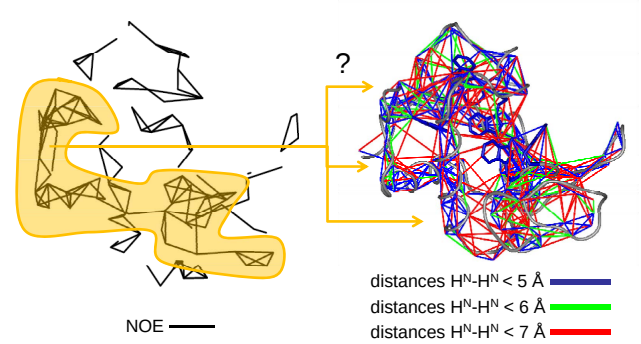
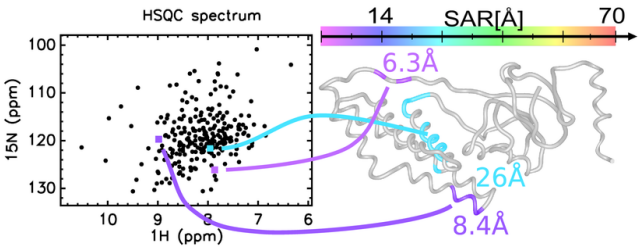
NOEnet 1.0 - documentation

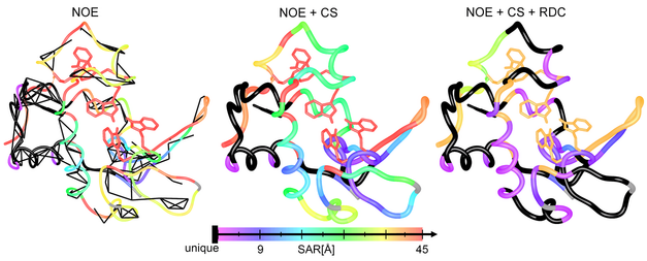
Dirk Stratmann[1], Eric Guittet[2], and Carine van Heijenoort[2]

[1]Institut de Minéralogie et de Physique des Milieux Condensés (IMPMC), Université Pierre et Marie Curie, Paris VI, France
[2]Laboratoire de Chimie et Biologie structurales, ICSN-CNRS, 91190 Gif-sur-Yvette, France

23rd August 2010







Contents

1 Introduction

1.1 What is NOEnet?

1.1.1 Aim

The NOEnet software aims to solve the assignment problem of protein backbone resonances in Nuclear Magnetic Resonance (NMR) by exploiting an available 3D structure of the protein. The general idea behind *structure-based* assignment approaches is that NMR can do more than just obtain the 3D structure of a protein. NMR is well suited to study the dynamics of the protein or its interaction with binding partners. Often a high-resolution 3D structure of the protein studied is already known from X-ray crystallography, so that the inverse way structure→assignment is possible in many cases, helping functional NMR studies. *Structure-based* assignment allows the use of alternative NMR data sets, compared to the usually used sequential J-coupling connectivities. At the same time, structure-based assignment requires sophisticated algorithms to exploit these alternative, but often sparse NMR data sets.

Compared to the standard NMR assignment approach which is based on sequential J-coupling connectivities involving ^{13}C nuclei, the *structure-based* assignment approach of NOEnet uses a completely independent data set, based on spatial NOE-connectivities among $H^N - H^N$ nuclei. The comparison of the NOE-network with the 3D structure yields already satisfactory assignment results, which are improved by the addition of CS and/or RDC data. One step further, the combination of the two orthogonal data sources - J-coupling and NOE - gives very good assignment results, even for large proteins or difficult cases. The *structure-based* assignment approach of NOEnet is therefore a promising alternative to the standard NMR assignment approach and at the same time a helpful complement to it.

Compared to other (structure-based) assignment approaches, NOEnet has the particularity to search for an *assignment ensemble* instead of a unique assignment. The assignment ensemble comprises all assignments that are compatible with the available experimental data. A search for a unique assignment for all NMR resonances would introduce assignment errors using sparse experimental data. NOEnet can handle a wide range of input data, it needs only unambiguous $H^N - H^N$ NOEs as a minimum data set and chemical shifts (CS) of different nuclei as well as residual dipolar couplings (RDC) can be added.

More information about NOEnet can be found in two publications: [1], [2].

1.1.2 Method

NOEnet samples all possible matches between the experimental NOE-network and the network of spatial connectivities obtained from the available 3D structure of the protein. Each matching respects at the same time the restraints imposed by CS and/or RDC data.

The sampling is limited by several threshold-parameters which have to be optimized for each individual data set. Tighter thresholds means less possibilities to sample, but they can also cause the removal of correct assignments from the list of assignment possibilities given for each peak (=resonance). Fortunately, too tight thresholds can be identified by "holes" in the assignment ensemble, i.e. some peaks have no assignment possibility left at all. The optimal thresholds are generally just a little bit less tight, than the ones that generate a small number of "holes". This property is exploited in the *parameter optimization protocol* of NOEnet that is explained in [2]. In the current version of NOEnet, this protocol can only be done manually, but it may be automated in future versions.

Due to the large number of matchings that have to be sampled, the runtime of NOEnet can vary from 30sec up to several days, depending mainly on the tightness of the threshold-parameters. Fortunately, runs that take much more time than 24 hours have usually non-

optimal parameters and can therefore be aborted. Nevertheless, several runs with different threshold-parameters have to be performed, in order to find the optimal parameter set. Therefore the requirements in computing resources are quite important for NOEnet and the use of a computer cluster is recommended. A single run of NOEnet is not parallelized, but the different parameters sets can be tested in parallel by running NOEnet in several instances.

The obtained assignment ensemble can be used as assignment constraints in other assignment programs. For example, we showed[2] with the large protein EIN (259 a.a.) that the assignment ensemble obtained from NOEnet can be used as input to the assignment program MARS[3]. MARS exploits sequential J-coupling connectivities to obtain assignment information. By combining NOEnet with MARS (or others), spatial NOE and sequential J-coupling connectivities can be used together to obtain a more complete and more confident assignment of the protein backbone resonances.

1.2 Installation and testing

NOEnet is a stand-alone command line executable without any graphical interface. Its installation is therefore straightforward:

1. Unpack the tar.gz archive: "tar -xvzf NOEnet_1.0.tgz"
2. A directory "NOEnet_1.0" will be created, containing the binary "NOEnet" and four sub-directories:
   (a) data/ → the input data of the given examples
   (b) doc/ → the documentation of NOEnet including the present PDF
   (c) par/ → the parameter files of the given examples
   (d) out/ → the output files of the given examples
3. NOEnet can be executed simply from inside the directory "NOEnet_1.0" by typing "./NOEnet *parameter-file*".

Three examples are given with the current distribution of NOEnet: Ubiquitin, Lysozyme and EIN. The correct installation of NOEnet can be tested by executing one or all three examples:

From a terminal window go into the installed directory "NOEnet_1.0" and execute one of the three commands:

1. "./NOEnet par/Ubiquitin_NOE.txt"
2. "./NOEnet par/Lysozyme_NOE_CS_RDC.txt"
3. "./NOEnet par/EIN_NOE_CScarbon_RDC.txt"

After a short runtime of about 30-60 seconds for the Ubiquitin- and Lysozyme-example and about 5-10 minutes for the EIN-example, NOEnet should have written all output files into the output directory "out/xxx" (xxx = Ubquitin, Lysozyme or EIN) and finished with the statement: "NOEnet has correctly finished the calculation :-)".

As detailed below in the "Input & Output" section, NOEnet writes a "*.log" file and several "*_N.txt" files, with "N" being the iteration number. The "*_N.txt" files contain the obtained assignment ensembles. After running NOEnet on the three examples, the obtained "*_N.txt" files (with the highest "N") can be compared to the given ones using "diff" or similar (sort the files by their date to distinguish between the given ones and the ones that you generated). Small differences between the assignment ensembles can occur, as the search path is chosen randomly with each new run, but most parts of the assignment ensembles should be the same.

To obtain more information about the syntax of the output files, please read the section "Output files" below.

2 Input & Output

2.1 Input files

The current distribution of NOEnet 1.0 contains the following four sub-directories:

1. data/
2. doc/
3. par/
4. out/

The "data" and the "out" subdirectories contain a subdirectory for each example, e.g. data/EIN, data/Lysozyme and data/Ubiquitin.

2.1.1 Input data

The data subdirectory contain the input data files for NOEnet. Some input files are always required, while others are optional, depending on whether you use the chemical shift (CS) filters and/or the residual dipolar coupling (RDC) filter. First the list of input files that are always required:

1. xxx_peaks.txt A three column data file, each line containing the ID of a peak of the HSQC and its ^{15}N and $^{1}H^{N}$ chemical shift values.

2. xxx_noes.txt A three column data file, the first two columns give the IDs of two HSQC peaks for which an unambiguous $^{1}H^{N} -^{1} H^{N}$ NOE has been observed experimentally on the protein that has to be assigned. The thrid column gives the intensity class of the NOE: 1 = weak, 2 = medium, 3 = strong

3. xxx.pdb The free form 3D structure of the protein that has to be assigned. The $^{1}H^{N}$ (in PDB-file: 'H') backbone protons and the $^{1}H^{\epsilon 1}$ (in PDB-file: 'HE1') tryptophan side-chain protons must be present in the PDB file. The tryptophan $^{1}H^{\epsilon 1}$ protons have to be included, as they cannot be distinguished from the $^{1}H^{N}$ resonances in an unassigned [^{1}H, ^{15}N] HSQC spectrum. You can use for example WHAT IF (http://swift.cmbi.ru.nl/ servers/ html/ index.html) or reduce (http://kinemage.biochem.duke.edu/ software/ reduce.php) to add all protons to X-ray PDB structures.

Predicted chemical shift values have to be given, if the $^{1}H^{N}$ and ^{15}N chemical shift filters are used (useCSfilter = 1 in the parameter file, see Table 2), which is usually the case as experimental [$^{1}H^{N}$, ^{15}N] chemical shift values are always available (xxx_peaks.txt input file).

4. xxx_CS_ShiftX.txt The predicted chemical shift values using the same free form 3D structure as the one given to NOEnet (see previous input file). The file-format must be in BMRB *.str format (http://www.bmrb.wisc.edu). You can use for example the ShiftX-server to predict the chemical shifts from the PDB 3D structure: http://redpoll. pharmacy. ualberta.ca/ shiftx. Choose "All Shifts" and "BMRB format" in the server options.

If experimental ^{13}C chemical shifts are available, the corresponding filter can be used using the xxx_CS_ShiftX.txt and the following input file:

5. xxx.tab The experimental chemical shift values in the same tabular format that is used by MARS[3]. The first column is the ID of a peak of the HSQC as defined in xxx_peaks.txt. The first row contains the chemical shift names. As NOEnet does not handle sequential connectivity data obtained by standard triple resonance experiments, only one chemical shift per carbon-nuclei is used, namely the one that corresponds to the previous (i-1) residue. Accordingly the chemical shift name must be written as 'CA-1' for $^{13}C^{\alpha}$, 'CB-1' for $^{13}C^{\beta}$ and 'CO-1' for $^{13}C'$. The 'CA', 'CB' and 'CO' values given for the EIN example (see data/EIN/1zym_cs_peakIDs.tab) are not used by NOEnet and are not needed in the xxx.tab input file. Also the 'N' and 'H' columns given in the example file for EIN are not needed, as these values are already in the xxx_peaks.txt input file. NOEnet searches in the first row of the xxx.tab file for the names 'CA-1', 'CB-1' or 'CO-1', so that their order or column number is not important. Also any subset of the three carbon types can be used. Inside the data (row > 1), missing values must be indicated by any non-numeric character like '-'.

If experimental residual dipolar couplings (RDC) of the $^{15}N - H^N$ bond vectors are available, the RDC-Filter can be applied using one input file per complete RDC data set. More than one RDC data set can be measured by the use of different types of alignment media. The current version of NOEnet handles up to two $^{15}N - H^N$ RDC data sets, but on request it could be extended easily to handle more than two. Below is the description of the data format of the one or two input data files for RDCs:

6. xxx_RDC_one.txt and/or xxx_RDC_two.txt These are two-columns text files without any header. The first column contains the IDs of the [$^1H^N$, ^{15}N] peaks and the second one the corresponding $^{15}N - H^N$ RDC values of one alignment medium. Missing RDC values can either be indicated by -999 (see Lysozyme example files) or the whole (peak-ID + RDC-value) entry can be skipped from the list.

Finally, you can find also the following file with the example input data:

7. xxx_assignment.txt The reference assignment: the first column contains the IDs of the peaks and the second column the corresponding residue numbers. **This is not an input file, but it is just given with the examples to verify the output of NOEnet, i.e. to compare the obtained assignment ensemble with the reference assignment.**

2.1.2 Input parameters

The par subdirectory contains the parameter file for NOEnet. Each parameter has an identifier and a value (integer, float or string) written on an individual line and separated by one or more space(s) or tabulator(s). The order of the parameters is not relevant for NOEnet, but they should be organized in a similar way as in the examples distributed with NOEnet.

The list of mandatory parameters are given in Table 1 and the parameters for the chemical shift and residual dipolar couplings filters are given in Table 2 and 3, respectively. The essentially constant parameters of the search algorithm of NOEnet are shown in Table 4.

Table 1: Mandatory parameters for NOEnet.

name	description	typical values
	Files and Directories	
outputDir	Output directory	out/EIN
peakFilename	List of HSQC peak IDs with ^{15}N- and $^{1}H^{N}$-CS	data/EIN/ EIN_peaks.txt
pdbFilename	PDB file of the protein 3D structure	data/EIN/ 1ZYM.pdb
noefile	NOE data	data/EIN/ EIN_noes.txt
	NOE-data	
noedistTheo(weak)	**Maximum $^{1}H^{N}$-$^{1}H^{N}$ distance d_{max}^{theo} for weak NOEs**	**7Å**
noedistTheo(medium)	**Maximum $^{1}H^{N}$-$^{1}H^{N}$ distance d_{max}^{theo} for medium NOEs**	**6Å**
noedistTheo(strong)	**Maximum $^{1}H^{N}$-$^{1}H^{N}$ distance d_{max}^{theo} for strong NOEs**	**5Å**
useNOEoutlierFilter	Switches the NOE outlier filter on(=1) or off(=0)	1
noedistOutlierRange	**Δd distance by which the maximum distance for non-outliers is reduced**	**1-1.5Å**
maxNOEoutlierNum	**Number of allowed outliers T_{NOE}**	**1-10**

The units are not written in the parameter file. The parameters shown in bold font have to be optimized, for the others the indicated typical value can be used in all runs. See [2] for the definition of the parameters and their optimization.

Table 2: Parameters for the chemical shift (CS) filters

name	description	typical values
	CS-data	
useCSfilter	Switches the chemical shift filter on(=1) or off(=0)	1
theoCSfile	Predicted chemical shifts from the 3D structure	data/EIN/ 1ZYM_CS_ShiftX.txt
csFilterDecayConstant	**Decay constant c_{CS} of RMSD filter**	**10-30 residues**
startCSmaxRMSD_N	Upper RMSD threshold u_{CS} for ^{15}N-CS	10ppm
endCSmaxRMSD_N	**Minimum RMSD threshold m_{CS} for ^{15}N-CS**	**3ppm**
startCSmaxRMSD_HN	Upper RMSD threshold u_{CS} for $^{1}H^{N}$-CS	2ppm
endCSmaxRMSD_HN	Minimum RMSD threshold m_{CS} for $^{1}H^{N}$-CS	0.8ppm
csFilterMinNodesChecked	Minimum number of assigned peaks required for the CS-filter	5
CSerrorN	The maximum relative error between predicted and measured CS for a given assignment of a single peak (in % of the range of experimental CS values). CSerror = $abs(100 * \frac{\delta_{exp}-\delta_{theo}}{\Delta\delta_{exp}})$. The current assignments outside this error margin are not rejected, but just tested at the end of the search tree, as they are less likely to be the correct assignment. Improves the runtime performance of NOEnet.	20%
CSerrorHN	See above for CSerrorN	60%
useCScarbonFilter	Switches the CS filter for ^{13}C-CS on(=1) or off(=0)	0
expCSfileCarbon	Experimental ^{13}C-CS	data/EIN/ 1zym_cs_peakIDs.tab
CSerrorCA	See above for CSerrorN	10%
CSerrorCB	See above for CSerrorN	10%
CSerrorCO	See above for CSerrorN	10%
startCSmaxRMSD_CA	Upper RMSD threshold u_{CS} for $^{13}C^{\alpha}$-CS	10ppm
endCSmaxRMSD_CA	**Minimum RMSD threshold m_{CS} for $^{13}C^{\alpha}$-CS**	**2ppm**
startCSmaxRMSD_CB	Upper RMSD threshold u_{CS} for $^{13}C^{\beta}$-CS	10ppm
endCSmaxRMSD_CB	**Minimum RMSD threshold m_{CS} for $^{13}C^{\beta}$-CS**	**2ppm**
startCSmaxRMSD_CO	Upper RMSD threshold u_{CS} for ^{13}CO-CS	10ppm
endCSmaxRMSD_CO	**Minimum RMSD threshold m_{CS} for ^{13}CO-CS**	**2ppm**

The units are not written in the parameter file. The parameters shown in bold font have to be optimized, for the others the indicated typical value can be used in all runs. See [2] for the definition of the parameters and their optimization.

Table 3: Parameters for the residual dipolar couplings (RDC) filters

name	description	typical values
	RDC-data	
useRDCfilter	Switches the RDC filter on(=1) or off(=0)	1
expRDCfileOne	Experimental $^{15}N-^{1}H^{N}$ RDC values of the first alignment medium	data/EIN/ EIN_RDC_one.txt
expRDCfileTwo	Experimental $^{15}N-^{1}H^{N}$ RDC values of the second alignment medium	data/EIN/ EIN_RDC_two.txt
rdcFilterDecayConstant	**Decay constant c_{RDC} of RMSD filter**	**10-30 residues**
rdcMaxRMSDstart	Upper RMSD threshold u_{RDC}	10Hz
rdcMaxRMSDend	**Minimum RMSD threshold m_{RDC}**	**3Hz**
maxTensorMargin	The maximum additional margin for the RDC threshold. The additional tensor margin ΔT_{RDC} decreases with the number of uniquely assigned peaks according to $\Delta T_{RDC}(n_{unique}) = maxTensorMargin * (1 - n_{unique}/N_{peaks})$	2Hz
useTemporaryTensor	Switches the use of temporary alignment tensors on(=1) or off(=0)	1
minSimplices	Minimum number of assigned simplices ($\approx$ peaks) required for the use of a temporary alignment tensor	15
rdcFilterMinNodesChecked	Minimum number of assigned peaks required for the RDC-filter (using a temporary or permanent alignment tensor)	5
RDCerrorTolerance	The maximum relative error between predicted and measured RDC for a given assignment of a single peak (in % of the range of experimental RDC values). RDCerror = $abs(100 * \frac{D_{exp}-D_{theo}}{\Delta D_{exp}})$. The current assignments outside this error margin are not rejected, but just tested at the end of the search tree, as they are less likely to be the correct assignment. Improves the runtime performance of NOEnet.	25%

The units are not written in the parameter file. The parameters shown in bold font have to be optimized, for the others the indicated typical value can be used in all runs. See [2] for the definition of the parameters and their optimization.

Table 4: Parameters for the search algorithm of NOEnet

name	description	typical values
TRPadditionalDist	Amount by which the distance thresholds are increased for tryptophan NH groups	0.5Å
initialNumPeaks	Maximum size of the subset fragment of the NOE network at the beginning of the search	30
maxSingleTrialCt	Maximum number of search steps	10^6
finalMaxSingleTrialCt	Maximum number of search steps with the entire NOE networks	10^7
maxSingleTrialCtMulti	Multiplier by which (final)maxSingleTrialCt are increased for the next iteration	10
retestStoppedInterval	Indicates the interval of the increase in the size of subset fragments for which the "stopped" assignment possibilities are marked as "not tested" in the assignment matrix, i.e. these possibilities are tested again.	10
retestFoundInterval	Indicates the interval of the increase in the size of subset fragments for which the previously found assignment possibilities are marked as "not tested" in the assignment matrix, i.e. these possibilities are tested again. With larger subset fragment sizes more assignment possibilities become impossible.	10
testOverlap	Switches the "overlap check" on(=1) or off(=0)	1
fastTest	**Switches the "fast mode" of the overlap check on(=1) or off(=0). The fast overlap check should be used for large proteins (> 150 amino acids).**	**0**
maxTestOverlapCt	Maximum number of DFS search steps of the overlap check. Not relevant for the "fast mode".	10^6
maxIterationNum	Number of full search iterations before (final)maxSingleTrialCt are increased by maxSingleTrialCtMulti	1
randomStart	Switches the random choice of the start simplex on(=1) or off(=0)	1
DFS	Switch between a deep first search (DFS) (=1) and a breadth first search (BFS) (=0)	0

The units are not written in the parameter file. The parameters shown in bold font have to be optimized, for the others the indicated typical value can be used in all runs. See [1] for the definition of the parameters.

2.2 Output files

2.2.1 Introduction

The output files of NOEnet are written the specified directory in the parameter file ("outputDir"), for example to out/protein-name. The name of each output file begins with "NOEnet", followed by the name of the parameter file (incl. the path, the slashes '/' are replaced by '_') and the start date (just the day without the month or year) and time of the run. By including the start time of the run in the output file names, name conflicts between parallel runs should be avoided.

The log-file "NOEnet_xxx.log" records the used parameters, all messages of the program, and the results.

After each iteration a "NOEnet_xxx_N.txt" file is written, with N the number of the iteration (beginning with zero). These files contain the assignment ensembles of the current iteration in form of a list of assignment possibilities for each peak. Each peak is indicated by its peak ID followed by a double point ':' and the comma-separated list of residues to which this particular peak can be assigned. Residue numbers > 10000 correspond to tryptophan side-chain [$^{1}H^{\epsilon 1}$, $^{15}N^{\epsilon 1}$] groups (see section 2.1.1), the numbers are generated by adding to the real residue number of the tryptophan-residue the value of 10000. This allows a direct identification of the real residue number, simply by subtracting 10000.

The spatial assignment range (SAR) values are also given with the list of assignment possibilities. At the end of the list is summarized the number of uniquely assigned peaks, as well as the number of peaks, which have a spatial assignment range (SAR) below 4Å, 6Å and 10Å, respectively.

2.2.2 Ranked and refined assignment ensemble

The last "NOEnet_xxx_N.txt" file contains the same assignment ensemble as the previous "NOEnet_xxx_(N-1).txt" file, but the assignment possibilities are scored and ranked individually according to the chemical shift (CS) or residual dipolar coupling (RDC) data, if available. Each line has the following syntax:

peakID [N_{data}]: resID1(score1), resID2(score2), ...

with N_{data} being the number of data values available for this peak (for example, ^{15}N-CS + $^{1}H^{N}$-CS $\rightarrow N_{data} = 2$). The assignment that has a smaller score, gives a better agreement between the experimental and calculated CS/RDC data point. The larger N_{data} is, the higher is the confidence in the ranking.

These individual peak rankings are used in an *individual peak assignment refinement* procedure, described in [2]. The refined assignment ensemble is written in the same file after the ranked assignment ensemble (after "AssignmentRanking::saveRefinement"). The syntax is the same as for the ranked assignment ensemble.

As the refined assignment ensemble retains less assignment possibilities, it may contain a limited number of assignment errors.

References

[1] Dirk Stratmann, Carine van Heijenoort, and Eric Guittet. NOEnet–use of NOE networks for NMR resonance assignment of proteins with known 3D structure. *Bioinformatics (Oxford, England)*, 25(4):474–481, February 2009.

[2] Dirk Stratmann, Eric Guittet, and Carine van Heijenoort. Robust structure-based resonance assignment for functional protein studies by NMR. *Journal of Biomolecular NMR*, 46(2):157–173, February 2010.

[3] Young-Sang Jung and Markus Zweckstetter. Mars – robust automatic backbone assignment of proteins. *J. Biomol. NMR*, 30:11–23, September 2004.