Biological Sciences: Biophysics Supporting Information

# Consistent blind protein structure generation from NMR chemical shift data

Yang Shen[*], Oliver Lange[†], Frank Delaglio[*], Paolo Rossi[‡,††], James M. Aramini[‡,††], Gaohua Liu[‡,††], Alexander Eletsky[§,††], Yibing Wu[§,††], Kiran K. Singarapu[§,††], Alexander Lemak[¶,††], Alexandr Ignatchenko[¶,††], Cheryl H. Arrowsmith[¶,††], Thomas Szyperski[§,††], Gaetano T. Montelione[‡,††], David Baker[†,‖], and Ad Bax[*,‖]

[*] Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892; [†] Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195; [‡] Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, and Robert Wood Johnson Medical School, Piscataway, NJ 08854; [§] Departments of Chemistry and Structural Biology, University at Buffalo, State University of New York, Buffalo, NY 14260; [¶]Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada M5G IL5, [††]Northeast Structural Genomics Consortium

[‖] To whom correspondence should be addressed. Email: dabaker@u.washington.edu; bax@nih.gov

**SI Text**

**Methods**

**Structural database and fragment searching.** A protein database was constructed using 5665 x-ray structures selected from the PISCES server (1) by using the criteria of: (*i*) minimum sequence length of 40, (*ii*) sequence identity <40%, and (*iii*) resolution of 2.4 Å or better. These criteria result in a highly diverse set of good quality fragments. Hydrogen atoms were added using the REDUCE program (2), and the backbone $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{15}N$, $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts were added to the database by prediction from the structure by means of SPARTA (3). The ROSETTA idealization routine (4, 5) was then applied to each full atom structure to regularize the protein backbone. The secondary structure classification was obtained by the DSSP program (6), and was used by the standard ROSETTA program for identifying strand-strand and helix-strand pairing terms (7).

Test proteins (Table 1) were selected to represent different structural classes, with chemical shifts ($^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{15}N$, $^{1}H^{\alpha}$, and $^{1}H^{N}$) available from the BMRB (8) or TALOS (9) database, and with published x-ray or NMR structures. For each consecutively overlapping 3-residue and 9-residue fragment in each query protein, an exhaustive search was conducted throughout the structural database by using the conventional MFR method (10) to find the fragments with the best matched chemical shifts and residue sequence patterns. The final score for a particular fragment was adapted from (10) by removing the term for RDC data:

$$E_{TOTAL} = c_{CS}E_{CS} + c_{HOMO}E_{HOMO} + c_{RAMA}E_{RAMA} \qquad [2]$$

with $c_{CS}$ = 0.1, $c_{HOMO}$ = 0.01, $c_{RAMA}$ = 0.02, and the modified chemical shift score:

$$E_{CS} = \sum_{i,j} c_i \times \left[ \frac{\delta_{i,j}^{\exp} - \delta_{i,j}^{calc}}{\sigma_{i,j}^{calc}} \right]^2 / N \qquad [3]$$

where $\delta_{i,j}$ stands for the chemical shifts of atom $i$ ($i$ = $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{15}N$, $^{1}H^{\alpha}$ and $^{1}H^{N}$) for residue $j$ in the fragment; <graphic6>is the experimental chemical shift in the target segment; <graphic7> and <graphic8> denote the SPARTA-derived chemical shifts and uncertainties, respectively, for the fragments in the protein structural database; $N$ is the total number of chemical shifts in the fragment; $c_i$ is the weighting factor for each atom type (1.0 for $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{\alpha}$; 0.9 for $^{1}H^{N}$ and $^{15}N$). $E_{HOMO}$ and $E_{RAMA}$ are terms previously introduced during MFR searching (10) which favor sequence homology and assign a penalty to disfavored regions of the Ramachandran map. Iterative adjustment of the weight factors resulted in a very low value for $c_{RAMA}$, reflecting the fact that structures in the database generally do not include such disfavored local geometries.

Fragments from proteins with homologous sequence (PSI-BLAST e-value <0.05) to the target protein were excluded from the structural database before MFR fragment searching. For each query protein segment, the 200 9-residue and 200 3-residue fragments with the lowest matching scores are kept and stored along with the idealized backbone torsion angles and secondary structure classification for each residue, and subsequently used as input for the ROSETTA Monte Carlo fragment assembly procedure.

**Blind protein structure generation.** The nine structural genomics targets used for blind testing the CS-ROSETTA method (Table 2) were produced by the NESG consortium following established procedure (11). The experimental NMR structures were solved by three NESG NMR groups using either conventional or high-throughput NMR structure determination protocols (12, 13).

**Identification and exclusion of flexible tails and loops.** Residues in the flexible N- and C-terminal tails and loops are identified by $\boldsymbol{S}^2 < 0.7$, with $\boldsymbol{S}^2$ derived by RCI analysis (14), and the absence of “good” predictions using the program TALOS (9). N- and C-terminal tails that had been positively identified as flexible on the basis of their RCI scores were excluded from the structure prediction. Moreover, when adjusting the all-atom energies of the predicted models for their agreement with the experimental chemical shifts (Eq 1), all terms involving long loops (≥9 residues) that had been positively identified as flexible, were excluded from the total energy.

1. Wang GL, Dunbrack RL (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19:1589-1591.

2. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* 285:1735-1747.

3. Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* 38:289-302.

4. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268:209-225.

5. Bradley P, Misura KMS, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868-1871.

6. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.

7. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 34:82-95.

8. Doreleijers JF, Nederveen AJ, Vranken W, Lin JD, Bonvin A, Kaptein R, Markley JL, Ulrich EL (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J Biomol NMR* 32:1-12.

9. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13:289-302.

10. Kontaxis G, Delaglio F, Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. *Methods Enzymol* 394:42-78.

11. Acton TB, Gunsalus KC, Xiao R, Ma LC, Aramini J, Baran MC, Chiang YW, Climent T, Cooper B, Denissova NG, Douglas SM, Everett JK, Ho CK, Macapagal D, Rajan PK, Shastry R, Shih LY, Swapna GVT, Wilson M, Wu M, Gerstein M, Inouye M, Hunt JF, Montelione GT (2005) Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods Enzymol.* 394:210-243.

12. Liu GH, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Lemak A, Bhattacharya A, Acton TA, Arrowsmith CH, Montelione GT, Szyperski T (2005) NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc. Natl Acad Sci USA* 102:10487-10492.

13. Grishaev A, Steren CA, Wu B, Pineda-Lucena A, Arrowsmith C, Llinas M (2005) ABACUS, a direct method for protein NMR structure computation via assembly of fragments. *Proteins* 61:36-43.

14. Berjanskii M, Wishart DS (2006) NMR: prediction of protein flexibility. *Nature Protoc* 1:683-688.

**Table 3**. Full survey of converged protein structures generated by CS-ROSETTA

| Protein name | PDB*/ BMRB ID | $N_\alpha/N_\beta$† | $N_{all}$‡ | $N_{cs}$§ | $RMSD^{mean}$¶ [Å] | | $RMSD^{exp}$‖ [Å] | |
|---|---|---|---|---|---|---|---|---|
| | | | | | *Backbone* | *All* | *Backbone* | *All* |
| GB3 | 2OED | 14/26 | 56(1-55) | 332 | 0.25±0.08 | 0.48±0.11 | 0.74±0.05 (0.69) | 1.43±0.05 (1.34) |
| CspA | *1MJC*/4296 | 0/33 | 70(4-70) | 405 | 0.96±0.23 | 1.44±0.19 | 1.43±0.29 (1.08) | 2.25±0.33 (1.74) |
| Calbindin | 4ICB/390 | 47/0 | 75(3-74) | 435 | 0.68±0.23 | 0.90±0.21 | 1.39±0.11 (1.20) | 2.13±0.07 (1.92) |
| Ubiquitin | 1D3Z | 18/25 | 76(2-72) | 426 | 0.34±0.11 | 0.76±0.12 | 0.82±0.06 (0.75) | 1.59±0.14 (1.40) |
| XcR50 | *1TTZ*/6363 | 28/16 | 76(3-73) | 352 | 0.98±0.32 | 1.37±0.39 | 1.67±0.27 (1.34) | 2.13±0.50 (2.06) |
| DinI | 1GHH | 36/21 | 81(1-77) | 463 | 0.90±0.24 | 1.16±0.25 | 1.73±0.25 (1.54) | 2.38±0.14 (2.07) |
| HPr | *1POH* | 29/23 | 85(2-83) | 419 | 0.95±0.32 | 1.28±0.35 | 1.30±0.43 (0.93) | 1.99±0.37 (1.54) |
| MrR16 | 1YWX/6799 | 23/35 | 88(2-81) | 514 | 0.73±0.18 | 1.03±0.19 | 1.77±0.22 (1.61) | 2.40±0.21 (2.17) |
| TM1112 | *1O5U*/5357 | 10/52 | 89(4-88) | 524 | 1.06±0.26 | 1.55±0.22 | 1.58±0.16 (1.16) | 2.30±0.14 (1.70) |
| PHS018 | 2GLW/7116 | 20/41 | 92(6-88) | 531 | 1.12±0.31 | 1.51±0.28 | 1.56±0.26 (1.08) | 2.27±0.20 (1.69) |
| HR2106** | *2HZ5*/6210 | 37/25 | 96(2-92) | 470 | 0.80±0.26 | 1.10±0.22 | 1.85±0.27 (1.47) | 2.58±0.23 (2.14) |
| TM1442 | 1SBO/5921 | 41/23 | 110(5-109) | 647 | 0.66±0.31 | 1.02±0.29 | 1.22±0.27 (1.01) | 1.90±0.20 (1.60) |
| Vc0424 | 1NXI/5589 | 55/25 | 114(2-112) | 679 | 0.88±0.16 | 1.34±0.17 | 1.74±0.09 (1.35) | 2.53±0.11 (2.04) |
| Spo0F | *1SRR*/5899 | 55/25 | 121(2-115) | 590 | 1.09±0.21 | 1.41±0.22 | 1.67±0.19 (1.26) | 2.30±0.13 (1.80) |
| Profilin | *1PRQ* | 41/41 | 125(2-123) | 595 | 1.04±0.31 | 1.46±0.35 | 2.26±0.35 (2.02) | 2.88±0.34 (2.49) |
| Apo_Ifabp | *1LFO*/4098 | 15/70 | 129(5-126) | 688 | 1.36±0.35 | 1.64±0.30 | 1.72±0.55 (1.12) | 2.33±0.43 (1.68) |

* Proteins for which experimental structures were obtained by X-ray diffraction are in *italic*; for proteins solved by NMR the first model of the NMR ensemble is used as the experimental reference structure.

† Number of residues in $\alpha$-helix and $\beta$-strand.

‡ Total number of residues. Numbers of the first and last residue involved in secondary structures are listed in parenthesis; these and all intervening residues were used to calculate the RMSD values of the predicted models relative to experimental structures. For cspA, residues 39 to 46 in the flexible loop are excluded for RMSD calculation.

§ Total number of the backbone chemical shifts used for the structure prediction; no $\delta^{13}C'$ available for XcR50, Hr2106 and Spo0F; no $\delta^1H^N$ available for Profilin.

¶ RMSD between the 10 lowest-energy models and the mean coordinates for all backbone $C^\alpha$, $C'$ and N atoms (referred as "Backbone"), and all non-hydrogen atoms ("All").

‖ RMSD between the 10 lowest-energy models and the experimental structure. The RMSD of the mean coordinates of the 10 lowest-energy models and the experimental structures are listed in parenthesis.

** Protein HR2106 is a homo-dimer, only the monomer conformation is predicted by CS-ROSETTA and used for comparisons.

**Table 4**. Survey of proteins for which CS-ROSETTA did not meet convergence criteria.

| Protein name | PDB*/BMRB code | $N_\alpha/N_\beta$† | $N_{all}$‡ | $N_{shifts}$§ | Cα RMSD [Å]¶ | |
|---|---|---|---|---|---|---|
| | | | | | *Lowest RMSD* | *Lowest Energy* |
| HI0719 | 1J7H/5606 | 40/30 | 130 (3-129) | 733 | 4.50‖ | 14.31‖ |
| MTH1598 | 1JW3/5165 | 32/47 | 140 (4-139) | 830 | 3.65** | 12.17** |
| HR1958 | *1TVG*/6344 | 8/73 | 140 (4-139) | 829 | 9.37†† | 16.29†† |
| CcR19 | 1T17/6120 | 37/59 | 148 (2-144) | 842 | 3.67 | 7.09 |
| YwlE | 1ZGG/6460 | 68/21 | 150 (2-145) | 851 | 3.72 | 9.37 |
| Flua | *1N0S*/5756 | 26/83 | 173 (2-163) | 1022 | 5.54 | 15.57 |
| nsp1 | 2GDT/7014 | 17/33 | 116 (2-112) | 609 | 5.16‡‡ | 5.16‡‡ |

* Proteins with reference X-ray structures are in *italic*; for proteins solved by NMR the first model of the NMR ensemble is used as the reference structure.

† Number of residues in α-helix and β-strand.

‡ Total number of residues. The first and last residue numbers of the secondary structures are listed in parenthesis; Numbers of the first and last residue involved in secondary structures are listed in parenthesis; these and all intervening residues were used to calculate the RMSD values of the predicted models relative to experimental structures.

§ Total number of backbone chemical shifts.

¶ $C^\alpha$ RMSD (relative to the experimental reference structures) for the models with the lowest RMSD and lowest energy.

‖ Residues 7 to 20 and 31 to 45, which are in flexible loops, are excluded for the RMSD calculation.

** Residues 39 to 47 and 104 to123, in flexible loops, are excluded for the RMSD calculation.

†† Flexible loop residues 17-38 are excluded for the RMSD calculation.

‡‡ Flexible loop residues 63-73 are excluded for the RMSD calculation.

**Table 5**. Survey of protein structures generated by CS-ROSETTA and independently by the NESG consortium

| Protein name | RpT7 | StR82 | RhR95 | NeT4 | TR80 | VfR117 | PsR211 | AtR23 | NeR45A[‡‡] |
|---|---|---|---|---|---|---|---|---|---|
| UniProt ID | Q6N4D8_RHOPA | Q04822_SALTY | Q3IZ23_RHOS4 | Q82V59_NITEU | RLX_METTH | Q5E7H1_VIBF1 | Q885L4_PSESM | Q8UEE9_AGRT5 | Q82VF2_NITEU |
| PDB/BMRB ID | 2jtv | 2jt1 | 2jvm | 2jv8 | 2jxt | 2jvw | 2jva | 2yja | 2jxn |
| Protein Size[*] | 65(2-63) | 69(5-69) | 72(22-68) | 73(3-66) | 78(5-77) | 80(15-75) | 100(2-100) | 101(2-78) | 147(16-143) |
| M.W [kDa][*] | 7.8 | 8.0 | 8.5 | 8.7 | 9.8 | 10.2 | 11.6 | 10.8 | 15.4 |
| $N_\alpha/N_\beta$[†] | 38/15 | 36/10 | 4/19 | 11/18 | 23/31 | 43/0 | 29/21 | 11/25 | 41/52 |
| $N_{CS}$ | 345 | 400 | 405 | 429 | 357 | 468 | 589 | 569 | 765 |
| Predicted models[‡] | | | | | | | | | |
| $RMSD_{bb}/RMSD_{all}$[§], [Å] | 0.73±0.10<br>1.25±0.18 | 0.24±0.09<br>0.53±0.13 | 0.68±0.26<br>1.26±0.26 | 0.47±0.15<br>1.05±0.15 | 0.44±0.11<br>0.84±0.11 | 0.68±0.16<br>1.15±0.22 | 1.34±0.27<br>1.72±0.24 | 1.19±0.67<br>1.73±0.65 | 0.83±0.17<br>1.29±0.14 |
| Ramachandran plot[¶,§], [%] | 98/2/0/0 | 98/2/0/0 | 95/5/0/0 | 90/10/0/0 | 96/4/0/0 | 96/4/0/0 | 95/5/0/0 | 96/4/0/0 | 95/5/0/0 |
| Procheck G-factor[§], Φ&Ψ/All | 0.20/0.38 | 0.47/0.56 | -0.26/0.11 | -0.13/0.21 | -0.1/0.16 | 0.50/0.56 | 0.11/0.27 | -0.12/0.20 | -0.01/0.21 |
| MOLPROBITY clash score[§] | 6.71 | 7.28 | 4.40 | 1.98 | 3.62 | 4.50 | 6.38 | 4.41 | 3.34 |
| DP score[§], [%] | 69 | 65 | 55 | 57 | 67 | 37 | 57 | 60 | 53 |
| NMR ensembles | | | | | | | | | |
| $RMSD_{bb}/RMSD_{all}$[§] [Å] | 0.32±0.05<br>0.97±0.09 | 0.50±0.09<br>1.02±0.10 | 0.50±0.11<br>0.91±0.11 | 0.42±0.07<br>0.94±0.09 | 0.42±0.08<br>0.87±0.08 | 0.59±0.10<br>1.17±0.11 | 0.58±0.10<br>0.96±0.10 | 0.42±0.08<br>0.89±0.09 | 0.70±0.08<br>1.22±0.07 |
| Ramachandran plot[¶,§], [%] | 97/3/0/0 | 97/3/0/0 | 92/7/1/0 | 85/13/1/1 | 92/8/0/0 | 94/6/0/0 | 93/7/0/0 | 90/10/0/0 | 90/10/0/0 |
| Procheck G-factor[§], Φ&Ψ/All | 0.20/0.07 | 0.14/0.12 | -0.44/-0.31 | -0.31/-0.32 | -0.31/-0.20 | 0.17/0.19 | -0.09/-0.16 | -0.32/-0.32 | -0.34/-0.35 |
| MOLPROBITY clash score[§] | 20.89 | 19.20 | 12.73 | 29.01 | 19.80 | 14.65 | 16.64 | 11.2 | 20.44 |
| DP score[§], [%] | 72 | 78 | 80 | 70 | 85 | 81 | 80 | 76 | 71 |
| Expert time[‖] [days] | 15 | 15 | 17 | 12 | 15 | 20 | 14 | 25 | 35 |
| $RMSD_{bb}$[**] [Å] | 0.64 | 0.57 | 0.66 | 0.70 | 0.69 | 0.60 | 2.07 | 1.10 | 2.03[‡‡] |
| $RMSD_{all}$[††] [Å] | 1.29 | 1.14 | 1.18 | 1.42 | 1.27 | 1.40 | 2.34 | 1.81 | 2.85[‡‡] |

[*] Total number of residues. The expressed proteins except RpT7 and NeT4 contained a C-terminal tag with sequence LEHHHHHH, NeT4 contained a C-terminal tag with sequence GS from cloning artifact, those residues are not counted in this table. The first and last residue numbers of the experimentally determined secondary structures are listed in parenthesis; Long flexible loops: RhR95, residues 7-18, 32-42; NeT4, 28-51; PsR211, 20-36; AtR23, 7-22, 31-47. Molecular weights do not include residues in RCI-identified disordered tails: AtR23, residues 80-101; NeR45A, 1-13, 143-147.

[†] Number of residues in α-helix and β-strand.

[‡] 10 lowest energy models.

[§] RMSD to the mean coordinates for backbone ($RMSD_{bb}$) and all-non-hydrogen ($RMSD_{all}$) atoms. Ordered regions as reported by NESG for experimentally determined entries: RpT7, residues 2-22, 26-42, 49-63; StR82, 5-19, 23-56, 66-69; RhR95, 22-32, 40-58, 61-68; NeT4, 3-18, 21-31, 35-36, 48-66; TR80, 4-79; VfR117, 15-37, 42-43, 46-75; PsR211, 1-21, 32-33, 35-101; AtR23, 2-9, 22-30, 32-33, 36-38, 50-78, NeR45A, 15-22, 28-48, 51-69, 73-92, 98-110, 114-144. Locally ordered flexible loops and minor differences between disorder of terminal residues in experimental and CS-ROSETTA structures were identified manually and excluded for obtaining the actual ranges used in this comparison: RpT7, residues 2-22,26-42,50-63; StR82, 5-19, 23-56, 66-68; RhR95, 21-24, 28-31, 43-46, 53-55, 62-65, 66-68; NeT4, 3-8, 11-16, 20-23, 26-27, 52-62; TR80, 7-18, 20-26, 33-46, 50-63; VfR117, 18-28,31-36,41-43,49-57,60-75; PsR211, 2-4, 8-10, 15-19, 37-42, 50-58, 69-74, 80-100; AtR23, 2-7, 22-29, 49-78; NeR45A, 16-20, 28-40, 42-44, 52-65, 68-69, 73-74, 76-89, 94-96, 100-108, 114-117, 120-142. The Ramachandran plots, Procheck G-factors, MOLPROBITY clash scores and DP scores in this table are also restricted to these ranges.

[¶] Most-favored regions/additional allowed regions/generously allowed regions/disallowed regions.

[‖] Total time (±3 days) for the side-chain resonance assignments, NOESY assignment and structure calculations.

** RMSD ($C^{\alpha}$, C′ and N) of the mean coordinates of 10 lowest-energy models to the mean coordinates of the experimental NMR structure.
†† RMSD (all non-H atoms) of the mean coordinates of 10 lowest-energy models to the mean coordinates of the experimental structure.
‡‡ The residues ranges in the disorder tails obtained from the preliminary NMR structure were released to Y.S. prior to CS-ROSETTA structure prediction, but RCI analysis identified the same residues for the disordered tails. The predicted models for this protein were sent to NESG before the final stage of NMR structure refinement had been completed, but this refinement of the experimental structure was conducted independent of those CS-ROSETTA models.

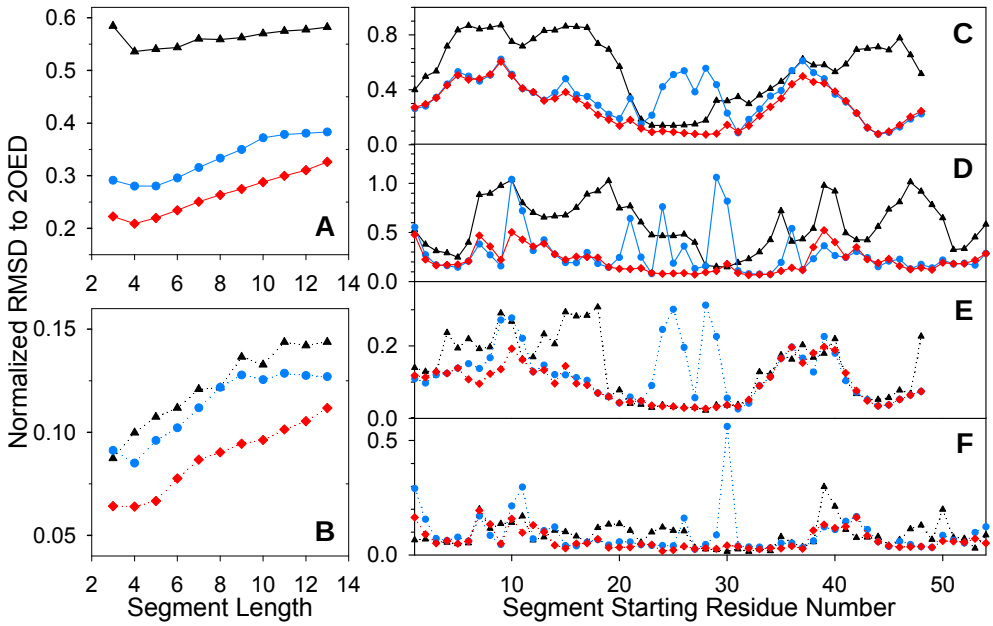


**Fig. 5.** Plots of fragment accuracy for GB3. For each specific GB3 segment, 200 fragment candidates were selected using either the standard ROSETTA procedure ("▲"), or from an MFR search of the 5665-protein structural database, assigned by the programs DC ("●") or SPARTA ("♦"). Like SPARTA, DC also can readily assign chemical shifts to a large database of protein structures, but the error in predicted chemical shift is on average slightly worse than for SHIFTX, and about 17% worse than SPARTA. For all panels, coordinate RMSDs (N, $C^{\alpha}$ and $C'$) between query segment and selected fragments are normalized with respect to randomly selected fragments (*i.e.*, the average RMSD between this target fragment and 1200 randomly selected fragments of the same length). The averaged RMSD of the 200 selected fragments is plotted as a solid line; dotted lines represents the lowest RMSD (best fragment out of 200). (A) Average and (B) lowest RMSD of 200 selected fragments, as a function of fragment size, relative to the NMR coordinates of the corresponding GB3 segment, averaged over all (overlapped) consecutive segments. (C, D) Average RMSD of 200 9-residue (C) and 3-residue (D) fragments relative to the X-ray coordinates, as a function of position in the GB3 sequence. (E, F) Lowest RMSD of any of these selected 9-residue (E) or 3-residue (F) fragments.

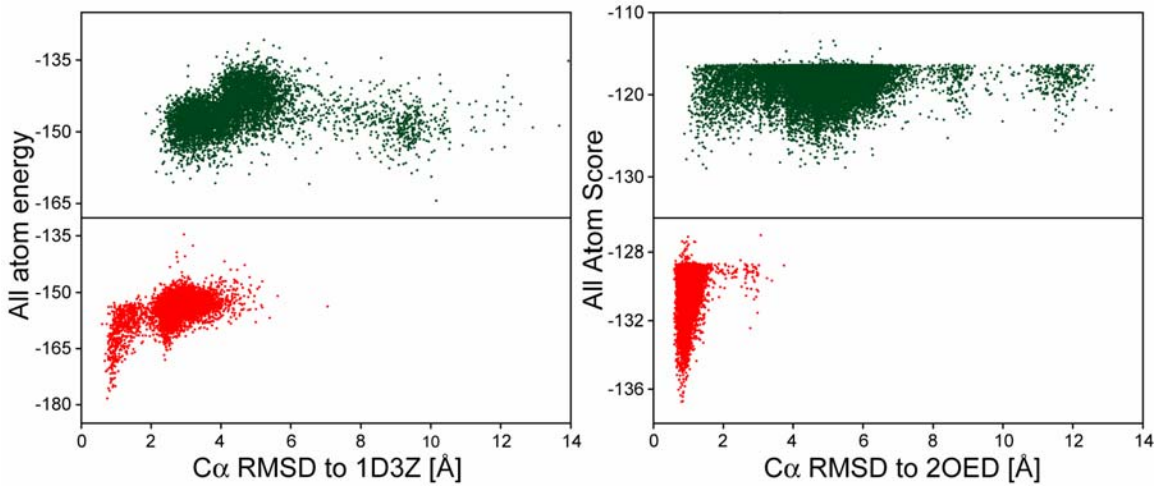


**Fig. 6.** Comparison of results obtained with standard ROSETTA and CS-ROSETTA for ubiquitin and GB3. All atom energy *versus* $C^{\alpha}$ RMSD of the ROSETTA models obtained using standard sequence based ROSETTA-selected fragments (top) and chemical shift based MFR-selected fragments (bottom) for ubiquitin (left) and GB3 (right). All-atom energies correspond to the raw ROSETTA energy score, prior to rescoring using experimental chemical shifts

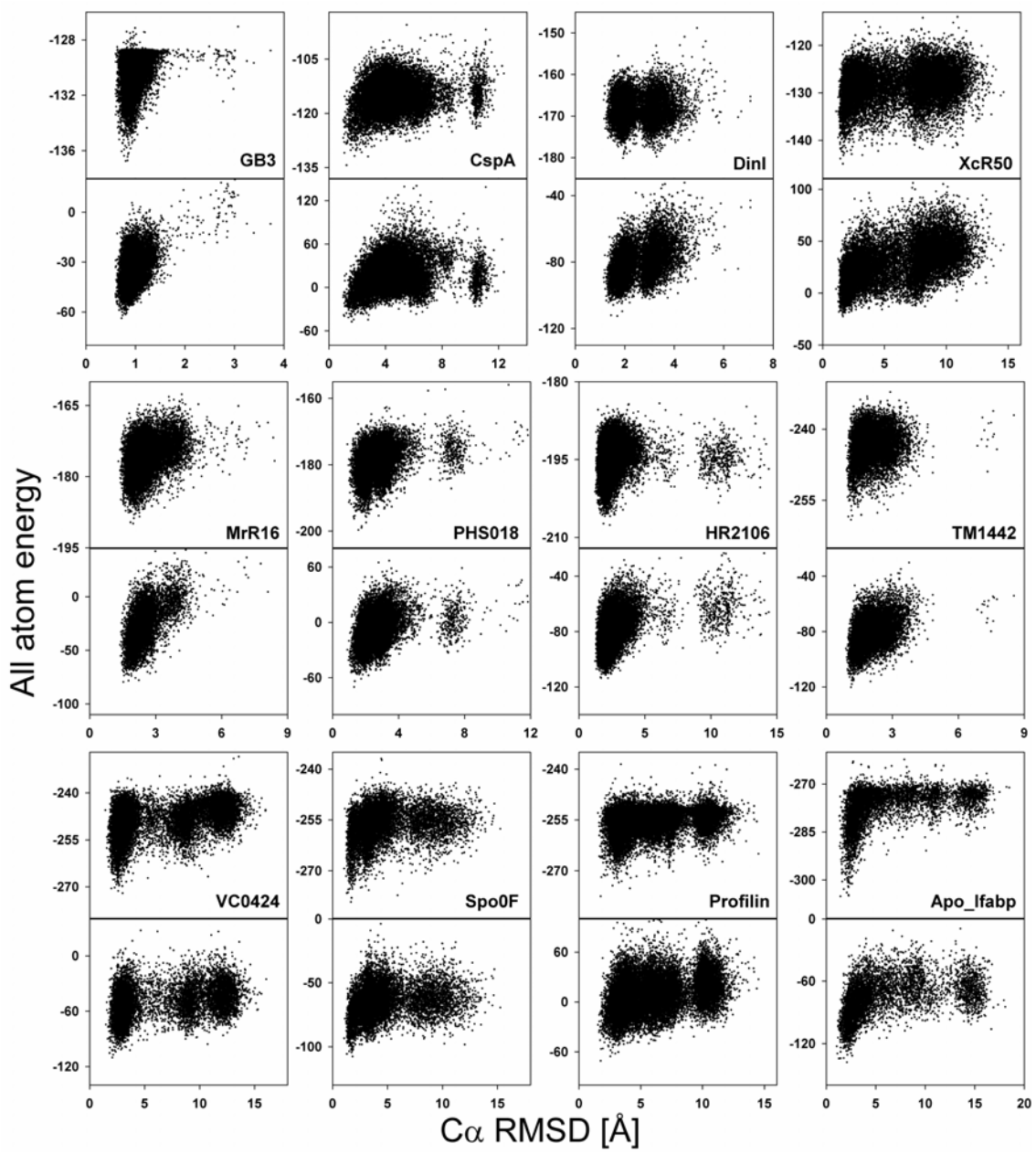


**Fig. 7.** Plots of ROSETTA all atom energy *versus* $C^{\alpha}$ RMSD relative to the experimental structures for proteins of Table 1, not presented in Figure 2. For each of these proteins, the upper plots show the standard ROSETTA all atom energy *versus* $C^{\alpha}$ RMSD from the experimental structures (see SI Table 3), and the lower plots show ROSETTA all atom energy rescored by using the experimental chemical shifts (*cf* eq 3).

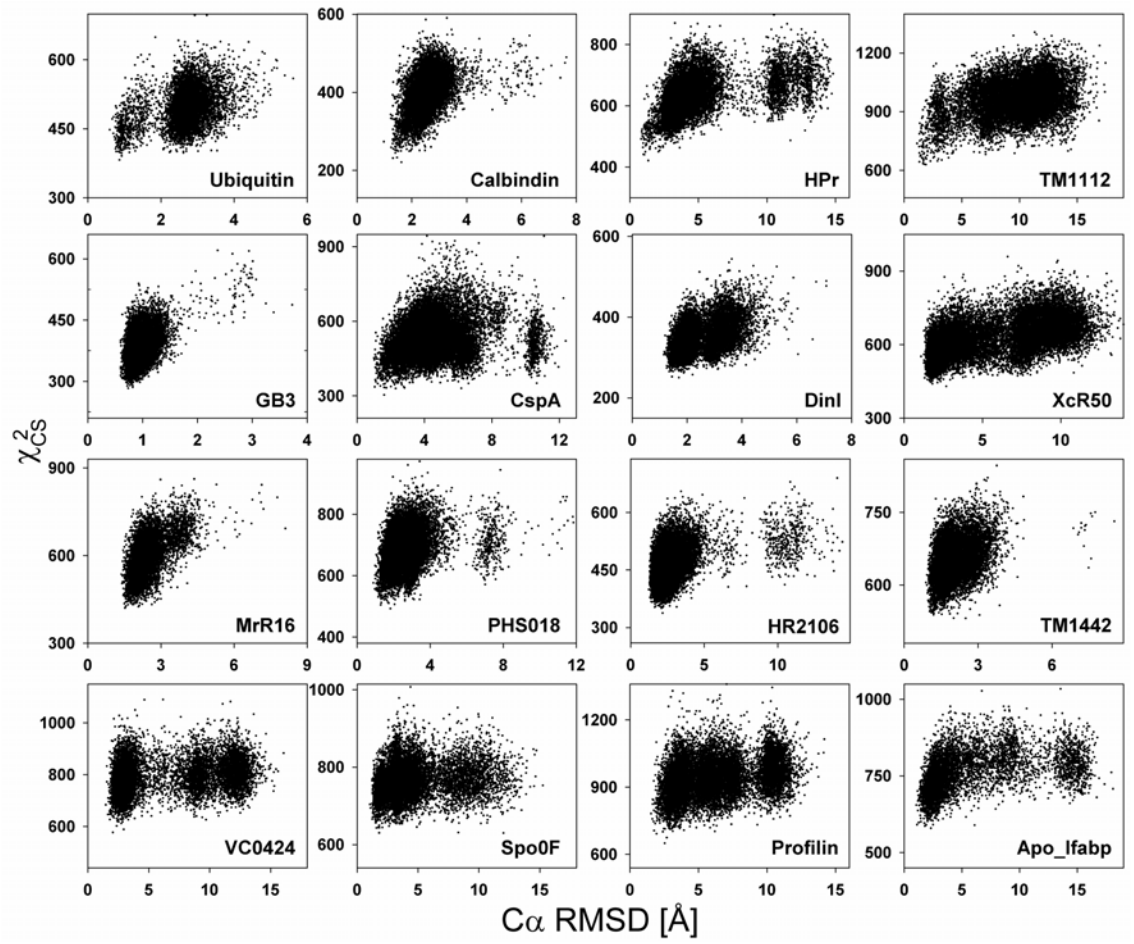


**Fig. 8.** Plot of $\chi^2_{cs}$ score (eq 3) of CS-ROSETTA models *versus* $C^{\alpha}$ RMSD relative to the experimental structures for proteins listed in Table 1.

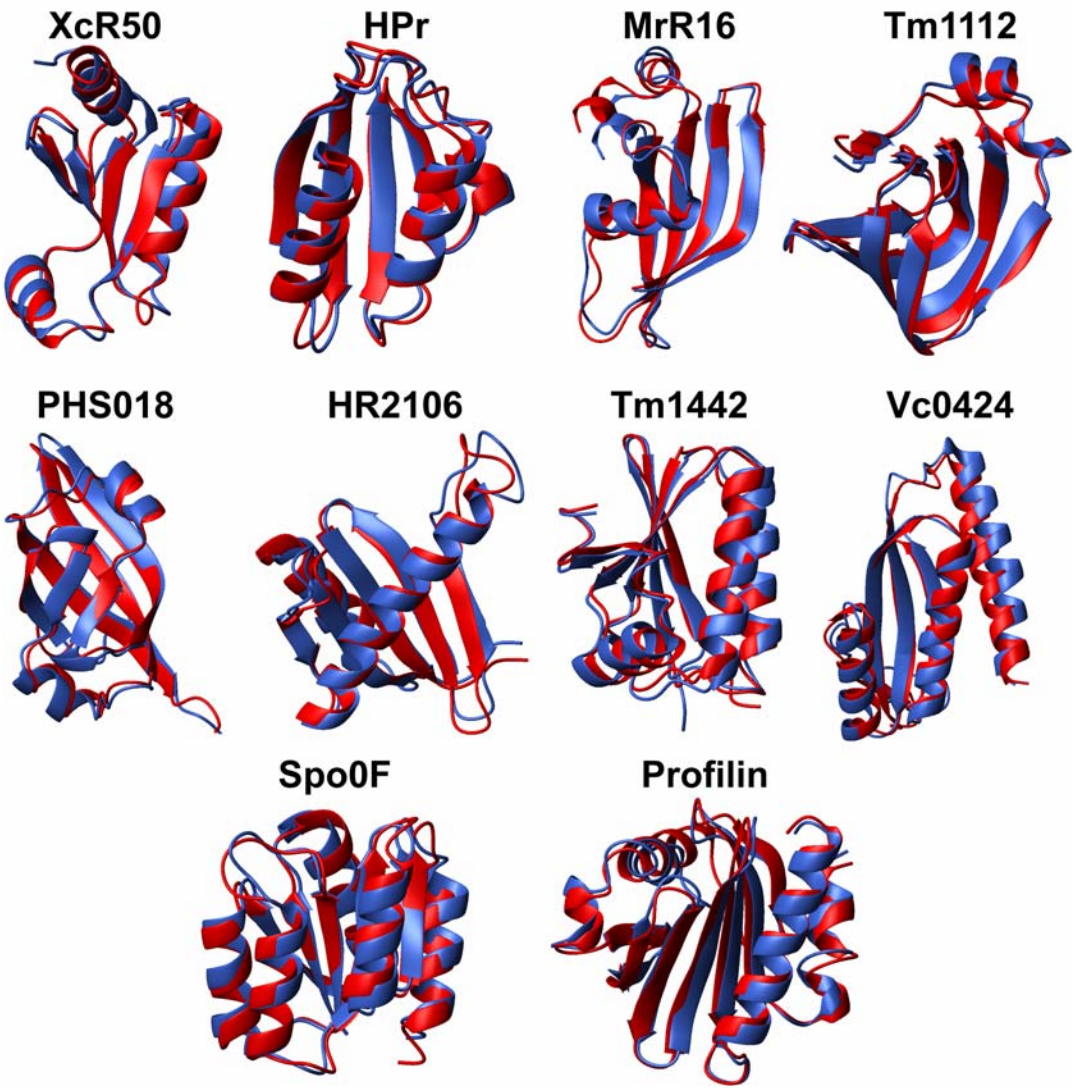


**Fig. 9.** Backbone ribbon representations of the lowest-energy CS-ROSETTA model (red), superimposed on the experimental Xray/NMR structures (blue) for the proteins listed in Table 1. Overlays of the 6 remaining structures are shown in Fig. 3.

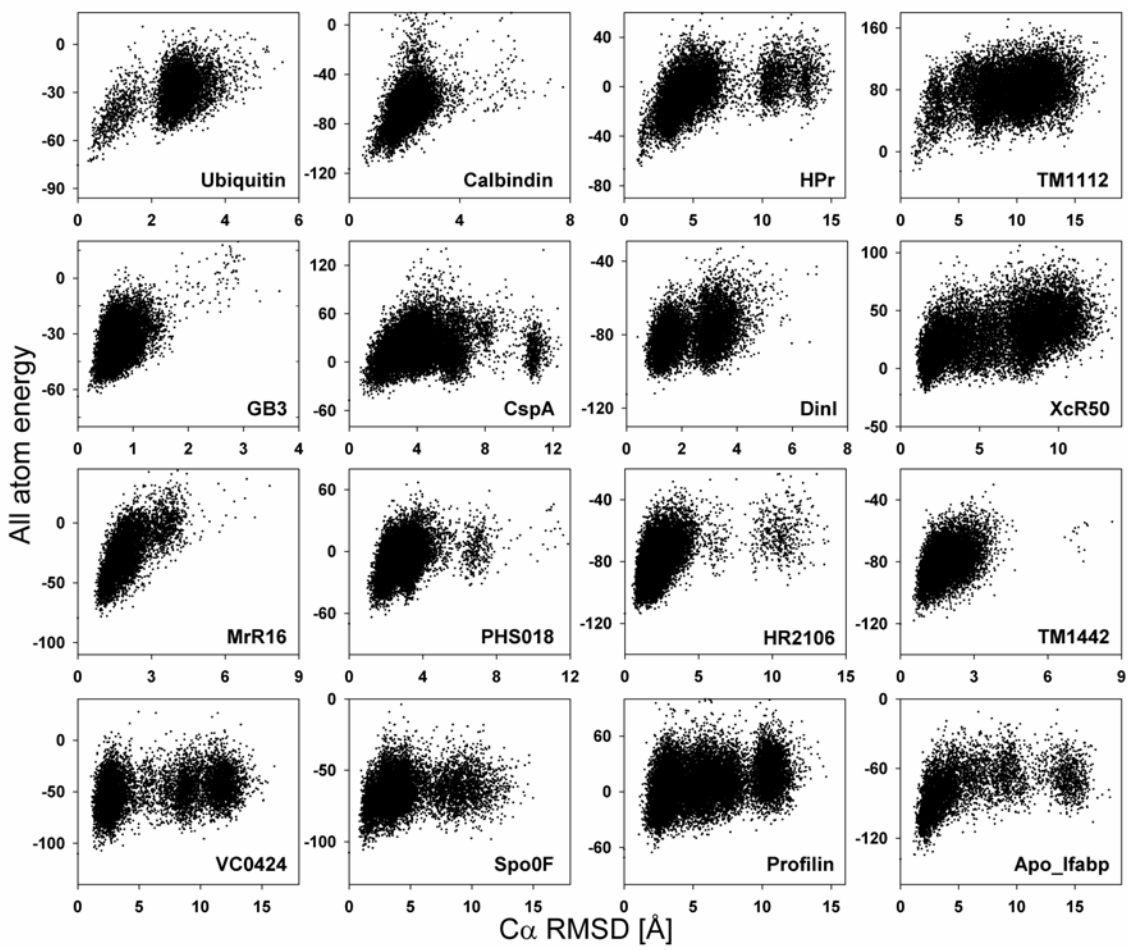


**Fig. 10.** Plots of ROSETTA all atom energy *versus* $C^{\alpha}$ RMSD of CS-ROSETTA models relative to the lowest-energy models for each of the 16 test proteins of Table 1.

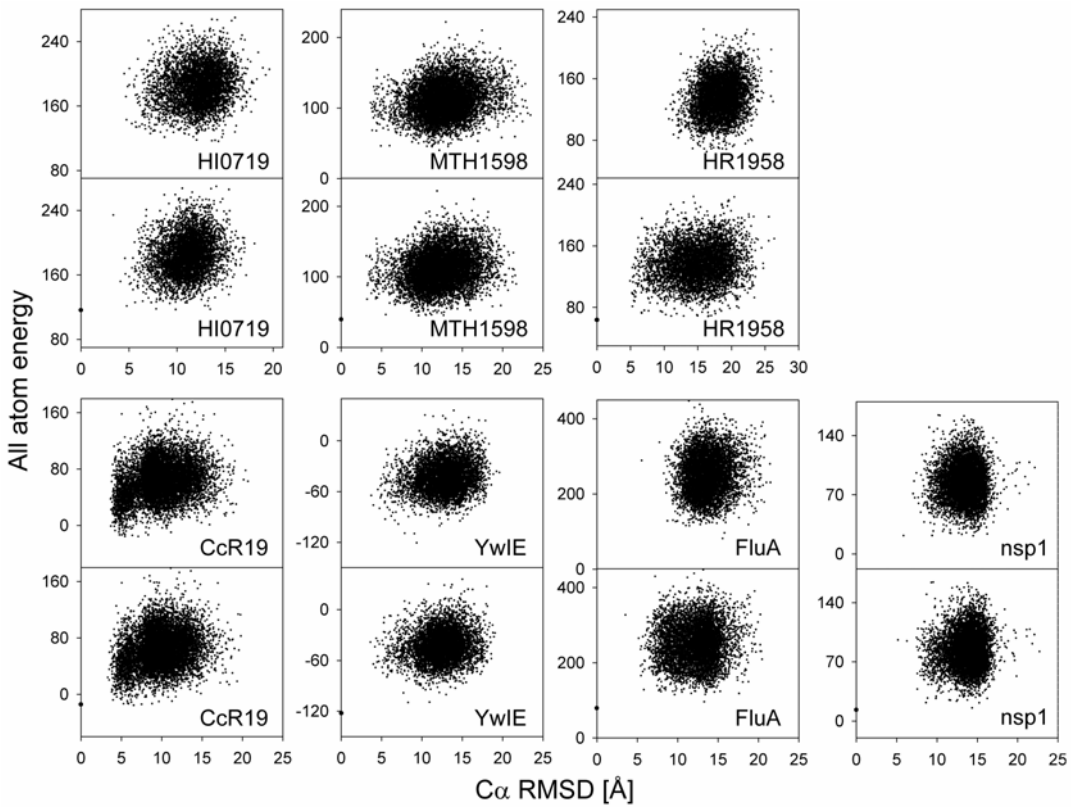


**Fig. 11.** Plots of ROSETTA all atom energy *versus* $C^{\alpha}$ RMSD of CS-ROSETTA models for the 7 proteins of SI Table 4. For each protein, the upper panel presents the chemical-shift-rescored ROSETTA all atom energy *versus* the $C^{\alpha}$ RMSD from the experimental structure; for the lower panels the $C^{\alpha}$ RMSD is calculated *versus* the coordinates of the lowest-energy model, whose energy is marked as a bold dot on the *y* axis. For nsp1 protein, the lowest-energy model is the only one out of 12,000 generated models that has the same topology as the experimental NMR structure, and even then it deviates considerably (backbone RMSD of 5.1 Å) from the experimental NMR structure.

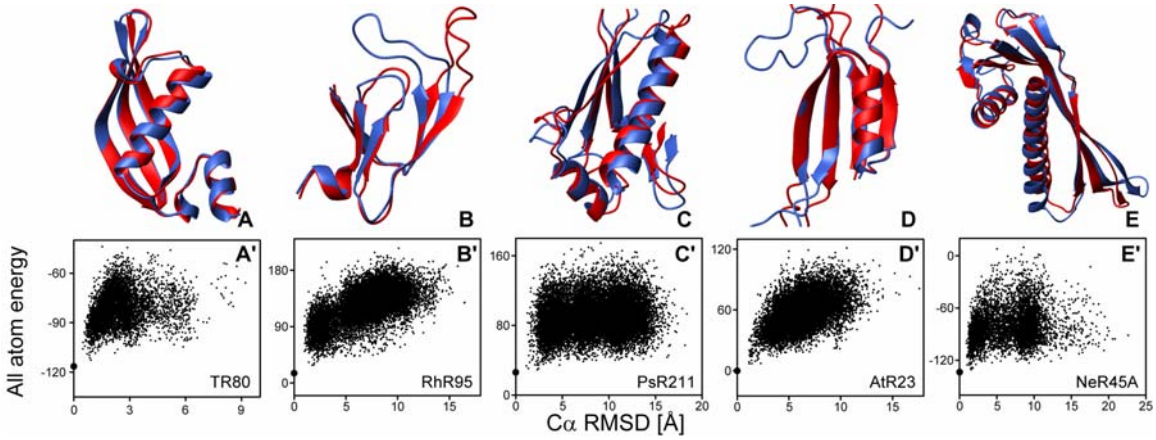


**Fig. 12.** CS-ROSETTA structures generated for five structural genomics targets (Table 2). The remaining four are shown in Fig. 4. (A - E) Superposition of lowest-energy CS-ROSETTA models (red) with experimental NMR structures (blue). (A′-E′) Plots of rescored (eq 3) ROSETTA all-atom energy *versus* $C^{\alpha}$ RMSD, calculated relative to the lowest-energy model (bold dot on *y* axis). (A, A′) TR80; (B, B′) RhR95; (C, C′) PsR211; (D, D′) AtR23; (E, E′) NeR45A.

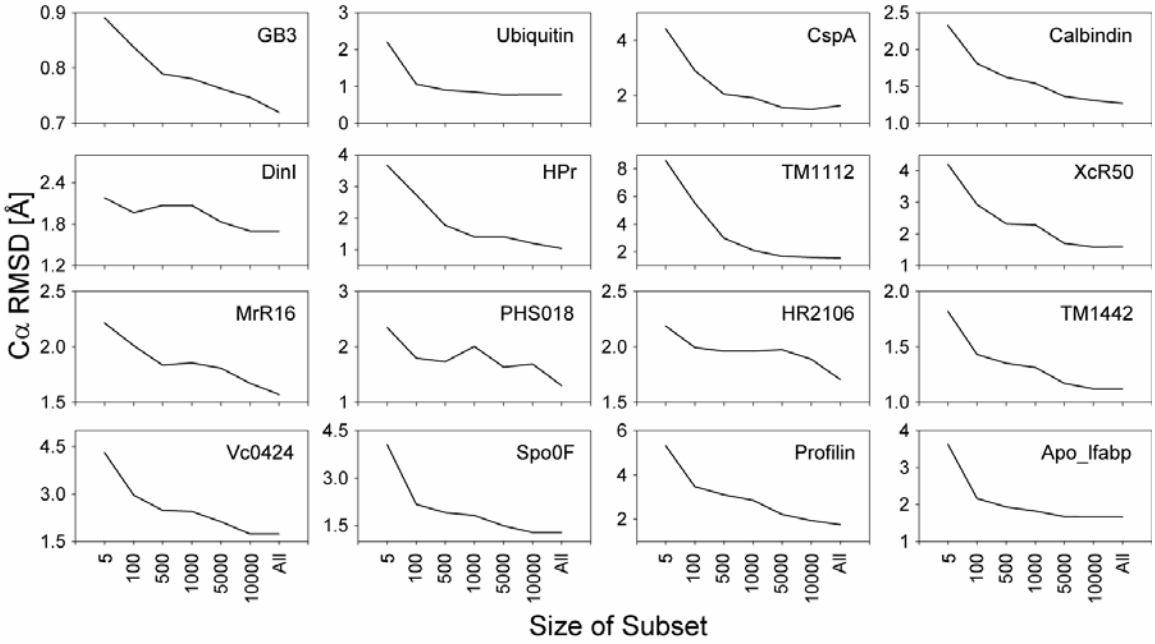


**Fig. 13.** Accuracy of the models in subsets randomly selected from the final ROSETTA all-atom models. For each protein (Table 1, SI Table 3), the $C^{\alpha}$ RMSD values (relative to the experimentally determined reference structure) of the lowest-energy models in 100 randomly selected 5-, 50-, 100-, 1000-, 5,000- and 10,000-sized subsets from the final ROSETTA all atom models were calculated and these averaged values are plotted against the size of the subsets. The figure shows that for 13 of the 16 proteins, generation of 5,000 ROSETTA full atom models suffices to yield a lowest-energy model that differs by ≤0.2 Å from the lowest-energy models obtained by using 10,000-20,000 ROSETTA predictions (Table 1).