Supplementary

Robust structure-based resonance assignment for functional protein studies by NMR

Dirk Stratmann · Eric Guittet · Carine van Heijenoort

Datasets

Simulated NOE data sets.

In order to test the program performance, we used the medium size protein lysozyme as a model with various simulated NOE data sets. The first case (NOE_1^{SIM}) simulates an ideal case. For that, the NOE interaction graph G^{exp} and the 3D structure contact graph G^{theo} were generated from the same 3D structure (PDB-code 193L (Vaney et al, 1996), X-ray, 1.33Å resolution) and with the same distance threshold $d_{max}^{exp} = 5$Å (283 NOEs and 283 distances). Unfortunately not all theoretical possible NOEs are usually measured, resulting in a graph G^{exp} sparser than G^{theo}. Ambiguous NOEs were identified as described in the article by identifying the superposed experimental [^{15}N, $^{1}H^{N}$] HSQC peaks (taken from the BMRB: bmr4831.str (Schwalbe et al, 2001) for ^{15}N-CS and bmr4562.str (Wang et al, 2000) for $^{1}H-CS$). Removing the ambiguous NOEs reduced the number of simulated NOEs from 283 to 253 (NOE_2^{SIM}). The sparseness of G^{exp} also comes from the decreasing completeness of NOEs with increasing proton-proton distances. A simple completeness function $c(d)$ with a cutoff distance of $d_{max}^{exp} = 6$Å was chosen (Figure S1), approximating experimental distributions (Doreleijers et al, 1999; Koharudin et al, 2003). The reduction in the number of edges in the graph G^{exp} in comparison to G^{theo} is already important, if only the completeness function $c(d)$ is applied without the removal of ambiguous NOEs: G^{exp}- 297 edges ($d_{max}^{exp} = 6$Å), G^{theo} - 385 edges ($d_{max}^{theo} = 6$Å) (NOE_3^{SIM}). Removing also the ambiguous NOEs yielded only 263 edges (NOE_4^{SIM}).

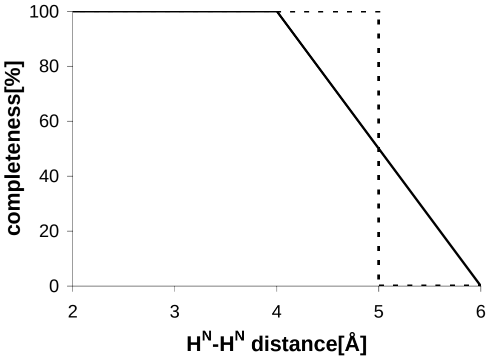


Fig. S 1 NOE distributions: for the 5Å threshold, all distances are used to generate a simulated NOE (dotted line). For the 6Å threshold, distances greater than 4Å are chosen randomly, while respecting the completeness function shown by the plain line.

To test the effect of imperfect matches between experimental NOEs and the reference 3D structure, we generated a more realistic simulated NOE data set (see table S5) using the NMR structure 1E8L (model 49) (Schwalbe et al, 2001) with a cutoff distance of $d_{max}^{exp} = 6$Å for the generation of G^{exp} and the X-ray structure 193L (Vaney et al, 1996) for the generation of G^{theo} with $d_{max}^{theo} = 7.5$Å. The RMSD over the heavy backbone atoms between 193L and 1E8L (model 49) is equal to 1.3Å. As for NOE_4^{SIM}, the ambiguous NOEs were removed and the completeness function $c(d)$ was applied, yielding the simulated NOE data set NOE_5^{SIM}. NOE classes were introduced here: a distance threshold of $d < 3$Å generated 79 simulated NOEs from the NMR structure, classified as *strong* NOEs, distance ranges of $3\text{Å} < d < 4\text{Å}$ and $4\text{Å} < d < 6\text{Å}$ yielded 42 medium and 159 weak NOEs, respectively.

Dirk Stratmann
NMR, Utrecht University, Padualaan 8
3584 CH Utrecht, the Netherlands

Eric Guittet · Carine van Heijenoort
Laboratoire de Chimie et Biologie Structurales
ICSN-CNRS; 1, av. de la terrasse
91190 Gif-sur-Yvette, France
E-mail: eric.guittet@icsn.cnrs-gif.fr

Parameter optimizations

The principle of the parameters optimization is described in the article. Tables S1 and S2 show all the trials performed for the thresholds optimization on lysozyme with realistic simulated (NOE_5^{SIM}) and experimental NOE data (see main text), respectively, without or with adding CS and RDCs experimental data. The trials for EIN are shown in table S3. The optimized threshold combinations that have been retained for the results section of the article are marked by a star.

Realistic NOE simulations for Lysozyme (table S1)

Even without the NOE outlier approach (trial #3), the standard values d_{max}^{theo}= (5Å, 6Å, 7Å) for short, medium and long distances were not compatible with the simulated NOE data set, as always one or more holes occurred in the assignment list. Hole-free assignment ensembles can be obtained with a distance threshold for the weak NOEs increased from 7Å to 7.5Å, after incrementing T_{NOE} from 3 to 10. We tried to obtain better results by the application of tighter CS- or RDC-filter thresholds, but almost all tighter thresholds tested resulted in assignment errors. The best optimization trials #11, #15, #20 and #22 were retained for further discussion, and are called cases 1, 2, 3 and 4, respectively, in the results section below and in table S5.

Sparse experimental NOE data in combination with RDC data on Lysozyme (table S2).

Hole free assignment ensembles can be obtained here with standard distance thresholds and only one NOE outlier. Additionally, the outlier-range could be increased to 1.2Å. The low value of $T_{NOE} = 1$ in comparison with the high value optimized for the simulated NOE data set ($T_{NOE} \geq 10$) can be explained by the fact that the simulated NOE data set contained a higher number of NOEs and thus a higher number of NOE-outliers in the same outlier range. The sparseness of the measured NOE data set brings out the use of tighter thresholds not only for the number of NOE outliers, but also for the CS- and RDC-data as shown in the optimizations of CS and RDC parameters. It should be noted that the precision of the assignment is anyway lower for sparser data. We could reduce the decay constants c_{CS} and c_{RDC} to 10 with the same end-point as for the simulated NOE data set (m_{CS} = 3ppm, m_{RDC}= 3Hz). Although CS- and RDC-data are the same for both cases (simulated and experimental NOEs), the occurrence of incompatibilities in the constraint framework depends on all available data taken together (NOE+CS+RDC). Contrary to CS data, the adding of RDC data yielded a significant increase of unique assignments (trials #6, #19 and #31). The RDC data set contains here two almost independent values per HSQC peak, thanks to the use of two different alignment media. The CS data set only contains one significant value per HSQC peak, namely the ^{15}N CS (the ^{1}H CS does not restrict efficiently the assignment). This can explain the different impact of RDC and CS data. The best optimization trials, #6, #19 and #31, were retained as results in the article, and correspond to cases 1, 2 and 3 in table 2, respectively.

The case of a larger protein : EIN (table S3).

We started with slightly different values for the theoretical distance threshold d_{max}^{theo}= (4.5Å, 6Å, 7.5Å). We chose 7.5Å as upper distance limit because of the long mixing time (170ms) employed for the NOESY experiment. We reduced the distance limit for strong NOEs from 5Å to 4.5Å due to their low number (36 out of 407 NOEs in total). For $T_{NOE} = 3$, we obtained an error-free result with 46 uniquely assigned peaks and 37 peaks with SAR < 10Å (trial #4). We obtained a significant improvement (trials #5 and #6) for $T_{NOE} = 2$, with the limited drawback that the correct assignment has been removed for two peaks (a swap of assignment possibilities: residue 207<->208) without being detected. A value of $T_{NOE} = 1$ resulted in more assignment errors and finally holes in the assignment list. We also tried to tighten the CS-filter thresholds without success (holes being always generated).

Adding the two simulated RDC data sets showed that $T_{NOE} = 2$ is too low, as even with highly relaxed RDC thresholds ($m_{RDC} = 7.5Hz$, trial #17) no hole-free assignment ensemble could be obtained. $T_{NOE} = 3$ corresponds better to the given NOE data set, as even with highly restricted RDC thresholds ($m_{RDC} = 1Hz, c_{RDC} = 5$, trial #26) an error-free and hole-free assignment ensemble could be obtained. The RDC thresholds should not be too close to those yielding holes in the ensemble (trial #28), otherwise small assignment errors may remain undetected (trial #27).

The addition of carbon chemical shifts $^{13}C\alpha(i-1),^{13}C_\beta(i-1),^{13}CO(i-1)$ improves the precision of the assignment ensemble significantly, especially if the *individual peak assignment refinement* procedure (see article in Methods) is applied on the obtained assignment ensembles (trial #31 and #34).

The best optimization trials, #4, #5, #6, #26, #30, #31, #33 and #34, were retained as results in the article, and correspond to cases 2 to 9 in table 3, respectively.

Table S 1 Tested threshold parameters on lysozyme with realistic simulated NOEs (NOE_5^{SIM}) and experimental CS and RDC data

		NOE		CS		RDC		Result					
#	runtime	T_{NOE}	Δd[Å]	c_{CS}	m_{CS}	c_{RDC}	m_{RDC}	Status	N_u	N_{10}	N_e	N_{eu}	SAR_{max}
$d_{max}^{theo} = (5\text{Å}, 6\text{Å}, 7\text{Å})$, NOE only													
1	1 s	3	2					hole	0	0	3	0	
2	1 s	"	1					hole	21	19	4	0	
3	10 min	0	0					hole	4	50	6	0	
$d_{max}^{theo} = (5\text{Å}, 6\text{Å}, 7.5\text{Å})$, NOE only													
4	10 s	3	1					hole	41	12	3	0	
5	1 min	4	"					hole	49	20	5	2	2.5Å
6	20 min	5	"					hole	28	67	5	3	4.5Å
7	20 min	6	"					hole	27	82	3	1	4.5Å
8	2 h	7	"					hole	36	83	7	1	2.7Å
9	15 h	8	"					hole	34	89	5	1	2.7Å
10	5 days	9	"					hole	42	81	6	3	4.5Å
11*	31 h	10	"					not finished	13	100	0	0	
12	48 h	11	"					not finished	11	101	0	0	
13	47 h	12	"					not finished	8	74	0	0	
NOE + CS													
14	65 h	10	1	30	3ppm			hole	65	58	8	4	4.5Å
15*	19 h	11	"	"	"			not finished	18	95	0	0	
16	22 h	12	"	"	"			not finished	17	95	0	0	
Optimize CS													
17	11 h	11	1	20	3ppm			not finished	40	82	3	1	2.7Å
18	7.5 h	"	"	10	"			hole	52	72	2	0	
NOE + CS + RDC													
19	4h30	10	1	30	3ppm	30	3 Hz	hole	82	42	8	4	4.5Å
20*	11 h	11	"	"	"	"	"	finished	115	10	8	6	4.5Å
21	12 h	12	"	"	"	"	"	not finished	50	73	3	1	2.7Å
22*	16 h	13	"	"	"	"	"	not finished	42	78	0	0	
23	17 h	14	"	"	"	"	"	not finished	41	79	0	0	
24	19 h	15	"	"	"	"	"	not finished	40	79	0	0	
Optimize CS													
25	3 h	11	1	20	3ppm	30	3 Hz	finished	116	9	7	6	4.5Å
26	7 h	13	"	"	"	"	"	not finished	50	73	0	0	
Optimize RDC													
27	1 h	11	1	30	3ppm	20	3 Hz	hole	13	66	3	0	
28	1 h	13	"	"	"	"	"	hole	13	49	3	0	
29	1 h	15	"	"	"	"	"	hole	11	26	2	1	2.6Å

Three separated parameter optimizations are shown in this table. The first optimization assumes that only NOE data (NOE_5^{SIM} described above) are available. The second one assumes that NOE and CS data are available and the last one assumes that NOE, CS and RDC data are available. The theoretical distance thresholds for the three NOE classes are here either $d_{max}^{theo} = (5\text{Å}, 6\text{Å}, 7\text{Å})$ or $d_{max}^{theo} = (5\text{Å}, 6\text{Å}, 7.5\text{Å})$, yielding $N_{dist} = (292, 98, 138)$ or $N_{dist} = (292, 98, 224)$ distances in each class (short, medium, long) using the X-ray structure 193L. The number of simulated NOEs is here $N_{NOEs} = (79, 42, 159)$ for strong, medium and weak NOEs, respectively. The number of HSQC peaks is here $N_{peaks} = 132$. The following parameters were optimized: T_{NOE} - maximum number of permitted NOE outliers for an arbitrary matching. Δd - the theoretical distance range $[d_{max}^{theo} - \Delta d, d_{max}^{theo}]$ in Å for which a NOE is considered as outlier. c_{CS} and c_{RDC} - decay constant in number of currently assigned peaks for the decreasing exponential threshold function. m_{CS} and m_{RDC} - minimum limit of the threshold function. *Status: hole* indicates the presence of peaks, which have no assignment possibility left in the assignment table; *finished* and *not finished* indicates whether the trial run converged or not for the given runtime. N_u- number of uniquely assigned peaks. N_{10} - number of peaks having a SAR-value below 10Å and which are not uniquely assigned. N_e - number of peaks for which the correct assignment is missing. N_{eu} - number of uniquely, but wrongly assigned peaks. SAR_{max} - the maximum distance to the correct residue among the N_{eu} peaks.

* Optimized parameter combinations that have been retained for further analysis in the results section of this supplementary material. Trials #11, #15, #20 and #22 correspond to cases 5, 6, 7 and 8 of table S5, respectively.

Results

Introduction.

The results obtained after the parameter optimization using experimental NOE data are analyzed in the article. In order to best delineate the potential of the structure-based method, the influence of data completeness and the impact of differences between the solution and reference 3D structures, we analyze here the results obtained using simulated NOE data sets of increasing realism on lysozyme.

Simulated NOE data on lysozyme using ideal conditions.

NOE*net* yields 95% of unique and correct assignments when using a distance threshold equal to 5Å for both experimental and 3D-structure graphs (case 1 in Table S4). This shows that there exists almost only one possibility to match graph G^{exp} onto G^{theo} if the two graphs are identical. In this case the subgraph monomorphism problem is reduced to an automorphism problem. Removing the ambiguous NOEs reduces the part of uniquely assigned peaks from 95% to 79% (case 2). The majority of the peaks with multiple assign-

Table S 2 Tested threshold parameters on lysozyme with experimental NOE, CS and RDC data

		NOE		CS		RDC		Result					
#	runtime	T_{NOE}	Δd[Å]	c_{CS}	m_{CS}	c_{RDC}	m_{RDC}	Status	N_u	N_{10}	N_e	N_{eu}	SAR_{max}
NOE only													
1	6 h	3	1					not finished	0	27	0	0	
2	5 h	2	"					not finished	0	27	0	0	
3	20 min	1	"					not finished	18	11	0	0	
4	19,5 h	"	"					not finished	18	11	0	0	
5	15 min	"	1,2					not finished	22	9	0	0	
6*	5 h	"	"					finished	22	9	0	0	
7	10 sec	"	1,5					hole	0	0	4	0	
NOE + CS													
8	60 min	3	1	30	3.5ppm			not finished	4	25	0	0	
9	6 h	"	"	"	"			not finished	4	25	0	0	
10	30 min	2	"	"	"			not finished	5	24	0	0	
11	5 h	"	"	"	"			not finished	5	24	0	0	
12	20 min	1	"	"	"			not finished	23	7	0	0	
13	12 h	"	"	"	"			not finished	23	7	0	0	
14	50 min	"	1.2	"	"			not finished	24	7	0	0	
15	2 h	"	"	"	"			finished	24	7	0	0	
16	5 sec	"	1.5	"	"			hole	0	0	4	0	
Optimize CS													
17	75 min	1	1.2	30	3ppm			finished	24	7			
18	30 min	"	"	20	"			finished	24	7	0	0	
19*	10 min	"	"	10	"			finished	27	17	0	0	
NOE + CS + RDC													
20	45 min	3	1	30	3.5ppm	30	3.5 Hz	not finished	5	24	0	0	
21	50 min	2	"	"	"	"	"	not finished	6	26	0	0	
22	11 h	"	"	"	"	"	"	not finished	6	26	0	0	
23	10 min	1	"	"	"	"	"	finished	38	46	0	0	
24	1 min	"	1.2	"	"	"	"	finished	41	46	0	0	
25	5 sec	"	1.5	"	"	"	"	hole	0	0	4	0	
Optimize CS													
26	1 min	1	1.2	30	3ppm	30	3.5 Hz	finished	42	45			
27	1 min	"	"	20	"	"	"	finished	41	46	0	0	
28	1 min	"	"	10	"	"	"	finished	47	41	0	0	
Optimize RDC													
29	1 min	1	1.2	10	3ppm	30	3 Hz	finished	47	41			
30	1 min	"	"	"	"	20	"	finished	47	44	0	0	
31*	30 sec	"	"	"	"	10	"	finished	58	41	0	0	

Three separated parameter optimizations are shown in this table. The first optimization assumes that only NOE data are available. The second one assumes that NOE and CS data are available and the last one assumes that NOE, CS and RDC data are available. The theoretical distance thresholds for the three NOE classes are here $d_{max}^{theo} = (5\text{Å}, 6\text{Å}, 7\text{Å})$, yielding $N_{dist} = (292, 98, 138)$ distances in each class (short, medium, long) using the X-ray structure 193L. The number of experimental NOEs is here $N_{NOEs} = (52, 55, 62)$ for strong, medium and weak NOEs, respectively. The number of HSQC peaks is here $N_{peaks} = 132$. See table S1 for the variable definitions.

* Optimized parameter combinations that have been retained for further analysis in the results section of the article. Trials #6, #19 and #31 correspond to cases 1, 2 and 3 of table 2, respectively.

ment possibilities still has a *spatial assignment range* (SAR, see methods in the article) value below 10Å so that 94% of the peaks can be used for applications that do not require a unique assignment for all peaks. The remaining 6% with SAR values above 10Å correspond to peaks without NOE constraints. Applying the decreasing completeness function $c(d)$ (Figure S1) for NOE_3^{SIM} and NOE_4^{SIM} makes G^{exp} even sparser in comparison to G^{theo}. This multiplies the matching possibilities of G^{exp} onto G^{theo} and so also the peak assignment possibilities. In case 4, only 40% of the peaks are uniquely assigned. On the other hand, the multiple assignments possibilities remain spatially restrained with 93% of all peaks having a SAR value below 10Å in case 4. Even with a high number of multiple peak assignments, the accuracy of the assignment ensemble is equal to 100% in all four cases.

Simulated NOE data on lysozyme using realistic conditions.

The first four test cases shown in the previous paragraph make the unrealistic assumption that G^{exp} and G^{theo} are generated from the same 3D structure. We simulated a more realistic case, where G^{exp} and G^{theo} are generated from different 3D structures (see Datasets section). The structure discrepancies require the use of higher theoretical distance thresholds, yielding a much higher number of edges in G^{theo} than G^{exp} (614 vs 280). Otherwise G^{exp} cannot be matched correctly onto G^{theo} (occurrence of holes in the assignment list). NOE intensities were classified and NOE outliers introduced through labeling of the graph edges (Stratmann et al, 2009). Only a few NOEs correspond to distances near to the maximum allowed distance d_{max}^{theo} and can thus be considered as *outliers* of the distance distribution. To reflect this, the

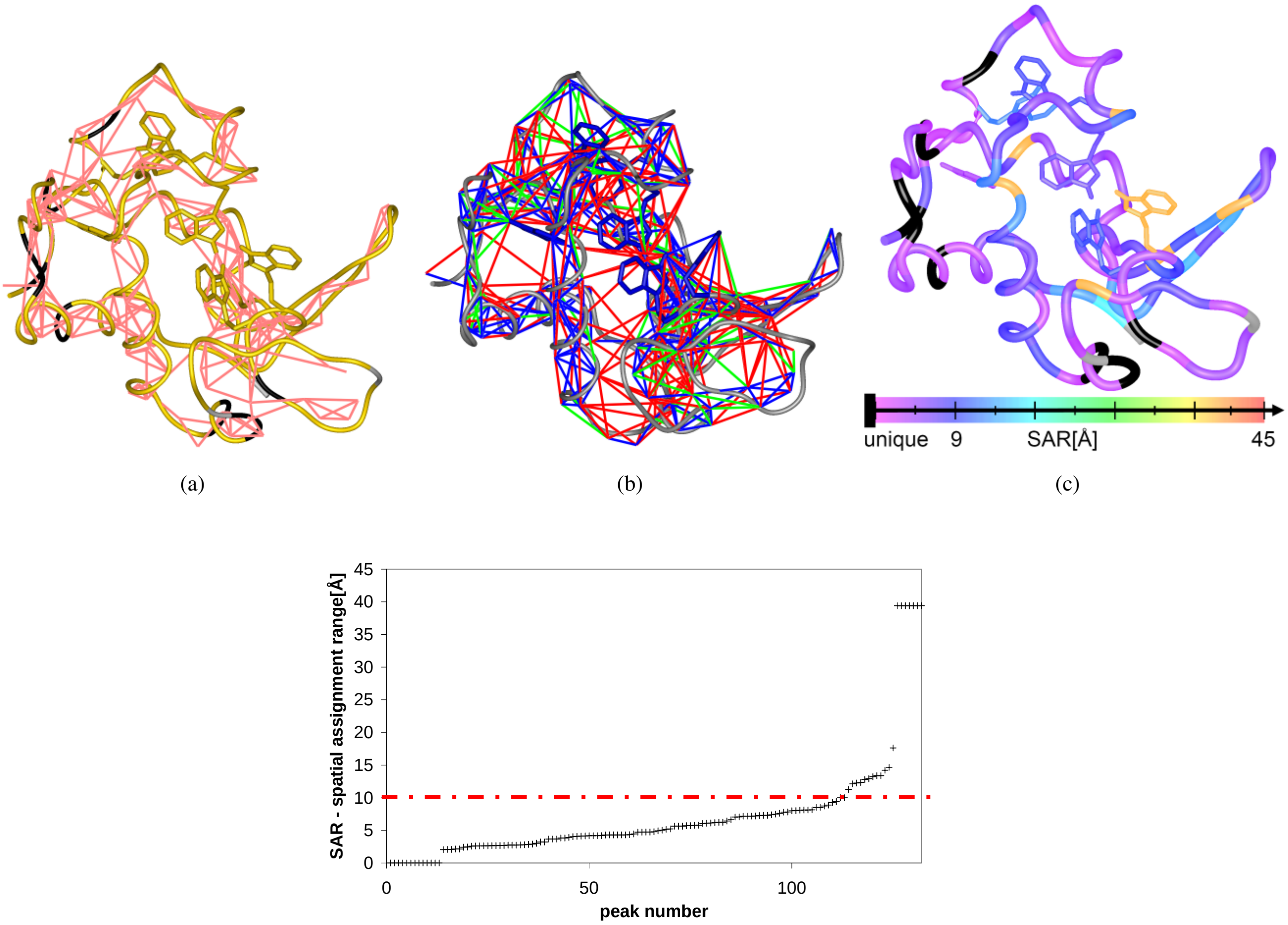
Table S 3 Tested threshold parameters on EIN with experimental NOE and CS data and with simulated RDC data.

		NOE		CS		RDC		Result					
#	runtime	T_{NOE}	Δd[Å]	c_{CS}	m_{CS}	c_{RDC}	m_{RDC}	Status	N_u	N_{10}	N_e	N_{eu}	SAR_{max}
NOE + CS													
1	13.5 h	5	1	30	3.5ppm			not finished	16	61	0	0	
2	11 h	4	"	"	"			not finished	19	59	0	0	
3	9 h	3	"	"	"			not finished	32	49	0	0	
4	107 h	"	"	"	"			not finished	46	37	0	0	
5*	6 h	2	"	"	"			not finished	73	56	one swap 207 <-> 208	2	
6*	6 days	"	"	"	"			finished	76	94	one swap 207 <-> 208	2	
7	15 min	1	"	"	"			hole	35	45	10	4	5.8Å
Optimize CS													
8	25 min	2	1	30	3ppm			hole	25	51	6	2	5.8Å
9	15 min	"	"	20	3.5ppm			hole	29	48	5	3	5.3Å
10	15 min	"	"	20	3ppm			hole	14	24	2	0	
11	10 min	"	"	10	3ppm			hole	12	21	4	1	1.6Å
12	3.5 days	3	1	30	3ppm			finished	73	10	2	2	4.3Å
13	15 min	"	"	20	3ppm			hole	24	32	0	0	
14	10 min	"	"	10	3ppm			hole	20	29	2	1	1.8Å
NOE + CS + RDC													
15	25 min	2	1	30	3.5ppm	30	3.5 Hz	hole	36	44	2	1	4.4Å
16	7 h	"	"	"	"	"	6.5 Hz	hole	80	90	0	0	
17	7.5 h	"	"	"	"	"	7.5 Hz	hole	77	93	0	0	
18	2.5 h	3	1	"	"	"	3.5 Hz	finished	78	94	0	0	
19	2 h	"	"	"	"	"	3 Hz	finished	76	97	0	0	
20	70 min	"	"	"	"	20	3 Hz	finished	81	96	0	0	
21	55 min	"	"	"	"	10	3 Hz	finished	90	106	0	0	
22	50 min	"	"	"	"	10	2 Hz	finished	97	100	0	0	
23	50 min	"	"	"	"	10	1 Hz	finished	121	77	0	0	
24	50 min	"	"	"	"	8	"	finished	136	65	0	0	
25	50 min	"	"	"	"	6	"	finished	142	63	0	0	
26*	50 min	"	"	"	"	5	"	finished	153	52	0	0	
27	45 min	"	"	"	"	4	"	finished	165	41	1	1	2.8Å
28	40 min	"	"	"	"	3	"	hole	175	31	1	1	2.8Å
NOE + CS + CScarbon($^{13}C_\alpha +^{13} C_\beta +^{13} CO$)													
					m_{CS} for ^{13}C-CS only								
29	10 min	3	1	30	2ppm			finished	91	93	0	0	
30*	5 min	"	"	"	1.5ppm			finished	103	91	0	0	
31*	1 sec	"	"	"	"			refined	160	41	1	0	
32	1 min	"	"	"	1ppm			hole	45	32	0	0	
NOE + CS + CScarbon($^{13}C_\alpha +^{13} C_\beta +^{13} CO$) + RDC													
					m_{CS} for ^{13}C-CS only								
33*	5 min	3	1	30	1.5ppm	5	1 Hz	finished	162	43	0	0	
34*	1 sec	"	"	"	"	"	"	refined	194	13	2	2	2.8Å

The parameter optimizations using NOE and CS data for EIN are shown in this table. The theoretical distance thresholds for the three NOE classes are here $d_{max}^{theo} = (4.5Å, 6Å, 7.5Å)$, yielding $N_{dist} = (391, 320, 323)$ distances in each class (short, medium, long) using the X-ray structure 1ZYM. The number of experimental NOEs is here $N_{NOEs} = (36, 208, 163)$ for strong, medium and weak NOEs, respectively. The number of HSQC peaks is here $N_{peaks} = 243$. See table S1 for the variable definitions.
* Optimized parameter combinations that have been retained for further analysis in the results section of the article. Trials #4, #5, #6, #26, #30, #31, #33 and #34 correspond to cases 2 to 9 of table 3, respectively.

number of NOEs which can be matched to distances labeled as outlier-distance is limited by the threshold T_{NOE}.

This organization of experimental NOEs reduced significantly the matching possibilities of G^{exp} onto G^{theo}, yielding 86% of all peaks with a SAR value below 10Å (case 5, Table S5 and Figure S2). Addition of $^1H^N$ and ^{15}N chemical shifts (CS) did not bring a significant improvement of the assignment ensemble (case 6 in Table S5, trial #15 of the optimization table S1). Compared to this, adding residual dipolar couplings (RDC) greatly improved the number of uniquely assigned peaks. However, just adding RDC data using the parameters previously optimized for NOE and CS data yielded undetected errors in the assignment (cases 8 / trial #20). This can be due to either a too low number of alowed NOE outliers or too tight thresholds for RDC data. Both cases can generate assignment constraints that are not compatible anymore with the correct assignment, but that can still be compatible with some uncorrect assignments and thus do not generate holes in the assignment list. Here, increasing the number of allowed NOE outliers to T_{NOE} = 13 without changing the thresholds for the RDC data yielded again a 100% accurate assignment ensemble (case 7 / trial #22). Using the actual number of outliers (T_{NOE} = 14) does not change notably the result. The huge increase in the num-



(a) (b) (c)

(d)

Fig. S 2 Result of NOE*net* on lysozyme with realistic simulated NOE data. Only case 5 of table S5 is shown, i.e. no CS or RDC data are included here. (a) The 280 simulated NOEs are shown by red lines on the lysozyme NMR structure 1E8L (Schwalbe et al, 2001). The residues having only one assignment possibility are shown in black and more than one in yellow. Proline residues are shown in gray. (b) The 614 theoretical contacts are represented on the X-ray structure 193L (Vaney et al, 1996) by blue, green and red lines corresponding to the three distance classes, short ($d < 5$Å), medium ($d < 6$Å) and long ($d < 7.5$Å), respectively. (c) The spatial assignment range (SAR) values are mapped on the NMR structure using the correct assignment and the indicated color code. Unique assignments are shown in black. (d) Spatial assignment range (SAR) for each peak. The peaks are ordered by increasing SAR values. A SAR-value of 10Å has been considered as the upper limit for the class of *exploitable* peaks.

ber of uniquely assigned peaks when just adding RDC data without changing T_{NOE} (from 18 to 115) appears to be a good indication of unreliable results. Borderline cases as trial 21 are more difficult to identify ($T_{NOE} = 12$, 50 uniquely assigned peaks, 4 correct assignments missed). However, in these cases, assignment errors appear again limited both in number and in space (small SAR_{max} values).

Influence of the completeness of NOEs on the assignment results

Taking the same sparse experimental NOE data of lysozyme as in described in the article, the number of unambiguous NOEs was increased by assuming higher resolution spectra with [tolN, tolH] equal to [0.1ppm, 0.01ppm] instead of [0.2ppm, 0.02ppm]. The resulting data set contains 183 unambiguous NOEs (NOE1 data set in figure S3) to be compared to 169 for the larger tolerances (data set NOE2 in figure S3). The gain brought by the 14 additional NOEs is notable mostly when all experimental data (NOE+CS+RDC) are used (Figure S3).

This indicates that the more constraints are available, the more the assignment ensemble is precise, especially if a critical quantity of assignment constraints is already achieved, so that a small number of additional constraints increases the number of unique assignment significantly. The combination of three orthogonal constraint sources (NOE, CS and

Table S 4 Simulated NOE data on lysozyme using ideal conditions.

case	N_{peaks}	runtime	data	ambiguous removed	d^{exp}_{max}	N_{NOEs}	d^{theo}_{max}	N_{dist}	$\frac{N_{unique}}{N_{peaks}}$	$\frac{N_{SAR<10A}}{N_{peaks}}$	accuracy
1	132	1 min	NOE^{SIM}_1	no	5Å	283	5Å	283	95%	100%	100%
2	132	1 min	NOE^{SIM}_2	yes	5Å	253	5Å	283	79%	94%	100%
3	132	16.5h	NOE^{SIM}_3	no	6Å	297	6Å	385	55%	100%	100%
4	132	21h	NOE^{SIM}_4	yes	6Å	269	6Å	385	40%	93%	100%

The same structure (PDB 193L, X-ray) has been used to generate the N_{NOEs} NOEs and the N_{dist} distances.

Table S 5 Simulated NOE data on lysozyme using realistic conditions.

case	trial # table S1	N_{peaks}	runtime	data	d^{exp}_{max}	N_{NOEs}	d^{theo}_{max}	N_{dist}	$\frac{N_{unique}}{N_{peaks}}$	$\frac{N_{SAR<10A}}{N_{peaks}}$	accuracy
5	11	132	31h	NOE^{SIM}_5	6Å	280	7.5Å	614	10%	86%	100%
6	15	132	19h	NOE^{SIM}_5+CS	6Å	280	7.5Å	614	14%	86%	100%
7	22	132	16h	NOE^{SIM}_5+CS+RDC	6Å	280	7.5Å	614	32%	91%	100%
8	20	132	11h	NOE^{SIM}_5+CS+RDC	6Å	280	7.5Å	614	87%	95%	94%

The NMR structure 1E8L has been used to generate the N_{NOEs} NOEs and the X-ray structure 193L for the N_{dist} distances. For all three cases the ambiguous NOEs have been removed as described in the datasets section above. NOE classes and NOE outliers have been used.

RDC) allows here to achieve such a critical quantity of assignment constraints.

Comparing the second column 'NOE2+CS+RDC' with the third column 'NOE1+CS' in figure S3, shows that RDC data can compensate for a lower completeness of NOEs.

Influence of the use of intensity-based classification of NOEs on the assignment result

The results shown in figure 3 of the article and figure S3 here were obtained considering the classification of the d_{NN} into weak, medium and strong NOEs proposed by (Schwalbe et al, 2001). This permitted the application of three different theoretical threshold distances, one for each class (Stratmann et al, 2009). The effect of the repartition of the NOEs in various NOE-classes is illustrated in Figure S4 for the previously described dataset NOE1.

Comparing columns one to four or five to eight, it appears that NOEs classification clearly improves the assignment precision. The identification of strong NOEs that can be assigned to short distances seems to be particularly relevant to obtain a high level of uniquely assigned peaks.

References

Doreleijers JF, Raves ML, Rullmann T, Kaptein R (1999) Completeness of NOEs in protein structure: a statistical analysis of NMR. J Biomol NMR 14:123–132

Koharudin LM, Bonvin AM, Kaptein R, Boelens R (2003) Use of very long-distance NOEs in a fully deuterated protein: an approach for rapid protein fold determination. J Magn Reson 163:228–235

Schwalbe H, Grimshaw SB, Spencer A, Buck M, Boyd J, Dobson CM, Redfield C, Smith LJ (2001) A refined solution structure of hen lysozyme determined using residual dipolar coupling data. Protein Sci 10:677–688

Stratmann D, van Heijenoort C, Guittet E (2009) NOEnet – Use of NOE networks for NMR resonance assignment of proteins with known 3D structure. Bioinformatics 25(4):474–481

Vaney MC, Maignan S, Riès-Kautt M, Ducruix A (1996) High-resolution structure (1.33 A) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. Acta Crystallogr D Biol Crystallogr 52:505–517

Wang Y, Bjorndahl TC, Wishart DS (2000) Complete 1H and non-carbonylic 13C assignments of native hen eggwhite lysozyme. J Biomol NMR 17:83–84

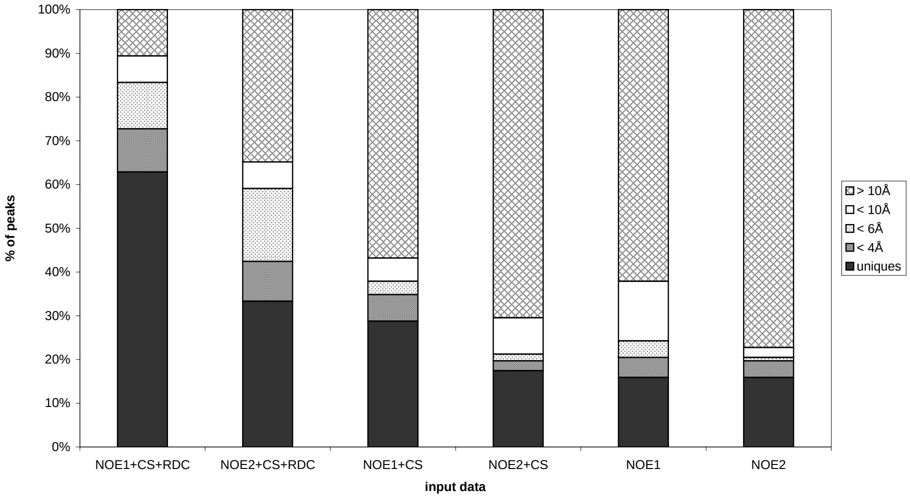


Fig. S 3 Assignment results on lysozyme using experimental data obtained by (Schwalbe et al, 2001) presented by spatial assignment range (SAR) classes, see legend on the right. NOE1 and NOE2 are defined in the text. NOE2 is the same NOE data set as in figure 3 of the article.

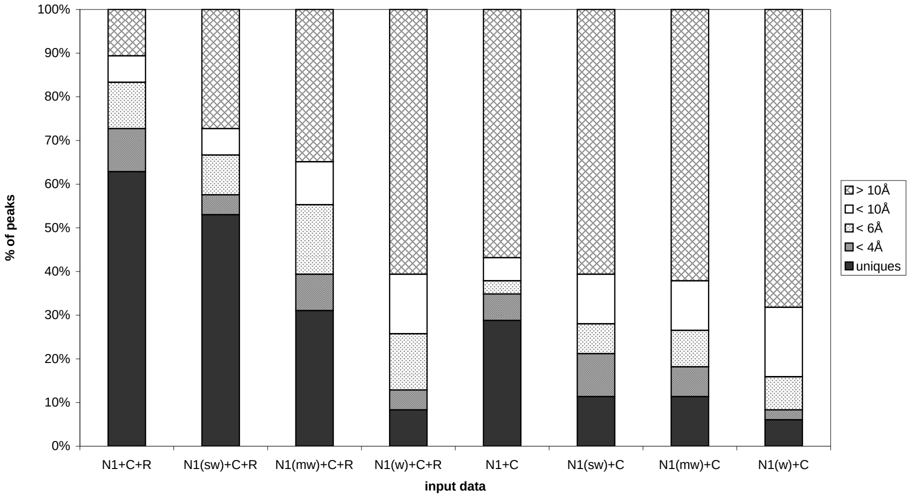


Fig. S 4 Assignment results on lysozyme presented by spatial assignment range (SAR). The input data are the same in Figure S3, only the number of NOE-classes varies. The abbreviations for the input data are: N1 = NOE1, C = CS, R = RDC, w = weak NOEs, m = medium NOEs and s = strong NOEs.The following combinations are reported, from left to right: 1) for comparison purposes: The three classes (strong, medium and weak) like in Figure S3. 2) All 64 medium NOEs are converted into weak NOEs (two classes, N1(sw)). 3) All 52 strong NOEs are converted into medium NOEs (two classes, N1(mw)). 4) All strong and medium NOEs are converted into weak NOEs, i.e. the theoretical distance threshold is the same for all 183 NOEs (one class, N1(w). 5-8) the same order of combinations without RDC-data.