

## Chapter 13

# Nuclear Magnetic Resonance-Based Modeling and Refinement of Protein Three-Dimensional Structures and Their Complexes



G. Fuentes, A.D.J. van Dijk, and A.M.J.J. Bonvin

AQ: Please provide first name for all the authors.

**Summary** Nuclear magnetic resonance (NMR) has become a well-established method to characterize the structures of biomolecules in solution. High-quality structures are now produced, thanks to both experimental and computational developments, allowing the use of new NMR parameters and improved protocols and force fields in structure calculation and refinement. In this chapter, we give a short overview of the various types of NMR data that can provide structural information, and then focus on the structure calculation methodology itself. We discuss and illustrate with tutorial examples both “classical” structure calculation and refinement approaches as well as more recently developed protocols for modeling biomolecular complexes.

**Keywords:** Docking; NMR; Refinement; Structure calculation; Validation of structures

## 1 Introduction

The first step of a structure determination by nuclear magnetic resonance (NMR) spectroscopy consists in the acquisition of NMR data, typically using heteronuclear multidimensional experiments that allow the assignment of the chemical shifts of all atoms/spins of a molecule ( $^1\text{H}$ ,  $^{15}\text{N}$ ,  $^{13}\text{C}$ ). Once the signals have been assigned,  $^{13}\text{C}$ - and  $^{15}\text{N}$ -edited three-dimensional (3D) nuclear Overhauser enhancement spectroscopy (NOESY) spectra are generally used to obtain interatomic distances from nuclear Overhauser effects (NOE); these provide the required structural information to define the 3D structure of the protein [1, 2]. In addition to distance restraints, other parameters, such as J-couplings [3] and residual dipolar couplings (RDCs) [4] can be measured, providing additional structural information to define the structure of a protein. The experimental NMR parameters are then typically used in restrained molecular dynamics simulations following some kind of simulated

annealing scheme (MD/SA) to generate 3D structures [5]. These are nowadays usually refined in explicit solvent (water), which has been shown to significantly improve the quality of the structures [6,7]. The resulting ensembles of structures should satisfy as many restraints as possible, together with general chemical properties of proteins (such as bond lengths and angles). The whole approach will converge, provided enough data of sufficient quality are available, allowing the determination of an ensemble of structures with a given fold.

In the last few years, a lot of attention is directed toward understanding biomolecular interactions, and NMR is playing an important role here [8], especially in its ability to detect weak and transient interactions [9]. When dealing with complexes, NMR suffers, however, because of the size limitation problem, and, therefore, complementary computational methods, such as docking, are becoming increasingly popular. Docking is defined as the modeling of the 3D structure of a complex from its known constituents, and its combination with a limited amount of (NMR-) data (so called *data-driven docking*) is extremely powerful and has found a wide range of applications [10].

In this chapter, we discuss first “classical” NMR structure calculation and refinement methods and then address the modeling of protein–protein complexes. These will be illustrated with tutorial examples making use of the program CNS [11] with ARIA-derived [12] scripts from the RECOORD [7] webpage and of the HADDOCK package [13].

## 2 Theory

### 2.1 NMR Structural Information Sources

Several NMR parameters providing structural information can be measured for use in structure calculations and refinement. These will be briefly reviewed in the following.

AQ: Please check the edit of the sentence.

#### 2.1.1 Nuclear Overhauser Effects

Classical protein structure determination by NMR relies on a dense network of distance restraints derived from NOEs between nearby hydrogen atoms in a protein [1, 2].

The NOE originates from cross-relaxation between dipolar-coupled spins through space spin–spin interactions that involve a transfer of magnetization from one spin to another. The NOE approximately scales with the distance  $r$  between the two spins as  $1/r^6$ . Because of this  $1/r^6$  dependency, NOEs are only detected between protons less than 5 to 6 Å away in space. They provide essential information for defining the tertiary structure of a protein.

### 2.1.2 Chemical Shifts

Although chemical shifts are very sensitive probes of the chemical environment of a spin, their dependency on the 3D structure is complex. Although several software packages exist that allow prediction of chemical shifts, such as ShiftX [14], SHIFTS [15], SHIFTCALC [16], and PROSHIFT [17], both computational and accuracy limits have prevented their common use as restraints in structure calculations, although direct refinement against chemical shifts has been described [18]. Their deviations from random coil values provide, however, valuable information regarding secondary structure preferences; they can be used to restrict the local conformation of a residue to a given region of the Ramachandran plot, either through torsion angle restraints [19] or by special database potential functions [20].

### 2.1.3 J-Couplings

Scalar or J-couplings are mediated through chemical bonds connecting two spins. The energy levels of each spin are slightly altered depending on the spin state of scalar coupled spins ( $\alpha$  or  $\beta$ ), resulting in splitting of the resonance lines. Particularly informative are the vicinal, three bonds scalar coupling constants,  $^3J$ , between atoms separated by three covalent bonds from each other, which are correlated to the enclosed torsion angle,  $\Theta$ , by an empirical correlation, the Karplus curve [21]. In particular,  $^3J(\text{HN-H}\alpha)$  and  $^3J(\text{H}\alpha\text{-H}\beta)$  give information regarding the  $\phi$ -angle and the  $\chi_1$  angle in an amino acid, respectively. The use of  $^3J(\text{H}\beta_2\text{-H}\beta_3)$  coupling does require stereospecific assignments of diastereotopic proton ( $\text{H}\beta_2/\text{H}\beta_3$ ) pairs.

The main difference with NOEs is that scalar coupling constants only provide information regarding the local conformation of a polypeptide chain. J-couplings have been used as dihedral angle restraints [1] or direct J-coupling restraints [22,23] in NMR structure calculations.

### 2.1.4 Hydrogen Bonds

Slow hydrogen exchange indicates that an amide proton is protected from the solvent, which is usually interpreted as involvement in a hydrogen bond [24]. The acceptor atom cannot, however, be identified directly, and one has to rely on NOEs around the postulated hydrogen bond or assumptions regarding regular secondary structures to define it. Hydrogen bond restraints should be used with caution, although they can be very useful in the case of large proteins when not enough NOE data is available yet. Note that hydrogen bonds can now also directly be detected from cross-hydrogen bond scalar coupling measured from constant time HNCO spectra [25,26]. These can provide useful restraints for structure calculations [27]. Hydrogen bond restraints are introduced into the structure calculation as distance restraints, typically by confining the donor hydrogen/acceptor distance to a given range.

### 2.1.5 Residual Dipolar Couplings

During the past years, RDCs have become an increasingly important source of structural information [28, 29]. They can be measured in solution by weakly aligning the molecule using a variety of methods [30]. RDCs provide angular information between the internuclear vector for which they are measured and a set of globally defined axes in the molecule, namely those of the alignment tensor. The measured RDCs are given by:

$$D^i(\beta^i \alpha^i) = 0.5D_0[A_a(3 \cos^2 \beta^i - 1) + \frac{3}{2}A_r(\cos 2\alpha^i \sin^2 \beta^i)].$$

Here,  $A_a$  is the axially symmetric part of the alignment tensor, equal to  $[A_{zz} - 1/2(A_{xx} + A_{yy})]$  and  $A_r$  is the rhombic component of the alignment tensor, equal to  $(A_{xx} - A_{yy})$ , where  $A_{xx}$ ,  $A_{yy}$ , and  $A_{zz}$  are the  $x$ ,  $y$ , and  $z$ -components of the alignment tensor, respectively;  $\alpha^i$  and  $\beta^i$  are the azimuthal and polar angles of the vector for which the RDC is reported, in the frame of the alignment tensor.  $D_0$  is the strength of the (static) dipolar coupling defined as:

$$D_0 = -\left(\frac{\mu_0}{4\pi}\right) \frac{\gamma_i \gamma_j h}{2\pi^2 r_{NH}^3},$$

which, in the case of N-NH RDCs is equal to 21.7 kHz.  $r_{NH}$  is the length of the NH vector,  $\mu_0$  is the magnetic permeability of vacuum,  $\gamma_i$  is the gyromagnetic ratio of spin  $i$ , and  $h$  is Planck's constant. The structural information is contained in the angles  $\alpha$  and  $\beta$ ; note that if an RDC is measured between two atoms that are not at a fixed distance from each other, there is also a distance dependence (via the  $r$  term in  $D_0$ ). RDCs can be added as orientational restraints to the target function of the structure calculation algorithm [31]. Usually, only RDCs measured for internuclear vectors with a fixed distance are used.

### 2.1.6 Diffusion Anisotropy

Diffusion anisotropy (relaxation) data contain orientational information comparable to RDCs [32]. NMR relaxation is characterized by relaxation times  $T_1$  and  $T_2$ , and the ratio  $T_1/T_2$  can be used to define diffusion anisotropy restraints in NMR structure calculations [33]. Again, the orientation information comes from the angles of internuclear vectors in an external frame, which, in the case of diffusion anisotropy data, corresponds to the orientational diffusion tensor frame.

### 2.1.7 Paramagnetic Restraints


If a paramagnetic metal ion is present in a protein, the NMR signals of the nuclei in a shell around it will be affected [34] by several effects, including contact and pseudocontact shifts, relaxation rate enhancements, and cross-correlation effects.

In principle, these can provide both distance and orientation information. They have been implemented as restraints in various structure calculation software packages [35,36].

## 2.2 Structure Calculation Software

The experimental information sources discussed above can be used as restraints in the calculation process. They have been implemented in several computer programs, among which are CNS [11], Xplor-NIH [37], CYANA [38], SCULPTOR [39], the SANDER module of AMBER [40], and even GROMACS [41]. The most commonly used are CYANA and Xplor/CNS.

Structure calculations are usually based on some molecular dynamic simulated annealing (SA) protocol performed in torsion angle and/or Cartesian space, followed by a final refinement phase in explicit solvent (water). A general feature of these protocols is that they use a “target function” that measures how well the calculated structure fits the experimental data and the chemical information; the lower this function, the better the agreement. The chemical information is defined in the force field that contains terms such as bond length, bond angles, van der Waals interactions, etc. Often, the description of long-range nonbonded interactions is simplified to increase the speed of the calculations by considering only repulsions between atoms and neglecting electrostatic interactions. A full nonbonded representation, including van der Waals (Lennard-Jones) and electrostatic (Coulomb) interactions, is typically reintroduced for final refinement in explicit solvent [6].

 AQ: Please specify where above.

## 2.3 Structural Statistics and Structural Quality

The first step in structure validation is the selection of NMR structures from a large ensemble of calculated structures. The most widely used structure selection procedure is based on the agreement with the experimental data (small number of violations) and a low energy of the structures; typically ensembles of approximately 20 lowest energy models are selected, although this number is arbitrary. Ideally, the selected ensemble should represent the available conformational space accessible to the structure while satisfying the experimental restraints. From this ensemble, a representative structure is usually defined; no real consensus exists, however, on how it should be selected. We recommend selection of the structure that differs the least from all other structures within the ensemble, i.e., the closest to the average structure.

The final ensemble is subsequently subjected to structural validation to obtain an indication of the quality and structural statistics. In practice, several quality indicators are often used to assess the quality of the NMR ensembles, such as:

- The goodness of fit to the experimental data, by analyzing restraint violations
- The precision of the ensemble, measured by positional root mean square deviation (RMSD)
- Several chemical and stereochemical quality indicators that are generally used to assess the local and overall quality of protein structures, many of them based on knowledge from high resolution x-ray structures [42]

Table 1 lists the most commonly used validation programs. The use of some of these programs will be described later.

AQ: Please specify where described later.

**Table 1** Internet resources of NMR-related programs and databases mentioned in this chapter

Software	Internet address	Purpose
CNS	<a href="http://cns.csb.yale.edu/v1.1">http://cns.csb.yale.edu/v1.1</a>	Multilevel hierarchical approach for the most commonly used algorithms in macromolecular structure determination (NMR, crystallography)
RECOORD	<a href="http://www.ebi.ac.uk/msd-srv/docs/NMR/recoord/scripts.html">http://www.ebi.ac.uk/msd-srv/docs/NMR/recoord/scripts.html</a>	Database of recalculated NMR structures with the CNS scripts used in the tutorial example
HADDOCK	<a href="http://www.nmr.chem.uu.nl/haddock/">www.nmr.chem.uu.nl/haddock/</a>  (installation notes: <a href="http://www.nmr.chem.uu.nl/haddock/installation.html">http://www.nmr.chem.uu.nl/haddock/installation.html</a> )	High ambiguity driven protein–protein docking based on biochemical and/or biophysical information
PDB	<a href="http://www.rcsb.org/pdb/Welcome.do">http://www.rcsb.org/pdb/Welcome.do</a>	An information portal to biological macromolecules structures
BMRB	<a href="http://www.bmrwisc.edu">http://www.bmrwisc.edu</a>	Biological Magnetic Resonance Data Bank
CCPN	<a href="http://www.ccpn.ac.uk">http://www.ccpn.ac.uk</a>	A collaborative computing project for NMR
PROCHECK	<a href="http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html">http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html</a>	Checks the stereochemical quality of a protein structure, producing a number of PostScript plots analyzing its overall and residue-by-residue geometry
PROCHECK_NMR	<a href="http://www.biochem.ucl.ac.uk/~roman/procheck_nmr/manual/manprochint.html">http://www.biochem.ucl.ac.uk/~roman/procheck_nmr/manual/manprochint.html</a>	PROCHECK-NMR is a suite of programs that have been derived from the PROCHECK programs to analyze ensembles of protein structures solved by NMR
PROCHECK_COMP	<a href="http://www.biochem.ucl.ac.uk/~roman/procheck_comp/procheck_comp.html">http://www.biochem.ucl.ac.uk/~roman/procheck_comp/procheck_comp.html</a>	Compares residue-by-residue geometry of a set of closely related protein structures
WHATIF	<a href="http://swift.cmbi.kun.nl/whatif/">http://swift.cmbi.kun.nl/whatif/</a>  web server: <a href="http://swift.cmbi.kun.nl/WIWWWI">http://swift.cmbi.kun.nl/WIWWWI</a>	Versatile molecular modeling package that is specialized on working with proteins and the molecules in their environment such as water, ligands, nucleic acids, etc.

**Table 1** Continued

Software	Internet address	Purpose
WHATCHECK	<a href="http://swift.cmbi.ru.nl/gv/whatcheck">http://swift.cmbi.ru.nl/gv/whatcheck</a>	The protein verification tools from the WHAT IF program
MOLPROBITY	<a href="http://molprobity.biochem.duke.edu">http://molprobity.biochem.duke.edu</a>	Structure validation and all-atom contact analysis for proteins, nucleic acids, and their complexes
QUEEN	<a href="http://www.cmbi.kun.nl/software/queen/index.spy?site=queen&amp;action=Home">http://www.cmbi.kun.nl/software/queen/index.spy?site=queen&amp;action=Home</a>	Quantitative evaluation of experimental NMR restraints
TALOS	<a href="http://spin.niddk.nih.gov/bax/software/TALOS/info.html">http://spin.niddk.nih.gov/bax/software/TALOS/info.html</a>	Protein backbone angle restraints from searching a database for chemical shift and sequence homology
profit	<a href="http://www.bioinf.org.uk/software/profit/">http://www.bioinf.org.uk/software/profit/</a>	Least square-fitting program that performs the basic function of fitting one protein structure to another
NACCESS	<a href="http://wolf.bms.umist.ac.uk/naccess/nacwelcome.html">http://wolf.bms.umist.ac.uk/naccess/nacwelcome.html</a>	Stand-alone program that calculates the accessible area of a molecule from a PDB format file
Xmgrace	<a href="http://plasma-gate.weizmann.ac.il/Grace">http://plasma-gate.weizmann.ac.il/Grace</a>	Plotting tool
molmol	<a href="http://hugin.ethz.ch/wuthrich/software/molmol">http://hugin.ethz.ch/wuthrich/software/molmol</a>	Molecular graphics program
Rasmol	<a href="http://www.umass.edu/microbio/rasmol/index2.htm">http://www.umass.edu/microbio/rasmol/index2.htm</a>	Molecular graphics program

## 2.4 NMR-Based Modeling of Complexes

In principle, the structural information sources discussed above apply as well for structure calculation of protein–protein complexes. Again, NOEs are the most important information source, and, when available, RDCs are very useful as well [8, 43]. However, it is often difficult to obtain intermolecular NOEs, especially in the case of weakly interacting and transient complexes. For those cases, however, NMR remains a powerful method that provides several ways of mapping the interface between the components of a complex.

In one approach, the so-called chemical shift perturbation (CSP) experiment,  $^1\text{H}^{15}\text{N}$ -HSQC spectra of one  $^{15}\text{N}$ -labeled component of the complex are recorded in the absence and presence of increasing amounts of its partner [9]. Changes in chemical shift after addition of the partner reveal residues that are possibly involved in the interaction. In cross-saturation or saturation transfer (SAT) experiments [44], the observed protein is  $^{15}\text{N}$  labeled and perdeuterated with its amide deuterons exchanged back to protons, whereas the “donating” partner protein is unlabeled. Saturation of the unlabeled protein leads, by cross-relaxation mechanisms, to signal



AQ: Please specify where discussed above.

attenuation (typically monitored by  $^1\text{H}^{15}\text{N}$ -HSQC spectra) of those residues in the labeled protein that are in close proximity. Finally, in the case of paramagnetic systems, several of the above-mentioned paramagnetic effects can also be used to map interfaces [45].

To make use of those interface mapping data, NMR-based docking approaches have been developed [13,46–49]. One of these is HADDOCK [13], which combines a limited amount of (NMR-)data with docking in so called *data-driven docking*. Data-driven docking in HADDOCK follows a three-stage procedure:

1. Rigid body energy minimization.
2. Semiflexible refinement following an SA protocol during which increasing amounts of flexibility are allowed:
  - (a) High temperature rigid-body search
  - (b) Rigid body simulated annealing (SA)
  - (c) Semiflexible SA with flexible side chains at the interface
  - (d) Semiflexible SA with fully flexible interface (both backbone and side chains)
3. Final refinement in explicit solvent (water or DMSO).

During the docking, the (NMR-)data are introduced as “ambiguous interaction restraints” (AIRS). These are defined between active and passive residues, active residues being the residues that, based on the experimental data, have been identified to be involved in the interaction, and passive residues being their surface neighbors. An AIR is defined between each active residue and all active and passive residues of the partner protein. This restraint is only fulfilled when the active residue will make contact with at least one of the active or passive residues of the partner protein, which means that the restraint will indeed drive the docking.

The ranking of the docking solutions is performed using a “HADDOCK-score,” which is a combination of several terms including restraint energies, intermolecular energies (van der Waals and electrostatic), desolvation energy, buried surface area, etc.

### 3 Methods

In this tutorial section, we describe the procedures to generate various type of NMR restraints and their use in structure calculation using CNS with the RECOORD scripts. This will be followed by a description of the steps to be followed for NMR-based modeling of biomolecular complexes using HADDOCK. As a convention, the commands to be executed are highlighted in grey using a Courier font. Information regarding the various programs and web pages used in this tutorial can be found in Table 1.



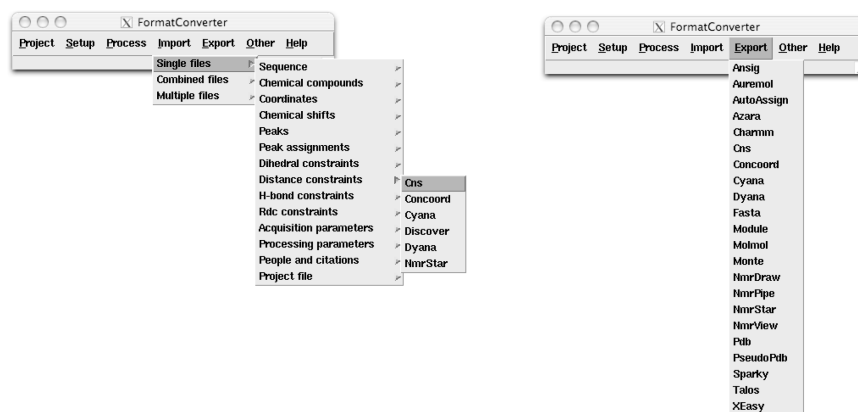
### 3.1 Restraint Generation

The keystone of NMR structure determination consists of three different types of experimental structure restraints: distance restraints, dihedral angle restraints, and orientational restraints.

#### 3.1.1 Distance Restraints

A cross-peak in a NOESY spectrum indicates special proximity between two nuclei. Thus, each peak can be converted into a maximum distance between the nuclei, usually between 1.8 and 6.0 Å. This distance can be obtained according to the intensity of the NOESY peak (proportional to the distance to the minus 6th power,  $1/r^6$ ). This intensity–distance relationship is only approximate, thus, usually a distance range is assumed. The assignment of the NOESY peaks to the correct nuclei based on the chemical shifts is of crucial importance. The manual detection of NOEs is an intensive and time-consuming job. Some programs, such as CANDID [50] and ARIA [12], can perform this task in an automated fashion coupled to the structure calculation protocol.

A common problem in NMR has been the limited availability of software allowing easy conversion between different data formats, which makes data exchange and use of different programs a tedious process. The CCPN Data Model for macromolecular NMR [51] is intended to cover all data needed for macromolecular NMR spectroscopy from the initial experimental data to the final validation. The ccpNmr FormatConverter application allows the import and export of data from and to a large variety of formats (Fig. 1).



**Fig. 1** Graphical user interface (GUI) layer of the CCPN FormatConverter (<http://www.ccpn.ac.uk>)

```

REMARK Ubiquitin input for TALOS, HA2/HA3 assignments arbitrary.
DATA SEQUENCE MQIFVKLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL
DATA SEQUENCE EDGRTLSDYN IQKESTLHLV LRLRGG
VARS  RESID RESNAME ATOMNAME SHIFT
FORMAT %4d %1s %4s %8.3f
      1 M      HA      4.23
      1 M      C      170.54
      1 M      CA     54.45
      1 M      CB     33.27
      2 Q      N     123.22
      2 Q      HA     5.25
      2 Q      C     175.92
      2 Q      CA     55.08
      2 Q      CB     30.76
      ...
     10 G      N     108.89
     10 G      HA2    4.35
     10 G      HA3    3.61
     10 G      C     174.07
     10 G      CA     45.46
      ...

```

Fig. 2 Example of the input shift table required in TALOS

### 3.1.2 Dihedral Angle Restraints

J-coupling and secondary chemical shifts can be used to define restraints on the torsion angles of the chemical bonds, typically the  $\phi$ ,  $\psi$ , and  $\chi_1$  angles can be generated and included in the protocol. They can be calculated applying the Karplus equation [21] to the measured J-couplings, or by using chemical shifts in programs such as TALOS [52] or CSI [53]. We will describe here the use of TALOS.

TALOS is a database system for empirical prediction of  $\phi$  and  $\psi$  backbone torsion angles using a combination of available chemical shifts ( $H\alpha$ ,  $C\alpha$ ,  $C\beta$ ,  $CO$ ,  $N$ ) for a protein sequence. To use TALOS, the following steps should be followed:

1. Create a directory for the predictions from where all the following commands will be executed.
2. Prepare the input table with the sequence and shift assignments in the required format. For preparing the input shift table required by TALOS (for example, see Fig. 2), we can again use the FormatConverter. In this case, we need a sequence file and the chemical shift table, and the program will export the table in the proper format for TALOS, taking into account naming conventions and shift referencing.
3. Run TALOS to perform the database search:

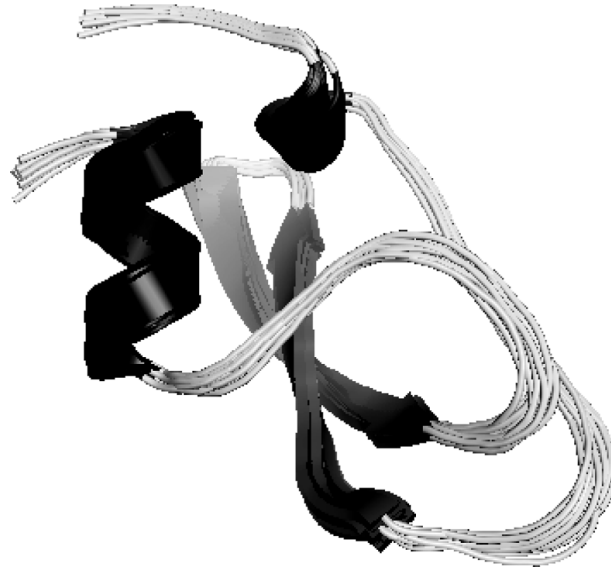
```
talos.tcl -in myshifts.tab
```

During the searching phase, a series of files will be created in "pred/res.\*.tab." Each of these files contains the 10 best matches in the database for a given residue. In addition, a file "pred.tab" is created, where a summary of the prediction results is stored.

4. Run VINA to summarize the results. This can be done with one of the following commands, depending whether a structure template is available or not:

```
vina.tcl -in myshifts.tab
```

```
vina.tcl -in myshifts.tab -ref mystructure.pdb
```



AQ: Please cite Figure 3 in text.



**Fig. 3** Ensemble of final water-refined structures for PDB entry 1bf0 recalculated with the RECOORD scripts

This step will adjust the individual prediction files to identify outliers in the detected matches and it will prepare a new summary file. This step is optional, because, in the previous step, a “pred.tab” was already created.

5. Run RAMA to inspect and adjust the predictions made by the program:

```
rama.tcl -in myshifts.tab  
rama.tcl -in myshifts.tab -ref mystructure.pdb
```

During the manual prediction, you will classify the results for a given residue as “Good,” “Ambiguous,” or “Bad.” For that purpose, you have to examine the  $\phi/\psi$  distributions of the detected matches and decide which ones should be included in the prediction and which ones are outliers. The prediction files will be overwritten to reflect any changes made interactively, and a final “pred.tab” will be created containing the classification and predictions (average and standard deviations) for the  $\phi$  and  $\psi$  angles for each residue.

To convert the TALOS predictions into CNS/Xplor restraints, we can use the perl script `talos2xplor.pl`, which can be obtained from the Biomolecular NMR laboratory at UAH (<http://daffy.uah.edu/nmr/analysis.html>). The script will ask you for an input a TALOS prediction file, the minimum  $\pm$  error (e.g.,  $20^\circ$ ) you want to include in the restraints, and an output CNS/Xplor restraint file.

### 3.2 NMR Structure Calculation and Refinement

For the structure calculation part, we are going to describe the use of the program CNS [11] with an SA protocol derived from ARIA [54], followed by refinement in explicit solvent [7]. All of the scripts mentioned in this section can be downloaded from the RECOORD webpage (see Table 1).

Start by creating a folder where you will run the calculations, download there the tar file containing the RECOORD scripts, and decompress it:

```
mkdir struct-calc
cd struct-calc/
wget http://www.ebi.ac.uk/msd-
srv/docs/NMR/recoord/files/RECOORDscripts.tgz
tar xzfv RECOORDscripts.tgz
```

In case the `wget` command does not work, use a web browser to download the scripts manually from the RECOORD webpage (see Table 1).

Before starting the calculations, you need to set up your current path for the scripts to work. To do this, you need to edit `changeScriptsDir.sh` found in the `RECOORDscripts/` folder, change the path for `newDir` in line 8 by your current path (you can find it by typing “`pwd`” in the shell) and execute it:

```
cd RECOORDscripts
nedit changeScriptsDir.sh #(change line 8 for pwd)
./changeScriptsDir.sh
cd ..
```

(`nedit` is a text editor; if not installed, use your preferred editor instead).

Most of the scripts use the CNS executable, so check that CNS is properly installed.

The last step is to setup a working directory, assigning a project name for the protein on which you are going to work. This project name will be used to generate the file names at the different stages of the protocol. We will use the `1bf0` structure [55] as an example, with the corresponding NMR restraints available for this entry from the BioMagResBank (BMRB) [56].

```
mkdir 1bf0
cd 1bf0
wget http://www.pdb.org/pdb/files/1bf0.pdb.gz
gunzip 1bf0.pdb.gz
```

The easiest way of obtaining the restraints in a format ready to be used in this protocol is to go to the BMRB from the PDB entry, select `4-filtered-FRED`

in the stage window, select the distance restraints in XPLOR/CNS format by clicking on it, and then click on “170823” in `mrblock_id` and copy and paste these restraints in a text file called `unambig.tbl` (see **Note 1**).

### 3.2.1 Generation of Molecular Topology Files

We can generate the molecular topology of the protein using the RECOORD script `generate.sh` (a modified version of the CNS script `generate_easy.inp`), either from the primary sequence or from a PDB coordinate file, depending on availability. Here, we will use the downloaded PDB file (see **Note 2**):

```
../RECOORDscripts/generate.sh 1bf0.pdb
```

If you give a name such as `1bf0.pdb`, a topology file called `1bf0_cns.mtf` will be generated. You should check the `ERRORS_generate` file created inside the `1bf0/` folder for possible errors. In this particular case, you can see that the script reported many errors but they are basically nomenclature errors and they can be ignored. A new `pdb` file called `1bf0_cns.pdb` is also generated with the proper CNS nomenclature (display the structure in your favorite molecule viewer to make sure that it looks reasonable and that the script worked properly).

### 3.2.2 Generation of Extended Starting Structure

The next step is the generation of an extended starting conformation, which will be used as input in the SA protocol. For this, use the RECOORD script, `generate_extended.sh`.

```
../RECOORDscripts/generate_extended.sh 1bf0_cns.mtf
```

Keeping the name given before, an extended structure called `1bf0_cns_extended.pdb` will be generated. In addition, it is advisable to check the `ERRORS_generate_extended` file in the same working directory for errors.

### 3.2.3 SA Stage

Use the script `annealing.sh` to start the structure calculation run. This script will generate a CNS parameter file (`run.cns`) with all details and specifications of the protocol that is used. The NMR restraints are contained in table files. We can use three different types of restraints, depending on their availability: `unambig.tbl` (NOE distance restraints), `hbonds.tbl`, and `dihedrals.tbl`. Note that the `annealing.sh` script should be run from a higher level than the previous two scripts.


```
cd ..
../RECOORDscripts/annealing.sh 1bf0
```

Individual job files will be generated and executed for each model you want to calculate. By default, two models will be generated in the created `str/` folder, with names similar to `1bof_cns_[1-2].pdb`. The CNS input and output files can be found in the directory `cnsRef/`, together with possible error files (see **Note 3**). The header of every PDB file generated contains information regarding violations and energy values.

### 3.2.4 Water Refinement Stage

Once the SA phase is finished and all resulting structures have been written into the `str/` directory, we can proceed to water refinement. For this purpose, we are going to use the script `re_h2o.sh`.

```
../RECOORDscripts/re_h2o.sh 1bf0
```

In the `str/` directory, a new directory called `wt/` will be created, that energy structures will be copied there and subsequently refined (see **Note 4**). 

## 3.3 Structure Validation and Quality Assessment

### 3.3.1 Restraint Violations

To obtain statistics regarding distance and dihedral angle violations for the water refined ensemble, use the scripts `calcViol.sh` and `analysViol.sh`, which analyze and summarize violations, respectively.

```
cd 1bf0
../RECOORDscripts/calcViol.sh 1bf0_cns str/wt/violations
0.3 convertOff 1bf0_cns.mtf unambig.tbl
```

where the input parameters correspond to the entry name (in this specific case, `1bf0_cns`), the directory in which the coordinate files can be found (in this case the directory containing the water refined structures `1bf0_cns_w_[1-25].pdb`), the violation distance cut-off (a frequently used value is  $0.3 \text{ \AA}$ ), the conversion switch to CNS format (it is optional and by default set to `convertOff`), the topology and restraint files (these are also optional, with, as the default, `'entryname'_cns.mtf` and `unambig.tbl`). Files with violations statistics will be created in the new `violations/` folder, with names as `viol_1bf0_cns_w_0.3`. Once the violations have been calculated, they can be analyzed with `analysViol.sh`:

```
../RECOORDscripts/analysViol.sh lbf0_cns violations
```

where `violations/` is the directory created previously with the `calcViol.sh` script. The results are summarized in the `violations` folder in the `viol_results` file.

### 3.3.2 Structural Validation

Various software tools are available to assess the stereochemical quality of the generated protein structures. Some of the most widely used packages are PROCHECK [57] and WHATIF [58]. Procheck provides a detailed graphical indication of the quality of a protein structure, giving an assessment of both the overall quality of the structure, as compared with well-defined structures of the same resolution, and of highlight regions that may need further investigation.

The command to run PROCHECK, once the program is properly installed, is:

```
procheck filename [chain] resolution
```

where `filename` indicates the coordinates file in Brookhaven format, `chain` is an optional one-letter chain-ID, in case several chains are included in the model, and `resolution` is a real number giving the resolution of the structure, to select the structures from the database to compare with our model.

Because PROCHECK only allows the analysis of a single structure at a time, it is worthwhile to also use PROCHECK\_COMP or PROCHECK\_NMR [59], a suite of programs that have been derived from the original PROCHECK programs, to compare, residue by residue, the geometry of a set of closely related protein structures, such as those in an NMR ensemble. To run PROCHECK\_COMP, you need to create a file, e.g., `lbf0.list`, containing the names of the structures you want to analyze, and then type the command:

```
procheck_comp lbf0.list
```

Both programs produce easily interpreted color postscript files that can be viewed using `ghostview` or similar programs. Type, for example, “`gs lbf0_01.ps`” to display the Ramachandran plot showing the  $\phi/\psi$  torsion angles for all residues in the structure. The coloring/shading on the plot represents the different regions: the darkest areas correspond to the “core” regions representing the most favorable combinations of  $\phi/\psi$  values. `lbf0_06.ps` shows various graphs and diagrams of protein geometrical properties as a function of the amino acid sequence, allowing you to possibly distinguish regions with normal geometry from those that might be poorly defined and present unusual geometry (see **Note 5**).

Another very useful protein validation tool is WHATCHECK, based on WHATIF, which also uses reference values from an x-ray database for most of the checks carried out. A great advantage of WHATCHECK is that the reference database of high-resolution protein structures is larger than in PROCHECK and continuously updated. Further, it provides many more checks and is more critical. WHATIF

is also available as a web server (see Table 1), where a variety of quality parameters can be obtained by uploading a PDB coordinates file to the server [42].

### 3.3.3 Precision of the Ensemble

The precision of a structure can be estimated by measuring the conformational variance over an ensemble of models. Usually, this variance has been expressed as the positional RMSD of the individual models from the mean structure. This parameter is useful for estimating the precision of the calculation, but does not report on the accuracy. The latter can only be calculated if a standard reference is available.

The positional RMSD from the mean and the prediction of secondary structure elements can be obtained using the molecular graphical program Molmol [60]. The least-square fitting program Profit can also perform the basic function of fitting a protein structure to another and allows for much more flexibility. It can be used to calculate the accuracy of structures, provided a reference structure is known. This program can be used in a direct interactive fashion in a terminal window or using scripts. A very simple script called here `profit.in` could be written as follows:

```
reference a.pdb
mobile b.pdb
! specifies the residues to fit on
in this case: 10-20 in the reference with 30-40 in
the mobile zone 10-20:30-40
! specifies the atom subsets for both
! fitting and RMS calculation.
atom CA,C,N
fit
! writes the fitted coordinates to a file
write b_fiton.a.pdb
quit
```

To execute it, simply type:

```
profit < profit.in
```

## 3.4 Modeling of Complexes by Data-Driven Docking Using HADDOCK

We describe here the use of the HADDOCK2.0 package (see Table 1) for the modeling of a protein-protein complex. In the following, we will use data from the



haddock2.0/examples/e2a-hpr directory. You should first copy this directory to the directory in which you are working (see **Note 6**):

```
cp -r $HADDOCK/examples/e2a-hpr?
```

### 3.4.1 Preparation of PDB Files and Input Data

If you are using an ensemble of structures, split the file such that each individual PDB file contains only one structure (see **Note 7**). As input data, you should combine CSP data (or other data indicating residues at the interface) and solvent accessibility data calculated with NACCESS; use only those residues that have both a high enough CSP and a high enough relative accessibility. In the example, the (average) per residue solvent accessibilities calculated with NACCESS are already provided in the files `e2a_1F3G.rsa` and `hpr/hpr_rsa_ave.lis` (the latter containing the average for the 10 starting structures for hpr). From these files, you can select the residues with high enough (e.g., >40–50%) accessibility (see **Note 8**). You could calculate the accessibility values yourself using the following command:

```
naccess e2a_1F3G.pdb
```

### 3.4.2 Definition of Active and Passive Residues

Passive residues are defined as the solvent-accessible surface neighbors of active residues. To define them you can display your molecule in a space-filling model using, for example, rasmol:

```
rasmol e2a_1F3G.pdb
```

and color the active residues, for example, in red. Then, filter out the residues having a low solvent accessibility and select all surface neighbors to define the passive residues (color them, for example, in green), which, again, you should filter with the solvent accessibility criterion. In the `e2a-hpr` example, several rasmol scripts are provided with the respective residues already colored according to this scheme:

```
e2a_rasmol_active.script, e2a_rasmol_active_passive.script
```

and similar for hpr.

You will use the active and passive residues for both molecules to generate AIRs; for this, go to the HADDOCK project setup section on <http://www.nmr.chem.uu.nl>, click on “generate AIR restraint file” and follow the instructions. You should save the resulting file as `ambig.tbl` in the working directory; note that, in the `e2a-hpr` example directory, `ambig.tbl` is already present (see **Note 9**).

### 3.4.3 Setup of a New Run: new.html

To set up a new run, return to the project setup page on <http://www.nmr.chem.uu.nl>, click on “start a new project” and follow the instructions. Depending on the experimental data you have available, you will input various data files, such as ambiguous restraints, unambiguous restraints, RDCs, etc. After saving the `new.html` file to disk, type “`haddock2.0`” in the same directory. This will generate a run directory containing all of the necessary information to run haddock. An example of a `new.html` file can be found in the `e2a-hpr` directory as `new.html-example` (see **Note 10**).

### 3.4.4 Run.cns

The next step is to define all parameters to perform the docking run. For this, enter the newly created directory:

```
cd run1
```

You will find a file called `run.cns` containing all the parameters to run the docking. You need to edit this file and define a number of project-specific parameters, such as the semiflexible segments at the interface or fully flexible segments and other parameters governing the structure docking (see **Note 11**). You can edit your `run.cns` file via “project setup” on <http://www.nmr.chem.uu.nl>. More information is available via the “run.cns” option in the manual section on <http://www.nmr.chem.uu.nl>.

### 3.4.5 Docking Run

To actually start the docking run with HADDOCK, in the directory containing the `run.cns` file (see **Note 12**) type:

AQ: Please check the edit of the sentence.

```
haddock2.0 >& haddock.out &
```

As more extensively explained in “The Docking” section in the HADDOCK manual, the entire protocol consists of four stages:

1. *Topologies and structures generation*: The resulting topologies (\*.psf) and coordinates (\*.pdb) files are written into the `begin/` directory (see **Note 13**).
2. *Randomization and rigid body energy minimization*: The generated docked structures are written into `structures/it0/`. When all structures have been generated, HADDOCK will write the PDB filenames sorted according to the criterion defined in the `run.cns` into `file.cns`, `file.list`, and `file.nam` in the `structures/it0` directory.
3. *Semiflexible SA*: The best 200 structures after rigid body docking (this number is defined in `run.cns` and can be modified) will be subjected to a semiflexible

SA in torsion angle space. The temperatures and number of steps for the various stages are defined again in the `run.cns` parameter file. The resulting refined structures are written into `structures/it1`. At the end of the calculation, HADDOCK generates the `file.cns`, `file.list`, and `file.nam` files containing the filenames of the generated structures sorted accordingly to the criterion defined in the `run.cns` parameter file (see Note 4). At the end of this stage, the structures are analyzed and the results are placed in the `structures/it1/analysis` directory (see Sects. 3.4.6 and 3.4.7).

4. *Flexible explicit solvent refinement.* The `re_h2o.inp` (or `re_dmsol.inp`, if the chosen solvent is DMSO) CNS script is used for this step. The resulting structures are written in the `structures/it1/water` directory. At the end of the explicit solvent refinement, HADDOCK generates the `file.cns`, `file.list`, and `file.nam` files containing the filenames of the generated structures sorted accordingly to the criterion defined in the `run.cns` parameter file. Finally, the structures are analyzed and the results are placed in the `structures/it1/water/analysis` directory (see Sects. 3.4.6 and 3.4.7).

AQ: Please check edited sentence.

AQ: Correct as edited.

### 3.4.6 Automatic Analysis

A number of analysis scripts are automatically run after the semiflexible and explicit solvent refinement stages, with the results placed into `structures/it1/analysis` and `structures/it1/water/analysis`, respectively. Here we discuss a few of the most relevant output files.

- `e2a-hpr_rmsd.disp`: Contains the pairwise RMSD matrix; this file is used as input for RMSD clustering.
- `noe.disp`: Contains the number of distance restraints violations per structure and averaged over the ensemble over all distance restraint classes and for each class (unambiguous, ambiguous, hbonds) separately. Comparable files are generated when you have RDC restraints (`sani.disp`) or relaxation data restraints (`dani.disp`).
- `energies.disp`: Contains the various energy terms per structure and averaged over the ensemble.
- `ana*.lis`: There is a set of files called `ana*.lis` where `*` can be `dihed-viol`, `dist-viol_all`, `hbond-viol`, `hbonds`, `nbcontacts`, `noe-viol_all`, `noe-viol_ambig`, or `noe-viol_unambig`. The “viol” refers to violations, and those files contain listings of violations, including the number of times a restraint is violated and the average distance and violation per restraint. In addition, `ana_hbonds.lis` gives a listing of hydrogen bonds, and `ana_nbcontacts.lis` gives a listing of nonbonded contacts.
- `ene-residue.disp`: Contains intermolecular energies for all interface residues.
- `nbcontacts.disp`: Contains nonbonded contacts.

### 3.4.7 Manual Analysis

An important part of the analysis needs to be performed manually. A number of scripts and programs are provided for this purpose in the `tools` directory. These allow collection of various statistics on the generated models and, more importantly, clustering of solutions and their analysis on a per-cluster basis.

- *Collecting statistics of the models with `ana.structure.csh`*: Copy this script from the `tools` directory into `structures/it1` or `structures/it1/water`. This script should be run once the `file.list` file has been created. It extracts from the various PDB files various energy terms, violation statistics, and the buried surface area, and calculates the RMSD of each structure compared with the lowest energy structure (if the location of ProFit is defined [see installation and software links on <http://www.nmr.chem.uu.nl/haddock>]). Several files called “`structures * .stat`” are created, which contain the same information but sorted in different ways. The most important file is `structures.haddock-sorted.stat`, which is sorted based on the HADDOCK-score. You can generate a plot of the HADDOCK-score as a function of the RMSD (using Xmgrace, for example). A script called `make_ene-rmsd_graph.csh` is provided in `$HADDOCKTOOLS` for this purpose. Specify two columns to extract data from and a filename:

```
$HADDOCKTOOLS/make_ene-rmsd_graph.csh 3 2
structures.haddock-sorted.stat
```

This will generate a file called `ene_rmsd.xmgr`, which you can display with `xmgrace`:

```
xmgrace ene_rmsd.xmgr
```

- *Clustering of solutions using `cluster_struct`*: The clustering is run automatically in `it1/analysis` and `it1/water/analysis` based on the criteria defined in the `run.cns` file. However, try using different cut-offs for the clustering because it is difficult to know *a priori* the best RMSD cut-off. This will depend on the system under study and the number of experimental restraints used to drive the docking (see **Note 15**).

`cluster_struct` reads the `e2a-hpr_rmsd.disp` file containing the pairwise RMSD matrix and generates clusters. The usage is (in the `analysis` directory):

```
cluster_struct [-f] e2a-hpr_rmsd.disp cut-off
min_cluster_size > cluster.out
```

Here, `cut-off` indicates the RMSD cut-off and `min_cluster_size` is the minimum number of structures in a cluster (typically a number like 4 or 5) (`-f` is optional, see **Note 16**).

The output looks like:

```
cluster 1 → 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
18 19 23 24 27 28 43
cluster 2 → 25 26 29 32 34 35 57 71 73 20 21 44 39 46
...
```

The numbers correspond to the structure number in the analysis file. For example, 2 corresponds to the second structure in analysis, i.e., the second structure in `file.list` in `it1` or `it1/water`.

- *Analysis of the clusters with `ana_cluster.csh`*: This script takes the output of `cluster_rmsd` to perform an analysis of the various clusters, calculating average energies, RMSDs, and buried surface area per cluster. To run it, type with as argument the output file of the clustering, e.g.:

```
$HADDOCKTOOLS/ana_cluster.csh [-best #]
analysis/cluster.out
```

The `-best #` is an optional argument to generate additional files with cluster averages calculated only on the best # structures of a cluster. The best structures are selected based on the criteria defined in `run.cns`, i.e., the sorting found in `file.list`. This allows removal of the dependency of the cluster averages based on the size of the respective clusters (see **Note 17**). The `ana_cluster.csh` script analyzes the clusters in a similar way as the `ana_structure.csh` script, but, in addition, generates average values over the structures belonging to one cluster. It creates a number of files for each cluster containing the cluster number `clustX` in the name (see **Note 18**). In addition, files containing various averages over clusters are created, `cluster_xxx.txt`; these contains the average and standard deviation of various terms such as intermolecular energy (`xxx = ene`) etc. In addition, files combining all of the above information and sorted based on various criteria are provided: `clusters.stat` that contains the various cluster averages, unsorted, and `clusters_xxx-sorted.stat`, where `xxx` is the energy term according to which the values are sorted (e.g., `xxx = ene` for intermolecular energy, etc.). The most relevant is `clusters_haddock-sorted.stat`.

- *Rerunning the HADDOCK analysis on a cluster basis*: Having performed the cluster analysis, you can now rerun the HADDOCK analysis for the best structures of each cluster to obtain various statistics on a “cluster bests” basis. For this, one needs the cluster-specific `file.nam_clust#`, `file.list_clust#`, and `file.cns_clust#` files. A script called `make_links.csh` is provided that will move the original `file.nam`, `file.list`, and `file.cns` files to `file.nam_all`, `file.list_all`, `file.cns_all`, and the same with the analysis directory. It will then create links to the appropriate files (`file.nam_clust#`, ...) and to a new `analysis_clust#` directory. For example, to rerun the analysis for the best 10 structures of the first cluster type in the water directory:

```
$HADDOCKTOOLS/make_links.csh clust1_best10
cd ../../..
haddock2.0
```

The `cd` command brings you back into the main run directory, from where you again start HADDOCK. Only the analysis of the best 10 structures of the first cluster in the water will be run. Once finished, go to the respective analysis directory and inspect the various files. The RMSD from the average structures should now be low (check `rmsave_disp`).

Having run the HADDOCK analysis on a cluster basis for each cluster, you should now have new directories in the water directory, called `analysis_clustX_best10`. Each analysis directory now contains cluster-specific statistics. You can also visualize the clusters. For Rasmol, first use the `joinpdb perl` script to concatenate the various PDB files into one single file:

```
$HADDOCKTOOLS/joinpdb -o e2a-hpr_clust1.pdb e2a-
hprfit_*.pdb
rasmol - nmrpdb e2a-hpr_clust1.pdb
```

In general, the cluster with the lowest HADDOCK score will be considered the best model. Scoring in docking is, however, a difficult problem and we recommend, if possible, the use of additional information for validation, such as, for example, mutagenesis data, if available. The selected model should explain as much as possible what is known about the system.

## 4 Notes

1. If dihedrals or any other types of restraints are available, they can be obtained in a similar way. The names assigned will be `dihedrals.tbl` and `hbonds.tbl`.
2. This only works if a PDB coordinates file is available. Otherwise, use `generate_seq.inp` and `generate_template.inp` from CNS to create such a PDB.
3. Once you have everything set up in a proper way to work, you can edit the script and make some changes for some protocol parameters. You can, for example, change the number of models to generate. It is set to 2 by default, but more common numbers would be 100 or 200. For systems that are more complex, you can switch to a longer annealing protocol, by doubling the number of steps to be carried out. Depending on whether you are going to use a cluster or your own computer, you should change the submit command. Remember also to change the sleeping time between submitting jobs, especially if you are not using a cluster and you do not want to have 100 jobs running on your computer at the same time! In such a case, choose a sleep time that matches the time needed for one structure calculation.

4. You should also edit this script and change the number of structures to refine because, by default, it is set to only 1. Increase this number to 25 models. They will be assigned names such as `1bf0_cns_w_[1-25].pdb`. The CNS input and output will be directed to the directory `cnsWtRef/`.
5. For visualizing these plots after running `procheck_comp`, the number tag is kept for the Ramachandran plot, however, for the residue properties plot, the number tag is now `_07.ps`.
6. The `$HADDOCK` environment variable should be defined if HADDOCK was properly installed.
7. Make sure that the format of the PDB files containing your starting structures is correct. There should be an END statement at the end, and there should be no SEGID (the SEGID is a four character long string at columns 73–76 in the PDB format) or ChainID (the ChainID is a chain identifier following the residue name in column 22). If you use a crystal structure, make sure that there are no missing residues.

Another point concerns ions; if proper care is not taken, they can give problems in torsion angle dynamics. To deal with this, the script `covalions.cns` defines artificial bonds to connect the ion to the protein. If you have another ion than is defined in the first line of the script, add it there. In addition, make sure that their name in the PDB file matches the ion name defined in the `ion.top` file in the `toppar` directory. To avoid having a N- or C-terminal patch applied to them, they should also be defined in the `topallhdg5.3.pep` file (look for the “first IONS” and “last IONS” statements).

8. The cut-off is not a hard limit; check the accessibilities and possibly include residues with lower accessibilities but functionally important groups.
9. Distance restraints can be used in HADDOCK in `ambig.tbl` or `unambig.tbl`. These are treated in the same way, except that the random removal option (`noecv=true`) only is applied to `ambig.tbl`. By default, one would use `ambig.tbl`; `unambig.tbl` could be used, for example, to provide extra NOEs or other data for which one wants to use different force constants.
10. An important setting in `new.html` is the value of `N_COMP`. This should be set to be equal to the number of components of the complex (two in case of a dimer, three for a trimer, etc.). Note that it can also be set to one, in which case, HADDOCK could be used for refinement instead of docking.
11. HADDOCK allows the definition of fully flexible regions: these are treated as fully flexible throughout all stages, except the initial rigid-body docking. This should be useful for cases in which part of a structure is disordered or unstructured or when docking small flexible molecules onto a protein. This option also allows the use of HADDOCK for structure calculations of complexes when classical NMR restraints are available to drive the folding.
12. This causes the HADDOCK program to run in the background. If, at some stage, HADDOCK stops producing new structures and the run is not yet finished, search for error messages in the output files: `gunzip xxx.out.gz` where `xxx.out.gz` is a particular output file, and look for ERR in this file.

Also, kill the current HADDOCK process:

```
ps -ef | grep haddock  
kill -9 id
```

Here, `id` is the process id that is returned by the `ps -ef` command.

13. The OPLS force field used by HADDOCK is a mixed united/all-atom force field; all atoms, including protons, are described; the later, however, do not have vdw parameters but are accounted for in the carbon parameters to which they are attached. From version 2.0 of HADDOCK, nonpolar hydrogen atoms are deleted by default to speed up the calculation; this does not really affect the resulting structures because the missing hydrogens are actually accounted for in the united atoms parameters. You can change this behavior by setting `delenph=true` in `run.cns`. This should be performed if classical NOE distance restraints are used.
14. A typical error would be that only one or two structures in `it1` are not successfully calculated. Often, you can cope with this by changing the random seed in `run.cns` (`iniseed`) and restart HADDOCK. Otherwise, try to decrease the `timestep` (e.g., 0.001 instead of 0.002). If none of this works, simply copy the missing structures from the `it0` directory so that the run can proceed.
15. For the RMSD calculation, the structures are superimposed on the interface backbone atoms of molecule A and the RMSD is calculated on the interface backbone atoms of molecule B; this might be called ligand interface RMSD. The resulting RMSD values are larger than would be obtained by fitting the whole molecule, which explains the large cutoff value that is used by default (7.5 Å). If only a small fraction of the structures do fall into clusters, try increasing the cut-off.
16. The `-f` option stands for full linkage, a method that generates larger clusters in which the structures within a cluster can, thus, differ more.
17. It is better to use a small number of structures (e.g., five) for comparison of the clusters than to use all structures of each cluster, because, in this way, the comparison will not depend on the cluster size.
18. The ordering of the structures in the `file.nam.clustXX` files comes from the clustering. The PDB files might, therefore, no longer be sorted accordingly to a defined criterion.

## References

1. Wüthrich, K., *NMR of proteins and nucleic acids*. Wiley: New York, 1986.
2. Neuhaus, D. and Williamson, M. P., *The nuclear Overhauser effect in structural and conformational analysis*. John Wiley & Sons: 2000.
3. Altona, C., Vicinal coupling constants & conformation of biomolecules. In *Encyclopedia of Nuclear Magnetic Resonance*, Harris, D. M. G. a. K. R., Ed. John Wiley, London: 1996; pp 4909–4922.



4. Bax, A., Kontaxis, G. and Tjandra, N. (2001) Dipolar couplings in macromolecular structure determination. *Methods in Enzymology* **339**, 127–174.
5. Guntert, P. (1998) Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics* **31**, 145–237.
6. Linge, J. P., Williams, M. A., Spronk, C. A. E. M., Bonvin, A. M. J. J. and Nilges, M. (2003) Refinement of protein structures in explicit solvent. *Proteins* **50**, 496–506.
7. Nederveen, A. J., Doreleijers, J.F., Vranken, W.F., Miller, Z., Spronk, C.A.E.M, Nabuurs, S.B., Güntert, P., Livny, M., Markley, J.L., Nilges, M., Ulrich, E.L., Kaptein, R., and Bonvin, A.M.J.J. (2005) Recoord: A recalculated coordinates database of 500+ proteins from the pdb using restraint data from the biomagresbank. *Proteins: Struct. Funct. & Bioinformatics* **59**, 662–672.
8. Bonvin, A. M. J. J., Boelens, R. and Kaptein, R. (2005) NMR analysis of protein interactions. *Current Opinion in Chemical Biology* **9**, 501–508.
9. Zuiderweg, E. R. (2002) Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* **41**, 1–7.
10. van Dijk, A. D. J., Boelens, R., and Bonvin, A. M. J. J. (2005) Data-driven docking for the study of biomolecular complexes. *Febs Journal* **272**, 293–312.
11. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., and Warren, G.L. (1998) Crystallography and NMR system (CNS): A new software system for macromolecular structure determination. *Acta Crystallogr. D Biol.* **54**, 905–921.
12. Linge, J. P., Habeck, M., Rieping, W. and Nilges, M. (2003) Aria: Automated NOE assignment and NMR structure calculation. *Bioinformatics* **19**, 315–316.
13. Dominguez, C., Boelens, R. and Bonvin, A. M. J. J. (2003) Haddock: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731–1737.
14. Neal, S., Nip, A. M., Zhang, H. Y. and Wishart, D. S. (2003) Rapid and accurate calculation of protein h-1, c-13 and n-15 chemical shifts. *Journal of Biomolecular NMR* **26**, 215–240.
15. Xu, X. P. and Case, D. A. (2001) Automated prediction of <sup>15</sup>N, <sup>13</sup>C $\alpha$ , <sup>13</sup>C $\beta$  and <sup>13</sup>C' chemical shifts in proteins using a density functional database. *Journal of Biomolecular NMR* **21**, 321–333.
16. Williamson, M. P., Kikuchi, J. and Asakura, T. (1995) Application of h-1-NMR chemical-shifts to measure the quality of protein structures. *Journal of Molecular Biology* **247**, 541–546.
17. Meiler, J. (2003) Proshift: Protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR* **26**, 25–37.
18. Clore, G. M. and Gronenborn, A. M. (1998) New methods of structure refinement for macromolecular structure determination by NMR. *Proceedings-National Academy of Sciences USA* **95**, 5891–5898.
19. Luginbuehl, P., Szyperski, T. and Wüthrich, K. (1995) Statistical basis for the use of <sup>13</sup>C $\alpha$  chemical shifts in protein structure determination. *Journal of Magnetic Resonance Series B* **109**, 229.
20. Kuszewski, J., Qin, J., Gronenborn, A. M. and Clore, M. G. (1995) The impact of direct refinement against <sup>13</sup>C $\alpha$  and <sup>13</sup>C $\beta$  chemical shifts on protein structure determination by NMR. *Journal of Magnetic Resonance Series B* **106**, 92.
21. Karplus, M. (1962) Vicinal proton coupling in nuclear magnetic resonance. *Journal American Chemistry Society* **84**, 2870–2871.
22. Kim, Y. P., J. H. (1990) Refinement of the NMR structures for acyl carrier protein with scalar coupling data. *Proteins* **8**, 377–385.
23. Torda, A. E., Brunne, R. M., Huber, T. and Kessler, H. (1993) Structure refinement using time-averaged j-coupling constant restraints. *Journal of Biomolecular NMR* **3**, 55.
24. Wagner, G. and Wüthrich, K (1982) Amide proton exchange and surface conformation of the basic pancreatic trypsin inhibitor in solution. *Journal Molecular Biology* **160**, 343–361.
25. Pervushin, K., Ono, A., Fernandez, C., Szyperski, T., Kainosho, M. and Wüthrich, K. (1998) NMR scalar couplings across Watson-Crick base pair hydrogen bonds in DNA observed by transverse relaxation-optimized spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14147–14151.

AQ: Author name seem to be missing. Please check.

26. Cordier, F., Rogowski, M., Grzesiek, S. and Bax, A. (1999) Observation of through-hydrogen-bond  $2h_{jhc}$  in a perdeuterated protein. *Journal of Magnetic Resonance* **140**, 510–512.
27. Bonvin, A. M. J. J., Houben, K., Guenneugues, M., Kaptein, R. and Boelens, R. (2001) Rapid protein fold determination using secondary chemical shifts and cross-hydrogen bond  $^{15}\text{N}$ - $^{13}\text{C}'$  scalar couplings ( $^3h_{jnc}$ ). *Journal of Biomolecular NMR* **21**, 221–233.
28. Bax, A. (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Science* **12**, 1–16.
29. Bax, A. and Grishaev, A. (2005) Weak alignment NMR: A hawk-eyed view of biomolecular structure. *Current Opinion in Structural Biology* **15**, 563–570.
30. Prestegard, J. H., Bougault, C. M. and Kishore, A. I. (2004) Residual dipolar couplings in structure determination of biomolecules. *Chemical Reviews* **104**, 3519–3540.
31. Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M. and Bax, A. (1997) Use of dipolar  $^1\text{H}$ - $^{15}\text{N}$  and  $^1\text{H}$ - $^{13}\text{C}$  couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Structural Biology* **4**, 732–738.
32. Fushman, D., Varadan, R., Assfalg, M. and Walker, O. (2004) Determining domain orientation in macromolecules by using spin-relaxation and residual dipolar coupling measurements. *Progress in Nuclear Magnetic Resonance Spectroscopy* **44**, 189–214.
33. Tjandra, N., Garrett, D. S., Gronenborn, A. M., Bax, A. and Clore, G. M. (1997) Defining long range order in NMR structure determination from the dependence of heteronuclear relaxation times on rotational diffusion anisotropy. *Nature Structural Biology* **4**, 443–449.
34. Bertini, I., Luchinat, C., Parigi, G. and Pierattelli, R. (2005) NMR spectroscopy of paramagnetic metalloproteins. *Chembiochem* **6**, 1536–1549.
35. Banci, L., Bertini, I., Cavallaro, G., Giachetti, A., Luchinat, C. and Parigi, G. (2004) Paramagnetism-based restraints for xplor-nih. *Journal of Biomolecular NMR* **28**, 249–261.
36. Bertini, I., Luchinat, C. and Parigi, G. (2002) Paramagnetic constraints: An aid for quick solution structure determination of paramagnetic metalloproteins. *Concepts in Magnetic Resonance* **14**, 259–286.
37. Schwieters, C. D., Kuszewski, J. J. and Clore, G. M. (2006) Using xplor-nih for NMR molecular structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy* **48**, 47–62.
38. Guntert, P., Mumenthaler, C. and Wuthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program dyana. *Journal of Molecular Biology* **273**, 283–298.
39. Hus, J. C., Marion, D. and Blackledge, M. (2000) De novo determination of protein structure by NMR using orientational and long-range order restraints. *Journal of Molecular Biology* **298**, 927–936.
40. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B. and Woods, R. J. (2005) The amber biomolecular simulation programs. *Journal of Computational Chemistry* **26**, 1668–1688.
41. Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. and Berendsen, H. J. C. (2005) Gromacs: Fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701–1718.
42. Spronk, C. A. E. M., Nabuurs, S. B., Krieger, E., Vriend, G. and Vuister, G. W. (2004) Validation of protein structures derived by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **45**, 315–337.
43. Clore, G. M. (2000) Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc Natl Acad Sci U S A* **97**, 9021–9025.
44. Takahashi, H., Nakanishi, T., Kami, K., Arata, Y. and Shimada, I. (2000) A novel NMR method for determining the interfaces of large protein-protein complexes. *Nature Structural Biology* **7**, 220–223.
45. Sakakura, M., Noba, S., Luchette, P. A., Shimada, I. and Prosser, R. S. (2005) An NMR method for the determination of protein-binding interfaces using dioxygen-induced spin-lattice relaxation enhancement. *Journal of the American Chemical Society* **127**, 5826–5832.
46. Clore, G. M. and Schwieters, C. D. (2003) Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from  $^1\text{H}/^{15}\text{N}$  chemical shift mapping and backbone  $^{15}\text{N}$ - $^1\text{H}$  residual dipolar couplings using conjoined rigid body/torsion angle dynamics. *J Am Chem Soc* **125**, 2902–2912.

47. Dobrodumov, A. and Gronenborn, A. M. (2003) Filtering and selection of structural models: Combining docking and NMR. *Proteins* **53**, 18–32.
48. Fahmy, A. and Wagner, G. (2002) Treedock: A tool for protein docking based on minimizing Van der Waals energies. *J Am Chem Soc* **124**, 1241–1250.
49. McCoy, M. A. and Wyss, D. F. (2002) Structures of protein-protein complexes are docked using only NMR restraints from residual dipolar coupling and chemical shift perturbations. *J Am Chem Soc* **124**, 2104–2105.
50. Herrmann, T., Guntert, P. and Wuthrich, K. (2002) Protein NMR structure determination with automated NOE assignment using the new software candid and the torsion angle dynamics algorithm dyana. *Journal of Molecular Biology* **319**, 209–227.
51. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J. and Laue, E. D. (2005) The ccpn data model for NMR spectroscopy: Development of a software pipeline. *Proteins* **59**, 687–696.
52. Cornilescu, G., Delaglio, F. and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR* **13**, 289–302.
53. Wishart, D. S. and Sykes, B. D. (1994) Chemical shifts as a tool for structure determination. *Methods In Enzymology*, 363.
54. Linge, J. P., O'Donoghue, S. I. and Nilges, M. (2001) Automated assignment of ambiguous nuclear Overhauser effects with aria. *Methods in Enzymology* **339**, 71–90.
55. Gilquin, B., Lecoq, A., Desne, F., Guenneugues, M., Zinn-Justin, S. and Menez, A. (1999) Conformational and functional variability supported by the bpti fold: Solution structure of the ca<sup>2+</sup> channel blocker calcicludine. *Proteins* **34**, 520–532.
56. Seavey, B. R., Farr, E. A., Westler, W. M. and Markley, J. L. (1991) A relational database for sequence-specific protein NMR data. *Journal Of Biomolecular NMR* **1**, 217–236.
57. Laskowski, R. A., Macarthur, M. W., Moss, D. S. and Thornton, J. M. (1993) Procheck—a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**, 283–291.
58. Vriend, G. (1990) What if—a molecular modeling and drug design program. *Journal of Molecular Graphics* **8**, 52–56.
59. Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R. and Thornton, J. M. (1996) Aqua and procheck-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR* **8**, 477–486.
60. Koradi, R., Billeter, M. and Wüthrich, K. (1996) Molmol: A program for display and analysis of macromolecular structures. *Journal Molecular Graphics* **14**, 51–55.

