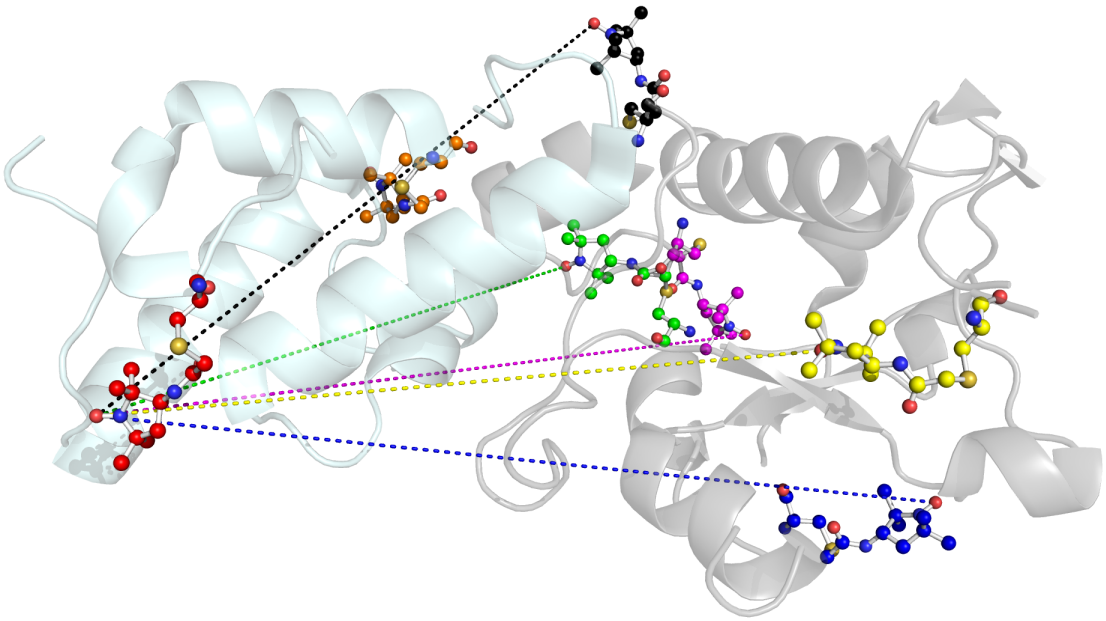
MISHAP

v13.08



Morgan Haydn Bye

This document was created during study for the degree of Doctor of Philosophy whilst at the University of East Anglia, UK.

17th August 2013

morganbye.net

...always accepting a worthy challenge

University of East Anglia
Henry Wellcome Unit for Biological EPR

"An expert is a man who has made all the mistakes which can be made, in a narrow field."

- Niels Bohr

Contents

1 General **1**
- 1.1 Overview . . . 1
  - 1.1.1 What is MISHAP? . . . 1
  - 1.1.2 Functionality . . . 1
  - 1.1.3 Interface . . . 1
  - 1.1.4 Help . . . 2
- 1.2 Getting started . . . 2
  - 1.2.1 Getting MISHAP . . . 2
  - 1.2.2 Installing MISHAP . . . 2
- 1.3 Change history . . . 3

2 User interface **5**
- 2.1 MMM . . . 5
  - 2.1.1 Save rotamers from MMM . . . 5
- 2.2 MISHAP Installer . . . 7
- 2.3 MISHAP distance distribution window . . . 11
  - 2.3.1 Loading a distance distribution . . . 11
  - 2.3.2 Multiple spin systems . . . 12
  - 2.3.3 Protein information . . . 12
  - 2.3.4 Output location . . . 12
  - 2.3.5 RUN! . . . 12
- 2.4 MISHAP PDB creator window . . . 15
  - 2.4.1 Load a PDB file . . . 15
  - 2.4.2 Load a MMM rotamer file . . . 16
  - 2.4.3 Output location . . . 17
  - 2.4.4 RUN! . . . 17

3 HADDOCK configuration **21**
- 3.1 Spin label definitions . . . 21
- 3.2 Protocols . . . 22
- 3.3 Topology parameters (toppar) . . . 23
- 3.4 Run settings . . . 23
  - 3.4.1 CNS address . . . 24
  - 3.4.2 Topology selection . . . 24
  - 3.4.3 Parallelisation . . . 24
  - 3.4.4 Initial seed number . . . 24
  - 3.4.5 MD time step . . . 25
  - 3.4.6 Annealing temperature . . . 25
  - 3.4.7 Protocols . . . 26
  - 3.4.8 Structure generation . . . 26
- 3.5 Run analysis . . . 26

4 An example docking experiment **29**
- 4.1 Example files . . . 29
- 4.2 MISHAP . . . 29
- 4.3 HADDOCK run generation . . . 32
- 4.4 HADDOCK run analysis . . . 33

5 Other tools **35**
- 5.1 Naccess . . . 35
- 5.2 HADDOCK processing tools . . . 35
  - 5.2.1 HADDOCK cluster processing . . . 35
  - 5.2.2 HADDOCK clusters to PyMOL . . . 36

6 Postface **37**
- 6.1 License . . . 37
- 6.2 Third-party software . . . 40
  - 6.2.1 Software within MISHAP . . . 40

6.2.2 MMM 40
6.2.3 HADDOCK 40
6.3 Acknowledgements 40
6.4 Hardware and computation time 41
6.5 References 41

List of Figures

2.1 Loaded protein model in MMM . . . 5
2.2 Select a residue in the command line . . . 6
2.3 Message board update showing the select residue(s) . . . 6
2.4 MMM menu showing the site scan option . . . 6
2.5 MMM site scan setup window . . . 7
2.6 MMM site scan warning dialogue . . . 7
2.7 MMM site scan labelling conditions window . . . 8
2.8 MMM site scan save dialogue window . . . 8
2.9 MMM menu showing MISHAP . . . 10
2.10 MISHAP distance distribution window . . . 11
2.11 MISHAP distance distribution load interface . . . 11
2.12 MISHAP distance distribution window showing a loaded distance distribution . . . 12
2.13 Flowchart of the MISHAP distance distribution window process . . . 13
2.14 MISHAP distance distribution and PDB creator windows . . . 15
2.15 MISHAP PDB creator load PDB file selection . . . 16
2.16 MISHAP PDB creator with loaded PDB file . . . 16
2.17 MISHAP PDB creator load MMM file selection . . . 16
2.18 MISHAP PDB creator load with PDB and MMM file loaded . . . 17
2.19 MISHAP PDB creator showing 2 proteins loaded with MMM rotamer files . . . 17
2.20 Flowchart of PDB handling within the MISHAP PDB creator . . . 18

3.1 Atom naming convention used in MMM and MISHAP for nitroxide spin labels . . . 22

4.1 MISHAP distance distribution window . . . 30
4.2 MISHAP distance distribution load interface . . . 30
4.3 MISHAP distance distribution window showing a loaded distance distribution . . . 31
4.4 MISHAP distance distribution and PDB creator windows . . . 31
4.5 MISHAP PDB creator load with PDB and MMM file loaded . . . 32

5.1 A typical figure from the `HADDOCK_CLUSTERS_PROCESSING` script showing HADDOCK score plotted against the RMSD of the structure from the average structure . . . 35
5.2 A typical figure from the `HADDOCK_CLUSTERS_TO_PYMOL` script showing chain B (in grey cartoon) plotted against the top 4 structures of the top 4 clusters . . . 36

1 General

1.1 Overview

1.1.1 What is MISHAP?

MISHAP is the open-source MMM *in silico* simulated spin label to HADDOCK program. It combines several MatLab scripts, functions, programs and user interfaces as well as a set of additional parameter files for HADDOCK.

MISHAP aims to take spin labelled proteins and the distance distributions between labels from *in silico* (MMM) calculations or experimental data (from pulsed electron-electron double resonance experiments which have been analysed with DeerAnalysis) and pass them to the biomolecular docking program HADDOCK as docking restraints.

1.1.2 Functionality

MISHAP provides the ability to generate HADDOCK compatible PDB files of single chains of amino acids where one or more residues have been substiuted for nitroxide spin label attached cysteine residues.

MISHAP can generate distance based docking restraints from pulsed EPR experiments (PELDOR/DEER) when they have been analysed with DeerAnalysis or MMM.

The generated PDB files can be used in conjunction with or without the MISHAP generated distance based docking restraints to help determine the docking of protein complexes.

Currently, MISHAP only supports the generation of models and restraints in proteins. Though the open-source nature will allow for the expansion of MISHAP to accept non-nitroxide spin labels, DNA, RNA and other common ligands.

1.1.3 Interface

The user interface of MISHAP initially consists of a simple installer that installs MISHAP as well as modifying any existing installations of MMM and HADDOCK.

After installation the user interface largely consists of 2 windows. The first handles the distance restraints whilst the second handles the generation of HADDOCK compatible PDB files.

The changes to HADDOCK allow for the user to continue using HADDOCK normally via the command line with the addition of being able to use the residues R1A (MTSL attached cysteine) and IA1 (IA-PROXYL attached cysteine).

1.1.4 Help

MISHAP is designed to be as intuitive as possible whilst still maintaining user control. With this in mind, all options are clearly labelled as well as extensively documented in this document as well as online at morganbye.net/mishap

1.2 Getting started

1.2.1 Getting MISHAP

MISHAP is available online from 2 locations.

MISHAP general releases (for most users) should be downloaded from:

https://sourceforge.net/projects/mishap/

MISHAP developer releases (constantly updated and should be considered unstable) can be obtained from the git repositary

https://github.com/morganbye/MISHAP

Alternatively, you can create a new git folder and pull the repository directly from the command line with:

```
git pull https://github.com/morganbye/MISHAP.git
```

1.2.2 Installing MISHAP

If MISHAP was downloaded from sourceforge then it will come as a ZIP file which includes everything required. Simply unpack the ZIP file into a directory of your choosing. The folder should contain several text files including; `LICENSE.txt`, `README.txt` and `UPDATE.txt`. There will also be `MISHAP.m`; the main MatLab file and 2 directories: `test_files` containing some example files (referenced later) and `_private` which contains all the files associated with MISHAP.

The extracted folder needs to be added to the MatLab path with subfolders to the; this can be done be navigating to the folder in the *Current folder* view, right clicking and selecting

Add to Path → with Selected folders and Subfolders

or by clicking from the main MatLab menu

File → Set Path

in the main MatLab window, then click *Add folder with subfolders* then *Save* and *Close*.

The MISHAP installer then needs to be run. This is detailed in §2.2, but simply in the MatLab command window type

```
MISHAP
```

This begins the installation of MISHAP.

1.3 Change history

Release v13.08 - BETA

v13.08 shows much improved error handling and allows for MISHAP generated PDBs be reloaded in the PDB creator and have additional spin labels attached.

Release v13.06 - BETA

v13.06 is the first publicly viable release of MISHAP and marks the first build from alpha. It shows far greater error handling, Windows compatibility and an improved graphical user interface. MMM and MatLab programming is complete but significant errors still occur in HADDOCK/CNS.

Build v13.05 - ALPHA

v13.05 is the first functional build of MISHAP that generates distance distribution constraints and MMM rotamer merged PDBs.

2 User interface

2.1 MMM

MMM is an open-source program for the visualisation, inspection and improvement of models of proteins and protein assemblies based on restraints from multiple experimental techniques, though primarily focussing on experimental techniques from electron paramagnetic resonance[1].

2.1.1 Save rotamers from MMM

Perhaps MMM's most useful feature is the ability to calculate the rotamer populations of nitroxide spin labels attached to proteins.

For MISHAP we require these spin label attached protein models. To generate a rotamer model first start MMM from the MatLab command window with the command

```
MMM
```

After the splash screen, load a PDB model using the *File* menu - more information available in the MMM manual §2.1.1.

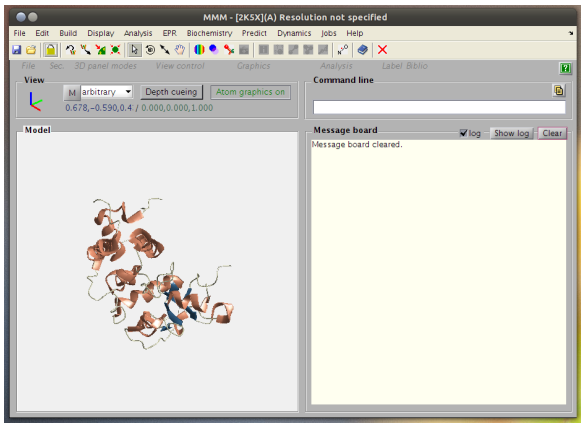


Figure 2.1 *Loaded protein model in MMM*

The protein model should be visualised in the Model panel whilst the Message board should update.

Next select a residue. To do this at the command line level type

```
select [structure](chain)residue
```

the strucutre option is not necessary depending upong the file (MMM §1.6). Residue selection can also be done graphically using:

[1] MMM - http://www.epr.ethz.ch/software/index

Figure 2.2 *Select a residue in the command line*

Display → Hierarchy (MMM manual §2.3 and 4.1.2)

However, the command line is superior as it allows the selection of multiple residues at once.

Figure 2.3 *Message board update showing the select residue(s)*

Once a residue is select the message board will update with the selection. If a selection is made incorrectly use:

```
unselect [structure](chain)residue
```

Once happy with the selection we label the selected residues with:

EPR → Site scan/selected residues (MMM manual §2.1.7)

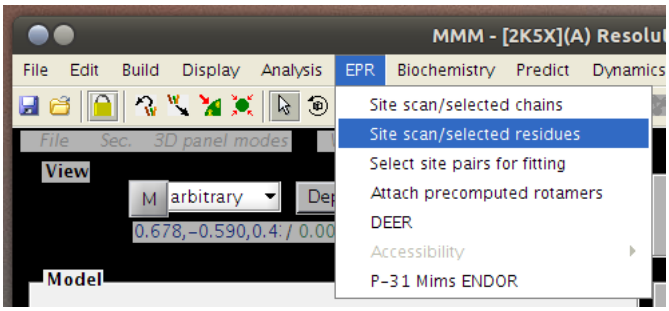


Figure 2.4 *MMM menu showing the site scan option*

This displays the Site scan setup window (MMM manual §4.1.2). It is imperative that *Save statistics* and *Save PDB rotamers* are selected.

Selecting *Save PDB rotamers* will display a warning message. Ensure that you click *Yes*.

Clicking *OK* will display the *Set labeling conditions for selected residues* window. Here select the appropiate nitroxide spin label and temperature.

Finally a prompt appears to set the save path of the rotamer analysis files

The rotamer analysis can be tracked in the message board and once complete a prompt appears to ask if the user wishes to view the results in a browser.

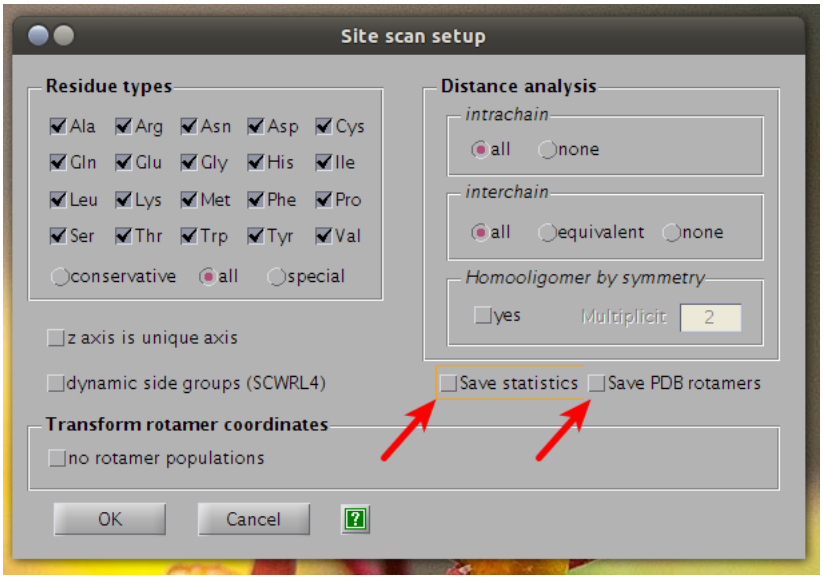


Figure 2.5 MMM site scan setup window

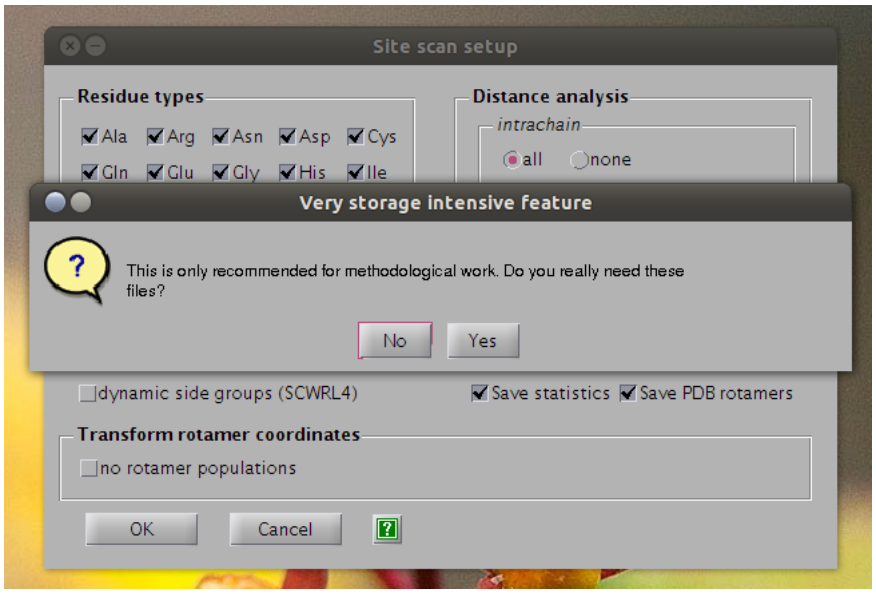


Figure 2.6 MMM site scan warning dialogue

2.2 MISHAP Installer

Installation of MISHAP is kept as much as possible within the MatLab environment for simplicity.

To install MISHAP, extract the MISHAP zip folder. Add the folder with subfolders to the MatLab path; this can be done be navigating to the folder in the *Current folder* view right clicking and selecting

Add to Path → with Selected folders and Subfolders

or by clicking from the main MatLab menu or by clicking

File → Set Path

in the main MatLab window, then click *Add folder with subfolders* then *Save* and *Close*.

In the MatLab command window now type

```
MISHAP
```

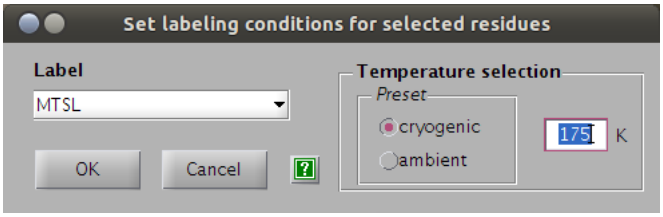


Figure 2.7 *MMM site scan labelling conditions window*

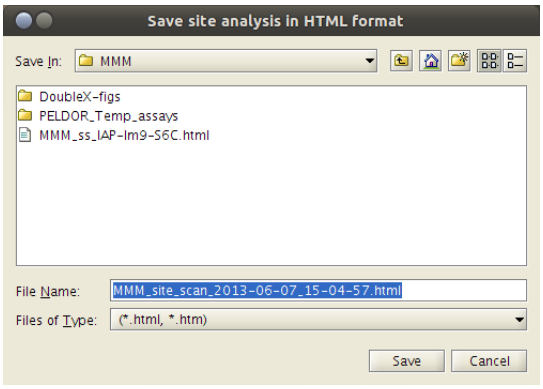


Figure 2.8 *MMM site scan save dialogue window*

This begins the installation of MISHAP. The installer addresses 3 different programs in turn. First the installer looks to see if MISHAP has previously been installed, if so then the installer aborts and simply loads MISHAP.

If MISHAP has not previously been installed then the installation begins looking first to see if the user has MMM installed (line 36). For safety the installer backs up `MMM_prototype.m` and `MMM_prototype.fig` files to `/.../MMM_path/old_files` before copying across new versions of `MMM_prototype`, these versions allow you to access MISHAP from MMM.

> At a later date this versions will be submitted to MMM and this part of the installer will be removed.

Next the installer asks the user if HADDOCK will be run locally, if so then the local copy of HADDOCK requires some new parameter files. Again exisiting files are backed up first to `/.../HADDOCK_path/old_files` before the new ones are copied across.

> As HADDOCK is outside the MatLab environment a system search is required and this may take some time. As a safety feature, the search for HADDOCK will time out in 10 seconds and ask the user for the location instead.

The command window output from a typical installation should look something like:

```
 __  __ _____  _____ _    _          _____
|  \/  |_   _|/ ____| |  | |   /\   |  __ \
| \  / | | | | (___ | |__| |  /  \  | |__) |
| |\/| | | | |  \___ \|  __  | / /\ \ |  ___/
| |  | |_| |_ ____) | |  | |/ ____ \| |
|_|  |_|_____|_____/|_|  |_/_/    \_\_|

                            by

                                    _                            _
                                   | |                          | |
 _ __ ___   ___  _ __ __ _  __ _ _ __| |__   _   _  ___   _ __   ___| |_
| '_ ` _ \ / _ \| '__/ _` |/ _` | '_ \| '_ \| | | |/ _ \ | '_ \ / _ \ __|
| | | | | | (_) | | | (_| | (_| | | | | |_) | |_| |  __/_| | | |  __/ |_
|_| |_| |_|\___/|_|  \__, |\__,_|_| |_|_.__/ \__, |\___(_)_| |_|\___|\__|
                      __/ |                   __/ |
                     |___/                   |___/


==========================================
STARTING INSTALLATION
==========================================

Installer version - 13.06
Build date        - 04th May 2013

This installer will install
MISHAP version    - 13.06
Release date      - 04th May 2013

System            - Linux - 64 bit

MatLab version    - 7.12.0.635 (R2011a)

==========================================

Finding MMM installation...
MMM was located at:
/opt/matlab/MyScripts/MMM_2013

Backing up MMM files to:
/opt/matlab/MyScripts/MMM_2013/old_files

Files backed up!

Copying across new files...
Files successfully copied!

Installing...

Configuring...
MMM configured!

==========================================

MISHAP parameters are not yet compatible with the HADDOCK webservers,
though this should change shortly, MISHAP must be run locally in the mean time

Are you going to be using HADDOCK locally? Y/N [Y]: y

Finding HADDOCK installation...
detection will abort if nothing is found in 10 seconds...
HADDOCK was located at:
/opt/haddock

Backing up HADDOCK files to:
```

```
/opt/haddock/old_files

Files backed up!

Copying across new files...
Files successfully copied!

Installing...

Configuring...
HADDOCK configured!

================================================

Writing changes to MatLab...
MatLab configured!

================================================

I hope you find MISHAP useful.

Documentation can be found at:
morganbye.net/mishap

Please address any feedback to:
morgan.bye@uea.ac.uk

================================================
INSTALLATION COMPLETE
================================================
```

With MISHAP installed, MISHAP can now be called from MMM using

Predict → Quaternary → HADDOCK → MISHAP

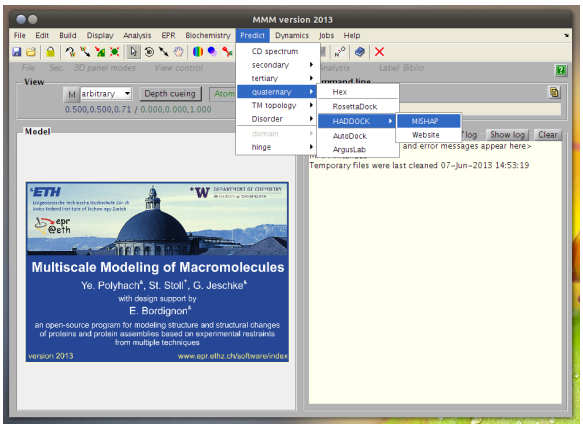


Figure 2.9 *MMM menu showing MISHAP*

Or without MMM from the command window with the command:

```
MISHAP
```

This will launch the primary MISHAP window, the distance distribution window.

2.3 MISHAP distance distribution window

2.3.1 Loading a distance distribution

Calling MISHAP will launch the primary MISHAP window, the distance distribution window.

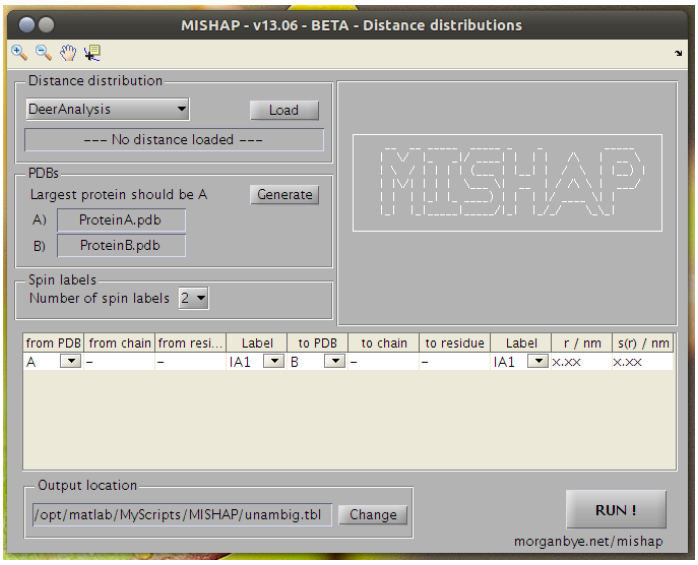


Figure 2.10 *MISHAP distance distribution window*

From the distance distribution panel select whether the distance distribtion is from *DeerAnalysis* or *MMM* and click *Load*.

If a DeerAnalysis or MMM window is open then MISHAP will try to pull the distance distribution information from the open window. If more than one is open then it will load whichever was opened first. If no window can be found then a user interface for the selection of a previously saved file is opened.

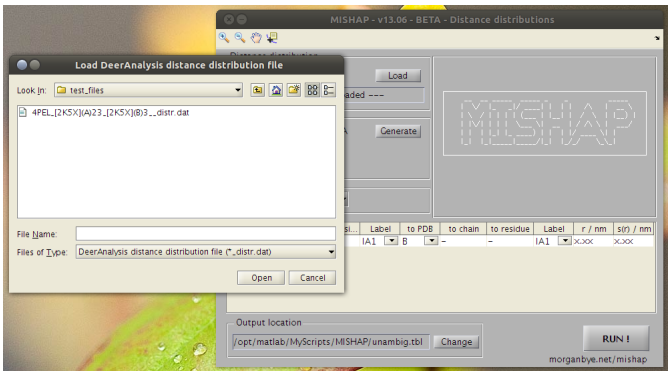


Figure 2.11 *MISHAP distance distribution load interface*

Once a file is selected MISHAP will automatically plot the distance distribution and search it for any peaks. Peaks are marked with circles. The largest peak is selected as the peak of interest and the peak location is updated into the *r* column of the table.

A *s(r)* of 0.5 nm is automatically set and the boundaries are plotted about the distance as dotted lines in the same colour as the peak. At a later date the *s(r)* will be detected and calculated from the plot using a half height analysis of a Gaussian fit to the peak.

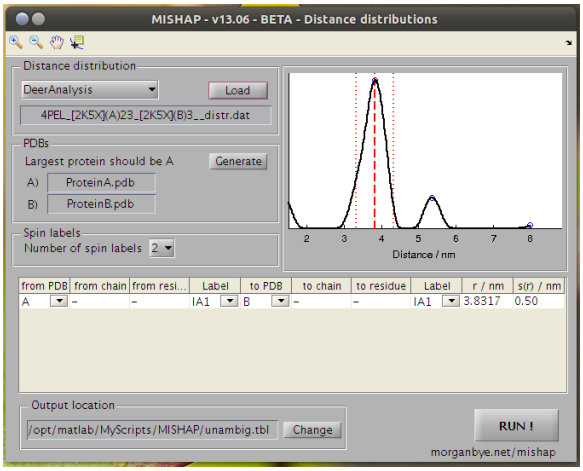


Figure 2.12 *MISHAP distance distribution window showing a loaded distance distribution*

2.3.2 Multiple spin systems

Currently MISHAP has support for systems with up to 4 nitroxide spin labels. By default it is set for only 2 spins (1 distance), selecting 3 or 4 spin labels adds more distances to the table (3 distances for 3 labels, 6 distances for 4 labels). The distance distribution plot displays each distance in a seperate coloured dashed line with the peak's *s(r)* dotted in the same colour.

2.3.3 Protein information

If the distance distribution tool is being used without the protein PDB creator then the table information needs to be completed with the appropiate information for the chain ID and residue number of the spin label in the PDB.

However, it is recommended to generate the PDBs for HADDOCK with MISHAP before proceeding further. To do so click *Generate* to launch the PDB Creator (see §2.4).

2.3.4 Output location

Once happy with the information displayed in the table finally check the output location box. This is the location and file name that the distance restraints file will be saved as.

By default the file name `unambig.tbl` will be used and the path will be that of the MatLab path when MISHAP was first opened.

If this is not suitable simply click *Change*.

2.3.5 RUN!

When the information in the table is correct and the output location is set click *RUN!*

All information is passed out to the MatLab command window. First the constraint generator checks that the table is complete, then it checks that it can write to the selected location. If

the folder does not exist or the user does not have permission to write to it then they will be prompted.

Finally the file is created. This process should be very fast even on a slow computer, ensure that *RUN!* is only pressed once.

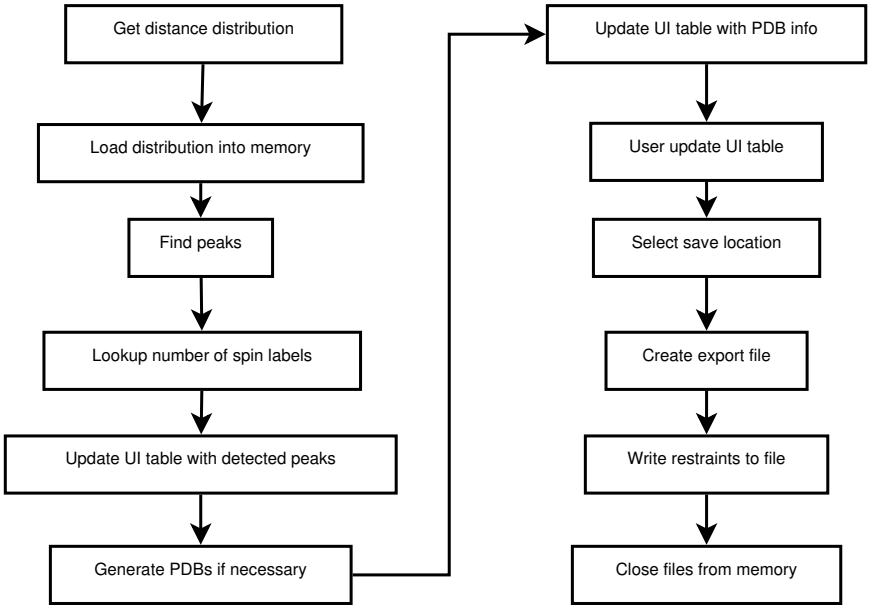


Figure 2.13 *Flowchart of the MISHAP distance distribution window process*

On the next page is an example output from the command window.

```
             __  __ _____  _____ _    _           _____
            |  \/  |_   _|/ ____| |  | |   /\    |  __ \
            | \  / | | | | (___ | |__| |  /  \   | |__) |
            | |\/| | | |  \___ \|  __  | / /\ \  |  ___/
            | |  | |_| |_ ____) | |  | |/ ____ \| |
            |_|  |_|_____|_____/|_|  |_/_/    \_\_|

      _____                 _             _       _
     / ____|               | |           (_)     | |
    | |     ___  _ __  ___| |_ _ __ __ _ _ _ __ | |_
    | |    / _ \| '_ \/ __| __| '__/ _` | | '_ \| __|
    | |___| (_) | | | \__ \ |_| | | (_| | | | | | |_
     \_____\___/|_| |_|___/\__|_|  \__,_|_|_| |_|\__|
                                                    _
                                                   | |
                    __ _  ___ _ __   ___ _ __ __ _| |_ ___  _ __
                   / _` |/ _ \ '_ \ / _ \ '__/ _` | __/ _ \| '__|
                  | (_| |  __/ | | |  __/ | | (_| | || (_) | |
                   \__, |\___|_| |_|\___|_|  \__,_|\__\___/|_|
                    __/ |
                   |___/

==========================================
STARTING Distance Constraint Creation
==========================================

MISHAP version    - 13.06
Release date      - 16th May 2013

System            - Linux - 64 bit

MatLab version    - 7.12.0.635 (R2011a)

==========================================

Checking the data table...
Checking complete

Setting label variables...

Saving file...
File saved as
/opt/matlab/MyScripts/MISHAP/test_files/unambig.tbl

==========================================

MISHAP - distance constraint generator completed in 0.00589 seconds
Thank you for using MISHAP
```

2.4 MISHAP PDB creator window

The MISHAP PDB creator window is launched from the MISHAP distance distribution window by clicking *Generate*.

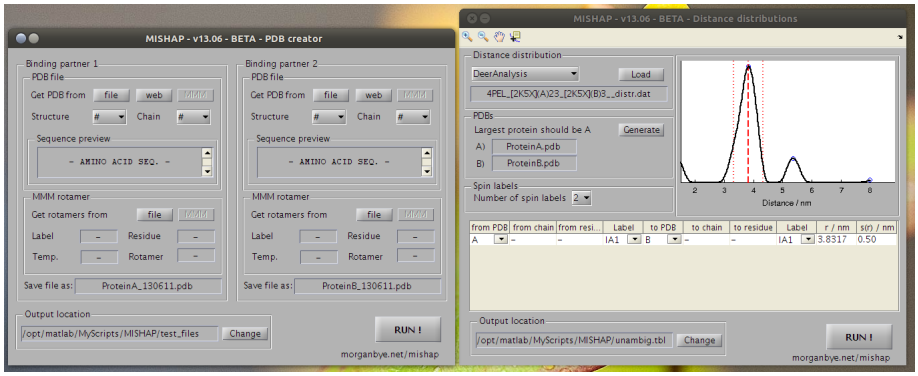


Figure 2.14 *MISHAP distance distribution and PDB creator windows*

The purpose of the PDB creator is 2 fold. First is to ensure that PDB files are generated that are compaitble with HADDOCK.

Unfortunately, the PDB format was first created in the 1970's and has been extensively developed, without any strict enforcement of the PDB format for many years meaning that it is quite common for PDB files generated in one program to not be compatible with a different PDB handling program. MMM version 2013 after a major rewrite of the PDB file writing functions now conforms fully. HADDOCK will only accept PDB files in a very explicit format, which differs significantly from the PDB standard[2].

The second purpose of the PDB creator is to combine a PDB file supplied from a local source, from the PDB website (using an access number) or from a loaded MMM model with a MMM rotamer analysis.

MMM provides some basic functionality for this but we require a single PDB file where an amino acid residue has been mutated to a spin label attached cysteine residue. Alternatively, a set of PDB files and a list file that can be inserted into HADDOCK as an ensemble of starting conformations - similar to an NMR ensemble.

2.4.1 Load a PDB file

For the first protein select the PDB file location. If the PDB is a local file click *file* this will open a user interface for selecting the file from your computer.

The PDB can also be obtained directly from rcsb.org in which case select *web* and enter the accession number when prompted.

If MMM is open, then the current model can be used by clicking *MMM*.

There will be a delay as the file is processed. Once processed the *Structure* box will update with the PDB accession code, the *Chain* dropdown box will update with the chains found in the PDB and the sequence of the selected chain will be shown in the *Sequence preview*.

[2] for more information see http://www.nmr.chem.uu.nl/haddock/pdb.html

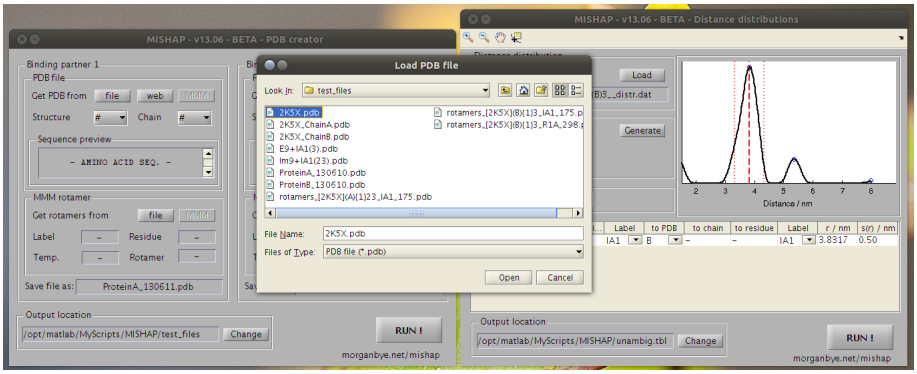


Figure 2.15 *MISHAP PDB creator load PDB file selection*

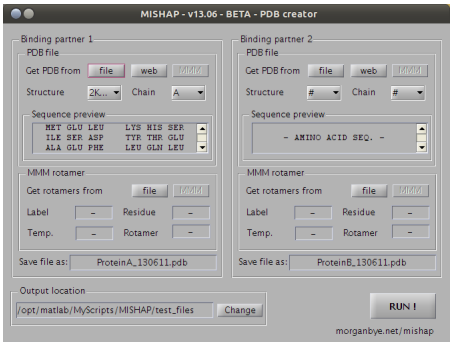


Figure 2.16 *MISHAP PDB creator with loaded PDB file*

2.4.2 Load a MMM rotamer file

For the first protein select the MMM rotamer file location. If the PDB is a local file saved during the MMM site scan then click *file* this will open a user interface for selecting the file.

In a later release, it will be possible to click *MMM* and pull the rotamer information from MMM.

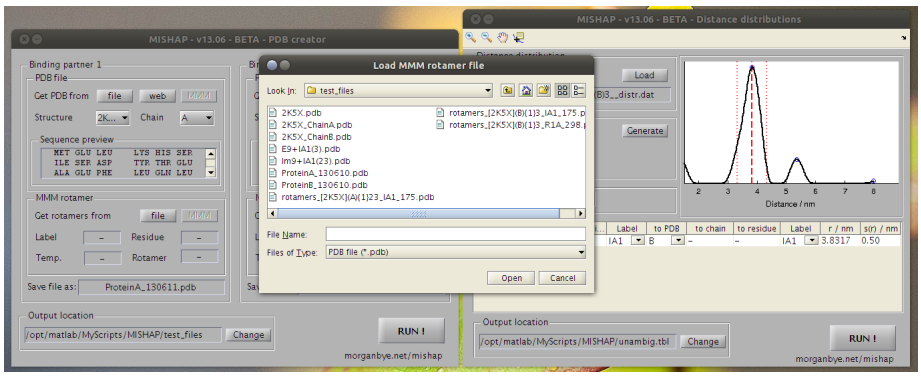


Figure 2.17 *MISHAP PDB creator load MMM file selection*

It will take a little time to process the rotamer information. Once processed the *Label* box will update with the 3 character amino acid residue code for the spin label, the *Temp.* box will update with the labeling temperature used in the site scan, the *Residue* box will update with the amino acid residue number of the labeling site and the *Rotamer* will update with the selected rotamer number from the MMM rotamer analysis.

By default, the *Rotamer* box will default to rotamer 1 though this can be changed to any rotamer. If a rotamer is selected that is outside the range of available rotamers then the box will automatically default back to rotamer 1. If an ensemble style rotamer analysis is desired

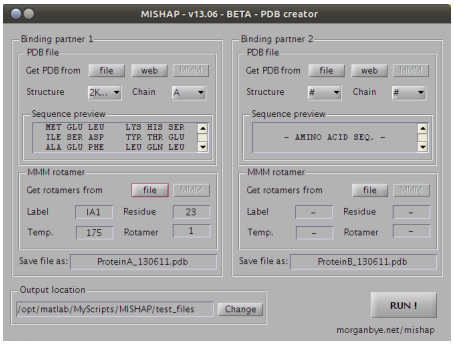


Figure 2.18 *MISHAP PDB creator load with PDB and MMM file loaded*

type in this box:

```
all
```

2.4.3 Output location

By default the protein PDB will be saved as:

```
Protein[A/B]_[YYMMDD].pdb
```

These can be changed freely with any operating system valid name. Do not include *.pdb* in the file name, otherwise you will generate a file called `ProteinA_YYMMDD.pdb.pdb`

The output path is defined in the *Output location* box, the path will be that of the MatLab path when MISHAP was first opened.

If this is not suitable simply click *Change*.

2.4.4 RUN!

When happy with the settings for *Binding Partner 1* the PDB creator can be run or a PDB and MMM rotamer analysis can also be loaded for *Binding Partner 2*.

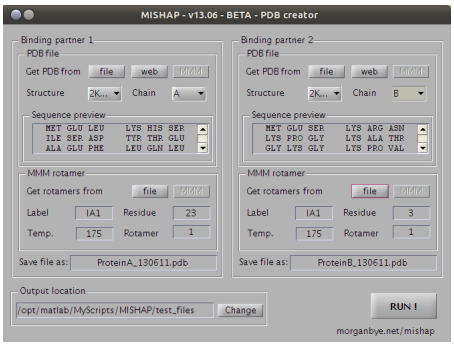


Figure 2.19 *MISHAP PDB creator showing 2 proteins loaded with MMM rotamer files*

When all the information is loaded and set click *RUN!*

All information is passed out to the MatLab command window. *Binding Partner 1* is addressed before *Binding Partner 2* but in both the creator checks that a PDB and MMM rotamer analysis is loaded, then it checks that it can write to the selected output location. If

the folder does not exist or the user does not have permission to write to it then they will be prompted.

Finally the creator merges the selected MMM rotamer(s) with the PDB file and the output file is created.

Briefly, the creator essentially splits the PDB file before and after the select residue. It then pastes in the rotamer residue and merges the residues before the residue, the rotamer residue and residues after the rotamer residue with appropiate linkers. All hydrogen atoms are then removed as these are prone to causing problems in HADDOCK and CNS. Finally, every atom is renumbered in the merged PDB so that read sequentially. The file is finally written out line by line with space padding so that each line is exactly 80 characters.

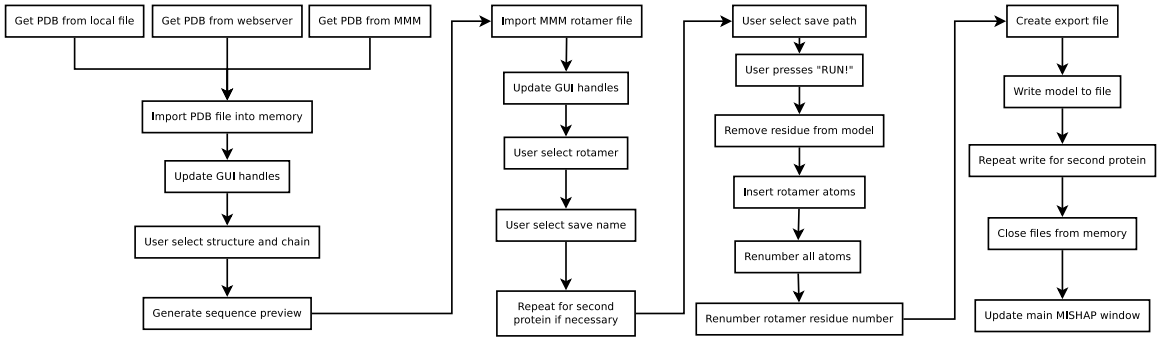


Figure 2.20 *Flowchart of PDB handling within the MISHAP PDB creator*

On the next page is an example output from the command window.

```
 __  __ _____  _____ _    _           _____
|  \/  |_   _|/ ____| |  | |   /\    |  __ \
| \  / | | | | (___ | |__| |  /  \   | |__) |
| |\/| | | | |  \___ \|  __  | / /\ \ |  ___/
| |  | |_| |_ ____) | |  | |/ ____ \| |
|_|  |_|_____|_____/|_|  |_/_/    \_\_|

 _____  _____  ____     _____                _
|  __ \|  __ \|  _ \   / ____|              | |
| |__) | |  | | |_) | | |     _ __ ___  __ _| |_ ___  _ __
|  ___/| |  | |  _ <  | |    | '__/ _ \/ _` | __/ _ \| '__|
| |    | |__| | |_) | | |____| | |  __/ (_| | || (_) | |
|_|    |_____/|____/   \_____|_|  \___|\__,_|\__\___/|_|

==========================================
STARTING PDB Creation
==========================================

MISHAP version    - 13.06
Release date      - 15th May 2013

System            - Linux - 64 bit

MatLab version    - 7.12.0.635 (R2011a)

==========================================
Binding partner 1
==========================================

Finding the selected PDB...
Loading file...
/opt/matlab/MyScripts/MISHAP/test_files/2K5X.pdb...
Using structure 2K5X  and chain A
PDB loaded!

Finding the selected MMM rotamer...
Loading file...
/opt/matlab/MyScripts/MISHAP/test_files/rotamers_[2K5X](A){1}23_IA1_175.pdb

Rotamers loaded!

Setting labelling conditions:
Label                  - IA1
Temperature            - 175 K
Attaching to residue   - 23
Using rotamer          - 1

Adding rotamer to PDB...
Rotamer successfully added!

Saving file...
File saved as
/opt/matlab/MyScripts/MISHAP/test_files/ProteinA_130610.pdb

==========================================
Binding partner 2
==========================================
```

```
58
59 Finding the selected PDB!
60 Loading file...
61 /opt/matlab/MyScripts/MISHAP/test_files/2K5X.pdb
62 Using structure 2K5X  and chain B
63 PDB loaded!
64
65 Finding the selected MMM rotamer
66 Loading file...
67 /opt/matlab/MyScripts/MISHAP/test_files/rotamers_[2K5X](B){1}3_IA1_175.pdb
68 Rotamers loaded!
69
70 Setting labelling conditions:
71 Label                  - IA1
72 Temperature            - 175 K
73 Attaching to residue   - 3
74 Using rotamer          - 1
75
76 Attaching rotamer to PDB...
77 Rotamer successfully added!
78
79 Saving file...
80 File saved as
81 /opt/matlab/MyScripts/MISHAP/test_files/ProteinB_130610.pdb
82
83 ========================================
84
85 MISHAP - PDB Creator completed in 0.79016 seconds
86 Thank you for using MISHAP
```

3 HADDOCK configuration

HADDOCK (the high ambiguity data driven biomolecular docking program) is an approach that makes use of biochemical and biophysical interaction data such as chemical shift perturbations from NMR titration experiments, mutagenesis data or bioinformatic predictions to study protein complexes.

Traditionally, most approaches to protein complex formation relied upon a combination of energetics and shape complementarity rather than experimental data. The previously mentioned experiments are presented to the HADDOCK program as ambiguous interaction restraints (AIRs) to drive the docking process. More formally, these are defined as an ambiguous distance between all residues shown to be involved in the interaction.

Since the original 2003 JACS publication HADDOCK has been extended to deal with a large varity of data and complexes including protein-DNA, protein-RNA, protein-oligosaccharides and protein-ligand complexes. HADDOCK has consistently performed well in the blind protein-protein complex CAPRI competitions[1].

HADDOCK remains a command line tool requiring the generation and configuration of many files before a run can commence. HADDOCK consists primarily as a collection of python scripts derived from ARIA written by Michael Nilges and Jens Linge and makes use for CNS as structure calculation.

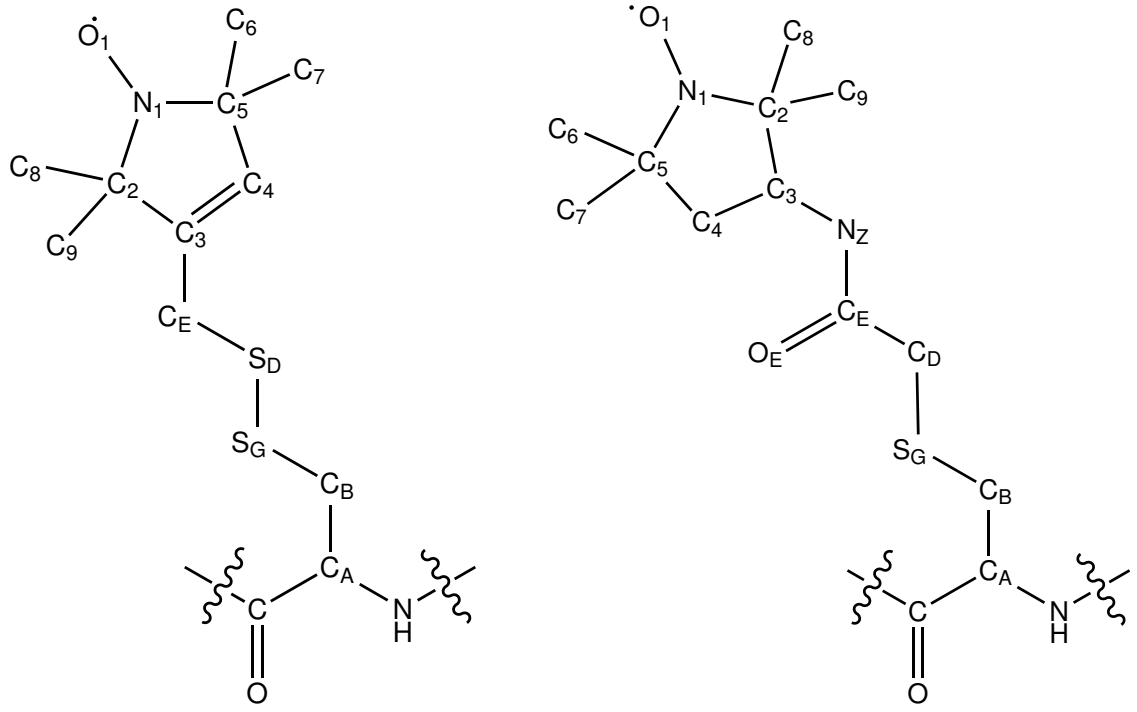
3.1 Spin label definitions

HADDOCK being designed to accept input data from several different sources already has the facility to use several unnatural amino acids such as phosphorylated CYS, HIS, SER, THR, several methyl variants such as mono-, di- and tri- methyl lysine, to name but a few.

Thus if correctly defined it is perfectly possible to add MTSL and IA-PROXYL labelled cysteines as an unnatural, but allowed, amino acid residue.

For an amino acid to be defined in HADDOCK and the underlying structural calculation program, CNS, requires explicit definitions of all atoms, charges, bonds, bond lengths and allowed bond angles (both dihedral and improper dihedral). When combined this gives the modelling software not only the atomic position and bonding of the atoms, but also allows for the rotation of bonds - which is essential in the final energy optimisation stage of HADDOCK before the final energy optimisation in water.

[1] http://capri.ebi.ac.uk



(a) MTSL *(b)* IA-PROXYL

Figure 3.1 *Atom naming convention used in MMM and MISHAP for nitroxide spin labels*

3.2 Protocols

With MMM there is an increasing amount of published data that shows that rotamer population prediction from a rotamer library accurately reflects the experimental situation. It is not known whether the bond rotation allowed in HADDOCK/CNS reflects well with that seen experimentally. Previously, the authors of HADDOCK when performing paramagnetic relaxation enhancement studies take the spin label and during the early stages of docking spawn 4 structures with the spin label in different conformations. Docking is then performed in an ensemble style analysis and it is hoped that during the allowed side chain freedom stage of energy minimisation that starting from 4 random positions will find the overall position(s) and not get caught in a local energy minima.

To this end, MISHAP provides the ability in the PDB creator to not only create a single PDB file but automatically generate an ensemble folder of PDB files, with the appropiate HADDOCK file list file, of rotamers from MMM with populations of greater than 1 %. Whilst a typical MMM analysis can generate up to 100 rotamers on a greatly solvent exposed site from the rotamer library most will have populations of below 1 %. In more rigidly located spin label positions the rotamer population quickly reduces in number with the remaining allowed rotamers covering a large proportion of the population, in some cases above 50 %, in these cases an ensemble analysis may not be necessary and a single rotamer PDB file should be sufficient for an accurate docking simulation.

3.3 Topology parameters (toppar)

HADDOCK contains a series of files which contain all of the necessary parameters for amino acid residues, ligands, DNA, etc. These parameters largely are default parameters for CNS that have been extensively modfied and added to for the purposes of HADDOCK. HADDOCK creates CNS scripts and then calls CNS in order to perform docking.

Note that HADDOCK must make use of topology version 5.4 for many of the features to work. HADDOCK however by default uses version 5.3 - to ensure as broad compatibility as possible the provided MISHAP files provide both 5.3 and 5.4 topology files. The following files from the `HADDOCK/toppar` folder require updating in order for HADDOCK to accept nitroxide spin labels:

`topallhdg5.3.pep` allows the labels to exist in a peptide

`topallhdg5.3.pep-noCter` allows the labels to exist in a peptide where the C-terminus does not exist

`topallhdg5.3.pep-noNter` allows the labels to exist in a peptide where the N-terminus does not exist

`topallhdg5.3.pro` defines the mass and charge of each atom, as well as each bond, angle, improper angle, dihedral angle and hydrogen bond acceptor/donor site.

`topallhdg5.3.pro-caro` defines the mass and charge of each atom, as well as each bond, angle, improper angle, dihedral angle and hydrogen bond acceptor/donor site.

`topallhdg5.4.pep` allows the labels to exist in a peptide, and defines how the peptide would bind to surrounding amino acid residues

`topallhdg5.4.pep-noCter` allows the labels to exist and bond to it's neighbours in a peptide where no C-terminus exists

`topallhdg5.4.pep-noNter` allows the labels to exist and bond to it's neighbours in a peptide where no N-terminus exists

`topallhdg5.4.pep-noter` allows the labels to exist and bond to it's neighbours in a peptide where neither a N nor C -termini exists

`topallhdg5.4.pro` defines the mass and charge of each atom, as well as each bond, angle, improper angle, dihedral angle and hydrogen bond acceptor/donor site

Whilst it is possible to create these topology parameter files by hand methodically, several tools exist that can speed up the process. For example the PRODRG tool[2] has a free web server[3] that can be used to generate CNS topology files. This tool is only designed to create ligands and so to incorporate a structure as an amino acid requires some modification.

3.4 Run settings

`run.cns` is the primary file that HADDOCK runs a docking experiment from. When a run is first generated it is imperative to change this file to reflect not only your own system but also to ensure the correct settings are used.

[2] Acta Cryst. (2004). D60, 1355-1363
[3] http://davapc1.bioch.dundee.ac.uk/prodrg/

3.4.1 CNS address

Despite what is set in the `new.html` when the run is generated the CNS address often does not update and is often defaulted to:

```
/sw/nmr/cns_solve_1.2-para/mac-intel-darwin/bin/cns
```

This can be editted by hand or with the command:

```
sed 's,'/sw/nmr/cns_solve_1.2-para/mac-intel-darwin/bin/
   cns', '/path/to/cns_solve_1.3/intel-x86_64bit-linux/
   bin/cns',' run.cns > tmp ; mv tmp run.cns
```

Where the `sed` command searches for the string in the first quote and replaces it with the second string. We then need to save to a temporary file and overwrite the original file with the temporary one.

3.4.2 Topology selection

HADDOCK will always by default use topology parameters v5.3, we require HADDOCK to use version 5.4 when using CNS v1.3.

Update the version selection with:

```
sed 's,'topallhdg5.3','topallhdg5.4',' run.cns > tmp ; mv
    tmp run.cns
```

3.4.3 Parallelisation

HADDOCK by default assumes only a dual cored processor. If you are using it upon a cluster or server ensure that you update the `run.cns` file to take this into account; replacing 'XXX' with the number available on your system.

```
sed 's,'cpunumber_1=2','cpunumber_1=XXX',' run.cns > tmp
   ; mv tmp run.cns
```

The number of CPU cores on a UNIX system can be easily found with the command:

```
grep "cpu cores" /proc/cpuinfo
```

or with the terminal command (available in most Linux distributions):

```
lscpu
```

3.4.4 Initial seed number

HADDOCK uses a random number generator to begin the structure formation in the rigid body stage. However, this number generator requires a starting point and should be recorded for reproducability. It is easiest to set this as a fixed number, though a random number generator could be piped into the argument.

```
sed 's,'iniseed=917','iniseed=999',' run.cns > tmp ; mv
   tmp run.cns
```

3.4.5 MD time step

The HADDOCK protocol uses a process of simulated heating and cooling for energy optimisation and minimisation. By default HADDOCK uses an molecular dynamics (MD) time step of 2 fs, however, docking runs will often fail with this setting. With such a time step, atoms that are moving toward each other in one time step can often jump on top of each other, when HADDOCK then calculates the energy there is a massive potential energy in these atoms. Entropy resolves this in the next time step by moving the atoms far apart, often outside the calculation co-ordinate space (> 99,999 $\AA$). This causes the structure to fail, if more than 20 % of a run's structures fail then HADDOCK will abort the run.

By decreasing the time step reduces the risk of this scenario, however, to ensure a good fit the total simulation cannot change, thus the number of steps increases, which increases the experiment time.

```
# Change MD time step
sed 's,'timestep=0.002','timestep=0.0005',' run.cns > tmp
    ; mv tmp run.cns

# Increase MD iterations to maintain total simulation
   time
sed 's,'initiosteps=500','initiosteps=2000',' run.cns >
   tmp ; mv tmp run.cns sed 's,'cool1_steps=500','
   cool1_steps=2000',' run.cns > tmp ; mv tmp run.cns
sed 's,'cool2_steps=1000','cool2_steps=4000',' run.cns >
   tmp ; mv tmp run.cns
sed 's,'cool3_steps=1000','cool3_steps=4000',' run.cns >
   tmp ; mv tmp run.cns
```

3.4.6 Annealing temperature

Energy optimisation in HADDOCK is conducted by raising the temperture in the system to a high value and then allowing it to cool *in vacuo*. Reducing the time step described in the previous section can help. However, if greater than 20 % of structures fail during it1 (energy optimisation) step then HADDOCK will abort and solvent optimisation will not occur. First, change the initial seed number and restart the run (this will require deletion of all it0 files, for the creation of new structures). If however, the runs continue to fail then the structures have too much energy and will fall to same problems discussed in the MD time step section.

The best solution is to lower the temperature that the system is raised to in each energy optimisation step. This again is defined in `run.cns` and can be quickly changed using the following commands.

```
sed 's,'tadhigh_t=2000','tadhigh_t=500',' run.cns > tmp ;
    mv tmp run.cns
sed 's,'tadinit1_t=2000','tadinit2_t=1000',' run.cns >
   tmp ; mv tmp run.cns
sed 's,'tadinit2_t=1000','tadinit2_t=500',' run.cns > tmp
    ; mv tmp run.cns
sed 's,'tadinit3_t=500','tadinit3_t=250',' run.cns > tmp
   ; mv tmp run.cns
sed 's,'tadfactor=8','tadfactor=4',' run.cns > tmp ; mv
   tmp run.cns
```

These temperatures can be dropped further, but should only be done with considerable care.

3.4.7 Protocols

The `run.cns` in the run root is not the same `run.cns` file that is called when protocols are run. To ensure that our changes are reflected throughout the entire HADDOCK run we need to copy the `run.cns` root and replace the `run.cns` file in the `protocols` folder.

```
cp run.cns protocols/run.cns
```

3.4.8 Structure generation

Each amino acid chain or ligand is generated as a seperate structure, so for a docking experiment the bear minimum we will have is structure A and B, whilst HADDOCK supports up to 6 structures.

For each structure to be used in the docking run requires that the structure generation script be forced to use topology parameters v5.4.

```
sed 's,'hdg5.3','hdg5.4',' generate_A.inp > tmp ; mv tmp
   generate_A.inp
```

Repeating this command but replacing `generate_A.inp` with `generate_B.inp`, ..., etc. for every structure.

Also check that the `generate_X.inp` file has the correct file location, all to often it will default to:

```
/home/abonvin/software/haddock/examples/e2a.pdb
```

3.5 Run analysis

If the HADDOCK run is not being performed upon the Utrecht University webserver then the run needs to be manually analysed (detailed in the documentation[4] and online tutorial[5]).

[4] http://www.nmr.chem.uu.nl/haddock/analysis.html
[5] http://www.wenmr.eu/wenmr/haddock-web-server-tutorial

With CNS and HADDOCK programs sourced in your terminal session navigate to the run then

```
cd structures/it1/water
```

In this directory run the HADDOCK tool `ana_structures.csh`

```
$HADDOCKTOOLS/ana_structures.csh
```

Move into the analysis directory

```
cd analysis
```

Now run:

```
cluster_struc RUN_NAME_rmsd.disp 7.5 4 > cluster.out
```

Where `RUN_NAME` is the name of the HADDOCK run, 7.5 is the RMSD distance cutoff to which structures within an RMSD of 7.5 $\AA$ are clustered, with a minimum cluster size of 4 structures.

Move back to the water (or other solvent) directory

```
cd ..
```

Now run `ana_clusters.csh` to perform cluster analysis on the structures (this requires ProFit to be defined in the `haddock_configure.csh` file). This script will only analyse the top 4 scoring structures in the cluster analysis, the best flag can be removed (but will increase calculation time) or the number changed.

```
$HADDOCKTOOLS/ana_clusters.csh -best 4 analysis/cluster.out
```

4 An example docking experiment

MISHAP comes with the folder `test_files` in the root directory. In this folder the user will find all the files necessary to run their first MISHAP/HADDOCK docking experiment, without performing any PELDOR experiments. It is assumed that the user is proficient in DeerAnalysis and MMM and would be able to follow the instructions in chapter 2 to generate their own files. If the user is not versed in these programs then it is recommended to read the freely available user manuals[1][2].

4.1 Example files

The `test_files` folder contains `2K5X.pdb` the published NMR energy minimised structure of colicin E9 DNase bound to it's immunity protein, Im9 from rcsb.org. There are also the files `2K5X_ChainA.pdb` and `2K5X_ChainB.pdb` files created with the EPR toolbox's[3] PDBSplitter function, which simply splits the original PDB file into seperate chains.

Additionally in this folder we have several files with the root `4PEL_[2K5X](A)23_[2K5X](B)3`, these files are saved DeerAnalysis results from a PELDOR experiment between the C23 position of the Im9 to the S3C position of colicin E9 DNase.

Finally we have the files:

`rotamers_[2K5X](A){1}23_IA1_175.pdb`

`rotamers_[2K5X](B){1}3_IA1_175.pdb`

These are the saved MMM site scan rotamer analysis files of the 2 sites using IA-PROXYL at 175 K.

4.2 MISHAP

Open MatLab and into the command window type

MISHAP

A splash screen will display briefly whilst MISHAP is loaded.

Initially only the MISHAP Distance Distribution window will be open. In the `Distance Distribution` box, select `DeerAnalysis` from the menu and click the `Load` button.

A file selection window will appear. Navigate to the MISHAP test files folder and select the DeerAnalysis file `4PEL_[2K5X](A)23_[2K5X](B)3__distr.dat`. Click `Open`.

[1] http://www.epr.ethz.ch/software/DeerAnalysis2013_manual.pdf
[2] http://www.epr.ethz.ch/software/MMM_2013_manual.pdf
[3] morganbye.net/eprtoolbox

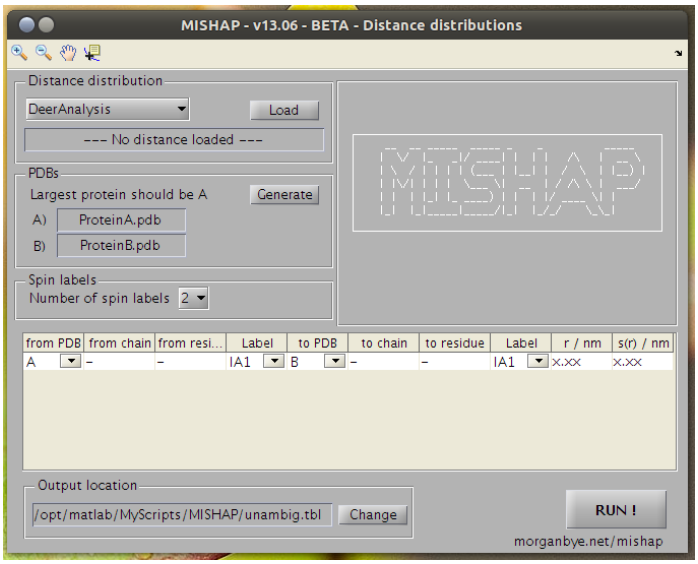


Figure 4.1 *MISHAP distance distribution window*

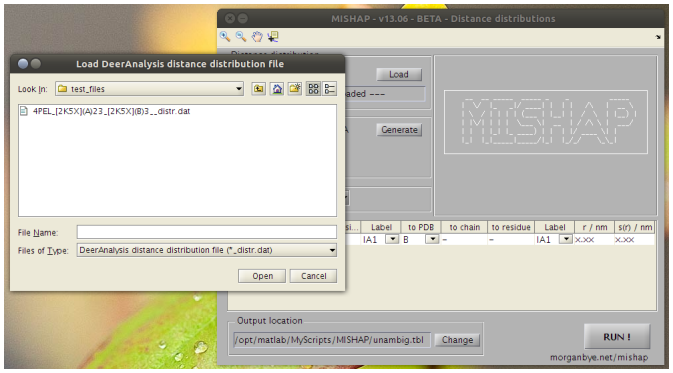


Figure 4.2 *MISHAP distance distribution load interface*

The distance distribution will be opened and displayed in the box in the top right of the window. The distribution will automatically be scanned for peaks and the best peak selected. Notice that the `r / nm` column of the table has been updated with the peak information.

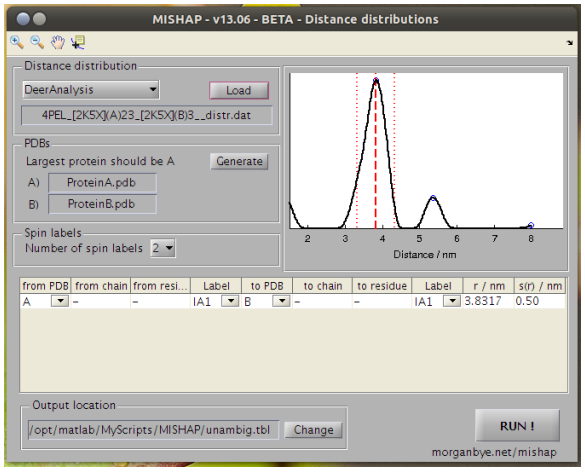


Figure 4.3 *MISHAP distance distribution window showing a loaded distance distribution*

In the `PDBs` box click the `Generate` button.

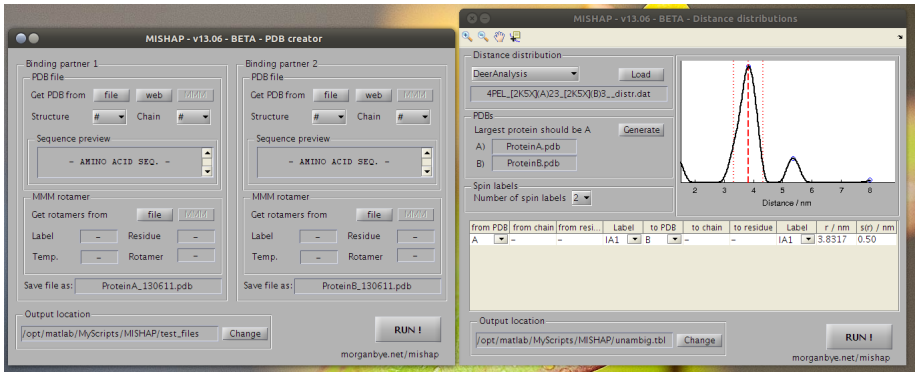


Figure 4.4 *MISHAP distance distribution and PDB creator windows*

This will open the MISHAP PDB creator window. First only consider the left hand side of the window.

Click the `file` button next to `Get PDB from`, again a file selection window will appear. Navigate to the MISHAP test files folder and select `2K5X.pdb`

After a few seconds the `Sequence preview` will display an amino acid sequence the `Structure` will update with `2K5X`, and `Chain` with `A`

In the `MMM rotamer` box click the `file` button, selecting `rotamers_[2K5X](A){1}23_IA1_175.pdb`

After a few seconds the boxes in `MMM rotamers` will update with the rotamer information that has just been loaded.

Now repeat the process on the right hand side for `Binding partner 2`. This time however, make sure to change the `Chain` with the dropdown menu to B and instead select the other MMM rotamer file: `rotamers_[2K5X](B){1}3_IA1_175.pdb`

Check that the `Output location` is set to a suitable location, it can be changed by clicking `Change`

Click `RUN`!

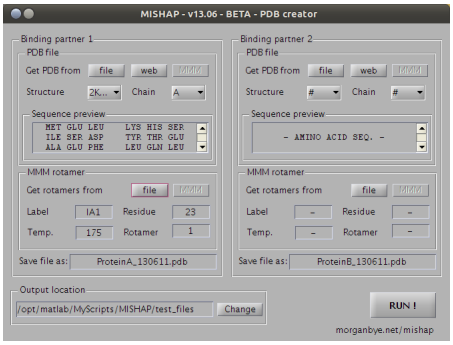


Figure 4.5 *MISHAP PDB creator load with PDB and MMM file loaded*

The command window will update as the PDBs are generated.

The PDB generation will also update the table in the Distance distribution window.

Now click `RUN!` in the Distance distribution window.

The command window will update as the distance restraints file is generated.

4.3 HADDOCK run generation

Following HADDOCK documentation create a new folder that contains the 2 protein PDBs (from MISHAP), `ambig.tbl` (NMR AIRs file, if available), `unambig.tbl` (from MISHAP) and a `new.html` file.

The new.html file needs to be editted to reflect your personal settings regarding folder location, HADDOCK location, run name, etc. Be sure to source HADDOCK with the MISHAP parameter files.

Open a terminal.

Per HADDOCK documentation source CNS and HADDOCK.

Move to run directory.

Initialise HADDOCK with:

```
haddock2.1
```

This will generate the HADDOCK run. Move into the run directory.

Edit `run.cns` as per requirements detailed in §3.4

Copy run.cns into /protocols

```
cp run.cns protocols/run.cns
```

Copy restraints files (`ambig.tbl` and `unambig.tbl`) into /structures/it0, /structures/it1 and /data/distances

Run HADDOCK with:

```
haddock2.1
```

4.4 HADDOCK run analysis

Perform the data analysis outlined in §3.5.

Open MatLab, change directory to the run folder. Move to /structures/it1/water.

Run `HADDOCK_CLUSTERS_PROCESSING.m` and `HADDOCK_CLUSTERS_TO_PYMOL.m` scripts if desired (detailed in §5.2).

5 Other tools

5.1 Naccess

Naccess is a program that calculates the exposed surface area of a protein by running a simulated sphere across the surface of a protein. By default the sphere radius is set to 1.4 $\mathring{A}$ the approximate diameter of a water molecule.

Within the `MISHAP/_private` folder is the `NACCESS` folder, this folder contains `MISHAP.radii`, a special file which allows Naccess to calculate the exposed surface area of a protein which includes the residues R1A/CYM and IA1.

5.2 HADDOCK processing tools

5.2.1 HADDOCK cluster processing

Included is a MatLab script, `HADDOCK_CLUSTERS_PROCESSING.m`, for the processing of HADDOCK clusters. It allows the user to select a cluster or to operate on all the final structures and present them as a figure plotting HADDOCK score against any of the HADDOCK parameters (bsa, dH, Edesolv, rmsd, rmsd-Emin (default), ener (by default this is Einter), ener-Einter, ener-Enb, ener-Evdw, ener-Eelec, ener-Eair, ener-Ecdih, ener-Ecoup, ener-Esani, ener-Evean, ener-Edani), as shown in figure 5.1. Each structure that completed the docking run is shown as a blue circle and the top 4 scoring structures from each cluster are averaged and shown as closed red circles.

Scores can be saved to the workspace, if the script is called with output variables.

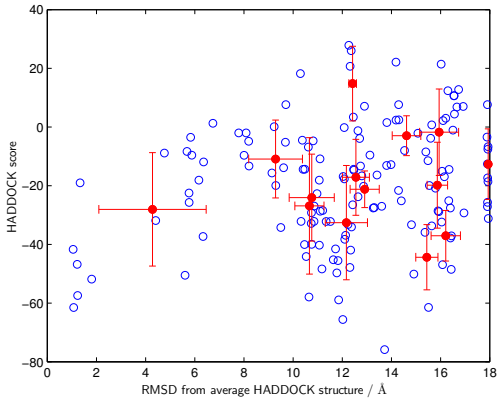


Figure 5.1 *A typical figure from the* `HADDOCK_CLUSTERS_PROCESSING` *script showing HADDOCK score plotted against the RMSD of the structure from the average structure*

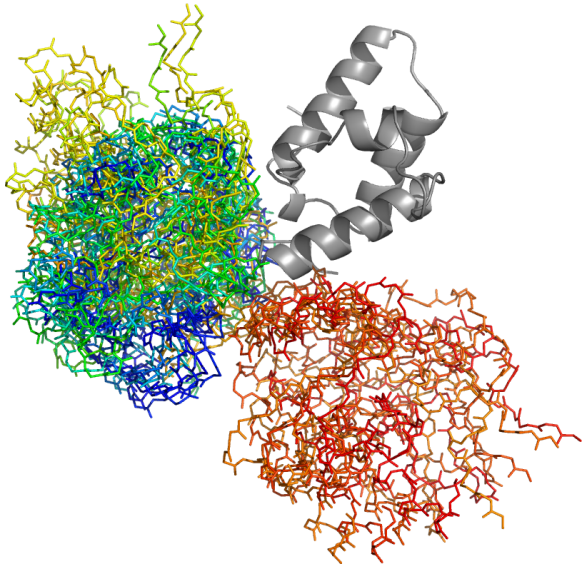


Figure 5.2 A typical figure from the `HADDOCK_CLUSTERS_TO_PYMOL` *script showing chain B (in grey cartoon) plotted against the top 4 structures of the top 4 clusters*

5.2.2 HADDOCK clusters to PyMOL

PyMOL is a long established program for the three dimensional visualisation of protein structures and their ligands. Due to the nature of the program, PyMOL can easily be called from the command line using Python code.

Included is a MatLab script, `HADDOCK_CLUSTERS_TO_PYMOL.m`, which when called from MatLab by default will take the top 4 structures from the top 4 clusters and align all of the chain B's using an RMSD refinement process. The top result from chain B is displayed as a cartoon model and all the chain A's backbones as colour coded lines, where blue is the best scoring model and red is the 4th structure from the 4th cluster.

This script can be called with different input arguments to adjust the number of clusters and structures from each cluster to be considered. This takes the general form:

```
HADDOCK_CLUSTERS_TO_PYMOL(number_of_clusters , structures)
```

For example,

```
HADDOCK_CLUSTERS_TO_PYMOL(1 , 16)
```

Would present the top 16 structures from only the top scoring cluster.

6 Postface

6.1 License

For a human-readable summary of the legal code below, please visit:

http://creativecommons.org/licenses/by-nc-sa/3.0/

This is a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To use MISHAP assumes agreement with this license, and is repeated in LICENSE.txt in the MISHAP root directory.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

"Adaptation" means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License. "Collection" means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(g) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined above) for the purposes of this License. "Distribute" means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership. "License Elements" means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, Noncommercial, ShareAlike. "Licensor" means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License. "Original Author" means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast. "Work" means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work. "You" means an individual or

entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation. "Publicly Perform" means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images. "Reproduce" means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium. 2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections; to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified."; to Distribute and Publicly Perform the Work including as incorporated in Collections; and, to Distribute and Publicly Perform Adaptations. The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved, including but not limited to the rights described in Section 4(e).

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(d), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(d), as requested. You may Distribute or Publicly Perform an Adaptation only under: (i) the terms of this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-NonCommercial-ShareAlike 3.0 US) ("Applicable License"). You must include a copy of, or the URI, for Applicable License with every copy of each Adaptation You Distribute or Publicly Perform. You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License. You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License. You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in con-nection with the exchange of copyrighted works. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and, (iv) consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(d) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit

required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties. For the avoidance of doubt:

Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License if Your exercise of such rights is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c) and otherwise waives the right to collect royalties through any statutory or compulsory licensing scheme; and, Voluntary License Schemes. The Licensor reserves the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License that is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c). Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise. 5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING AND TO THE FULLEST EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO THIS EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above. 8. Miscellaneous

Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

6.2 Third-party software

6.2.1 Software within MISHAP

MISHAP makes use of the following freely available scripts from MatLab Central

Peak detection

fpeak.m by Geng Jun, 2003

Wrapping of strings

linewrap.m by Steve Eddins, 2006

6.2.2 MMM

MMM (Multiscale Modeling of Macromolecular systems) is an open-source program for the visualization, inspection, and improvement of models of proteins and protein assemblies based on restraints from multiple experimental techniques. Authored by Yevhen Polyhach and Gunnar Jeschke, with design support by Enrica Bordignon from ETH Zurich, Switzerland with additional development by Stefan Stoll, University of Washington, USA.

Available from: http://www.epr.ethz.ch/software

6.2.3 HADDOCK

HADDOCK (High Ambiguity Driven biomolecular DOCKing) is an information-driven flexible docking approach for the modeling of biomolecular complexes. HADDOCK distinguishes itself from ab-initio docking methods in the fact that it encodes information from identified or predicted protein interfaces in ambiguous interaction restraints (AIRs) to drive the docking process. Authored by Alexandre Bonvin, Utrecht University, NL.

Available from: http://www.nmr.chem.uu.nl/haddock/

With public webservers available at: http://haddock.science.uu.nl/

6.3 Acknowledgements

Morgan Bye - primary software writer, testing, documentation

Fraser MacMillan - head of Henry Wellcome Unit for Biological EPR

6.4 Hardware and computation time

The vast majority of testing of MISHAP has been conducted upon a 64-bit Linux system (Ubuntu 12.04 LTS), MatLab R2011a.

This system comprises an Intel Core 2 processor @ 2 $\times$ 2.40 GHz with 4 GB of RAM.

With this system generation of a distance restraint file is ~5 ms, PDB file load time is ~3 s (for a ~25 kDa protein with ~2500 atoms), MMM rotamer file load time is ~3 s, the PDB creator running 2 proteins (~ 15 kDa) is ~1 s.

These timings will not vary much with newer machines as MatLab by default will not parallelise tasks. If your version of MatLab is set up to use parrallelisation these times will drop significantly.

The MISHAP MatLab interface has been briefly tested upon a Windows XP machine (MatLab 2010a) and is shown to be functional. Though with the specifications of this machine (Intel Celeron @ 1 $\times$ 2.00 GHz with 1 GB of RAM) loading of PDB files can take as long as 10 s.

HADDOCK runs were performed on a local computing cluster, a 16 core Intel Xeon server node with 32 GB of RAM would complete a run in ~4 - 18 h. Whilst HADDOCK runs are possible on a desktop computer the calculation time will be approximately a week and in which time the machine will be otherwise unusable.

6.5 References

S.J. de Vries, M. van Dijk and A.M.J.J. Bonvin "The HADDOCK web server for data-driven biomolecular docking." Nature Protocols, 5, 883-897 (2010).

T.A. Wassenaar, M. van Dijk, N. Loureiro-Ferreira, G. van der Schot, S.J. de Vries, C. Schmitz, J. van der Zwan, R. Boelens, A. Giachetti, L. Ferella, A. Rosato, I. Bertini, T. Herrmann, H.R.A. Jonker, A. Bagaria, V. Jaravine, P. Guntert, H. Schwalbe, W.F. Vranken, J.F. Doreleijers, G. Vriend, G.W. Vuister, D. Franke, A. Kikhney, D.I. Svergun, R. Fogh, J. Ionides, E.D. Laue, C. Spronk, S. Jurka, M. Verlato, S. Badoer, S. Dal Pra, M. Mazzucato, E. Frizziero and A.M.J.J. Bonvin "WeNMR: Structural Biology on the Grid." J. Grid. Comp., 10, 743-767 (2012).