Supporting Information

High-Resolution Solid-State NMR Structure of a 17.6 kDa Protein

Ivano Bertini[1,2], Anusarka Bhaumik[1], Gaël De Paëpe[3‖], Robert G. Griffin[3], Moreno Lelli[1], Józef R. Lewandowski[3§], Claudio Luchinat[1,4]*

[1]Magnetic Resonance Center, CERM, University of Florence, Via L. Sacconi, 6 - 50019 Sesto Fiorentino, Italy; [2]Department of Chemistry, University of Florence, Via della Lastruccia, 3 - 50019 Sesto Fiorentino, Italy; [3]Department of Chemistry and Francis Bitter Magnet Laboratory, Massachusetts Institute of Technology - Cambridge, Massachusetts 02139, USA; [4]Department of Agricultural Biotechnology, University of Florence, Via Maragliano, 75-77 - 50144 Florence, Italy

‖ Current address: Laboratoire de Chimie Inorganique et Biologique, UMR-E3 (CEA/UJF), CEA/DSM/INAC - 38054 Grenoble, France.
§ Current address: Université de Lyon, CNRS/ ENS Lyon/ UCB-Lyon 1, Centre RMN à Très Hauts Champs, 5 rue de la Doua, 69100 Villeurbanne, France.

Analysis of the spectra

The PDSD spectra were acquired with different mixing times ranging from 15 to 400 ms, but the spectrum with the largest number of resolved cross-peaks was the one with 200 ms of mixing time. Longer mixing times rapidly produce crowded spectra, where the significant overlap avoids a reliable cross-peaks assignment, giving only a limited contribution in terms of distance restraints. The 200 ms PDSD spectrum (Figure S1) yields 909 cross-peaks[i] in the aliphatic region from semi-automated peak-picking (see more details about peak-picking in the following section), among which 80 cross-peaks can be unambiguously assigned to inter-residue correlation, while up to additional 181 structurally unambiguous cross-peaks were assigned on the basis of the refined structure by using the iterative procedure described in the "structure calculations" section in the main text.

The DARR spectra generally show more intense spectra with minor differences with respect to PDSD. The spectrum with the largest number of resolved cross-peaks was the spectrum with 100 ms mixing time, which yields 824 cross-peaks in the aliphatic region from semi-automated peak-picking, providing 86 unambiguously assigned inter-residue correlations, and additional 138 structurally unambiguous correlations. A large portion of these assigned cross-peaks is common to the PDSD spectra; overall, DARR and PDSD spectra (with different mixing times) provided 95 unambiguous and 217 structurally unambiguous cross-peaks.

Further ^{1}H-^{1}H distance restraints can be indirectly determined by acquiring the signal of the proton-attached heteronuclei (CHHC and NHHC experiments). We limited our investigation to the more resolved CHHC spectra, the most structurally informative being those with 150 μs and 300 μs mixing time (Figure S2). These two CHHC spectra contain 574 and 692 aliphatic cross-peaks (for 150 μs and 300 μs, respectively), largely in

[i] Only cross-peaks showing diagonal-symmetric correlations were accounted for this value.

common, and most of them are concentrated in the region between 20 and 70 ppm. In these spectra, up to 64 unambiguous and 157 structurally unambiguous cross-peaks were assigned.

The largest number of restraints was obtained from ^{15}N-^{13}C PAIN-CP and aliphatic ^{13}C-^{13}C PAR spectra, providing overall 458 distance restraints. The PAIN-CP experiment was acquired with a mixing time of 15 ms and showed up to 563 N-C cross-peaks (Figure S3), with 55 unambiguous and 43 structurally unambiguous cross-peaks. The PAR spectra were acquired with mixing times of 15 ms at 21.1 T (900 MHz of ^{1}H Larmor frequency), 15 ms and 20 ms at 20.0 T (850 MHz of ^{1}H Larmor frequency); these three spectra showed 667, 747 and 1016 cross-peaks, respectively (all in the aliphatic region), and a very large part of the cross-peaks observed in the 15 ms spectra were also observed in the 20 ms spectrum (Figure S4). Up to 92 cross-peaks in total were unambiguously assigned, and additional 268 cross-peaks were structurally unambiguously assigned.

Table 1 summarizes all the assigned connectivities: excluding the common cross-peaks, overall 240 unambiguous and 537 structurally unambiguous cross-peaks were assigned. For each of the analyzed spectra the upper-distance limits used in the structural calculation are reported. In analogy with the NOESY spectra routinely used in solution NMR for protein structural determination, the detection of long-range contacts is in some way related to the signal to noise ratio (S/N) in the spectrum: a higher S/N allows one to observe weak, long-distance contacts. Of course, S/N is not the only factor to be taken into account to establish the upper-distance limits of the weakest cross-peaks since, as described in the main text, other effects, such as relayed magnetization transfer, were taken into account in choosing the used values. In particular, for PDSD and DARR experiments, the relayed transfer mechanism is so efficient that we often observed intense inter-residue cross-peaks between nuclei at 8 Å from one another; thus, even using spectra with less S/N of those used here, it is preferable not reduce the present upper distance limit of 9 Å. On the other side, for PAR spectra (as well as PAIN-CP or CHHC)

we did observe that increasing the S/N cross-peaks of nuclei being farther and farther away from one another were detected. Nevertheless, relayed magnetization transfer effects, even if much less effective than PDSD, are still present in these spectra. Due to that and other factors such as involvement of multiple protons in polarization transfer between two ^{13}C nuclei and complex geometry dependence it was not possible to accurately categorize all the inter-nuclear restraints into different upper-distance limits on the basis of the cross-peak intensity (or volume).

The PAR experiment is the experiment that provides the largest number of unambiguous cross peaks, especially if compared to the CHHC spectra. The PAR experiment has generally a higher S/N than CHHC: Figures S2 and S4 shows some 1D traces of the CHHC and PAR experiments.[ii] Comparing the intensities of the intra-residue correlations with the inter-residue correlation of the same residue, in the PAR spectra it possible to see inter-residue cross-peaks as weak as 20 % of the intra-molecular correlations, while in the CHHC experiment the weaker correlations have around 30 % of the intensity of the intra-molecular ones.

This is not the only reason why PAR is richer in cross-peaks than CHHC; another important point is that a large part of the PAR cross-peaks involve methyl ^{13}C nuclei, which have a narrower linewidth with respect to the other ^{13}C cross-peaks,[iii] and thus a higher number of unambiguous cross-peaks can be assigned.

Of the 92 unambiguously assigned cross-peaks in the PAR spectra, 40 are common to PDSD/DARR spectra, while only 12 are common to CHHC. Overall, the PAR/PAIN-CP

[ii] In order to improve the signal-to-noise ratio, the CHHC spectrum reported in figure S2 has been processed with a different windows function with respect to the PAR spectrum in figure S4.

[iii] The ^{13}C-^{13}C J-couplings give a significant contribution to the observed linewidth in the ^{13}C resonances in this sample. The methyl ^{13}C nuclei are coupled to only one carbon nucleus, with a J coupling ≈ 35 Hz, thus they show narrower cross-peaks than other nuclei such as Cα. Indeed, Cα nuclei are coupled to two ^{13}C nuclei with J coupling of 55 Hz (Cα-C') and 35 Hz (Cα-Cβ).

experiments provide 135 additional restraints with respect to those that could be already assigned through PDSD/DARR/CHHC spectra. Since in the CHHC spectra the methyl contacts are weak and hardly observed, PAR provides complementary structural information to CHHC (see also Figure S5 and the discussion section of the main text). From PAR, PAIN-CP and CHHC up to 199 unambiguous restraints can be collected, and PDSD/DARR only provide 41 additional unambiguous contacts.

Table 2 summarizes the composition of all the 777 unambiguous restraints in terms of sequential, medium-range and long-range contacts. While the large majority of these restraints are sequential contacts, the PAR and CHHC provide the largest number of long-range interactions.

Peak-picking procedure

The number of overall cross-peaks peak-picked in the above reported spectra was determined essentially using an automated peak-picking procedure (SPARKY software), while all the assigned unambiguous and structurally unambiguous cross-peaks were uniquely manually assigned. An accurate automated peak-picking procedure in SSNMR spectra is not a trivial task, since the sizable peak-overlap prevents the determination of the number and the exact frequencies of all the observable resonances. Even if the use of an accurate automated peak-picking routine is a necessary step for developing automated structural calculation protocols for SSNMR, this is far from the purpose of this work and has not been attempted here.

To avoid the significant number of errors due to the purely automated routine procedures, the automated peak-picked lists of each spectrum were manually corrected by removing cross-peaks that were too close to the noise level or that did not have a corresponding cross-peak across the diagonal. Figure S6 reports the cross-peaks peaked on the PAR (with 20 ms mixing) and the PAIN-CP (15 ms mixing) spectra, highlighting some regions

with different threshold levels. It is apparent how it is possible to assign a large number of cross-peaks even in the spectral regions with partial cross-peak overlap.

The possible errors, still remaining after using this semi-automated peak-picking approach as well as the errors due to peak overlap, could be some of the reasons for the difficulties in obtaining an accurate protein structure with the ARIA software.

For these reasons, unambiguous restraints as well as structurally unambiguous restraints were uniquely manually assigned, manually checking the reliability of each cross-peak and the presence of the corresponding peak across the diagonal. Whenever possible, the reliability of the assigned cross-peak was checked also on spectra acquired with different mixing times.

ARIA Calculations

For proteins smaller than 100 AA, solid-state NMR high-resolution structures can be obtained by using software designed for the automated assignment of the ambiguous restraints.[1,2,3,4,5] Starting from a limited number of unambiguous restraints, these software tools are able to manage hundreds or even thousands of ambiguous restraints solving, where possible, the structural ambiguity or handling together ambiguous and unambiguous restraints in the simulated annealing protocol.

For such a reason, we investigated the feasibility of a structural calculation starting from the 240 unambiguously assigned cross-peaks and using the software ARIA 2.2.2 [iv,6] to increase the number of restraints instead of using PCS.

The ARIA routine can, in principle, extract the structural information implicitly included in the PDSD/DARR, CHHC, and PAR once the number of unambiguously assigned

[iv] ARIA 2.2.2 is a not-yet released β-version of ARIA, especially designed for SSNMR. We thank M. Nilges and B. Bardiaux for allowing us to test this software in the present case.

restraints is large enough to allow the software to fix an enough large extent of structurally unambiguous restraints.

Thus, the ARIA calculation protocol was started by using two data sets of restraints: the first including the 240 unambiguous restraints directly obtained from the protein sequential assignment, and the second set obtained including all the remaining 2098 cross-peaks picked on the spectra and not yet assigned.

After 8 iterations in simulated annealing minimization steps (as detailed in the Materials and Methods session), and after the water-refinement step, a family of structures with 7.3 ± 1.0 Å RMSD to the mean is produced, with an accuracy to the X-ray structure of 8.6 Å, with 705 structurally unambiguous and 305 ambiguous restraints used. The precision of this structure family is still quite poor; indicating the initial number of unambiguous restraints per residue is still insufficient to allow ARIA to converge to a high-resolution structure.

On the other hand, these calculations highlight the potential of the ARIA routine in handling structures even larger than 100 AA, making it possible to reduce the RMSD within the family of about 2 Å with respect to calculations performed directly using the unambiguous diamagnetic restraints.

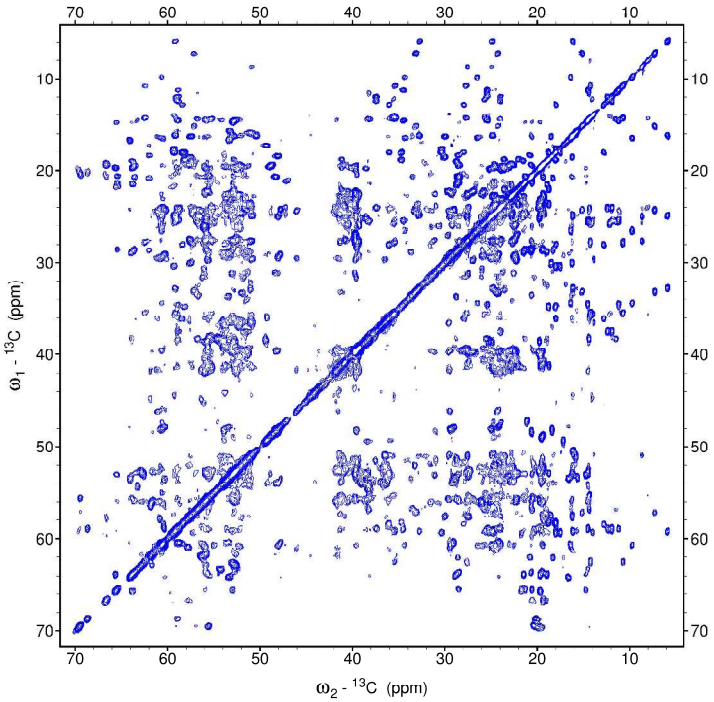

Figure S1. Aliphatic region of the DARR spectrum acquired on the Zn(II)-MMP-12 protein in the microcrystalline state (16.4 T, $\omega_R/2\pi$ = 11.5 kHz, mixing time 100 ms).

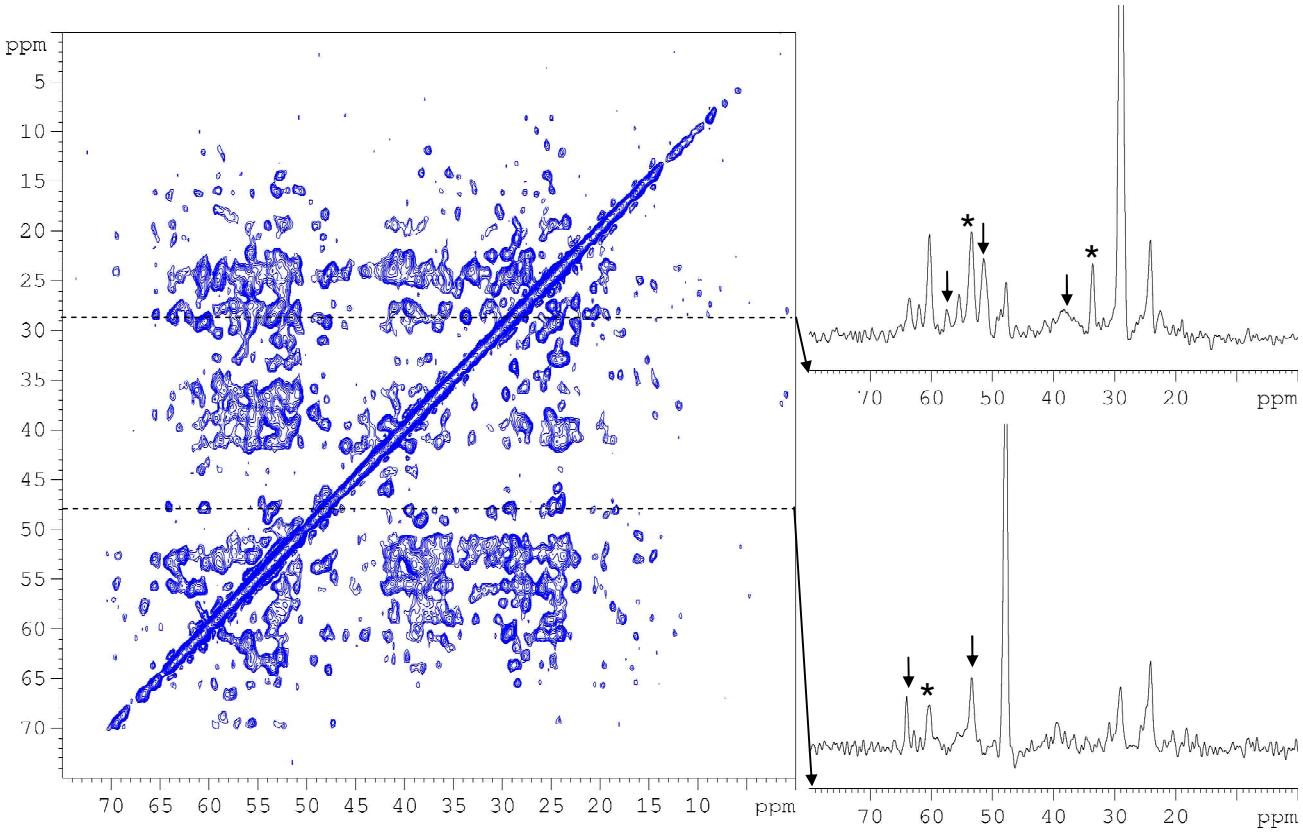

Figure S2. ^{13}C-^{13}C CHHC spectrum acquired on the Zn(II)-MMP-12 protein in the microcrystalline state (16.4 T, $\omega_R/2\pi$ = 11.5 kHz, mixing time 300 μs). Two one-dimensional slices are shown. Asterisks indicate assigned *intra*-residue peaks, and arrows indicate assigned *inter*-residue peaks.

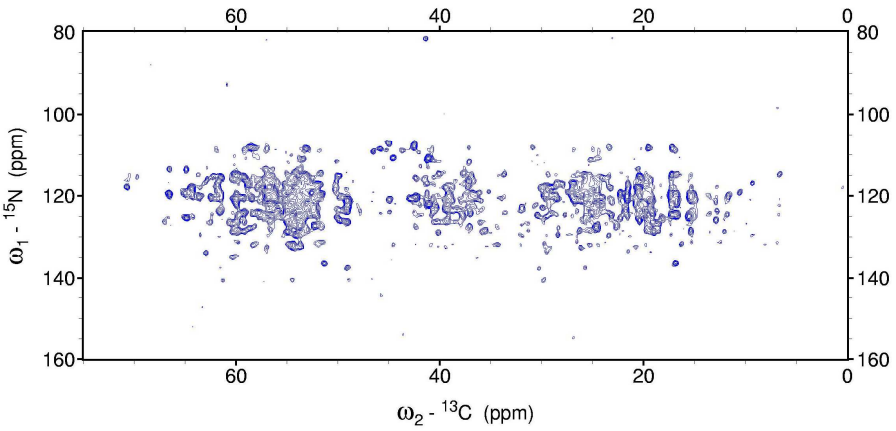

Figure S3. Aliphatic region of the ^{15}N-^{13}C PAIN-CP spectrum acquired on the Zn(II)-MMP-12 protein in the microcrystalline state (21.1 T, $\omega_R/2\pi$ = 20.0 kHz, mixing time 15 ms).

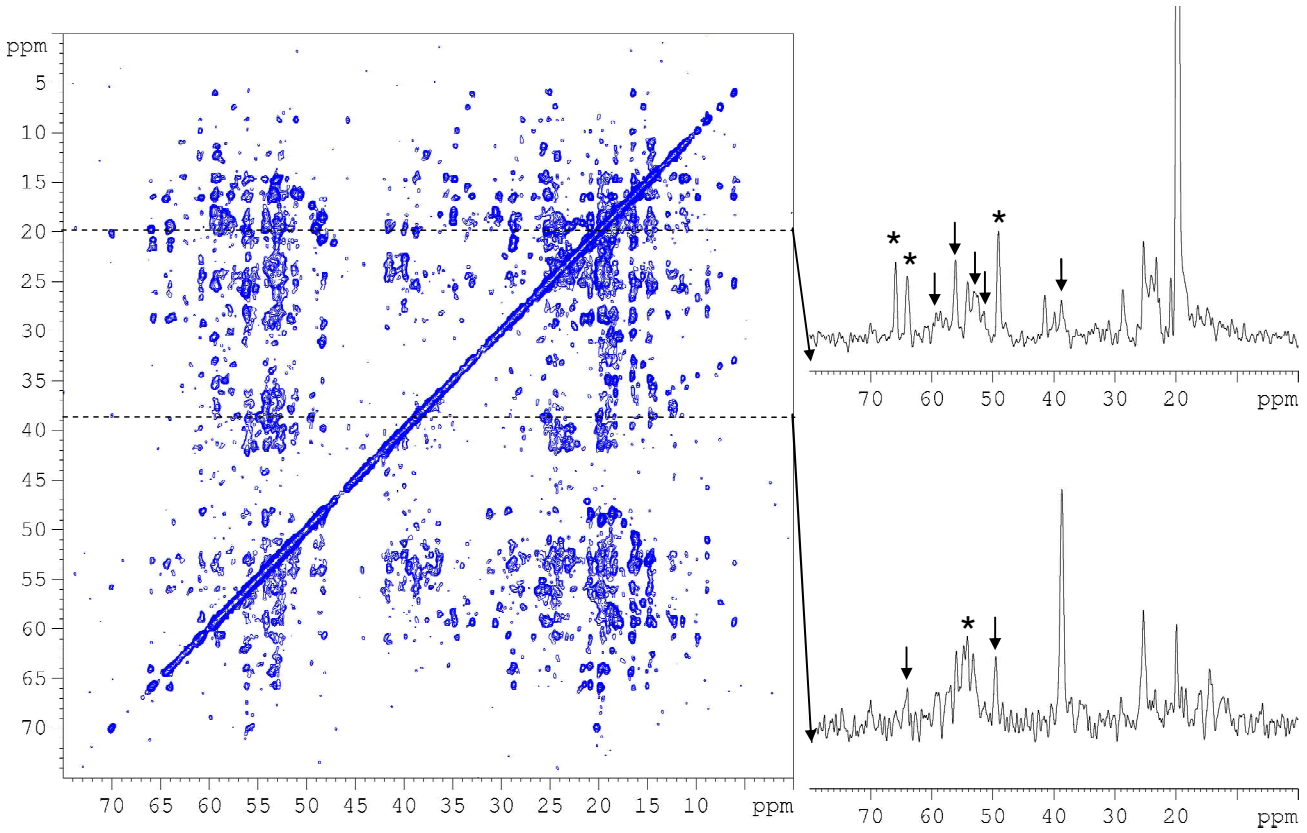

Figure S4. Aliphatic ^{13}C-^{13}C PAR spectrum acquired on the Zn(II)-MMP-12 protein in the microcrystalline state (20.0 T, $\omega_R/2\pi$ = 19.0 kHz, mixing time 20 ms). Two one-dimensional slices are shown. Asterisks indicate assigned intra-residue peaks, and arrows indicate assigned *inter*-residue peaks.

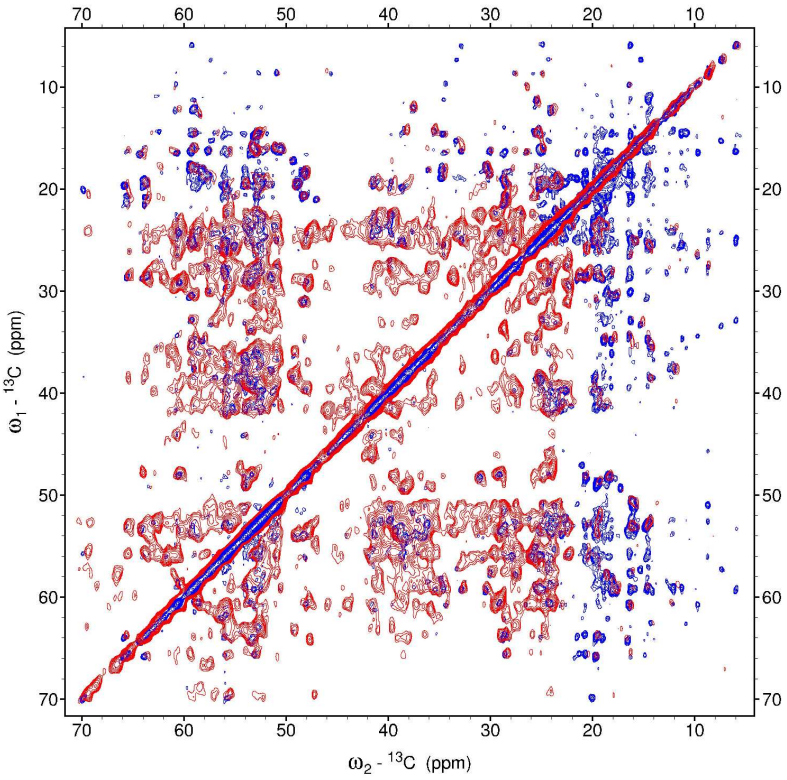

Figure S5. Superposition of the aliphatic regions of CHHC (Red, 16.4 T, $\omega_R/2\pi$ = 11.5 kHz, mixing time 300 μs) and PAR (Blue, 20.0 T, $\omega_R/2\pi$ = 19.0 kHz, mixing time 20 ms) acquired on the Zn(II)-MMP-12 protein in the microcrystalline state. It is apparent the larger number of correlation in the methyl region for PAR, while CHHC shows more correlations in the region from 20 to 70 ppm.

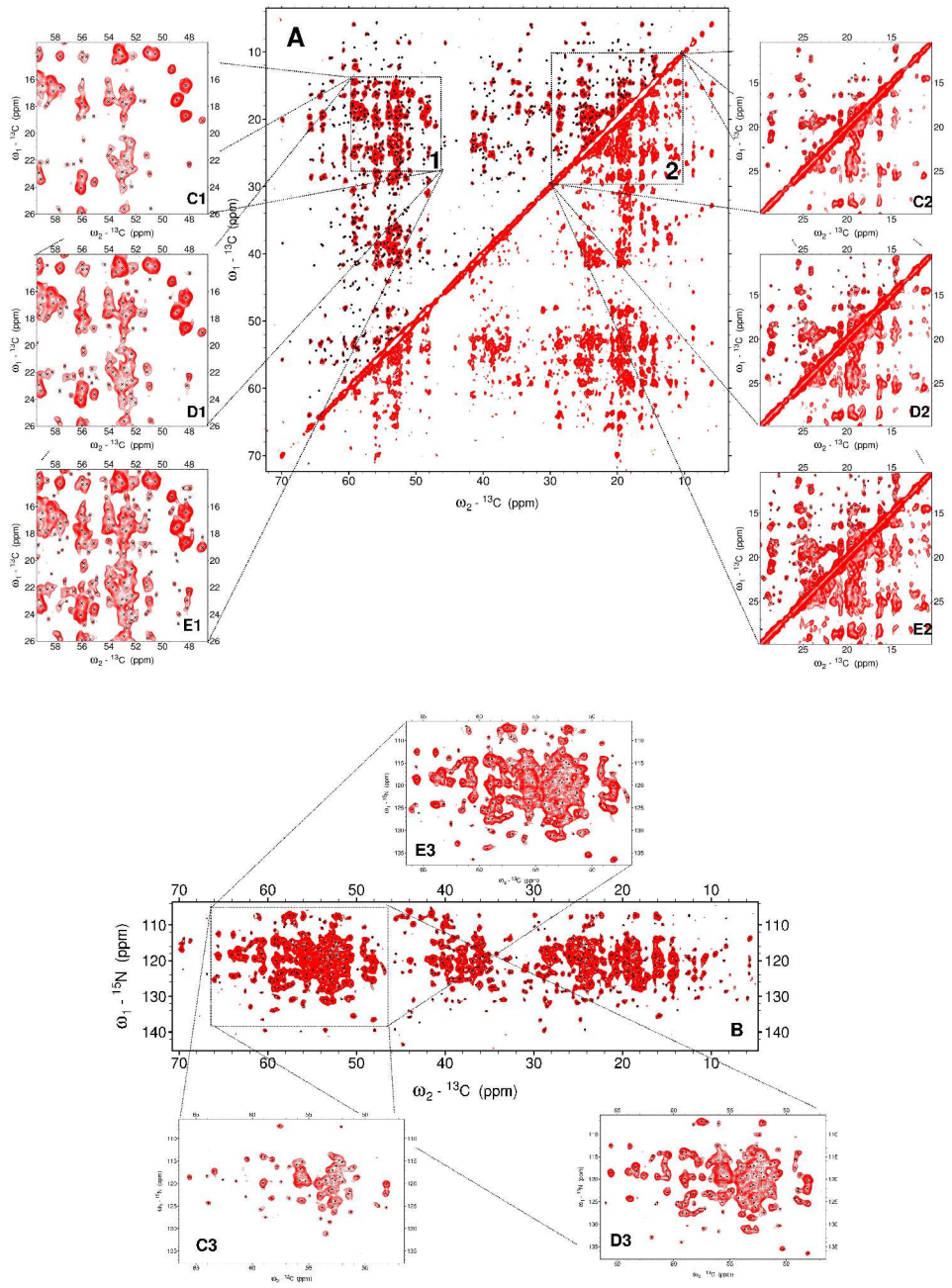

Figure S6. Aliphatic ^{13}C-^{13}C PAR spectrum (A) and aliphatic region of the ^{15}N-^{13}C PAIN-CP spectrum (B) acquired on the Zn(II)-MMP-12 protein in the microcrystalline state. The figure reproduces the same spectra shown in Figures S3 and S4 but with different threshold levels to illustrate the number of observed cross-peaks. The black spots are the peak-picked peaks. To evidence the number of peaks that can be assigned in crowded regions, the regions 1, 2, and 3 are reproduced with different threshold levels in the expansion C, D, and E.

References

[1] Manolikas, T.; Herrmann, T.; Meier, B.H. *J. Am. Chem. Soc.* **2008,** *130,* 3959–3966.

[2] Loquet, A.; Bardiaux, B.; Gardiennet, C.; Blanchet, C.; Baldus, M.; Nilges, M.; Malliavin, T.; Boeckmann, A. *J. Am. Chem. Soc.* **2008,** *130*, 3579–3589.

[3] Korukottu, J.; Schneider, R.; Vijayan, V.; Lange, A.; Pongs, O.; Becker, S.; Baldus, M.; Zweckstetter M. *PLoS ONE* **2008,** *3,* e2359.

[4] Nilges, M.; O'Donoghue, S.I. *Prog. NMR Spect.* **1998,** *32,* 107–139.

[5] Linge, J.P.; O'Donoghue, S.I.; Nilges, M. *Methods in Enzymology* **2001**, *339,* 71–90.

[6] Rieping, W.; Habeck, M.; Bardiaux, B.; Bernard, A.; Malliavin, T.E.; Nilges M. *Bioinformatics* **2007,** *23,* 381–382.