Influence of the completeness of chemical shift assignments on *de novo* protein structure generation

Yang Shen[1], Robert Vernon[2], David Baker[2], Ad Bax[1]

[1] Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520

[2] Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

Email addresses: shenyang@niddk.nih.gov; bax@nih.gov

Not for publication:
Contact: Ad Bax,
Building 5, room 126
NIH, Bethesda, MD 20892-0520
USA
Ph 301 496 2848; Fax 302 402 0907

Performance of CS23D for examples shown in main text

A subset of the structure calculations described in the main text were also carried out using the CS23D program. As can be seen from the results presented in table S1, CS23D performs very well for the test datasets used in our study. The major strength of CS23D is that it takes optimal advantage of sequence homologues present in the database during fragment selection. Such homologues were present in the structural database for all six proteins studied in our work (see Table S2), but were excluded from the database for the CS-Rosetta testing. The current implementation of CS23D allows exclusion only of the model(s) with "exact matching structure", and performance of the CS23D program in the absence of such models therefore could not be evaluated for the proteins studied. We note that based on results described by Wishart (http://busby1.cs.ualberta.ca/CS23D/documentation.html), the success rate of CS23D is considerably lower for proteins which have no homologues in its NR PDB database.

Table S1. Comparison between CS-ROSETTA and CS23D

Dataset	$RMS^{MFR\,\#}$	$RMS^{Hybrid\,*}$	$RMS^{CS23D\,\|\|}$	Dataset	$RMS^{MFR\,\#}$	$RMS^{Hybrid\,*}$	$RMS^{CS23D\,\|\|}$
MrR16				TM1442			
Ii	2.39/2.97	2.22/2.83	1.63/2.44	Ii	1.76/2.40	1.51/2.19	2.15/2.64
Ij	1.52/2.28	2.40/3.24	2.03/2.68	Ij	1.09/1.88	1.08/1.74	1.90/2.45
IIe	X[¶]	2.08/2.57	1.77/2.50	IIe	X	2.31/2.98	1.92/2.44
IIIb	2.46/3.19	2.04/2.76	1.78/2.46	IIIb	X	1.65/2.25	1.97/2.55
GB3	0.71/1.28	0.73/1.70	0.78/1.31	Ubiquitin	0.69/1.22	0.86/1.49	0.85/1.40
Calbindin	X	1.50/2.10	2.53/3.20	Ferredoxin	X	2.06/3.54	2.24/3.85

[#] RMSD value between the lowest energy CS-ROSETTA model (obtained with a MFR fragment search method) and experimental structure for backbone/heavy atoms.

[*] RMSD value between the lowest energy CS-ROSETTA model (obtained with a hybrid fragment search method) and experimental structure for backbone/heavy atoms.

[‖] RMSD value between the lowest energy CS23D model and experimental structure for backbone/heavy atoms.

[¶]X: not converged

Table S2. Number of sequence homologues in the PDB and CS-ROSETTA database

Protein Name	N^{NR} #	N^{CS} *
MrR16	2	0
TM1442	32	3
Gb3	104	1
Ubiquitin	190	6
Calbindin	167	30
Ferredoxin	82	7

N^{NR}: number of homologues in the NR PDB database, most of which are used by CS23D fragment search (only the one with "exact matching structures" is excluded).

* N^{CS}: number of homologues in the CS-ROSETTA database, all of which are excluded from the CS-Rosetta fragment search in this work.

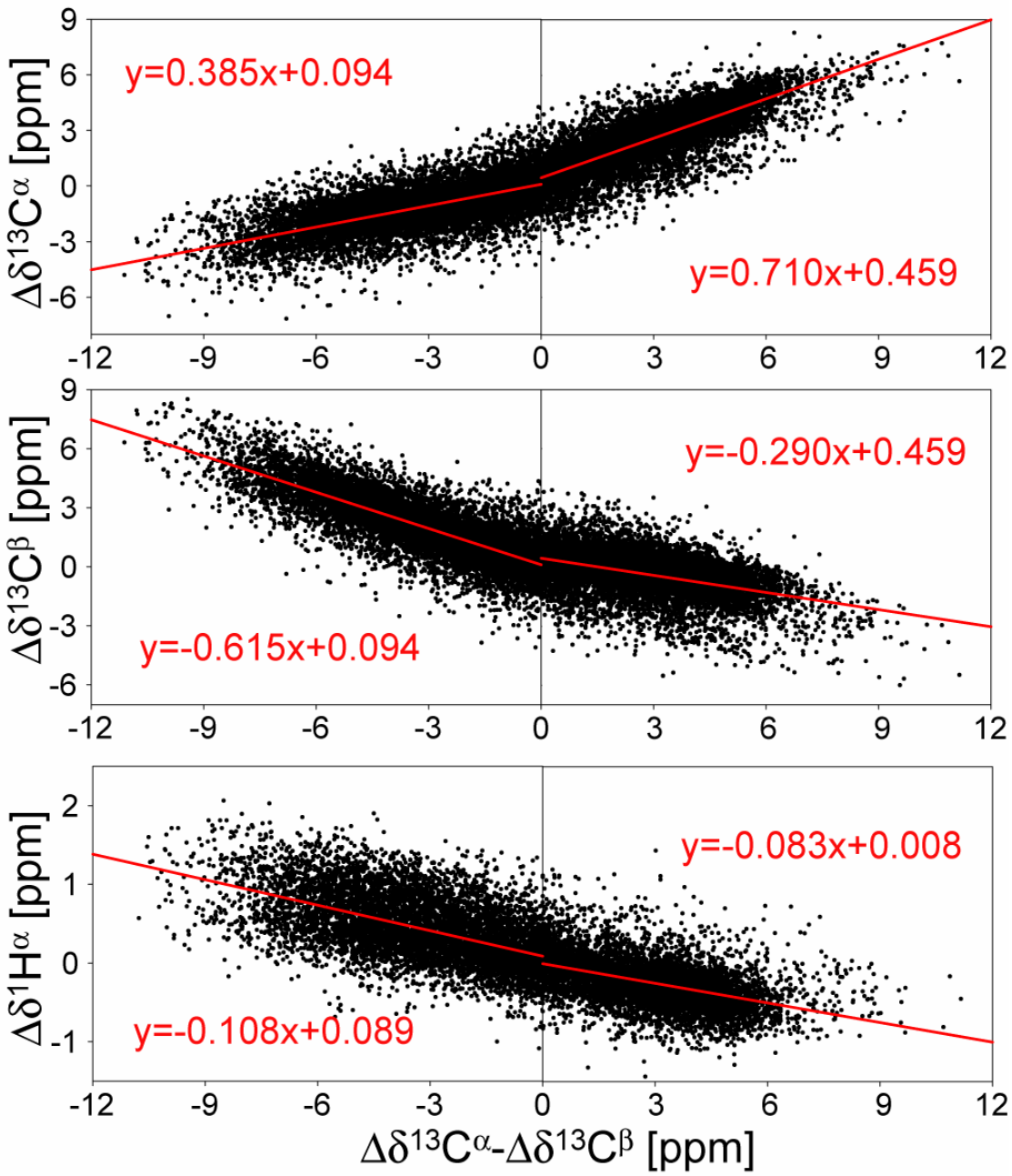

Figure S1. Correlation plots between protein backbone secondary chemical shifts. The experimental chemical shift data chosen from the SPARTA database contains 21,338 $\delta^{13}C^{\alpha/\beta}$ and 21,338 $^{1}H^{\alpha}$. The $\Delta\delta^{13}C^{\alpha}$, $\Delta\delta^{13}C^{\beta}$ and $\Delta\delta^{1}H^{\alpha}$ are plotted against $\Delta\delta^{13}C^{\alpha}$-$\Delta\delta^{13}C^{\beta}$; the best fitting are calculated and labeled for positive and negative $\Delta\delta^{13}C^{\alpha}$-$\Delta\delta^{13}C^{\beta}$, respectively, and plotted with red lines.

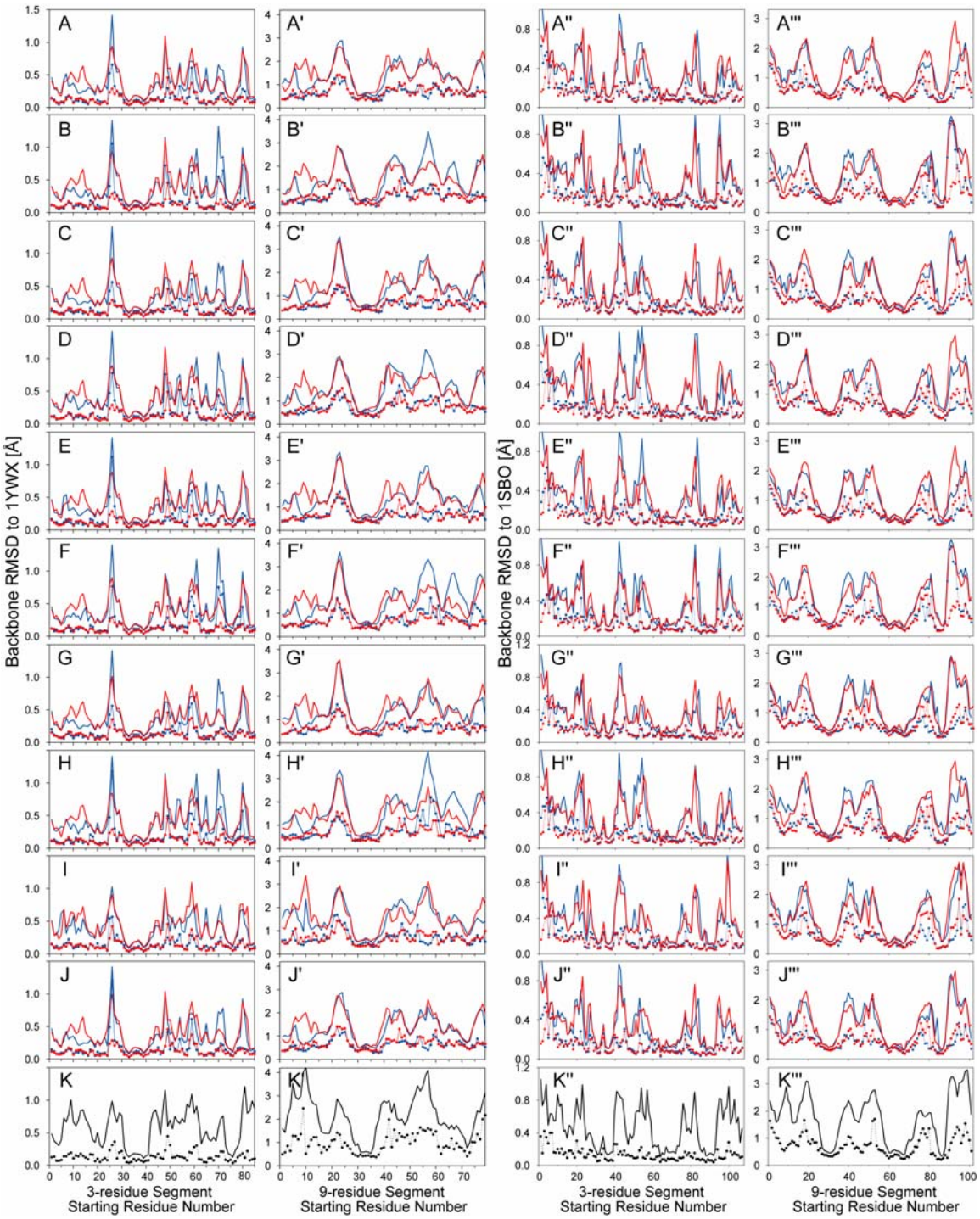

Figure S2. Fragment selections for proteins with missing chemical shifts of certain nucleus types. For MrR16 (A-K and A′-K′) and TM1442 (A″-K″ and A‴-K‴), 200 fragment candidates were selected using the MFR and the hybrid fragment selection

methods, respectively, for each overlapping segment in the proteins. (A-K) Plots of the lowest (lines with dots) and average (bold lines) backbone coordinate rmsds (N, C^{α} and C') between query segment and 200 3-residue fragment candidates, selected using the MFR (blue), the hybrid method (red), or the standard Rosetta method (black) with the inputs of the simulated chemical shift assignment datasets *Ia*-*Ik* as listed in Table 1, as a function of starting position in the sequence of MrR16. (A′-K′) same as (A-K) but for the 9-residue fragment candidates. (A″-K″ and A‴-K‴) same as (A-K and A′-K′) but for the fragment candidates of protein TM1442. Higher resolution figures can be downloaded from http://spin.niddk.nih.gov/bax/software/CSROSETTA/index.html

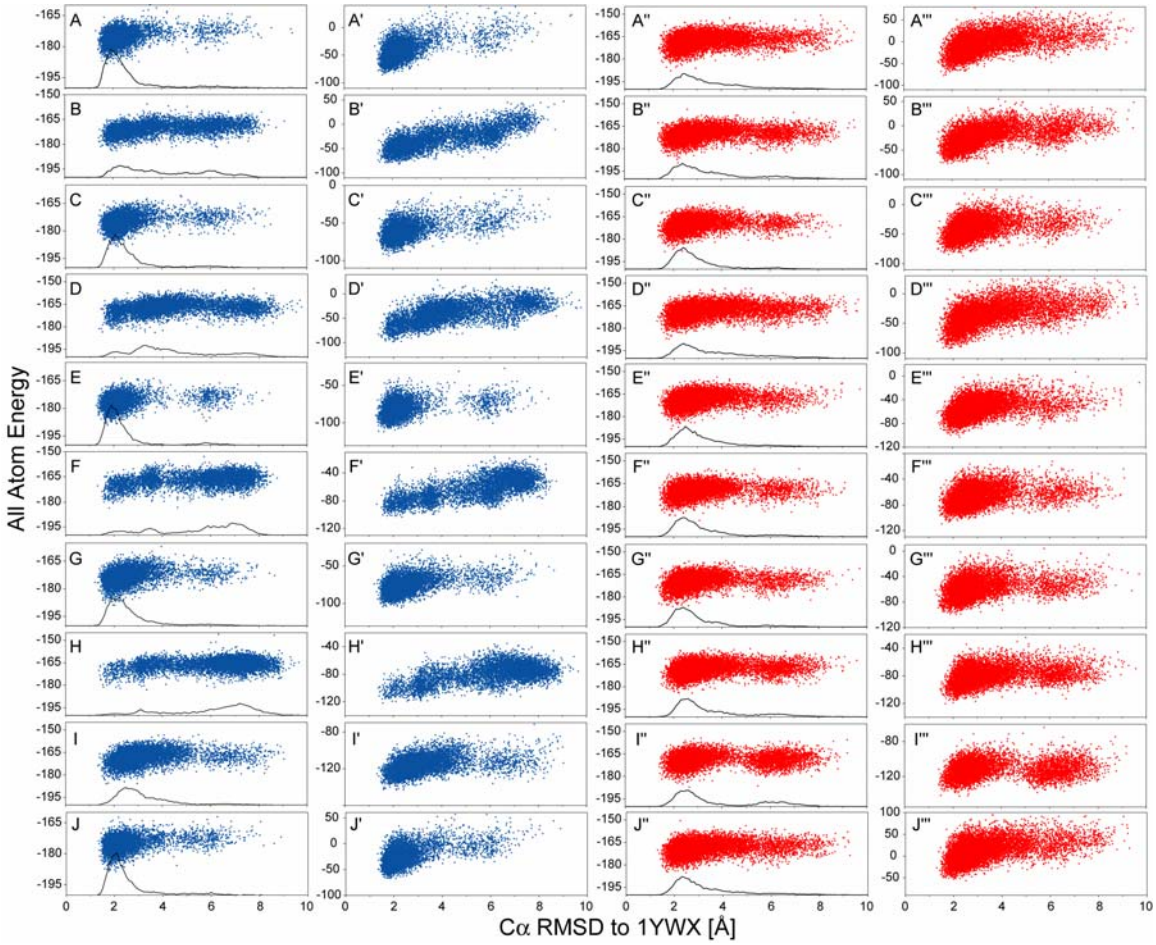

Figure S3. CS-Rosetta structure generation of protein MrR16 with missing chemical shifts of certain nucleus types, using either the MFR fragment selection (blue) or hybrid fragment selection (red) method. (A-J) Plots of Rosetta all-atom energy *versus* C^{α} rmsd relative to the experimental MrR16 structure for the CS-Rosetta models generated by using the MFR-selected fragment candidates with the inputs of the simulated chemical shift assignment datasets *Ia-Ij* as listed in Table 1. The (normalized) number of structures found for a given C^{α}-rmsd is plotted at the bottom of each panel. (A′-J′) Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in the datasets *Ia-Ij*), *versus* C^{α} rmsd relative to the experimental MrR16 structure for the CS-Rosetta models generated by using the MFR fragment selection method. (A″-J″ and A‴-J‴) same as (A-J and A′-J′) but for the CS-Rosetta all-atom models generated using the hybrid fragment selection method.

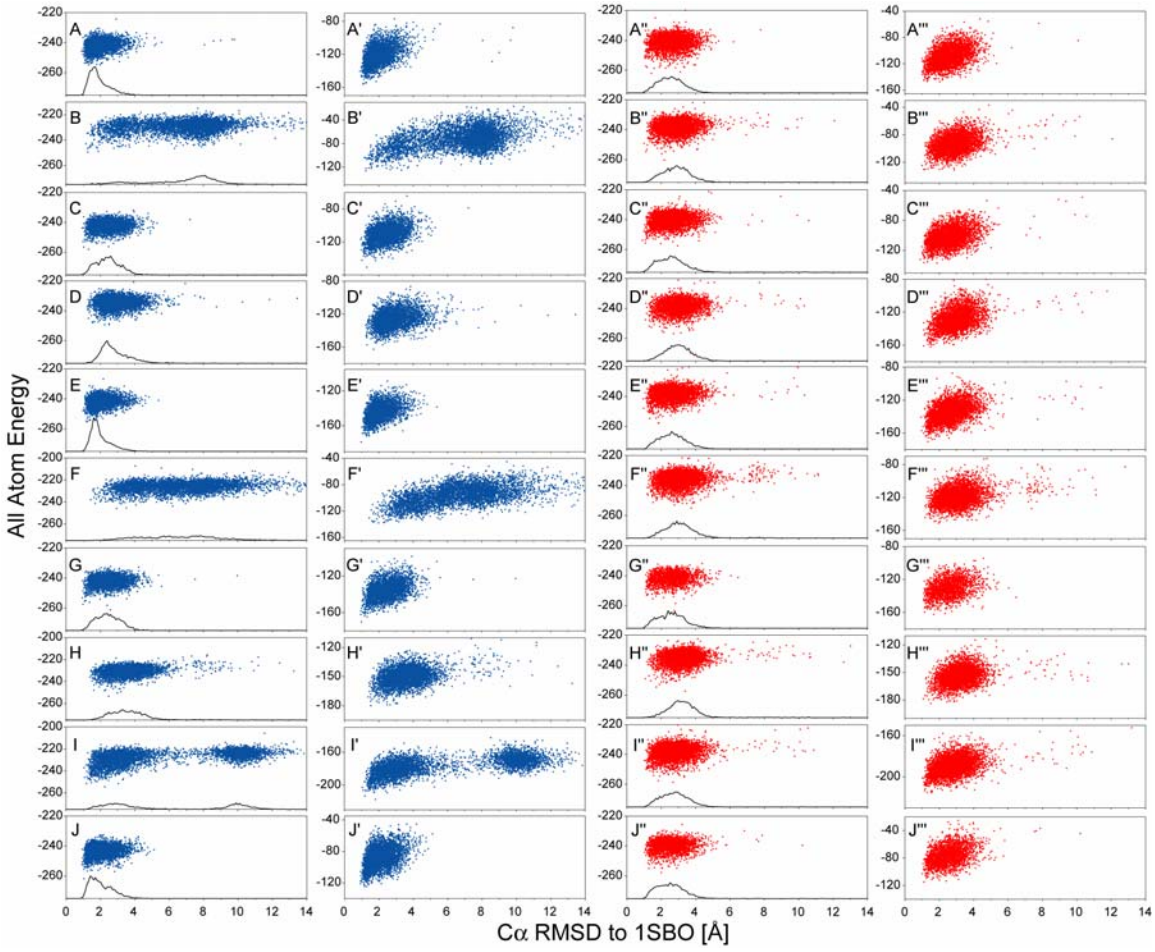

Figure S4. CS-Rosetta structure generation of protein TM1442 with missing chemical shifts of certain types of nuclei. (A-J) Plots of Rosetta all-atom energy *versus* C^{α} rmsd relative to the experimental TM1442 structure for the CS-Rosetta models obtained by using the MFR-selected fragment candidates with the inputs of the simulated chemical shift assignment datasets *Ia-Ij* as listed in Table 1. The (normalized) number of structures found for a given C^{α}-rmsd is plotted at the bottom of each panel. (A′-J′) Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in the datasets *Ia-Ij*), *versus* C^{α} rmsd relative to the experimental TM1442 structure for the CS-Rosetta models generated by using the MFR fragment selection method. (A″-J″ and A‴-J‴) same as (A-J and A′-J′) but for the CS-Rosetta all-atom models generated using the hybrid fragment selection method.

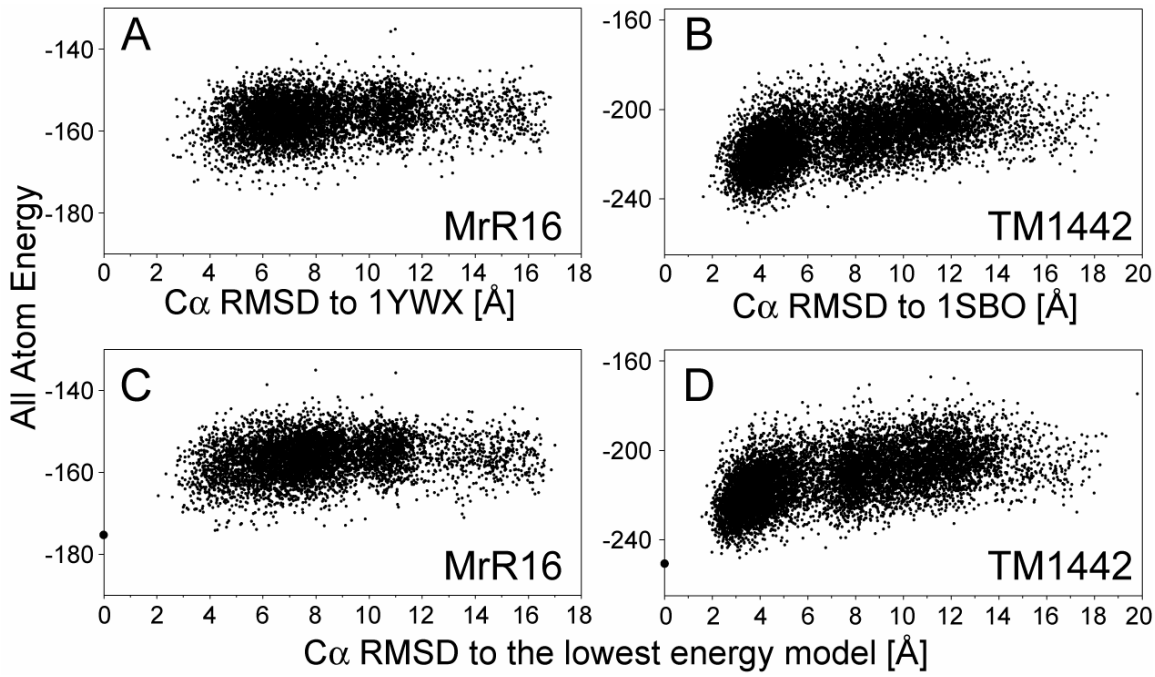

Figure S5. Reference Rosetta structure generation for two test proteins, in the absence of any chemical shift information. (A,B) Plots of Rosetta empirical energy *versus* C^{α} rmsd relative to the experimental NMR structure for MrR16 (A) and TM1442 (B) for 10,000 Rosetta all-atom models. (C,D) Plots of Rosetta empirical energy *versus* C^{α} rmsd relative to the model with the lowest Rosetta energy (shown as bold dot on the vertical axis) for the MrR16 (C) and TM1442 (D) models. For both proteins, the lowest-energy Rosetta folds are roughly correct, but the MrR16 results do not meet convergence criteria, and the TM1442 only meets more relaxed convergence criteria (10 lowest energy models within 4 Å C^{α} rmsd from the lowest energy model).

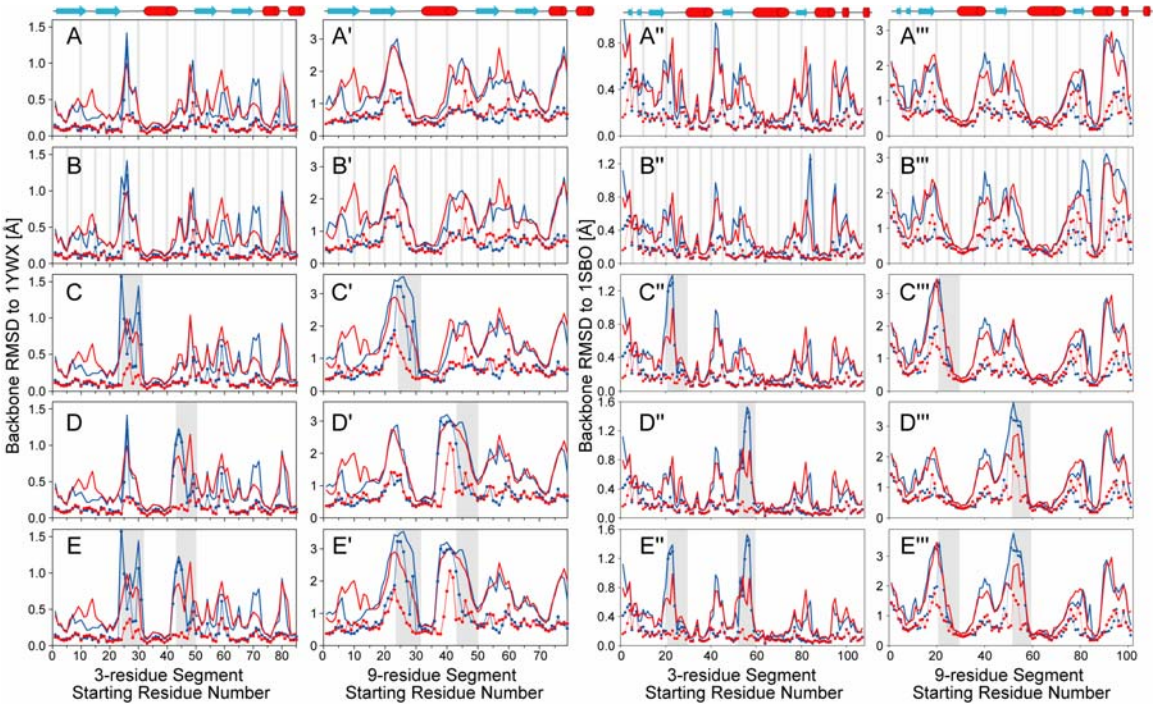

Figure S6. Fragment selections for proteins with missing chemical shift assignments of certain residues. For MrR16 (A-E and A′-E′) and TM1442 (A″-E″ and A‴-E‴), 200 fragment candidates were selected using the MFR and the hybrid fragment selection methods, respectively, for each overlapping segment in the proteins. (A-E) Plots of the lowest (lines with dots) and average (bold lines) backbone coordinate rmsds (N, C^{α} and C') between query segment and 200 3-residue fragment candidates, selected using the MFR (blue) and the hybrid (red) methods with the inputs of the simulated chemical shift assignment datasets *IIa-IIe* (see Method), as a function of starting position in the sequence of MrR16. The regions corresponding to the "unassigned" residues are shaded; the secondary structure elements are displayed at the top of each column. (A′-E′) same as (A-E) but for the 9-residue fragment candidates. (A″-E″ and A‴-E‴) same as (A-E and A′-E′) but for the fragment candidates of protein TM1442. Higher resolution figures can be downloaded from http://spin.niddk.nih.gov/bax/software/CSROSETTA/index.html

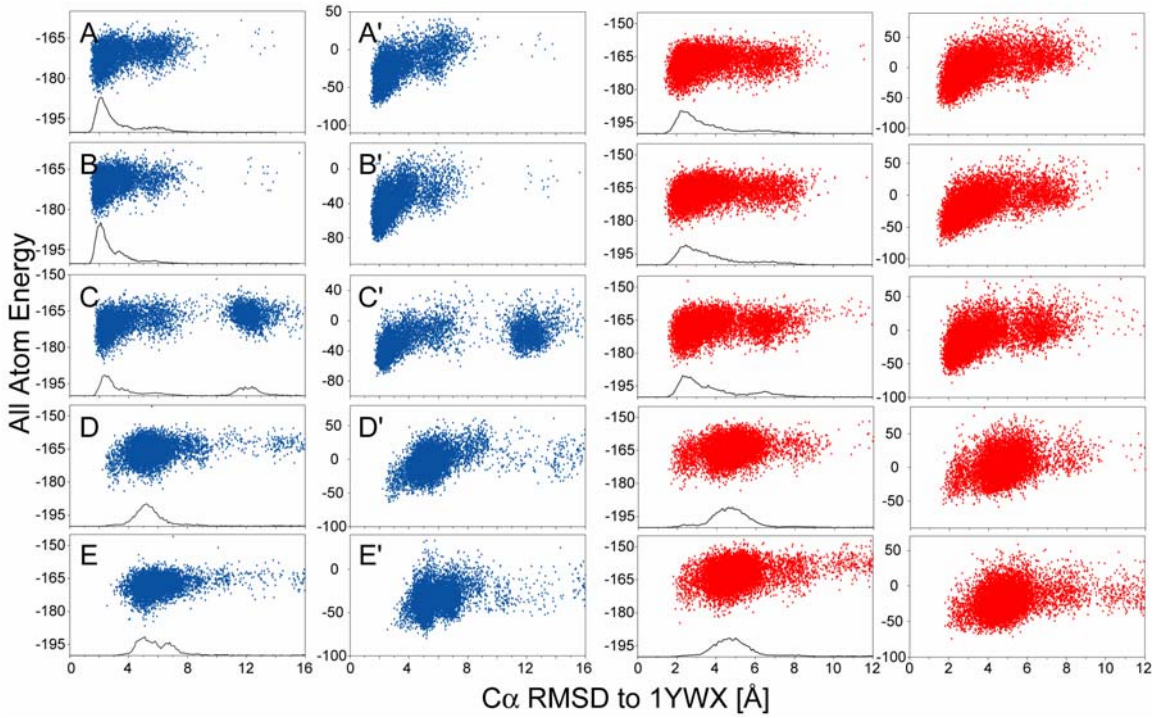

Figure S7. CS-Rosetta structure generation of MrR16 with missing chemical shifts of certain residues. (A-E) Plots of Rosetta all-atom energy *versus* C^{α} rmsd relative to the experimental MrR16 structure for the CS-Rosetta models obtained by using the MFR-selected fragment candidates with the inputs of the simulated chemical shift assignment datasets *IIa-IIe* (see Method). The (normalized) number of structures found for a given C^{α}-rmsd is plotted at the bottom of each panel. (A′-E′) Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in the datasets *IIa-IIe*), *versus* C^{α} rmsd relative to the experimental MrR16 structure. (A″-E″ and A‴-E‴) same as (A-E and A′-E′) but for the CS-Rosetta models generated using the hybrid fragment selection method.

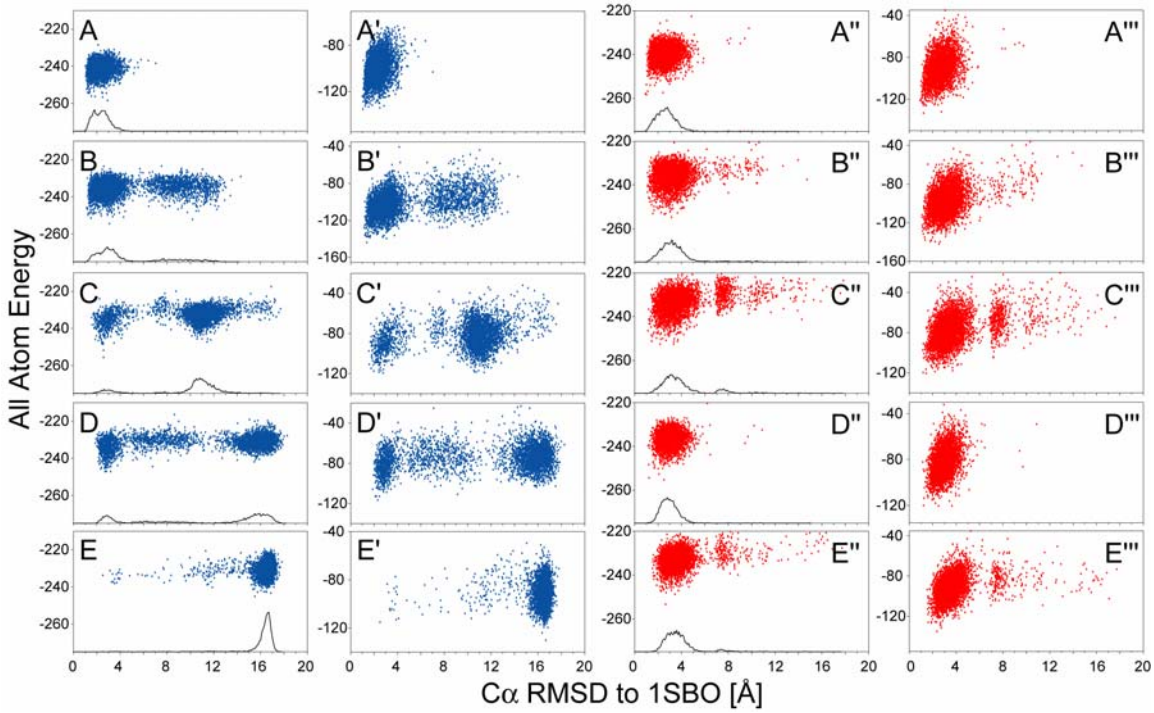

Figure S8. CS-Rosetta structure generation of protein TM1442 with missing chemical shifts of certain residues. (A-E) Plots of Rosetta all atom energy *versus* C^{α} rmsd relative to the experimental TM1442 structure for CS-Rosetta models obtained by using the MFR-selected fragment candidates with the inputs of the simulated chemical shift assignment datasets *IIa-IIe* (see Method). The (normalized) number of structures found for a given C^{α}-rmsd is plotted at the bottom of each panel. (A′-E′) Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in the datasets *IIa-IIe*), *versus* C^{α} rmsd relative to the experimental TM1442 structure. (A″-E″ and A‴-E‴) same as (A-E and A′-E′) but for the CS-Rosetta models generated using the hybrid fragment selected method.

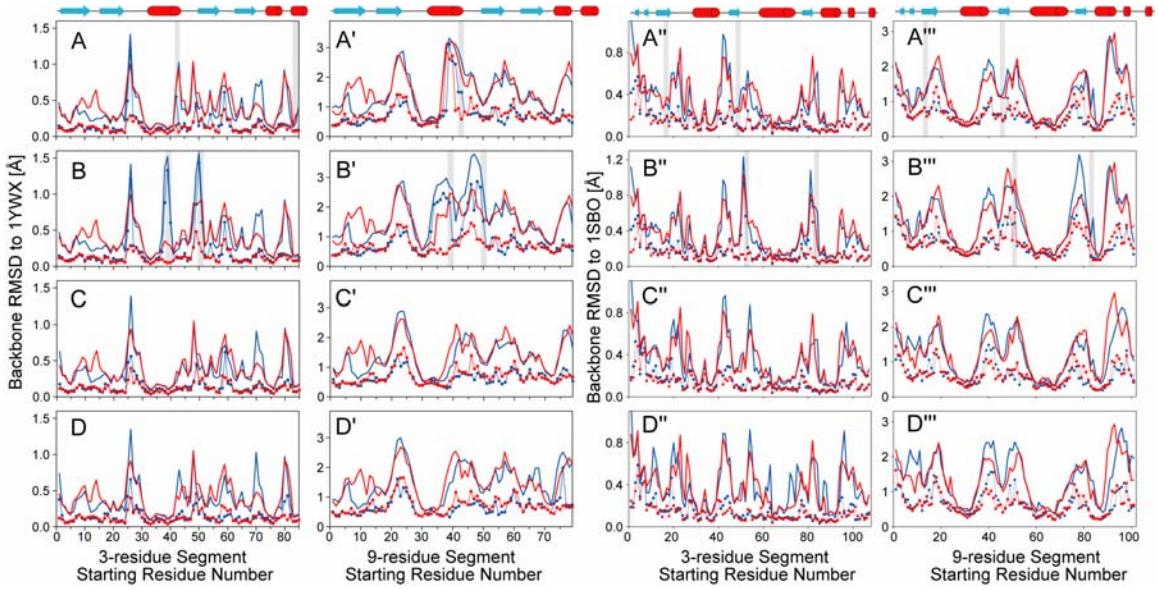

Figure S9. Fragment selections for proteins with chemical shift errors. For proteins MrR16 (A-D and A′-D′) and TM1442 (A″-D″ and A‴-D‴), 200 fragment candidates were selected using the MFR and the hybrid fragment selection methods, respectively, for each overlapping segment in the proteins. (A-D) Plots of the lowest (lines with dots) and average (bold lines) backbone coordinate rmsds (N, C^{α} and C') between query segment and 200 3-residue fragment candidates, selected by using the MFR (blue) and the hybrid (red) methods and with the inputs of the simulated chemical shift assignment datasets *IIIa-IIId* (see Method), as a function of starting position in the sequence of MrR16. The regions corresponding to the "miss-assigned" residues are shaded; the secondary structure elements are displayed at the top of each column. (A′-D′) same as (A-D) but for the 9-residue fragment candidates. (A″-D″ and A‴-D‴) same as (A-D and A′-D′) but for the fragment candidates of protein TM1442. Higher resolution figures can be downloaded from http://spin.niddk.nih.gov/bax/software/CSROSETTA/index.html

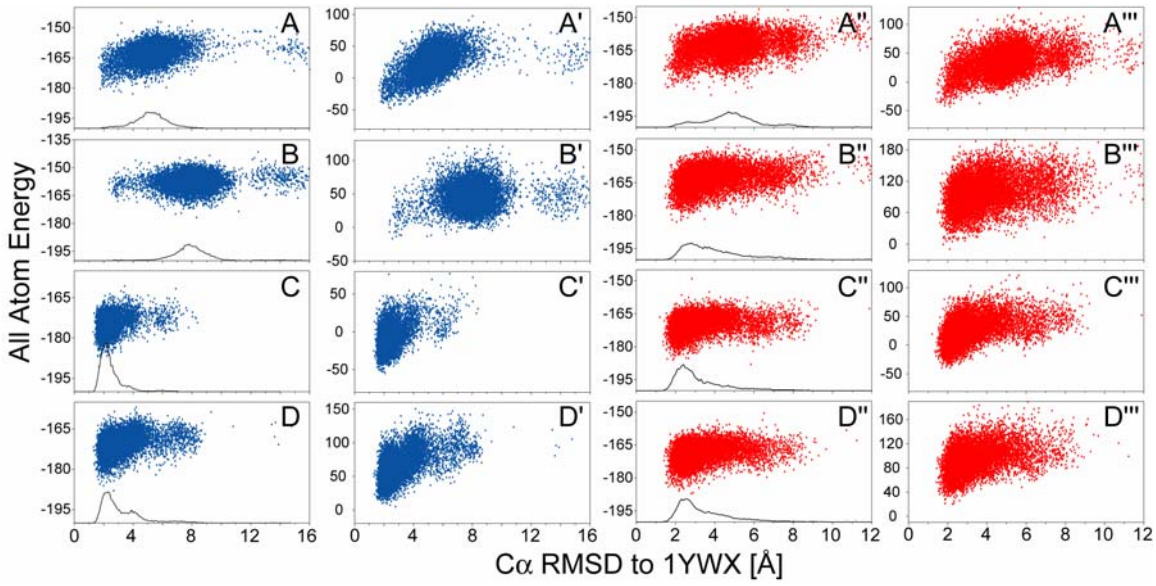

Figure S10. CS-Rosetta structure generation of protein MrR16 with chemical shifts errors. (A-D) Plots of Rosetta all atom energy *versus* C^{α} rmsd relative to the experimental MrR16 structure for the CS-Rosetta models obtained by using the MFR-selected fragment candidates with the inputs of the simulated chemical shift assignment datasets *IIIa-IIId* (see Method). The (normalized) number of structures found for a given C^{α}-rmsd is plotted at the bottom of each panel. (A′-D′) Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in the datasets *IIIa-IIId*), *versus* C^{α} rmsd relative to the experimental MrR16 structure. (A″-D″ and A‴-D‴) same as (A-D and A′-D′) but for the CS-Rosetta models generated using the hybrid fragment selected method.

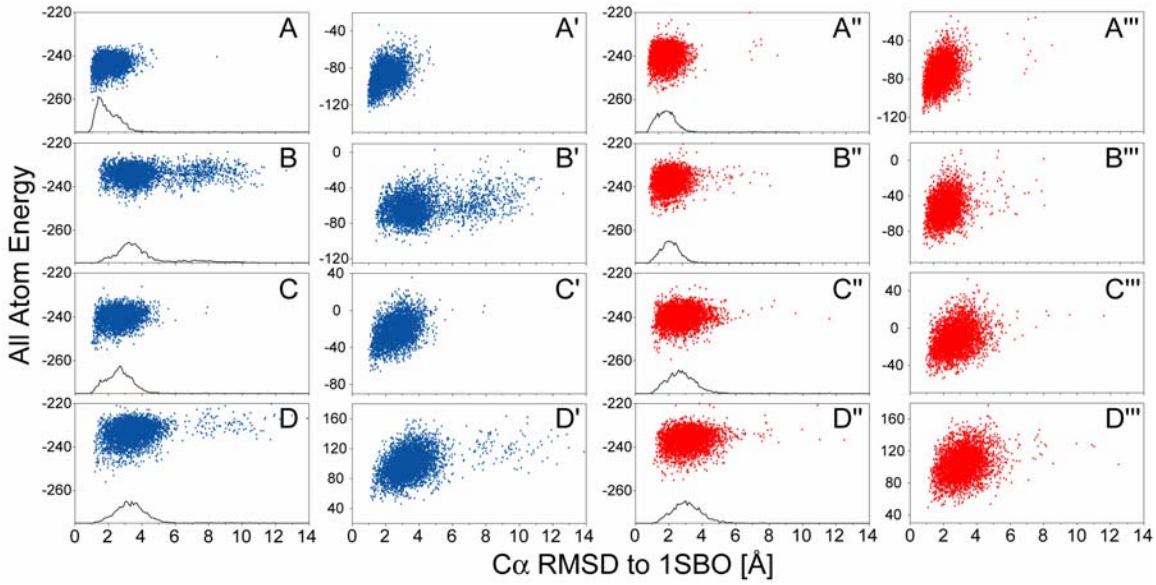

Figure S11. CS-Rosetta structure generation of protein TM1442 with chemical shifts errors. (A-D) Plots of Rosetta all atom energy *versus* C^{α} rmsd relative to the experimental TM1442 structure for the CS-Rosetta models obtained by using the MFR-selected fragment candidates with the inputs of the simulated chemical shift assignment datasets *IIIa-IIId* (see Method). The (normalized) number of structures found for a given C^{α}-rmsd is plotted at the bottom of each panel. (A′-D′) Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in the datasets *IIIa-IIId*), *versus* C^{α} rmsd relative to the experimental TM1442 structure. (A″-D″ and A‴-D‴) same as (A-D and A′-D′) but for the CS-Rosetta models generated using the hybrid fragment selected method.

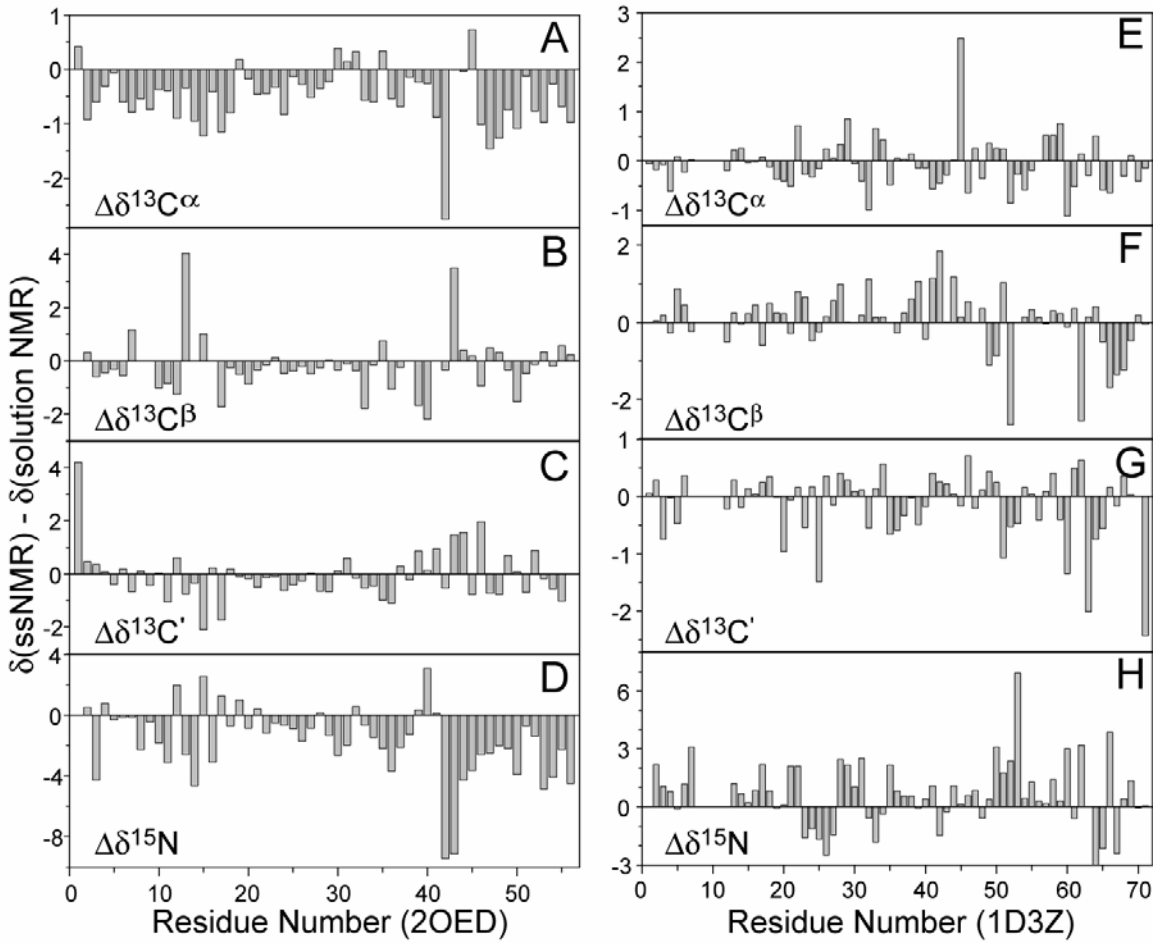

Figure S12. Difference between chemical shifts obtained using solid-state and solution NMR spectroscopy for protein GB3 (Left) and Ubiquitin (Right). For each protein, the differences of $\delta^{13}C^{\alpha}$ (A,E), $\delta^{13}C^{\beta}$ (B,F), $\delta^{13}C'$ (C,G) and $\delta^{15}N$ (D,H) are plotted.

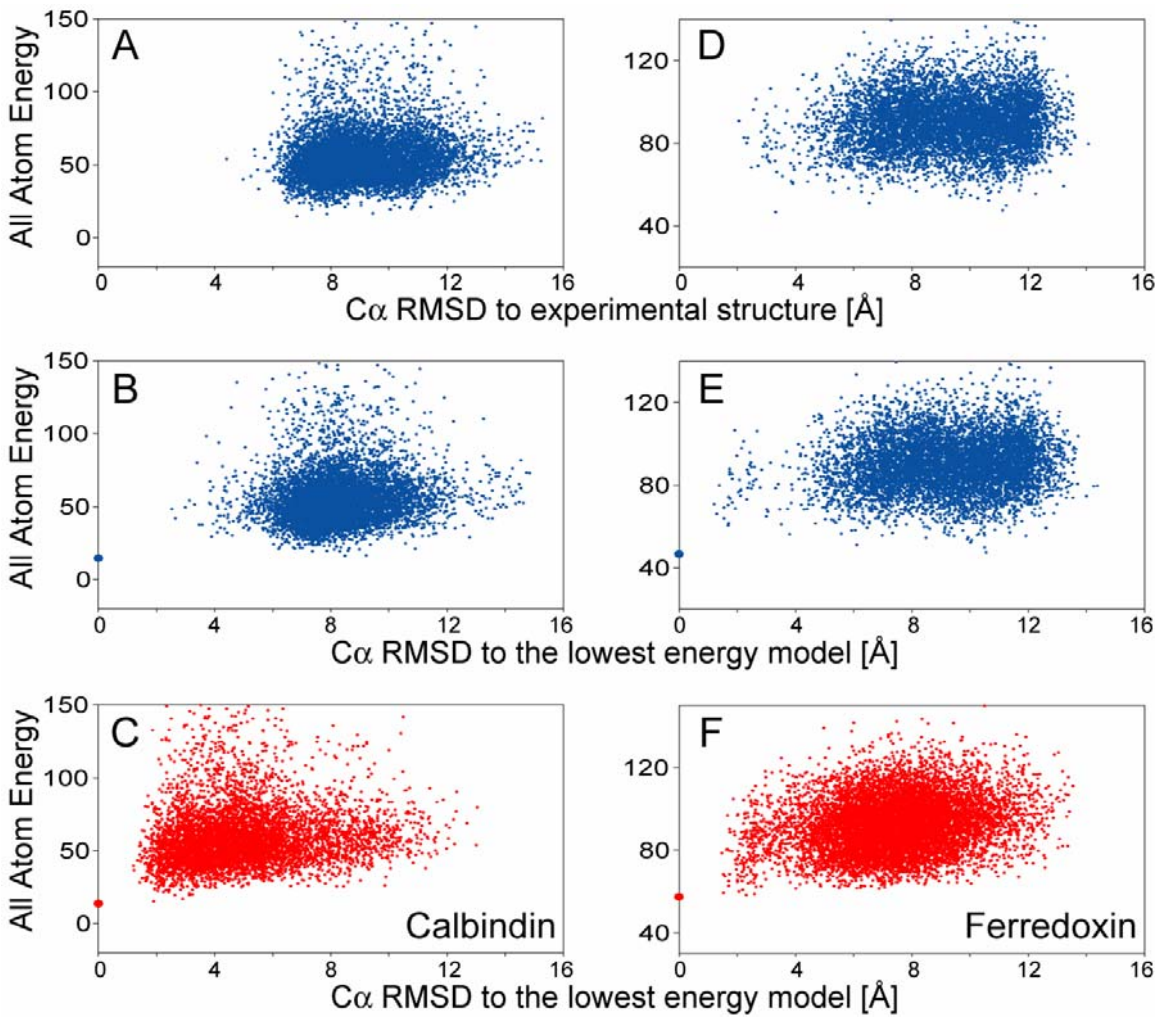

Figure S13. CS-Rosetta structure generation for paramagnetic proteins. For proteins calbindin (A,B,C) and ferredoxin (D,E,F), the all-atom models were generated by using a CS-Rosetta protocol with the MFR and the hybrid fragment selection methods separately, and their Rosetta all atom energy are plotted in blue (A,B,D,E) and red (C,F), respectively, with respect to their quality. (A,D) Plots of Rosetta all-atom energy, rescored by using the experimental NMR chemical shifts, *versus* C^{α} rmsd of all-atom models relative to the experimental structures. (B,C,E,F) Plots of Rosetta all-atom energy, rescored by using the experimental NMR chemical shifts, *versus* C^{α} rmsd of all-atom models relative to the model with the lowest energy (shown as a bold dot on the vertical axis). C^{α} rmsd values are calculated for the residues in secondary structure only, which contain residues 3-14, 25-40, 46-53 and 63-74 for calbindin, 4-11, 15-22, 27-34, 54-56, 71-75 and 91-93 for ferredoxin, respectively.