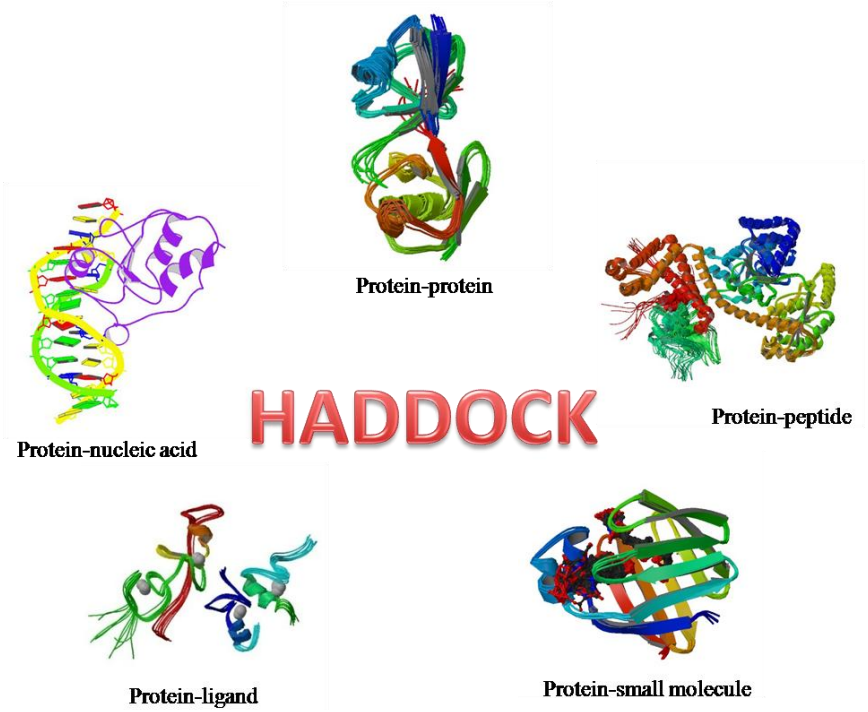

Figure 1: Examples of the different biomolecular complexes successfully modelled using HADDOCK data driven docking and deposited to Protein Data Bank (PDB): Protein-nucleic acid complex (PDB ID:2k7f)[84]; Protein-ligand complex (PDB ID:2l6z)[83] ; Protein-small molecule (PDB ID:2lfo)[87]; Protein-peptide complex (PDB ID:2vda)[86]; Protein-protein (PDB ID:2bgf)[85].

An overview of the experimental data supported by HADDOCK is provided in *Table 2* together with the ones that are currently under development (e.g. shape information coming

from cryo-EM). It can be noticed that the majority of the data "under development" comes from the low resolution methods, which are recently broadly applied to challenging biomolecular complexes, e.g. cryo-EM whereas, a broad repertoire of data coming from solution NMR is already successfully incorporated. In our study, we tested HADDOCK's ability to model complexes under the guidance of EPR and sm-FRET data for the first time. Thus, it was particularly intriguing to see HADDOCK's performance with the distance restraints derived from these techniques, as described in detail in *Section 4.*

Table 2: Overview of the types of the experimental data currently used or under the development in HADDOCK.

Data type	Information content	Incorporation in HADDOCK
NMR Spectroscopy:		
CSP	Interface residues	Active and passive residues
Cross-saturation	Interface residues	Active and passive residues
H/D exchange	Interface residues	Active and passive residues
RDCs	Orientation restraints	Directly
15N relaxation rate (R1,R2)	Orientation restraints	Directly
NOEs	Distance restraints	CNS restraints
PRE	Distance restraints	CNS restraints
Dihedral angles	Conformational restraints	Directly
Hydrogen bonding	Conformational restraints	Directly
MS :		
Chemical cross-linking	Distance distributions	CNS restraints
RP-MS	Interface residues	Active and passive residues
H/D exchange	Interface residues	Active and passive residues
ESI-MS: CCS	Shape	Under development
SAXS data	Shape	Under development
EM -electron crystallography -cryo EM -electron tomography	Shape	Under development
EPR	Long range distance restraints	**Not tested**
Sm-FRET	Long-range distance restraints	**Not tested**

*tested in this study as given in *Section 4*

Abbreviations used in *Table 2*: CSP- Chemical Shift Perturbation, RDCs- Residual Dipolar Couplings, NOEs-Nuclear Overhauser Effects, PRE- Paramagnetic Relaxation Enhancement, PS-Pseudocontact Shifts, RP-MS-Radical Probe Mass Spectrometry, ESI-MS- Electro Spray Ionization Mass Spectrometry, CCS-Collision Cross Section, NOEs-Nuclear Overhauser Effect, EM-Electron Microscopy

3. EPR and sm-FRET techniques-New data sources for biomolecular modeling

In *Section 1* we have discussed the information content of low-resolution methods, concluding that it varies from method to method (*Section 1.2*). In this respect, EPR and sm-FRET are of particular interest for structural biology, as the long-range distance information, derived from these methods, can be used to highlight the structural features of the current challenging macromolecular complexes, e.g. complexes of high molecular weights, involving multi- or flexible components or formed by transient and weak interactions. In fact, for some of these challenging systems, if several long-range restraints are provided, features such as e.g. domain arrangements, complex formation, structural changes upon ligand binding or global structure elements can already be revealed[49]. Furthermore, recent improvements in both of the techniques and increasing number of their applications make them perfect targets for exploring their potential in structural modeling.

3.1 Electron-paramagnetic resonance techniques on spin-labelled biomolecules

In order to apply EPR technique to address the structural features of the biomolecular complexes, biomolecules of interest are "spin-labelled"[50] : the paramagnetic centres (spins), i.e. the unpaired electrons of R1 organic radical side chains, most commonly nitroxide, are introduced at specific sites (e.g. cysteine residues on protein) by site-directed chemistry. All EPR methods rely on the magnetic dipole-dipole interaction between the magnetic moments μ_A and μ_B of two electron spins A and B, which is inversely proportional to the cube of the distance between two spins[1,50] (*Eq. 1*). This allows the calculation of the distance distribution between the spins[50].

$$E= \mu_A \mu_B/ R^3 - 3 (\mu_A R) (\mu_B R)/R^5 \quad (Eq.\ 1)$$

Eq.1 describes the magnetic dipole-dipole interaction energy *E*, in classical energy expression, where μ_A and μ_B are the magnetic moments of electron spins A and B, respectively, and R is the distance separating them. The dipolar Hamiltonian can be derived by relating the electron spin magnetic moment (μ) to electron spin operator *S*

$$\mu= -\gamma_e \hbar S \quad (Eq.\ 2)$$

,where γ_e stands for the gyro-magnetic ratio of the electron spin, and $\hbar$ for the reduced Planck's constant. For the nitroxide spin labels the expression for the dipolar Hamiltonian (H_{dip}) can be simplified by the following assumptions: 1) the *g* values of the two spins are only weakly anisotropic, and thus the electron spins quantization axes are parallel to the external field (z direction); and 2) the dipolar coupling is small compared to the splitting of the electron spin states by the Zeeman effect[50,51,110]. In this case H_{dip} is given by

$$H_{dip}= [g^A g^B \beta_e^2/ R^3] [S_{Az} S_{Bz} (1-3 \cos^2\theta)] \quad (Eq.\ 3)$$

, where g^A, g^B are the g values of the two spins (approximated to the isotropic value $g^A=g^B\approx$ 2.006), β_e is the Bohr magneton, $S_{Az} S_{Bz}$ is product of the electron spin operators (for the *z* axis direction), and θ is the angle between the external field axis and the spin-to-spin vector [50,51,110]. Taken together, the dipolar frequency ω_{dip} can be written as

$$\omega_{dip} = [D_{dip}/ R^3] (1\text{-}3 \cos^2\theta) \quad (Eq.\ 4)$$

, where D_{dip} is the splitting constant. Importantly, for the inter-spin distances shorter than 10 Å, the exchange interaction *J* has to be introduced to the Hamiltonian expression in *Eq 3*.

Different EPR methods have been developed to reveal the dipolar interaction, depending on the spectral and dynamical properties of the spin labels and the distance between them[51]. Here, we discuss two methods relevant for the biological samples in terms of their basic principles. These are: the continuous wave (CW) used to probe the distances up to 20-25 Å[52] , and double electron-electron resonance (DEER), used for the transition to larger distances, up to ~60-80 Å[53,54].

Continuous wave-EPR technique allows direct observation of the dipolar coupling in the EPR spectra[55], when two paramagnetic centres are separated not more than 20 Å and when the intrinsic line width of at least one of them is smaller than the dipolar splitting. If these conditions are not met, the result is the line width broadening and deconvolution methods are needed to extract the dipolar coupling contribution within other spin interactions[56]. Despite these difficulties, considering the shorter distance range, continuous wave experiments can be very useful as they can be conducted at the physiological temperatures and are usually easily experimentally accessible. However, to address the structural questions, it is often required to extend the measurable distance (R) between the paramagnetic centres above the 20 Å limit of CW-EPR. Under such conditions pulsed EPR methods can be employed. The pulsed EPR method recovers the dipolar coupling by increasing the spectral resolution by ”cancelling“ the broadening of the line width caused by *g* tensor anisotropy and the hyperfine couplings.

Different pulsed EPR strategies have been successfully applied to biomolecular samples with nitroxide spin labels[57]. For instance, Hahn-echo-like sequence (*Figure 2A*) to refocus all static inhomogeneous line width contributions[1], double quantum coherence build-up during the echo[58,59]or solid echo to refocus all of the spin interactions apart from the dipolar coupling[60]. When Hahn-echo-like experiment is used, the dipole-dipole interaction is essentially also refocused and is reintroduced by an inverse pulse to the second spin[50] (*Figure 2*). Typically, on the biomolecular samples, pulsed electron-electron double resonance (PELDOR) is performed and double electron-electron resonance (DEER) is the most widely used experiment (*Figure 2B*)[50].

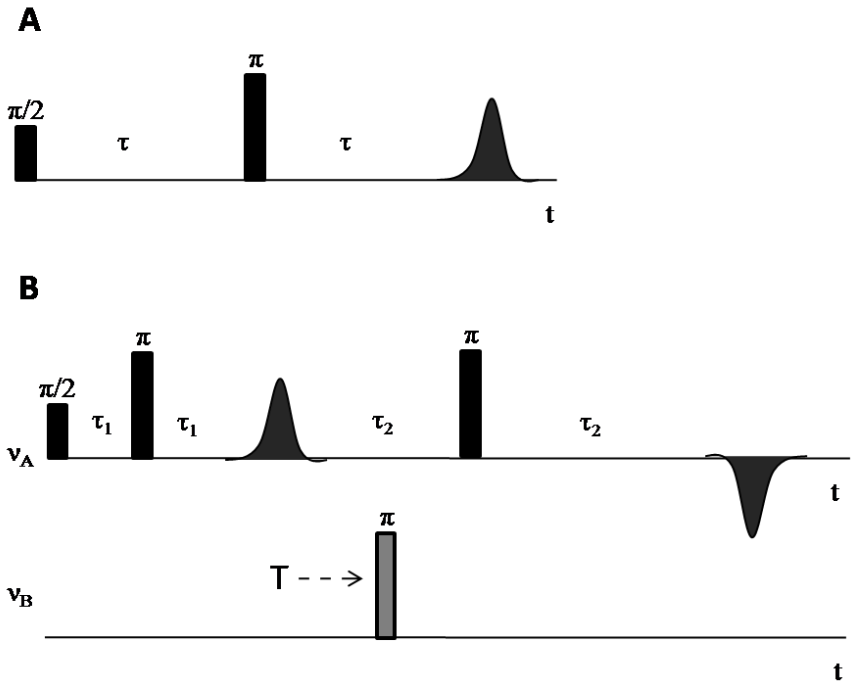

Figure 2: A) Hahn-Echo pulse sequence, the basis of the pulsed EPR experiments **B)** 4-pulse PELDOR (DEER) sequence. π/2 and π-pulses applied on spin A and B, τ_1, τ_2 –inter pulse delays, T- a microwave pulse applied at the frequency resonant with second spin (spin B).

Potentials and limitations of EPR in accurate distance determination

The maximum "extractable" EPR distance primarily depends on the length of the recorded time trace due to Fourier arguments[51]. In order to maximize the sensitivity and to optimize the pulsed EPR experiment in regard to the transverse relaxation time (T_2), the longitudinal relaxation time (T_1) and the polarization of the spin transitions, parameters such as the applied pulse lengths, spin radical concentrations and the experimental temperature, should be considered[50]. For the short distance ranges, the CW experiments can be useful, however, when deconvolution methods are needed, the following should be considered. First, given only the continuous wave EPR spectrum, it is impossible to distinguish the line width broadening caused by static dipolar coupling from that of the dipolar relaxation. Second, at short distances, the exchange *J* interaction might contribute to splitting, which can lead to wrong distance interpretation[51]. Finally, high signal-to-noise ratios are needed in order to avoid artefacts. Thus, although CW is used from time to time for the quantitative interpretation of the distances in the 10-20 Å range, it is more commonly applied in a qualitative fashion, for indicating the proximity of the spin labelled biomolecular sites[51].

In contrast to CW methods, pulsed EPR methods have been highly optimized for the quantitative distance measurements on biomolecules[50]. However, while applying pulsed EPR on biomolecular samples, difficulties arise from the high concentration of hydrogen atoms in sample's aqueous or lipid matrix. During echo, the modulation of these matrix protons is acquired in addition to that of the selected spin-spin dipolar coupling, thus, contributing to the line width[50].No matter how strictly the spectroscopic selection rules are applied during the

pulses, this effect cannot be completely eliminated. In order to maximally suppress it, the inter-pulse delays are varied (*Figure 2*) and echoes for the different delays summed[50].

In order to facilitate the use of pulsed EPR on biomolecular samples, DeerAnalysis2011 package[61] for the analysis of DEER experiments was created. It allows the extraction of the distance distributions from constant-time and variable-time four-pulse DEER[62,63]. The output of the analysis is distance distribution characterized by mean distance value $<R>$ and the width (standard deviation σ_r)[64]. This analysis tool provides several different approaches for the extraction of the distance distributions and allows the systematic error analysis of experimental noise and uncertainties. It is especially useful for defining the error bars of the data points[61].

3.2 Förster resonance energy transfer (FRET) - „a spectroscopic ruler on a nanometer scale“[66,67]

During a FRET experiment an energy transfer occurs between two flourophore molecules, when they are in close proximity, such that the emission spectrum of one flourophore overlaps with the absorption spectrum of the other flourophore in its near proximity[68]. Since its discovery, FRET phenomenon has been applied to biology, biochemistry, biophysics and, most recently, structural biology. Some examples of the biomolecular research questions which have been investigated using sm-FRET technique are given in *Figure 3.*

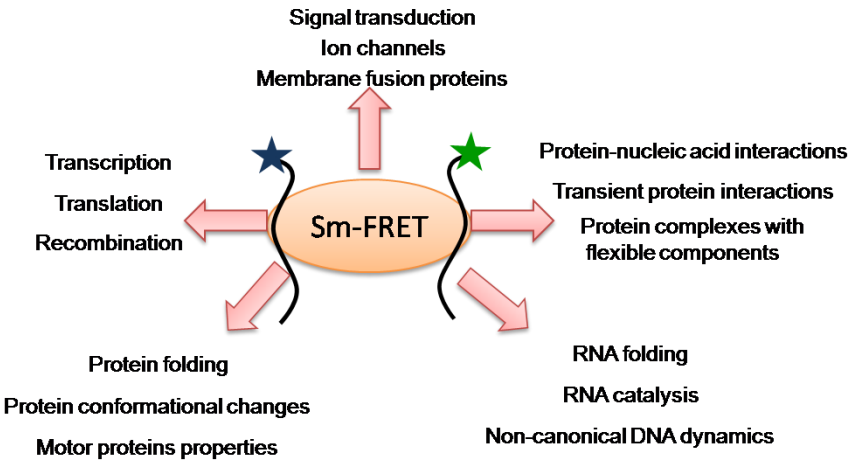

Figure 3: Examples of varies applications of sm-FRET in molecular biology, biochemistry and biophysics. The techniques is used to address important biological processes such as signal transduction or protein translation as well as to address different molecular properties such as folding of the RNA or interaction of proteins with nucleic acids (*see Section 4.*).

Single-molecule FRET (sm-FRET) is a special structural biology application of FRET. For running sm-FRET experiments, donor and acceptor fluorescent dyes (flourophores) are attached at the specific sites on the biomolecules by using site-directed chemistry techniques. The final output of the FRET measurement is FRET efficiency (E_{FRET}), which is used to calculate the inter-flourophore distance R (*Eq1*).

$$E_{FRET} = 1/1+(R/R_o)^6 \quad (1)$$

R_o (Förster radius) in *Eq.1* is a donor-acceptor dye pair specific constant, which combines several important FRET parameters: donor quantum yield (φ_D), donor-acceptor spectral overlap ($J(\lambda)$) and the relative orientation of the flourophores (κ^2) (*Eq. 3*).

$$R_o = 0.529\ \kappa^2\ \varphi_D\ J\ (\lambda) / N\ n^4 \quad (2)$$

Important to note is that κ^2 orientation factor can be critical for the correct inter-flourophore distance calculation, as discussed in the following section.

E_{FRET} itself is not directly measured by FRET experiment. It can be derived by measuring either donor and acceptor fluorescence intensities, or alternatively donor lifetime in the presence and absence of the acceptor (*Eq. 3*)

$$E_{FRET} = \left[1 + \gamma\ (I_D/I_A)\ \right]^{-1} = 1 - \tau_{D(A)}/\ \tau_{D(0)} \quad (3)$$

,where I_D, I_A and $\tau_{D(A)}$, $\tau_{D(0)}$ are donor(D), acceptor(A) fluorescence intensities and donor flourophore lifetime in the presence(A) and absence(0) of the acceptor, respectively. γ factor depends on the flourophore properties and the instrumentation. It is another factor that can critically influence accuracy of FRET derived distances (*Eq. 4*):

$$\gamma = \varphi_A\eta_A/\ \varphi_D\eta_D (4)$$

,where φ_A, φ_D is flourophore quantum yield and η_A, η_D, the detection efficiency of acceptor (A) and donor (D) flourophores, respectively[71].

If multiple sm-FRET experiments are run on differently labelled samples, a network of sm-FRET distances can be obtained, which is particularly useful when high resolution structures (or model) of individual, flourophore labelled components are available. In such cases, using computational approaches, the network of sm-FRET derived distances can be used to model three-dimensional model of the entire assembly[29].

Correct determination of γ and κ^2 factor influences the accuracy of sm-FRET derived distances

As introduced in previous section, for accurate conversion of the measured donor and acceptor intensities into the inter-flourophore distances, correct determination of γ and κ^2(and thus R_o) is critical (*Eq. 1,Eq. 3*). When using sm-FRET to address a particular structural question, experimentalist should keep in mind that this determination varies according to the choice of the experimental methodology[49]. For instance, to determine γ, three different approaches can be used, if the system under investigation contains diffusing or immobilized single molecules[49,72,73,74,75]. In the case of diffusing single molecules, empirical values for φ and η (*Eq.4*) are used[49,73,74]. For the immobilized single molecules it is better to determine γ from the donor and acceptor intensities in a photo bleaching experiment[72]. Finally, γ can also be determined from the difference in the FRET efficiency calculated using I_D,I_A or $\tau_{D(A)}$,$\tau_{D(0)}$

(*Eq.3*), however, in this case, specialized instrumentation for measuring all of these parameters must be available[75].

The accurate calculation of κ^2 is not trivial, as it cannot be directly measured and requires information about the flourophores' dynamics with the respect to attached molecule. In this respect, several studies have addressed the potential methods to accurately determine κ^2 [29,76,77,78,79,80]. From the aspects of structural modeling the assumption of the "freely rotating flourophores" can be readily accepted. Namely, if the flourophores reorient isotropically in a time much shorter than the lifetime of the donor's excited state, κ^2 can be approximated to the value of 2/3 [29,76,77]. The error of the sm-FRET derived distances, introduced by this approximation, is within ~10-20% of the distance value and thus smaller than the error boundaries used, for instance, in docking calculations[78,81].

Potentials and limitations of sm-FRET in accurate distance determination

Sm-FRET technique has recently experienced major technical improvements in terms of instrumentation, sample preparation and data analysis, which has led to its successful applications in structural modeling[29]. However, the technique, as every other, suffers from some inherent limitations and in order to extract the maximum out of the sm-FRET measurement, certain considerations should be taken during the experimental set-up, data analysis and data interpretation.

When it comes to the experimental set-up, the most important to consider is:

1) The efficiency of the site-directed labelling.

2) The sensitivity range of FRET, which is 0.5 R_o-to-2 R_o [29]. Distances lower than 0.5 R_o or higher than 2 R_o are experimentally difficult to measure.

3) The photo stability of the fluorescent dyes. Both donor and acceptor fluorescent dyes should be sufficiently photo stable to account for adequate signal-to-noise ratio and collection of enough data points before the photobleaching[70].

Furthermore, in the data analysis step the experimentalist should account for:

1) The correction of the raw fluorescent intensity data (I_D and I_A) [29,70].

2) Correct determination of γ and κ^2 factor (see previous section).

Finally, for the structural modeller, the data interpretation step is the most important. When interpreting the data following should be considered:

1) The definition of the distance error boundaries in respect to the experimental strategy used for the estimation of the orientation factor κ^2.

As discussed in previous section, κ^2 value can be approximated to 2/3. In this respect, the error boundaries to the given mean sm-FRET distance (<*R*>) should be at least 10-20% of the corresponding distance value[78].

2) The uncertainty of the flourophore positions due to the long and flexible linkers. It is important to keep in mind that sm-FRET derived distances are not the distances between the flourophore dyes' attachment points, but rather between the average dyes' positions.

3) The environmental influences on the fluorescence properties of the flourophores (R_o factor) and the E_{FRET} .These influences are difficult to consider, especially for the structural modellers, who make use of the processed data. However, awareness of the environmental influences should provoke critical interpretation of the "absolute distance" information coming from sm-FRET experiments[66]. In fact, particular molecules can show more influence on the flourophore dyes than the others. For instance, sm-FRET study performed on the tethered double-stranded DNA (ds-DNA) has demonstrated the dependency of FRET efficiencies on the additional factor apart from the increasing flourophores' separation, which was presumably coming from the stacking interactions between the flourophores and the DNA bases[94]. Thus, especially in case of the double-stranded nucleic acids, modeller should critically assess the sm-FRET study to make sure that the control experiments and considerations of environmental influences have been taken.

4. Challenging HADDOCK with sm-FRET and EPR data

4.1 Modeling of the Syt1 C2A and C2B domain complex by using EPR data in HADDOCK

Number of studies, in which EPR distances are used to model biomolecular complexes, is increasing. However, to our knowledge, only for one[31] both experimental measurements and high-resolution structural information of the complex[95] are available. Therefore, this study is a perfect candidate for a proof-of-the-concept study, in which HADDOCK's potential to deal with EPR distances can be tested.

In the study of *Lai et al.*[31], the interaction of the soluble fragments of Synaptotagmin 1 (Syt1) (C2A and C2B domains) and the neuronal core soluble N-ethylmaleimide-sensitive factor attachment receptor (SNARE) was probed using the combination of the site-directed spin labelling with CW-EPR techniques (see *Section 3*). The distances between the two C2 domains in the presence of Ca^{2+} ions and the absence or the presence of SNARE, were obtained using double electron-electron resonance (DEER) experiment[31]. We used the distances from the experiment with the C2A-C2B domains in the absence of SNARE as unambiguous restraints to dock the two domains in HADDOCK, as discussed below. The crystal structure of the domains (PDB ID: 2R83)[95] were available, which allowed the comparison and validation of the generated models.

Docking Protocol

We performed the docking using Guru-web interface of HADDOCK 2.2[33] for all test cases. Our approach is schematically shown in *Figure 4* and described in more detail in the steps bellow:

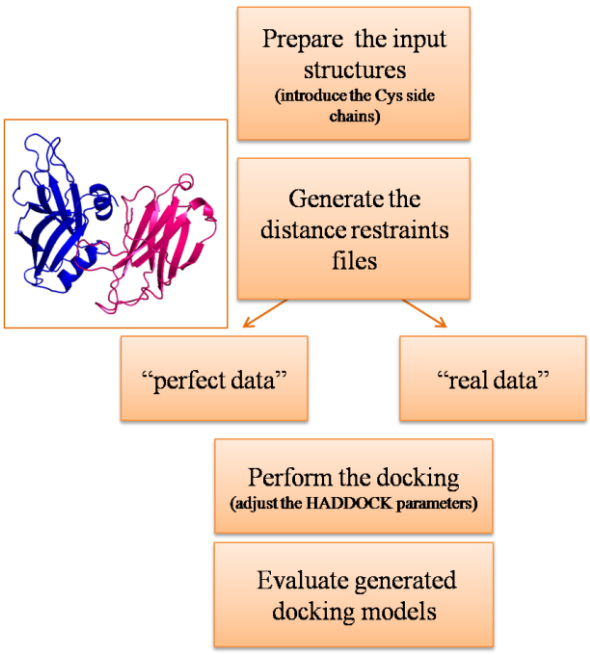

Figure 4: Overview of the approach used to introduce the EPR derived distance information into HADDOCK docking protocol.

1) *Preparation of the input structures for docking*

In study of Lai *et al.*[1] single cysteine mutants were introduced to C2A and C2B domains to allow for the incorporation of the nitroxide spin-labels using site-directed spin labelling (SDSL) techniques[31]. We have introduced the cysteine side chains to the relevant sites in C2A-C2B structure files[95] (PDB ID: 2R83) following their study. Afterwards, mutant C2A and C2B domains were separated at the linker connecting them (residues 266-269).

2) *Generation of the distance restraints files*

We have prepared two distance restraint files. The first file contained "perfect data", namely the distances measured on the available crystal structure between the C_β atoms, of the indicated side chains[31,95] (*Table 3*). Second file contained the "real data", namely, the experimentally determined EPR distances[31,61]. These distances were defined as the pair-wise distances between the Sγ atoms (introduced to the structure files with the cysteine side chains, see *step 1*) (*Table 3*).

In the case of perfect data, 0.5 Å was used as the upper and lower distance boundary. For the restraints derived from EPR experiment, the distance error boundaries obtained by DEER experiment analysis[31] were used (*Table 3*).

Table 3: Distances and distance boundaries derived from the DEER experiment analysis for the pairs of the mutated residues in C2A-C2B domains of Syt1.

C2A/C2B mutated residues	Distances-DEER experiment (Å)	Inter-atomic distances (C_β-C_β, Å)[95]	Upper(+)/Lower(-) distance boundary-DEER experiment (Å)
M173R1/V304R1	39	42.9	± 9
M173R1/K332R1	45	38.9	± 10
R199R1/V304R1	37	43.5	± 16
R199R1/K332R1	46	39.5	± 9
K244R1/K327R1	41	40.5	± 11
K189R1/K327R1	43	29.0	± 12
K213R1/V283R1	60	40.8	± 18

3) *The docking*

Two consecutive "bound" docking runs were prepared. In the first one perfect data were used to drive docking, whereas in the second one real data were applied. For both, distance restraint files were supplied as unambiguous distance restraints and the number of structures for *it0*, *it1*, and *water* were increased to 10000, 400, and 400, respectively. Furthermore, the residues between which the pairwise distances were

defined were set to be fully flexible. Also, to maintain the connectivity between starting structures, a distance restraint of 10 Å was imposed during *it0* and *it1*, whereas during the *water*, this distance was reduced to real peptide bond distance. The resulting solutions were clustered with a 7.5 Å RMSD threshold. The clustered conformations were evaluated on the basis of the interface root mean square deviation (*i*-RMSD) in comparison to the crystal structure and HADDOCK score[95]. A cluster was considered to be acceptable if the i-RMSD of one it's top 4 members is below 4Å.

Evaluation of the docking models

Figure 5 and *Tables 4* and *5* provide a comprehensive overview of the results generated for the perfect-case and real-case docking runs. The results revealed that in the case of perfect data, HADDOCK could generate a top ranking acceptable cluster (with i-RMSD 1.8±0.7Å, *Figure 3*). However for the real-case run, the best model was at the border of acceptable, with i-RMSD= 4.6±0.2Å (*Figure 5*) and ranked at the 5th position.

Table 4: Evaluation of the models generated in HADDOCK docking, using the distances measured on the crystal structure[95].

Cluster ID	i-RMSD (Å, mean±std)	Cluster Rank	HADDOCK score (avg. ± s.d.)
Cluster3	1.8±0.7	1	-111.4 +/- 5.6
Cluster1	6.3±0.2	2	-49.4 +/- 6.0
Cluster2	12.4±0.2	3	-28.5 +/- 5.3

Table 5: Evaluation of the models generated in HADDOCK docking using EPR derived distances[31]. Shown is the comparison of the i-RMSD with HADDOCK rank

Cluster ID	i-RMSD (Å, mean±std)	Cluster Rank	HADDOCK score (avg. ± s.d.)
Cluster6	4.6±0.2	5	-31.0 +/- 14.6
Cluster2	6.9±0.2	2	-51.9 +/- 9.5
Cluster8	7.3±0.4	9	-19.8 +/- 14.3
Cluster1	7.3±0.7	3	-50.0 +/- 5.9
Cluster4	8.3±0.2	6	-28.2 +/- 7.3
Cluster9	9.0±0.4	8	-21.9 +/- 10.8
Cluster3	10.2±0.2	1	-71.8 +/- 10.4
Cluster7	10.6±0.4	7	-22.1 +/- 7.2
Cluster5	10.8±0.4	4	-43.4 +/- 16.2

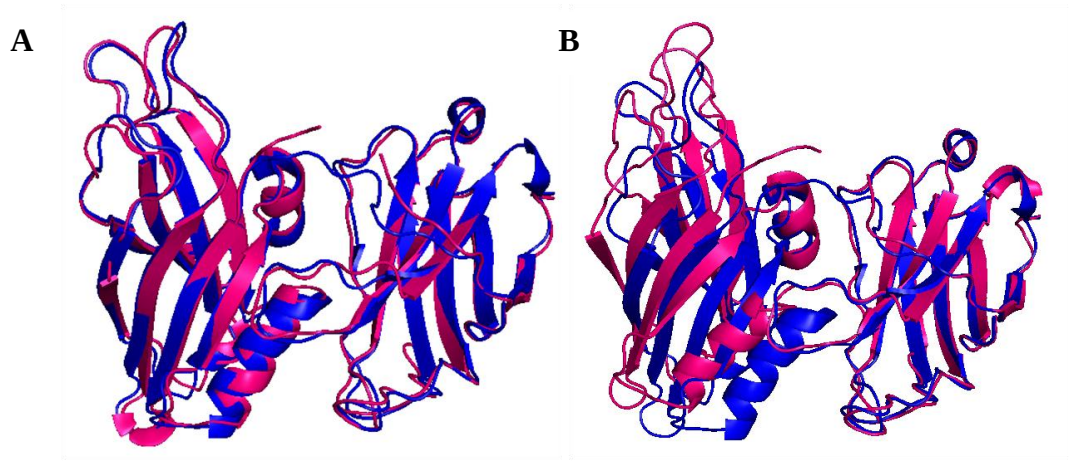

Figure 5: Overlay of theC2A-C2B Syt1 crystal structure[95] (PDB ID:2R83) (in blue) with the models generated in the docking in HADDOCK (in magenta) using A) "perfect data":7 inter-atomic distances measured on crystal structure; B) "real data": 7 DEER experiment derived distances and distance boundaries[31]. Models are given in the secondary structure representation.

4.2 Modeling of HIV1 RT-ds DNA complex by using sm-FRET data in HADDOCK

Due to the rapid development and the previously mentioned advantages (see *Introduction* and *Section* 3), sm-FRET has been used in an increasing number of studies to model high-resolution biomolecular structures [29,30,79,81,96-102]. Very recently, sm-FRET data has been also used as restraints to drive the rigid-body-docking[29,81,100]. A prominent example of such protocol was introduced by Kalinin *et al.* and presented as a "comprehensive toolkit and pipeline for FRET-restrained high precision structural modeling"[30]. In this work, they performed a benchmark study on the complex of a human immunodeficiency virus type 1 (HIV-1) reverse transcriptase (RT) with a DNA primer-template[30]. Even though this is not the first study to demonstrate the successful application of sm-FRET derived distance information in structural modeling, it is performed on a complex for which independent high-resolution structural data was available, thus allowing for the quantitative evaluation of their docking performance. Motivated by the high accuracy of the modeling demonstrated in this study (0.5Å RMSD from the crystal structure, calculated for all dsDNA atoms)[30] , as well as the previous successful applications of HADDOCK on protein-DNA complexes[103], we decided to use this case to assess the performance of HADDOCK with sm-FRET data.

Docking Protocol

Like in the case of EPR-driven docking, we have performed the docking using Guru-web interface of HADDOCK 2.1[33] for all test cases. In this case, however, we tried consecutive docking runs with varying amount of the data points (distance restraints). The approach is summarized schematically in *Figure 6* and described in more detail below:

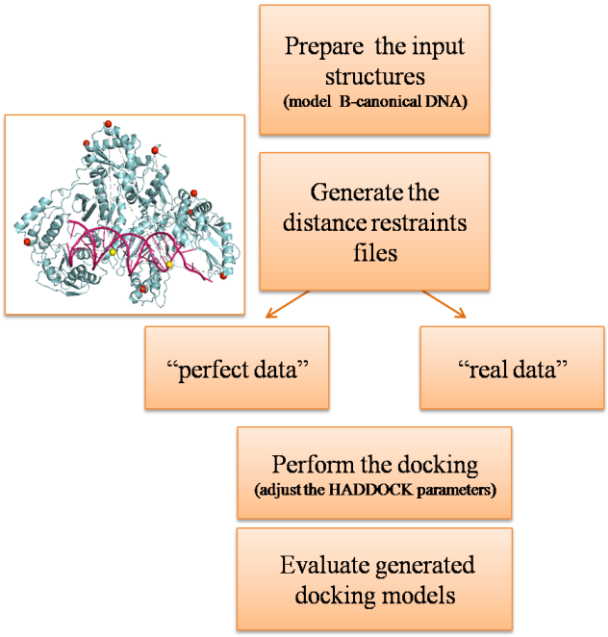

Figure 6: Overview of the approach used to introduce the sm-FRET derived distance information into HADDOCK docking protocol.

1) *Preparation of the input structures for docking*
   In the study of Kalinin *et al.*[30], fluorescent labels were attached to Cβ atoms of the specific residues on HIV-1 RT and to C5 atoms on deoxiribonucleotides, respectively[30] (*Table 6*). In the case of the nucleobases dp (1), phosphorous atom (P) of the DNA backbone was used as the dye attachment point.
   Starting structure for the HIV-1 RT was extracted from the available complex crystal structure[104] (PDB ID: 1R0A). To generate the input structure for the double stranded-DNA (ds-DNA), a canonical B-DNA was created by running 3D DART[106] web-server on the basis of 19 base-pair long ds-DNA sequence used in the study of Kalinin *et al.*[30] (see *Discussion*).

2) *Generation of the distance restraints files*
   Three sets of distance restraint files were prepared containing 5, 12 and 20 pairwise distances. In the "perfect scenario", crystal distances (measured on PDB ID:1R0A)[104] were imposed, whereas in the "real-case scenario" distances determined by the sm-FRET measurements[30] * were applied. The specifications of the inter-atomic distances used in both scenarios can be found in *Table 6*. For the "perfect" case, 0.5 Å was used as the upper and lower distance boundary. For the real data cases, lower boundary was defined by the sum of the lengths of the corresponding flexible dye-linkers attached to particular acceptor/donor atom pairs (*Figure 7*). Upper distance boundary was defined to be as 15 Å (see *Discussion*).

---

* Mean inter-fluorophore distances $<R_{DA}>$E(Å) as determined by Kalinin *et al.*[30]

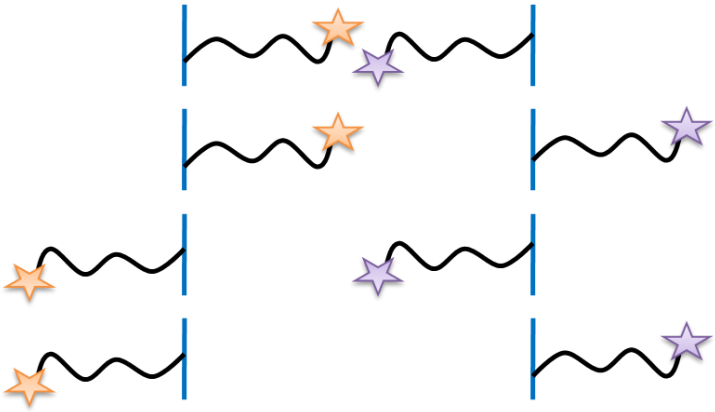

Figure 7: Schematic representation of the flourophore dyes displacements from their attachment points on biomolecules, the difficulty for accurate distance derivation in sm-FRET techniques. Four possibilities of the dye positions are given. It should be noted that the dyes can take any of the positions in between the depicted ones.

Table 6: Overview of the distances measured between the fluorescently labelled sites on HIV-1 RT and ds-DNA in the sm-FRET experiment[30] and the distances measured on the complex crystal structure[104] (PDB ID: 1R0A).

Dye position pairs	Mean inter-fluorophore distances $<R_{DA}>E$(Å) as determined by Kalinin *et al.*[30]	Inter-atomic distances (measured on PDB ID: 1R0A) (Å)
RT (p66Q6C):dp(19)/dt	46	37.1
RT(p66T27C):dp(19)/dt	30	34.8
RT(p66K287C):dp(19)/dt	44	21.8
RT(p51Q6C):dp(19)/dt	71	57.2
RT(p51K173C):dp(19)/dt	60	72.9
RT(p66E194C):dp(19)/dt	41	37.9
RT(p51E194C):dp(19)/dt	70.6	72.1
RT(p51K281C):dp(10)/dt	81	41.8
RT(p66Q6C):dp(10)/dt	46	48.1
RT(p66E194C):dp(10)/dt	44	59.3
RT(p51K173C):dp(10)/dt	62	58.7
RT(p66E194C):dp(10)/dt	68	56.5
RT (p66Q6C):dp(1)/dt	73	71.3
RT (p66T27C):dp(1)/dt	73	80.2
RT(p66E194C):dp(1)/dt	83	86.8
RT(p66K287C):dp(1)/dt	45	48.8
RT(p51Q6C):dp(1)/dt	65	61.1
RT(p51K281C):dp(1)/dt	35	20.2
RT(p51K173C):dp(1)/dt	67	60
RT(p51E194C):dp(1)/dt	55	57.6

→5 distance restraints
\+ → 13 distance restraints
\+ \+ → 20 distance restraints

3) *The docking*

For all cases, distance restraint files were supplied as unambiguous distance restraints and the number of structures for *it0, it1*, and water were increased to 10000, 400, and 400, respectively. Furthermore, the residues between which the pairwise distances were defined were set to be fully flexible. The resulting solutions were clustered with a 20 Å RMSD threshold to allow for the sampling of the larger conformational space. Epsilon constant for the electrostatic energy term was set to 78 (standard used for the protein-DNA docking)[103] and additional changes in the advanced sampling parameters were given based on the protein-DNA docking protocol described in the study of van Dijk *et al*.[103]. The clustered conformations were evaluated on the basis of the i-RMSD in comparison to the crystal structure and HADDOCK score[95]. A cluster was considered to be acceptable if the i-RMSD of one it's top 4 members is below 4Å[35].

Evaluation of the docking models

Figure 8 and *Tables 7* and *8* provide a comprehensive overview of the selected results generated for the perfect-case and real-case docking runs. The results showed that, when provided with perfect 12 distances, HADDOCK could generate a single, top ranking, acceptable cluster (with a mean i-RMSD of 1.1 Å, *Figure 8*). In the real-case runs with five or 13 distances, HADDOCK could not generate an acceptable solution. However, it should be noted that 13 distances were sufficient to reproduce the binding mode (*Figure 8*). Finally, when supplied with 20 real case distances, the best cluster, ranked at the top was an acceptable one, with a mean i-RMSD of 3.9.

Table 7: Evaluation of the models generated in the "proof-of-the-principle" test, using 12 crystal structure distances (PDB code: 1R0A). Compared is i-RMSD, HADDOCK rank and average score

Cluster ID	i-RMSD(Å, mean±std)	Rank	HADDOCK score (avg. ± s.d.)
Cluster1	1.1±0.1	1	-287.0 +/- 17.9

Table 8: Evaluation of the models generated using 20 sm-FRET derived distances. Compared is i-RMSD, HADDOCK rank and average score

Cluster ID	i-RMSD(Å, mean±std)	Rank	HADDOCK score (avg. ± s.d.)
Cluster1	3.9±0.2	1	-89.3 +/- 28.1
Cluster2	6.8±0.1	2	-37.1 +/- 8.5

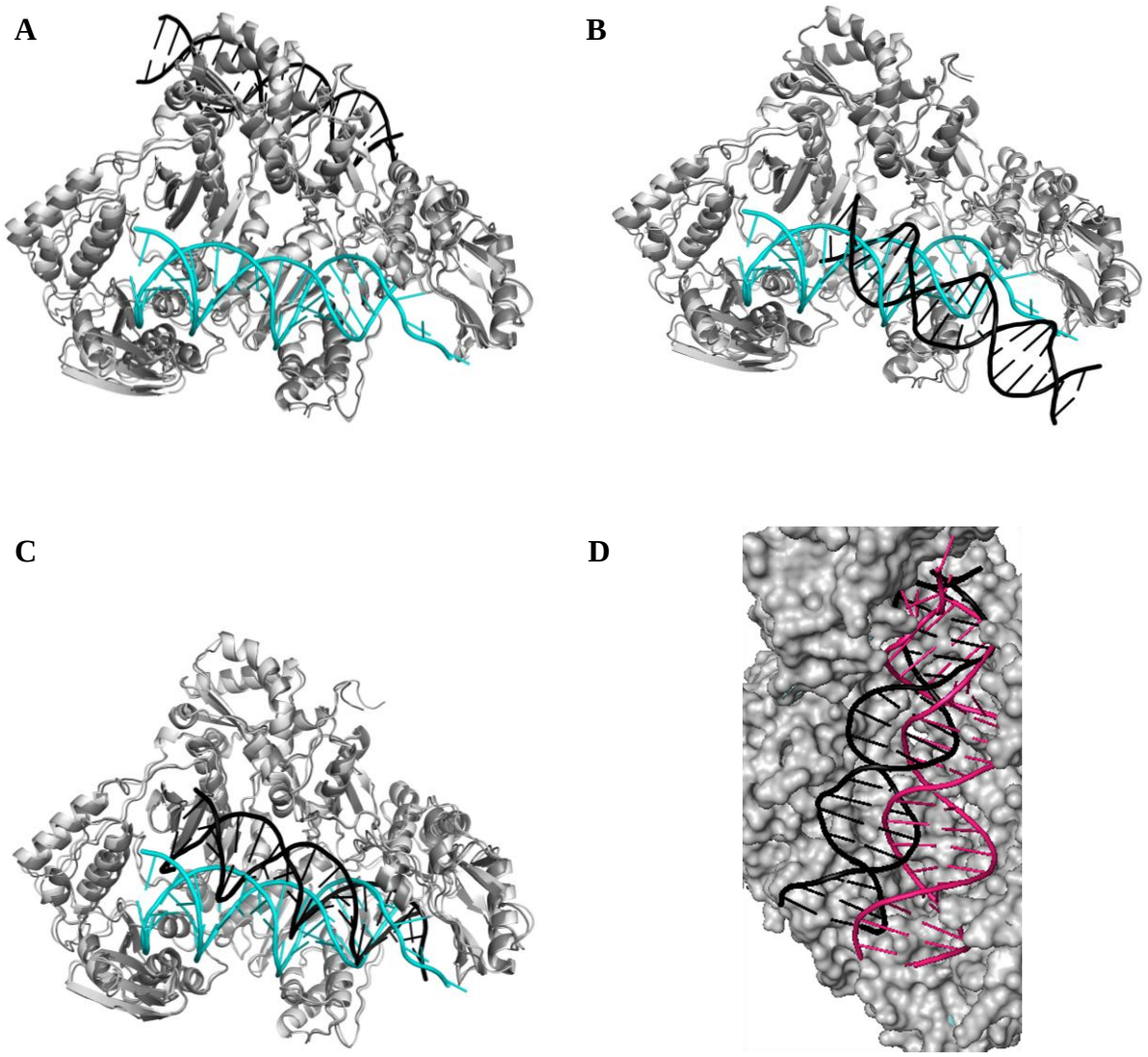

Figure 8: Models of the HIV-1 RT:dsDNA complex generated in HADDOCKing with sm-FRET derived distance restraints. A) 5 real-case distances B) 13 real-case distances C) 20 real-case distances D) Close up view of the interface of the model in C). Structures are given in cartoon (A,B,C), and surface representation (D) for HIV-1 RT (PDB ID:1r0a)[104]. In cyan (A,B,C) and magenta (D) the canonical B-DNA created by running 3D DART. In dark and light gray HIV-1 RT from crystal structure of the complex (PDB ID:1r0a)[104], and docked by HADDOCK, respectively. The HADDOCK models correspond to the top ranked clusters (A,C,D); and the second ranked cluster (B) (for the details, see *Section 4.2*).

5. Discussion

At the moment, state-of-the-art high-resolution structural biology techniques are streaming to overcome the challenges due to high-molecular weight, flexible, inherently disordered, transient or membrane associated complexes[12]. Meanwhile, the number of accurate low-resolution methods are rapidly increasing and information on even the most challenging macromolecular complexes is accumulating[16]. In order to obtain high-resolution insight on such complexes, one of the prominent ways is to translate low-resolution data by data-driven modeling approaches into structural information. Among those approaches, HADDOCK is particularly prominent, which can make use of wide range of available experimental data simultaneously to drive the prediction of the complex structure[33].

Recently, the interest in using data coming from EPR and sm-FRET techniques for structural modeling has been increasing[29-31,51].Both methods have experienced major technical improvements, allow the measurements in solution, support the variation of the experimental conditions and can complement high resolution techniques by providing long range distances for the challenging biomolecular complexes[29,51].Considering the successful application of HADDOCK data-driven docking approach to a variety of biomolecular complexes[33,34,103,106,107] and its abbility to drive the docking based on the long range distance restraints (e.g. those from MS-crosslinking experiments), we were particularly interested in testing its performance on EPR and sm-FRET data. Our choice of the sm-FRET and EPR studies was based on the modeling of biomolecular complexes for which high-resolution structural information was available[30,31,95,104].

In the case of EPR-driven docking, when crystal distances were applied, an acceptable top ranking cluster was generated. However, when provided with the real distances from DEER experiment on Syt1 C2A-C2B domains, the best solution generated by HADDOCK was on the edge of the acceptability and ranked 5th (*i-RMSD* of 4.6±0.2, *Figure 5*). This quality and rank difference can be due to the distance error bounds obtained in DEER experiment analysis (9-18 Å). In the original study, the authors concluded that there is a considerable flexibility and conformational variation of the two domains in solution[31]. So, applied distance distributions might not be restrictive enough to pull the two domains to a binding mode depicted by the crystal structure[95].

In the modeling of HIV-1 RT:ds DNA primer/template complex, an important observation was the increase in the accuracy of the docking as a function of the increase in the number of the distance restraints, both for the perfect and the real data sets (*Figure 8 A,B,C*). Importantly, when provided with the set of 13 real-case distances, HADDOCK could reproduce the binding mode, whereas 20 real distances generated a top ranking, acceptable solution (i-RMSD 3.9±0.2, *Figure 8 C,D*). In case of the perfect data sets, 5 distances were already enough to reproduce the correct binding mode (data not shown), whereas 12 distances resulted in a single, top ranking, acceptable solution (i-RMSD 1.1±0.1) in HADDOCK.

Furthermore, we have found that the definition of the error bounds and the mean sm-FRET distances (<*R*>)[30] are critical determinants. Regular HADDOCK protocol does not deal with flexible linkers attached to the atoms and/or pseudo atoms introduction as to account for the effect of the dyes and their flexible linkers. Therefore, in order to impose the sm-FRET inter-dye distances as the inter-atomic distances between flexible dyes attachment points we defined the distance boundaries considering the maximal possible effect of the dyes' dispositions, as illustrated in *Figure 6*. Defining the lower distance boundary as the sum of the lengths of the flexible dyes linkers, and the upper distance as 15 Å, was restrictive enough to drive the docking of the ds DNA and HIV-1 RT to the correct binding mode. It should be noted that 15 Å upper distance boundary is strongly case dependant and was defined by considering the inter-atomic distances measured on the crystal structure (PDB ID: 1R0A) to ensure the correct distance values for all the restraints. In the study of Kalinin *et al.*[30], which has introduced the "toolkit" for the sm-FRET restrained docking, the correct calculation of the dye position distributions ("explicit modelling of dye behaviour")[30] was defined crucial to account for use of sm-FRET distances in a quantitative manner. We determined that, for the use of these distances in HADDOCKing, the displacements of the dyes from their attachment points due to the maximal extensions of the flexible linkers must be taken into account.

Finally, the study of Kalinin *et al.*[30] reported the improvement in the agreement of the model with the crystal structure (PDB ID:1R0A) upon the relaxation of the docked B-DNA by MD simulations[30]. Namely, as the rigid body docking of "straight" B-DNA cannot account for the DNA bending, this B-DNA structure does not fit perfectly with the sm-FRET distances[30]. Thus, for sm-FRET restraints, the improvement in HADDOCK performance can be expected if the optimal "bent" ds-DNA is used as an input DNA structure for docking.

6. Conclusions and perspectives

According to the current literature and as supported by our current study, inter-atomic distances derived from sm-FRET and pulsed EPR techniques can be used in restrained docking of biomolecular complexes[29-32,49]. Assuming that the experimental data have been correctly processed and the data quality is adequatly high, the important considerations for a modeler should be the confidence intervals of the measured distances that account for the uncertainties introduced by the dye-linkers or spin-labels, for sm-FRET or EPR, respectively.

Our study demonstrates the performance of HADDOCK on the distance restraints derived from sm-FRET and EPR data, for the first time. Considering the number, the range and the confidence of the distance restraints provided, as well as the fact that the HADDOCK protocol has not been individually optimized for EPR or sm-FRET data, the demonstrated performance can be qualified as highly promising. Importantly, there is a room for future improvement. For instance, the optimizations of the sampling parameters in the protocol and the adjustments of the data-weight (*w*), could improve both the sampling and the scoring of the solutions in HADDOCK. Furthermore, in case of sm-FRET restrained docking, the direct improvement in the accuracy of the solutions could be achieved by docking the optimal "bent" form of ds DNA (see *Discussion*). Finally, the expansion of HADDOCKing to a

number of new biomolecular complexes, for which both the high-resolution information and sm-FRET and EPR data will be available, is needed. These could be generated, for instance, through the collaboration of the experimentalist and the modellers, allowing for the simultaneous refinement of both of the approaches. Such study could pave the way for including the EPR and sm-FRET distance restraints to the rich repertoire of the experimental data already driving HADDOCKing and could, thus, allow the HADDOCK to highlight some of the “dark parts” of the interactome.

7. References:

1. Tocilj A, Schlünzen F, Janell D, Glühmann M, Hansen HAS, Harms J, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A. 1999. The small ribosomal subunit from *Thermus thermophilus* at 4.5 Å resolution: Pattern fittings and the identification of a functional site. *Proc. Natl. Acad. Sci.* USA 96: 14252-14257.

2. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* 289(5481):905-20

3. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. *Science* 289(5481):920-30

4. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF. 2001. Crystal Structure of the Ribosome at 5.5 Å Resolution. *Science* 292(5518): 883-896

5. Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD. 2001. Structural Basis of Transcription: An RNA Polymerase II Elongation Complex at 3.3 Å Resolution. *Science* 292(5523):1876-1882

6. Cramer P, Bushnell DA, Kornberg RD. Structural Basis of Transcription: RNA Polymerase II at 2.8 Ångstrom Resolution. *Science* 292 (5523):1863-1876

7. Singleton MR, Dillingham MS, Gaudier M, Kowalczykowski SC, Wigley DB. Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature* 432: 187-193

8. Loll B, Kern J, Saenger W, Zouni A, Biesiadka J.Towards complete cofactor arrangement in the 3.0 Å resolution structure of photosystem II. *Nature* 438: 1040-1044

9. Wuthrich K. NMR of proteins and nucleic acids (Wiley, New York 1986)

10. Zhang S, Meier BH, Ernst RR. 1992. Polarization echoes in NMR. *Phys. Rev. Lett*. 69: 2149–2151

11.Bonvin AM, Boelens R, Kaptein R. 2005. NMR analysis of protein interactions. *Curr. Opin. Chem. Biol*. 9:501- 508

12. Stein A, Mosca R, Aloy P. 2011. Three-dimensional modeling of protein interactions and complexes is going ‘omics, *Curr. Opin. Struct. Biol*. 21: 200–208

13. Melquiond A, Karaca E, Kastritis PL, Bonvin A. 2011. Next challenges in protein-protein docking: from proteome to interactome and beyond. *WIREs Comput. Mol. Sci*. 2: 642–651

14. Baldus M. 2007. Magnetic resonance in the solid state: applications to protein folding, amyloid fibrils and membrane proteins.*Eur.Biophys. J.*36 (Suppl 1):S37–S48

15. Renault M, Tommassen-van Boxtel R, Bos MP, Post JA, Tommassen J,Baldus M. 2012. Cellular solid-state nuclear magneticresonance spectroscopy. *Proc. Natl. Acad. Sci.* USA 109(13): 4863–4868

16. Karaca E, Bonvin AMJJ. 2012. Advances in integrative modeling of biomolecular complexes.*Methods*doi: http://dx.doi.org/10.1016/j.ymeth.2012.12.004

17. Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P et al. 2011. EMDataBank.org: unified data resource for CryoEM *Nucl. Acids Res.* 39(suppl 1): D456-D464 doi:10.1093/nar/gkq880

18. Mitra K, Frank J. 2006. RIBOSOME DYNAMICS: Insights from Atomic Structure Modeling into Cryo-Electron Microscopy Maps.*Annu. Rev. Biophys. Biomol. Struct*. 35:299–317

19. Frank J. 2006. Three-Dimensional Electron Microscopy of Macromolecular Assemblies (Oxford: Oxford Univ. Press)

20. van Heel M, Gowen B, Matadeen R, Orlova EV, Finn R, et al. 2000. Single-particle electron cryo-microscopy: towards atomic resolution. *Q. Rev. Biophys.* 33:307–69

21. Henderson R. 2004. Realizing the potential of electron cryo-microscopy. *Q. Rev. Biophys.* 37:3–13

22. Alber F, Forster F, Korkin D, Topf M, Sali A. Integrating diverse data for structure determination of Macromolecular Assemblies. *Annu. Rev.Biochem.*77:443-477

23. Alber F, Dokudovskaya S, Veenhoff L, Zhang W, Kipper J, et al. 2007. The molecular architecture of the nuclear pore complex. *Nature* 450:695–701

24. Alber F, DokudovskayaS,Veenhoff L, Zhang W, Kipper J, et al. 2007. Determining the architectures of macromolecular assemblies.*Nature* 450:683–94

25. Sharon M, Robinson CV. 2007. The role of mass spectrometry in structure elucidation of dynamic protein complexes.*Annu. Rev. Biochem.* 76:167–93

26. Sharon M, Taverner T, Ambroggio XI, Deshaies RJ, Robinson CV. 2006.Structural organization of the 19S proteasome lid: insights from MS of intact complexes.*PLoS. Biol.*4(8):e267

27. Koch MH, Vachette P, Svergun DI. 2003.Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution.*Q. Rev. Biophys.*36:147–227

28. Karaca E, Bonvin AMJJ. On the usefulness of Ion Mobility Mass Spectrometry and SAXS data in scoring docking decoys (submitted for publication)

29.Brunger AT, Strop P, Vrljic M, Chu S, Weninger KR. 2011. Three dimensional molecular modeling with single molecule FRET. *J Struct Biol.* 173:497-505

30. Kalinin S, Peulen T, Sindbert S, Rothwell PJ, Berger S, Restle T, Goody RS, Gohlke H, Seidel CAM. 2012. A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nature Methods* 9(12): 1218-1224

31.Lai AL, Huang H, Herrick DZ, Epp N, Cafiso DS. 2011.Synaptotagmin 1 and SNAREs form a complex that is structurally heterogeneous. *J Mol. Biol.* 405(3):696-706

32. Choi UB, Strop P, Vrljic M, Chu S, Brunger AT, Weninger KR. 2010. Single-molecule FRET-derived model of the synaptotagmin 1-SNARE fusion complex. *Nat. Struct. Mol. Biol.*17:318–324

33. de Vries SJ, van Dijk M, Bonvin AMJJ.2010. "The HADDOCK web server for data-driven biomolecular docking." *Nature Methods* 5(5):883-897

34. Janin J, Henrick K, Moult J, Ten Eyck L, Sternberg MJE, Vajda S, Vasker I, Wodak SJ. 2003. CAPRI: A Critical Assessments of Predicted Interactions. *Proteins* 52: 2-9

35. Mendez R, Leplae R, De Maria L, Wodak SJ. 2003. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* 52: 51-67

36. Bordner AJ, Gorin AA. 2007.Protein docking using surface matching and supervised machinelearning.*Proteins*68:488–502

37. Schneidman-Duhovny D, Inbar Y, NussinovR,Wolfson HJ. 2005.Geometry-based flexible and symmetric protein docking.*Proteins*60:224–31

38. Fernandez-Recio J, Abagyan R, Totrov M. 2005. Improving CAPRIpredictions: optimized desolvation for rigid-body docking. *Proteins* 60:308–313

39. Wiehe K, Pierce B, Mintseris J, Tong WW, Anderson R, et al. 2005. ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5.*Proteins*60:207–13

40. Kozakov D, Brenke R, Comeau SR, Vajda S. 2006. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*65:392–406

41. Christen M, van Gunsteren WF. 2008. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems. *Curr.Opin.Struc.Biol.* 15:144-150

42. Zhang Y. 2008. Progress and challenges in protein structure prediction. *Curr.Opin.Struc.Biol* 18:342-348

43.Kowalsman N, Eisenstein M. 2007. Inherent limitations in protein–protein docking procedures. *Bioinformatic*23:421–26

44. Tobi D, Bahar I. 2005. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state.*Proc. Natl. Acad. Sci.* USA102:18908–13

45. Kastritis PL, Bonvin AMJJ. 2010. Are Scoring Functions in Protein-Protein Docking Ready To Predict Interactomes? Cluesfrom a Novel Binding Affinity Benchmark.*J. Proteome Res.* 9:2216-2225

46. van Dijk AD, Boelens R, Bonvin AM. 2005. Data-driven docking for the study of biomolecular complexes*FEBS J.*272:293–312

47.Melquiond ASJ, Bonvin AMJJ. 2010. Data-driven docking:using external information to spark the biomolecular randez-vous, in Protein-Protein Complexes: Analysis, Modeling and Drug Design, Imperial College Press 183-209

48. Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P et al. 2011. EMDataBank.org: unified data resource for CryoEM *Nucl. Acids Res.* (2011) 39(suppl 1): D456-D464 doi:10.1093/nar/gkq880

49.McCann JJ, Choi UB, Zheng L, Weninger K, Bowen ME. 2010. Optimizing Methods to Recover Absolute FRET Efficiency from Immobilized Single Molecules. *Biophys. J.* 99: 961-970

50. Jeschke G, Polyhach Y. 2006. Distance measurements on spin labelled biomacromolecules by pulsed electron paramagnetic resonance. *Phys. Chem. Chem. Phys.* 9:1895-1910

51. Schiemann O, Prisner TF. 2007. Long range distance determinations in biomacromolecules by EPR spectroscopy. *Q Rev Biophys* 40: 1-53

52. Rabenstein M, Shin YK. 1995. Determination of the distance between two spin labels attached to a macromolecule. *Proc. Natl. Acad. Sci. USA*, 1995, 92, 8239–8243

53. Jeschke G, Bender A, Paulsen H, Zimmermann H,Godt A. 2004. Sensitivity enhancement in pulse EPR distance measurements.*J. Magn. Reson.* 169(1): 1–12

54. P. P. Borbat, J. H. Davis, S. E. Butcher and J. H. Freed, J. Am.Chem. Soc., 2004, 126, 7746–7747

55. Husted EJ, Smirnov AI, Laub CF, Cobb CE, Beth AH. 1997. Molecular distances from dipolar coupled spin labels: the global analysis of multifrequencycontinuous wave electron paramagnetic resonance data. *Biophys. J.* 74: 1861-1877

56. Berliner LJ, Eaton SS, Eaton GR. 2001. Distance measurements in biological systems by EPR in *Biological Magnetic Resonance*Vol. 19

57. A. Schweiger and G. Jeschke, Principles of pulse electron paramagneticresonance, Oxford University Press, Oxford, 2001.

58. Saxena S, Freed JH, 1997. Absorption lineshapes in two-dimensional electron spin resonance and the effects of slow motions in complex fluids. *Journal of Chem. Phys.*107:1317-1340

59. Borbat PP, Freed JH. 1999. Multiple-quantum ESR and distance measurements. *Chem.Phys.Lett*. 313:145-154

60. Jeschke G, Pannier M, Godt A, Spiess HW. 2000. Isotope selection in distance measurements between nitroxides. *Chem. Phys.Lett.*331: 243–252

61. Jeschke G, Chechik V, Ionita P, Godt A, Zimmermann H, Banham J. et al. 2006. DeerAnalysis2006-a comprehensive software package for analyzing pulsed ELDOR data. *Appl. Magn. Reson.* 30:473-498

62. Pannier M, Veit S, Godt A, Jeschke G, Spiess HW.2011.Addendum to the paper "Dead-time free measurement of dipole-dipole interactions between electron spins" in *J. Magn. Reson.*142: 331–340; *J. Magn. Reson*. 2011. 213(2):326-8

63. Jeschke G, Bender A, Paulsen H, Zimmermann H,Godt A. 2004. Sensitivity enhancement in pulse EPR distance measurements.*J. Magn. Reson*. 169(1): 1–12

64. G. Jeschke, G. Panek, A. Godt, A. Bender, H. Paulsen. 2004. Data Analysis Procedures for Pulse ELDOR Measurements of Broad Distance Distributions *Appl. Magn. Reson*. 26:223-244

65. Jeschke G, M. Sajid, M. Schulte, A. Godt. 2009. Three-spin correlations in double electron-electron resonance. *Phys. Chem. Chem. Phys*. 11(31): 6580–6591

66. Stryer, L., Haugland, R.P., 1967. Energy transfer: a spectroscopic ruler. *Proc. Natl. Acad. Sci. USA*58 (2): 719–726

67. Stryer L. Fluorescence energy transfer as a spectroscopic ruler.1978.*Annu. Rev. Biochem*. 47:819–846

68. vonFörster T. 1948. ZwischenmolekulareEnergiewanderung und Fluoreszenz. *Ann.Phys*. 2:55-57

69. Deniz AA, Mukhopadhyay S, Lemke EA. 2008. Single-molecule biophysics:at the interface of biology, physics and chemistry. *J.R.Soc.Interface* 5:15-45

70. Roy R, Hohng S, Ha T. 2008. A practical quide to single molecule FRET. *Nature Methods* 5:507-516

71. Brunger AT, Strop P, Vrljic M, Chu S, Weninger KR. 2011. Three dimensional molecular modeling with single molecule FRET. *J. Struct. Biol.* 173:497-505

72. T. Ha, A.Y. Ting, S. Weiss et al. 1999. Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proc. Natl. Acad. Sci. USA* 96 : 893–898

73. Lee NK, Kapanidis AN, Weiss S et al. 2005. Accurate FRET measurements within single diffusing biomolecules using alternating-laser excitation. *Biophys. J.* 88: 2939–2953

74. R.B. Best, K.A. Merchant, W.A. Eaton et al. 2007. Effect of flexibility and cis residues in single-molecule FRET studies of polyproline.*Proc. Natl. Acad. Sci. USA* 104: 18964–18969

75. Merchant KA, Best RB, Eaton WA. 2007. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proc. Natl. Acad. Sci. USA* 104:1528–1533.

76. Dale RE, Eisinger J, Blumberg WE. 1979. The orientational freedom of molecular probes. The orientation factor in intramolecular energy transfer. *Biophys. J.* 26:161–193.

77.dosRemedios CG, Moens PD. 1995. Fluorescence resonance energy transfer spectroscopy is a reliable ‘‘ruler” for measuring structural changes in proteins. Dispelling the problem of the unknown orientation factor. *J. Struct. Biol*. 115:175–185

78.Ivanov V, Li M, Mizuuchi K, 2009. Impact of emission anisotropy on fluorescencespectroscopy and FRET distance measurements. *Biophys. J.* 97:922–929

79. Muschielok A, Andrecka J, Jawhari A, Bruckner F, Cramer P, Michaelis J. 2008.A nano-positioning system for macromolecular structural analysis. *Nature Methods*5(11):965-71

80. Wozniak, A.K., Schroder, G.F., Grubmuller, H., Seidel, C.A., Oesterhelt, F. 2008.Single-molecule FRET measures bends and kinks in DNA. *Proc. Natl. Acad. Sci. USA*105: 18337–18342

81. Choi UB, Strop P, Vrljic M, Chu S, Brunger AT, Weninger KR. 2010.Single-molecule FRET-derived model of the synaptotagmin 1-SNARE fusion complex. *Nature Struct. Mol. Biol.*17:318-324

82. van der Meer BW, 2002. Kappa-squared: from nuisance to new sense. *J. Biotechnol.* 82:181–196

83. Wilkinson-White L, Gamsjaeger R, Dastmalchi S, Wienert B, Stokes PH, Crossley M, Mackay JP, Matthews JM. 2011. Structural basis of simultaneous recruitment of the transcriptional regulators LMO2 and FOG1/ZFPM1 by the transcription factor GATA1. *Proc.Natl.Acad.Sci.USA* 108: 14443-14448

84. Kobayashi M, Ab E, Bonvin AM, Siegal G. 2010. Structure of the DNA-bound BRCA1 C-terminal region from human replication factor C p140 and model of the protein-DNA complex.*J. Biol. Chem.* 285: 10087-10097

85. van Dijk AD, Fushman D, Bonvin AM. 2005.Various strategies of using residual dipolar couplings in NMR-driven protein docking: application to Lys48-linked di-ubiquitin and validation against 15N-relaxation data.*Proteins*60(3):367-81.

86. Gelis I, Bonvin AMJJ, Keramisanou D, Koukaki M, Gouridis G, Karamanou S, Economou A, Kalodimos CG. 2007. Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell(Cambridge,Mass.)* 131: 756

87. Tomaselli S, Assfalg M, Pagano K, Cogliati C, Zanzoni S, Molinari H, Ragona L. 2012.A disulfide bridge allows for site-selective binding in liver bile acid binding protein thereby stabilising the orientation of key amino acid side chains.*Chemistry* 18(10):2857-66

88. Smith GJ, Sosnick TR, Scherer NF, Pan T. 2005. Efficient fluorescence labelling of a large RNA through oligonucleotide hybridization. *RNA* 11: 234–239

89. Dorywalska, M. et al. 2005. Site-specific labelling of the ribosome for single molecule spectroscopy. *Nucleic Acids Res*. 33:182–189

90. Deniz, A.A. et al. 2000. Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsininhibitor 2. *Proc. Natl. Acad. Sci. USA* 97:5179–5184

91. Jager, M., Nir, E. & Weiss, S. 2006. Site-specific labeling of proteins for single molecule FRET by combining chemical and enzymatic modification. *Protein Sci.* 15:640–646

92. Gopich I, Szabo A. 2005. Theory of photon statistics in single-molecule Förster resonance energy transfer. *J. Chem. Phys*. 122(1):14707

93. Schuler B, Lipman EA, Steinbach PJ, Kumke M, Eaton WA. 2005. Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence. *Proc. Natl. Acad. Sci. USA* 102: 2754–2759

94. Igbal A, Arslan S, Okumus B, Wilson TJ, Giraud G, Norman DG, Ha T, Lilley DM. 2008. Orientation dependence in fluorescent energy transfer between Cy3 and Cy5 terminally attached to double-stranded nucleic acids. *Proc. Natl. Acad. Sci. USA* 105:11176-11181

95. Fuson KL, Montes M, Robert JJ, Sutton RB. 2007. Structure of human synaptotagmin 1 C2AB in the absence of Ca2+ reveals a novel domain association. *Biochemistry*. 46(45):13041-8

96. MeklerV, KortkhonjiaE, MukhopadhyayJ, KnightJ, RevyakinA, KapanidisAN, NiuW, EbrightYW, LevyR, EbrightRH.2002. Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. *Cell* 108(5):599–614

97. Margittai M, Widengren J, Schweinberger E, Schröder GF, Felekyan S, Haustein E, König M, Fasshauer D, Grubmüller H, Jahn R, Seidel CA. 2003. Single-molecule fluorescence resonance energy transfer reveals a dynamic equilibrium between closed and open conformations of syntaxin 1. *Proc. Natl. Acad. Sci. USA* 100: 15516–15521

98. Andrecka J, Lewis R, Brückner F, Lehmann E, Cramer P, Michaelis J. Single-molecule tracking of mRNA exiting from RNA polymerase II. *Proc. Natl. Acad. Sci. USA* 105 (1):135–140 (2008)

99. Sabir, T., Schroder, G.F., Toulmin, A., McGlynn, P. & Magennis, S.W. Global structure of forked DNA in solution revealed by high-resolution single-molecule FRET. *J. Am. Chem. Soc*. 133, 1188–1191 (2011)

100. McCann, J.J., Zheng, L.Q., Chiantia, S. & Bowen, M.E. Domain orientation in the N-terminal PDZ tandem from PSD-95 is maintained in the full-length protein.*Structure.*19, 810–820 (2011)

101. Balci, H., Arslan, S., Myong, S., Lohman, T.M. & Ha, T. Single-molecule nanopositioning: structural transitions of a helicase-DNA complex during ATP hydrolysis. *Biophys. J*. 101, 976–984 (2011)

102. Boura E. et al. 2011. Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. *Proc. Natl. Acad. Sci. USA* 108: 9437–9442

103. vanDijk M et al. 2006. Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acid Res*. 34: 3317-3325

104. Peletskaya EN, Kogon AA, Tuske S, Arnold E, Hughes SH. 2004. Nonnucleoside inhibitor binding affects the interactions of the fingers subdomain of human immunodeficiency virus type 1 reverse transcriptase with DNA.*J Virol.* 78(7):3387-97

105. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC

106. van Dijk M, Bonvin AMJJ. 2009. "3D-DART: a DNA structure modelling server" *Nucl. Acids Res.* 37 (Web Server Issue):W235-W239

107. Karaca E., Bonvin AMJJ. 2011. A multidomain flexible docking approach to deal with large conformational changes in modelling of biomolecular complexes. *Structure*. 19(4):555-565

108. Fernandez-Recio J, Abagyan R, Totrov M. 2005. Improving CAPRI predictions: optimized desolvation for rigid-body docking. *Proteins* 60:308–313

109. van Dijk, A.D.J. et al. 2005. Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins* 60: 232–238

110. Slichter CP. 1980. Principles of Magnetic Resonance. Berlin: Springer Verlag