Automated Assignment of NMR Chemical Shifts Based on a Known Structure and 4D Spectra

Manual

Matthias Trautwein and Thomas Exner

8.5.2016

1 Introduction

This manual describe the usage of the program for the automated assignment of NMR chemical shifts in proteins based on a known 3D structure and 4D [^{1}H,^{15}N]-HSQC-NOESY-[^{1}H, ^{15}N]-HSQC spectra developed by Matthias Trautwein in the group of Thomas Exner at the University of Tuebingen, Department of Pharmacy. More information can be obtained from the paper M.Trautwein, K. Fredriksson, H.M.Moeller and T.E.Exner, in preparation. The prepository contains the source code, an executable for the Linux operation system and the input files for the Ubiquitin test case described in the paper. The following files should be available:

file	Usage
README	a short introduction
assignment	executable
header.h	source code
assignment.cc	source code
prepare.cc	source code
results.cc	source code
routines.cc	source code
logfile	an example logifle
fixed	an example file for the pre-assignment
NOESY.peaks	An artificial 4D spectrum
exp.nmr	chemical shifts for Ubiquitin, downloaded from the BMRB
1D3Z.pdb	3D structure of the Ubiquitin, downloaded from the PDB

The provided executable uses the input files for the Ubiquitin test case as standard parameters. Therefore, no parameters are needed to run this test and the program can be executed just by:

```
> ./assignment
```

An assignment for the Ubiqutin will be calculated and the results will be stored in the file 'logfile'. Please note that the logfile will be overwritten without warning if already present from a previous run. Therefore please move this file to keep previous results before executing the program again.

2 Compilation

Depending on the Linux version and the installed libraries, it might be necessary to recompile the programm. Since no special packages were used, compiling on any system with the gcc-standard compiler should be straight forward. Uisng the g++ compiler, just type the following to recompile the program:

```
> g++ -O2 -o assignment assignment.cc
```

The O2-flag is optional producing optimized code for faster execution. Any additional flags are not necessary.

3 Usage and Parameters

The following parameters are available to specify the input and adapt the program for other case studies:

Parameter	**Usage**
-h	This help screen
-p <file>	Pdb-file (default: 1D3Z.pdb)
-n <file>	NOESY peak list (default: NOESY.peaks)
-l <file>	Log-file (default: logfile)
-c <file>	Chemical shifts file (optional parameter: no default value)
-s <float>	Safety distance added to restraints (default: 0.2)
-f <file>	Fix chemical shifts (default: no default value)
-r <bool>	Call referencing routine, 0 = FALSE, 1 = TRUE (default: 0)
-d <bool>	Call identify double peaks routine, 0 = FALSE, 1 = TRUE (default: 0)

You can get this help screen by typing -h

```
> ./assignment -h
```

Only parameters, for which the default value need to be adapted, have to be specified. As already described above, all default values are set for the Ubiquitin test case but many can be directly applied to other case studies.

- -p: This parameter specifies the pdb-file with the 3D structure of the protein. The selected pdb-file and the executable have to be in the same directory. Only HEADER and ATOM entries are needed. All other entries will be ignored. Positions of the amid hydrogen atoms have to be present in the file. If this is not the case, please use other modeling tools to add them.

- -n: This parameter specifies the file with the peak volumes of the 4D-spectrum. Also this file has to be in the same directory as the executable. The format is as follows:

  ```
  #reference
  0.000 000.00 0.000 000.00   2000   2000 0
  #peaks
  8.859 123.26 8.859 123.26 819200 819200 0
  8.859 123.26 8.227 115.29    790    790 0
  ```

  The number of decimals is arbitrary since all shift values are read as floating point variables. The first line is a comment that indicates that the next line is the reference. The second line is the reference. It has four entries of zero and twice the reference value. The reference value, is the peak volume for a peak that originates from two signals that have a spatial distance of 4 Å. If the optimal value for this volume is not known, which will be the case in most application, it has to be determined by slowly increasing it as described in the paper (see -r parameter) . The third line is a comment that indicates that the experimental volumes follow. Signals have the following format: First the chemical shift values for 1H and 15N for the origin peak, followed by the chemical shift values for 1H and 15N for the NOESY peak. The fifth and sixth entry is the peak volume. One or zero in the last column indicates that the signal is and is not a double peak, respectively.

- -l: This parameter sets the name of logfile. All the output produced by the program will be stored in the logfile. The logfile will be produced in the same folder as the executable and an existing logfile will be overwritten without any warning!

- -c: If this parameter is set, it will be tried to remove any remaining ambiguities by comparison with calculated chemical shifts. The format for the file with the predicted shifts should look like this:

  ```
  891 75 GLY N    111.039
  892 75 GLY CA    45.3
  893 75 GLY HA2    3.92
  ```

  The first number is the entry number. It is not evaluated. The second number is the residue number, in the example lines it indicates that the shifts belong to residue 75. The following strings are the name of the residue and the atom respectively. The last value is a floating point, that contains the chemical shift in ppm.

- -s: This value is added to all calculated distances as a safety margin. Only distances that are longer than this safety margin plus the calculated distance from the 3D-structure are considered as violated.

- -f: This parameter is used to specify known, i.e. pre-assigned, chemical shifts of amid groups in the 3D-structure. The assignment that should be kept fixed are read from a file. As an example see the file 'fixed'. The format for the file with the assignments looks like this:

  ```
  1 47 GLY 102.58 8.10
  2 51 GLU 123.16 8.35
  ```

  The first number is the entry number. It is not evaluated. The second number is the residue number (residue 47 and 51 in the example above). The following two floating points contain the chemical shifts in ppm for the nitrogen and the hydrogen of the amide group, respectively.

- -r: If set true (1), the program ignores the reference value in the NOESY peak list (see -n parameter) and calls a routine to find a reference by its own. Be careful when setting this parameter: This may considerably increase the calculation time!

- -d: If set true (1), the program ignores the last character (0 or 1) in the NOESY peak list and tries to identify double peaks automatically. Be careful: Calculation time scales exponentially with the number of double peaks to be identified.