NMR structure calculation with ARIA

EMBO Practical Course, 2013, Basel

A. Using Xeasy data

Benjamin Bardiaux and Michael Nilges
Unité de Bio-Informatique Structurale, Institut Pasteur, Paris, France
bardiaux@pasteur.fr; nilges@pasteur.fr

Abstract

In this practical, we will calculate the structure of the Tudor domain (Selenko P, Sprangers R, Stier G, Buhler D, Fischer U, Sattler M., 2001, Nat Struct Biol 8:27-31) with ARIA2. The data comprise two NOESY specta, torsion angles from coupling constants, hydrogen bond restraints, and residual dipolar couplings.

1 Start ARIA

ARIA is written in the python language and need a valid python version with some particular packages as Numpy. For this practical, this installation is already done, but you can get more details about it on the ARIA web site `http://www.aria.pasteur.fr`.

First, setup the environment for ARIA:

1. Open a terminal window
2. Type the following command in the terminal

```
source setup-aria.sh
```

3. Unpack the archive containing the data for the tutorial by typing

```
tar -xzf nilges.tgz
```

4. And go to the `example/` directory

```
cd example
```

2 Data conversion

To create an ARIA project, you need:

- The sequence of your protein
- Chemical shifts assignments
- NOESY peaks-lists
- Additional NMR data if available

ARIA can import NMR data from various NMR formats (requires a conversion step) or directly from a CCPN project. The first option is the one described in this tutorial, by first converting the data from various format to the internal ARIA XML format, and then import them into an ARIA project.

Since the original data are not stored in ARIA XML format, the first thing that needs to be done is to convert the NOE spectrum files from the original XEASY format to the ARIA XML format. The conversion routines check the data for consistency, and convert the atom names to strict IUPAC convention. ARIA uses the XML format to describe and store most of the crucial input files.

1. To do the data conversion, you first need to create a template conversion XML file.
2. Type the following ARIA command in a terminal to create such template

```
aria2 --convert -t conversion.xml
```

3. Open this file (`conversion.xml`) with your favorite text editor (e.g. `gedit`) to specify the missing information.
4. Specify the name of your project and the name of the project file (e.g., `aria_project.xml`)

```
<project name="tudor">
   <output
     filename="aria_project.xml"/>
</project>
```

5. Specify the details of the molecule as follow

```
<molecule molecule_type="PROTEIN"
  molecule_name="tudor"
  molecule_segid="A"
  first_residue_number="12">
  <input
    filename="./data/sequence/tudor.seq"
    format="seq"
    naming_convention=""/>
  <output
    filename="./xml/tudor.xml"/>
</molecule>
```

6. Specify the details of the first available peak-list (13C 3D NOESY) as follow

```
<spectrum
  spectrum_name="13C"
  spectrum_type=""
  spectrum_ambiguity="intra"
  segids="A">
  <chemical_shifts>
    <input
      filename="./data/spectra/13Cnoesy.prot"
      format="xeasy"/>
    <output
      filename="./xml/13C_ppm.xml"/>
  </chemical_shifts>
  <cross_peaks>
    <input
      filename="./data/spectra/13Cnoesy.peaks"
      format="xeasy"
      proton1="1"
      hetero1="2"
      proton2="3"
      hetero2=""/>
    <output
      filename="./xml/13C_peaks.xml"/>
  </cross_peaks>
</spectrum>
```

7. Since we also use a 15N 3D NOESY peak-list, duplicate the block <spectrum> ... </spectrum> and modifiy the relevant input fields. Note that proton dimensions are here inverted

```
<spectrum
  spectrum_name="15N"
  spectrum_type=""
  spectrum_ambiguity="intra"
  segids="A">
  <chemical_shifts>
    <input
      filename="./data/spectra/15Nnoesy.prot"
      format="xeasy"/>
    <output
      filename="./xml/15N_ppm.xml"/>
  </chemical_shifts>
  <cross_peaks>
    <input
      filename="./data/spectra/15Nnoesy.peaks"
      format="xeasy"
      proton1="3" # Proton 1 is now column 3
      hetero1="1"
      proton2="2" # Proton 2 is now column 2
      hetero2=""/>
    <output
      filename="./xml/15N_peaks.xml"/>
  </cross_peaks>
</spectrum>
```

8. save the file `conversion.xml`.
9. create the `xml/` directory where converted files will be stored

```
mkdir xml/
```

10. start the data conversion by running ARIA with

```
aria2 --convert conversion.xml
```

Running this command will create 6 XML files (5 in the `xml/` directory and 1 in the same directory where you executed the command):

- the sequence file `tudor.xml` (in the `xml/` directory).
- the 13C spectrum files `13C_ppm.xml` and `13C_peaks.xml` (in the `xml/` directory).
- the 15N spectrum files `15N_ppm.xml` and `15N_peaks.xml` (in the `xml/` directory).
- the ARIA project file `aria_project.xml` (in the current directory).

The data conversion needs to be done once for every structure calculation project and does not need to be repeated for repeated runs.

3 Create your ARIA project

ARIA provides a Graphical User Interface (GUI) to fill in the specified information (Input data and location of the output). The GUI indicates where information are missing by a red exclamation mark. You also need to set parameters for ARIA and, if you want, to modify parameters for CNS (Figure 1).

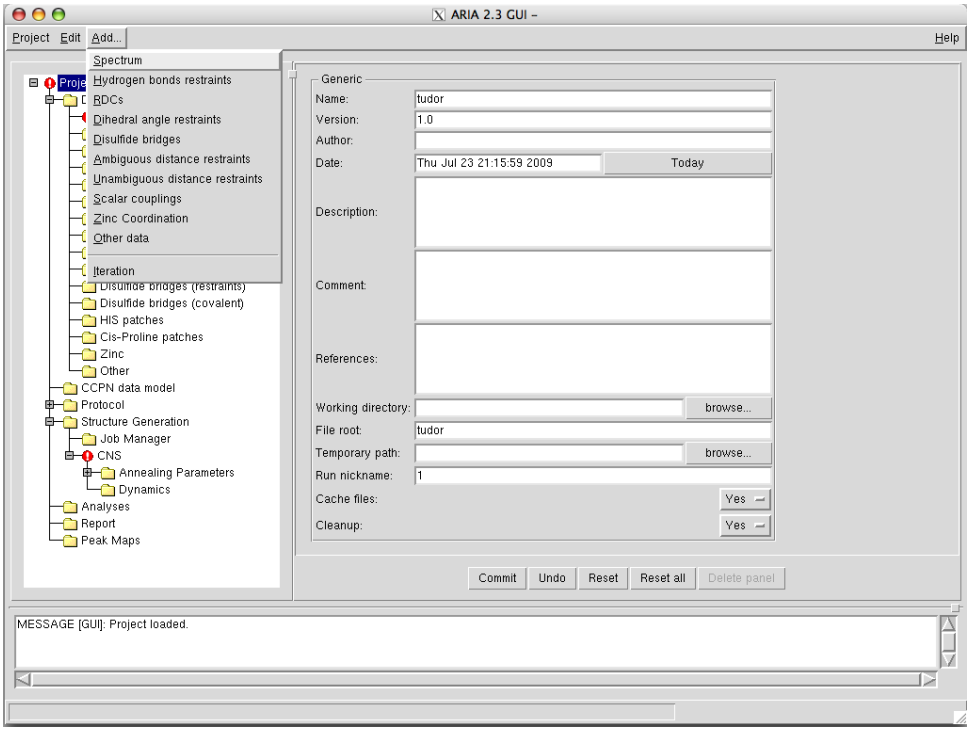

Figure 1: ARIA **G**raphical **U**ser **I**nterface (GUI).

1. Start the ARIAl user Interface (GUI) to edit the project file `aria_project.xml`, with the command

   ```
   aria2 -g aria_project.xml
   ```

2. Fill in the missing information in the `Project` panel
3. Enter `tudor` in the `File Root` field
4. Set `Working directory` and `Temporary path` to the `example/` directory in your home directory.
5. To specify the chemical shift tolerances for the different NOE spectra, click on the respective spectrum (`Data`→`Spectra`→`#xxx`).
6. Leave the the default value (0.5 ppm) for the heteronuclear dimensions.
7. For the proton dimensions, use 0.04 ppm for the indirect dimension and 0.02 ppm for the direct dimension.

   Note: The direct dimension corresponds to proton2/hetero2 for the 13C spectrum and proton1/hetero1 for the 15N spectrum.

8. For each spectra, turn off the use of the manual assignments by setting `Use manual assignment` to `No`. Otherwise, leave `Use manual assignment` to `Yes` if you want ARIA to use the NOE assignment already present in the peak list. You may also want to remove the diagonal peaks by enabling the `Filter diagonal peaks` option.

9. To add the remaining experimental data (hydrogen bonds, dihedral angle restraints and RDC) to the project file, use the menu **Add...** (Fig. 1). The data we use in this tutorial are stored in the following files:

   Hydrogen bonds `./data/hbonds/hbonds.tbl`

   Dihedral angles `./data/dihedrals/3J.tbl`

   RDCs `./data/rdcs/rdc2.tbl`

10. Specify the alignment tensor for the RDC data. Go to `Structure Generation` → `CNS` → `Annealing Parameters` → `RDCs` → `class 1`. Rhombicity is 0.22 and magnitude 7.5.

11. Go to node `Structure Generation` → `CNS` and fill in with the path to the CNS program which is used to perform the actual structure calculation.

```
/Bis/home/bardiaux/bin/cns1.21.exe
```

12. Save this project into the file `aria_project.xml` in the `example/` directory (Menu `Project`→`Save as...`)

13. Quit the GUI (Menu `Project`→`Exit`)

Analysis parameters

- Structures quality can be analyzed with `Whatif`, `Procheck`, `Prosa` and `Molprobity`. One could enter the corresponding paths in the node `Analyses`. To save time, these analysis have been already performed for this example.
- To see a 2-dimensional "Peak Map" in the GUI after the structures calculation, you need to enable the `python pickle output` in the node `Report`.

Do not forget to **save** the project as `aria_project.xml` in the `example/` directory.

4 Running ARIA

1. Setup the ARIA project directory by typing

```
aria2 --setup aria_project.xml
```

   This command will create the working directory (`run1/`) and copy all files needed to perform structures calculation.

2. To **speed up the calculation for the practical**, copy the pre-calculated structures to your directory tree:

```
cp -r ../results/structures ./run1/
```

3. Then start ARIA with

```
aria2 aria_project.xml
```

First, the data are filtered for errors and inconsistencies. ARIA then proceeds with creation of the molecular topology. Afterwards a initial assignment of NOE cross-peaks is derived from the chemical shift lists. Finally, ARIA determines which files are missing and calculates them.

Time for a break ?

5 Analysing ARIA output

5.1 PDB files

For each iteration, you will find the PDB files of the generated structures in their respective directories `run1/structures/itxxx`.

Visualise the ensemble of solvent-refined structures with the program **Pymol***:*
`pymol run1/structures/refine/fitted_tudor_water.pdb`

To show secondary structures, type the command **dss** in Pymol and then choose S → as → cartoon in the right menu. You can visualise all superimposed structures with **Movie** → `Show all states`.

5.2 Text output

ARIA generates various output files to report analysis results. For every iteration, the program creates the following report files:

1. `noe_restraints.unambig` and `noe_restraints.ambig`
   These files tabulate unambiguous and ambiguous restraints, respectively. Restraints discarded by the merging procedure are excluded. For every restraint, information is given on its reference cross-peak, restraint bounds, the average distance found in the ensemble and the result of violation analysis.
2. `noe_restraints.violations`
   Lists all violated restraints sorted with respect to their upper bound violations and contains the same information as (1).
3. `noe_restraints.assignments`
   Lists primarily restraint-wise assignments and gives information whether the assignment(s) stem from fully, partially or unassigned cross-peaks.
4. `noe_restraints.merged`
   Reports all restraints that have been discarded by the merging procedure.
5. `report`
   Summarises analyses of the restraint lists and the structure ensemble.

Look at the report for the last iteration `run1/structures/it8/report`.

- *How many peaks remain ambiguous ?*
- *What is the precision of the final structure ensemble ?*

Quality checks. ARIA uses the programs WHAT IF, PROCHECK, PROSA II and MOLPROBITY to evaluate the quality of both the final set of structures and the solvent-refined ensemble. For every program separate report files, `quality checks.*`, are stored in the directories of the respective ensembles.

Compare the overall quality scores (`quality_checks`*) for the last iteration (*`run1/structures/it8/`*) and for the water refinement (*`run1/structures/refine/`*).*

Miscellaneous analyses. Several CNS scripts calculate restraint energies, ensemble RMSDs, and an average structure. Results are stored in /analysis/.

Take a look at the restraint violations statistics for distance restraints (`noes.disp`*), dihedrals angles (*`dihedrals.disp`*) and RDCs (*`sani.disp`*) in the* `refine/analysis/` *directory.*

5.3 Graphics output

In the `it8/graphics` directory, you can find two PostScript files (use `evince` to open them).

- The first one, shows the RMS violation by residue in 1- and 2-dimensional plots: `run1/structures/it8/graphics/rms_analysis.ps`
- The second one shows different WHAT-IF Z-scores along the protein sequence: `run1/structures/it8/graphics/whatif_profiles.ps`

Regions of the structures with lower quality can be easily detected from this plots

5.4 GUI output

This option is available only if you enable the `python pickle output` of the `Report` node. Open your `aria_project.xml` project file with the ARIA Gui (`aria2 -g aria_project.xml`) and open the node `Peak Maps` → `it8`.

You can see a contact map based on the NOE assignments and get details on inter-residue constraints by clicking on the crosses of the map.

6 Advanced exercises

1. Go to the `advanced/` directory.

```
cd advanced/
```

6.1 Spot the difference

The directory `run2/` contains the result of an ARIA calculation for the tudor domain, with the same data as in the tutorial. Here, a parameter has been changed in the definition of the ARIA protocol.

- *Compare this result with the one obtained in* `run1` *(violations/precision/structure quality)*
- *What parameter could explain the differences ?*

6.2 Obvious mistake

The directory `run3/` contains the result of another ARIA calculation for the tudor domain. Yet, the result obtained in this case may not be very satisfactory :

- *Why is this result not satisfactory ?*
- *If you think something went wrong, could you point it out ?*
- *What would be the possible origin(s) for this (deliberate) error ?*

6.3 Symmetric dimer

The directory `dimer/run1` contains the result of an ARIA calculation for a symmetric dimer.

- *Analyze the violations and quality parameters of* `dimer/run1`
- *Visualize the ensemble of structures. Does this result look reasonable ?*

Now, move to the the directory `dimer/run2` which contains another result from ARIA for the same symmetric dimer with the same input data.

- *Compare the ensembles of structures of* `dimer/run1` *and* `dimer/run2`
- *What would be the more plausible solution ? Why ?*
- *Can you identify the incriminated restraint(s) ?*

Note: You can check the violations in the file `noe_restraints.assignments` (lines with `viol: yes`)

Links

ARIA `http://aria.pasteur.fr`

CCPN `http://www.ccpn.ac.uk`

CNS `http://cns-online.org`

WHAT IF `http://swift.cmbi.ru.nl/whatif/`

PROCHECK `http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/`

MolProbity `http://molprobity.biochem.duke.edu/`

Further reading

[1] B Bardiaux, A Bernard, W Rieping, M Habeck, TE Malliavin, and M Nilges. Graphical analysis of NMR structural quality and interactive contact map of NOE assignments in ARIA. *BMC Struct Biol*, 8(1):30, Jun 2008.

[2] B Bardiaux, A Bernard, Wolfgang Rieping, M Habeck, T Malliavin, and M Nilges. Influence of different assignment conditions on the determination of symmetric homodimeric structures with aria. *Proteins*, 75(3):569–585, Sep 2008.

[3] Aymeric Bernard, Wim F Vranken, Benjamin Bardiaux, Michael Nilges, and Thérèse E Malliavin. Bayesian estimation of NMR restraint potential and weight: A validation on a representative set of protein structures. *Proteins: Structure, Function, and Bioinformatics*, January 2011.

[4] M. Habeck, W. Rieping, J. P. Linge, and M. Nilges. *NOE assignment with ARIA 2.0: the nuts and bolts.*, volume 208 of *Methods in Molecular Biology*, pages 379–402. Humana Press, Totowa, NJ 07512, USA, 08 2004.

[5] J P. Linge, M A Williams, C A Spronk, A M Bonvin, and M Nilges. Refinement of protein structures in explicit solvent. *Proteins Struct. Funct. Genet.*, 20(3):496–506, 2003.

[6] Michael Nilges, Aymeric Bernard, Benjamin Bardiaux, Thérèse Malliavin, Michael Habeck, and Wolfgang Rieping. Accurate NMR structures through minimization of an extended hybrid energy. *Structure*, 16(9):1305–1312, September 2008.

[7] W Rieping, M Habeck, B Bardiaux, A Bernard, TE Malliavin, and M Nilges. ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3):381–382, 2007.

[8] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, and E. D. Laue. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, 59(4):687–696, Jun 2005.

7 Answers for advanced exercises

7.1 Spot the difference

Here, we used a **log-harmonic potential** instead of the classical *flat bottom harmonic wall* potential for the distance restraints. In addition, the optimal weight applied on the distance restraints was determined automatically. The final average weight for the NOE restraints is 10.6 $kcal.mol^{-1}$ (where it was 50 $kcal.mol^{-1}.\AA^{-2}$ with the flat potential). The general quality of the structures is better with fewer clashes. See Nilges *et al.* Structure, 2008 for details.

7.2 Obvious mistake

Here, we deliberately exchanged assignments of Cys20 and Cys45. Moreover, we asked ARIA to use the manual assignment and to keep them until the end of the calculation. The resulting structures failed to converge due to the incoherent restraints with the two cysteines. One can detect such errors by looking at the file `rms_analysis.ps` for instance (large peaks for residues 20 and 45).

7.3 Symmetric dimer

In `run1`, helix $\alpha 1$ is swapped from one monomer to the other. We actually used the network-anchoring with stringent thresholds and for the first 3 ARIA iterations. Even if the the structures appear reasonable according to violations, precision and global structure quality, one can detect the error when looking at the WHAT IF scores along the sequence and in particular the backbone and bumps checks (residues 32 to 36). The incorrect fold comes from the mis-assignment of intra-molecular cross-peaks as inter-molecular. For instance, VAL43HB–LEU30HA and TYR21HB#–VAL38HG2# (peaks 251 and 1407 of 13C NOESY). The correct expected structures are in `run2`. See the following reference for details: Bardiaux *et al.*, Influence of different assignment conditions on the determination of symmetric homodimeric structures with ARIA. Proteins (2008) vol. 75 (3) pp. 569-585