Journal of Biomolecular NMR (2005) 31: 59–64

Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling

Jianhan Chen**, Hyung-Sik Won**, Wonpil Im, H. Jane Dyson & Charles L. Brooks III*
Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, LaJolla, CA 92037, U.S.A.

Received 23 August 2004; Accepted 28 October 2004

Key words: generalized born, implicit solvent, NOE assignment, replica exchange, structure determination, structure refinement

Abstract

Determining an accurate initial native-like protein fold is one of the most important and time-consuming steps of *de novo* NMR structure determination. Here we demonstrate that high-quality native-like models can be rapidly generated from initial structures obtained using limited NOE assignments, through replica exchange molecular dynamics refinement with a generalized Born implicit solvent (REX/GB). Conventional structure calculations using an initial sparse NOE set were unable to identify a unique topology for the zinc-bound C-terminal domain of *E. coli* chaperone Hsp33, due to a lack of unambiguous long range NOEs. An accurate overall topology was eventually obtained through laborious hand identification of long range NOEs. However we were able to obtain high-quality models with backbone RMSD values of about 2 Å with respect to the final structures, using REX/GB refinement with the original limited set of initial NOE restraints. These models could then be used to make further assignments of ambiguous NOEs and thereby speed up the structure determination process. The ability to calculate accurate starting structures from the limited unambiguous NOE set available at the beginning of a structure calculation offers the potential of a much more rapid and automated process for NMR structure determination.

Abbreviations: GBSW – generalized born with a simple switching function; NMR – nuclear magnetic resonance; NOE – nuclear overhauser effect; NOESY – nuclear overhauser enhanced spectroscopy; REX – replica exchange; RMS – root-mean-square; RMSD – root-mean-square deviation.

Determination of a three-dimensional protein structure by NMR spectroscopy relies primarily on distance restraints (Wüthrich, 1986), which are mainly derived from cross-peaks of two-dimensional or higher-dimensional NOESY spectra. For this, chemical shift values of all possible nuclei need to be assigned and are then used to identify nuclei that give rise to NOESY cross peaks. However, due to limited spectral resolution and chemical shift accuracy, only a limited number of NOESY cross peaks can be unambiguously assigned based on the chemical shift values alone. Further NOE assignment typically relies on a recursive process whereby preliminary structures computed from previous assignments are used to resolve ambiguous cross peaks and correct wrong assignments. Fully automated procedures such as CANDID (Herrman et al., 2002) and PASD (Kuszewski et al., 2004) have been developed in the last few years. These procedures

*To whom correspondence should be addressed. E-mail: brooks@scripps.edu
**Authors contributed equally to this work.

seem to be able to dramatically reduce the human intervention and have substantial tolerance of errors in NOESY peak-picking and chemical shift assignments. Nevertheless, interactive manual procedures seem to be still prominent in NMR structure determination (Herrmann et al., 2002), sometimes used in combination with semi-automated tools such as ARIA (Nilges et al., 1997) and SANE (Duggan et al., 2001).

A particularly difficult problem arises from assigning sufficient long-range NOEs to obtain a low-resolution initial fold. While information from primary sequence and secondary structure data can be used, in addition to chemical shift fitting, to resolve ambiguities for intra-residue, sequential and medium range NOEs, ambiguities in the long-range NOEs can often only be resolved with knowledge of the structure. When the initial assignments are not sufficient to define the protein fold, errors and ambiguities in the initial restraints have to be reduced by either human intuition or exhaustive automated computer evaluation. Once a reliable low-resolution fold is determined, iterative refinement procedures can be used effectively to obtain a high-resolution final structure (Nilges et al., 1997; Duggan et al., 2001). Furthermore, the accuracy and efficiency of a fully-automated procedure can be significantly improved when more correct manual assignments are incorporated and/or manually generated initial structures with a native-like fold are used as templates. Therefore, a reliable estimation of the native fold in the early stage of NMR structure determination can speed up the whole process dramatically, whether a manual or automated strategy is used.

Conventional methods for NMR structure calculation still rely completely on the experimental data to identify the protein fold, even though there have been considerable efforts in developing methods to predict the protein structure with sparse NMR data (Skolnick et al., 1997; Bowers et al., 2000; Li et al., 2003). With the continual improvement of molecular force fields, especially recent advances in generalized Born (GB) implicit solvent models (Still et al., 1990; Roux and Simonson, 1999; Bashford and Case, 2000; Lee et al., 2002; Im et al., 2003; Feig and Brooks, 2004), it is possible to generate more native-like folds using limited initial assignments by combining information from both experimental data and theoretical models. Recently we showed that replica exchange (REX) refinement in a GB implicit solvent model with a simple switching function (GBSW) (Im et al., 2003) can significantly improve the quality of structures and the radius of convergence when the experimental data is limited (Chen et al., 2004). The REX method (Sugita and Okamoto, 1999), also known as *parallel tempering*, is an advanced sampling technique that has been shown to offer generally better conformational sampling than simulated annealing procedures in applications such as protein folding and unfolding studies (Hansmann and Okamoto, 1999; Mitsutake et al., 2001). Multiple copies (replicas) of the system are simulated at different temperatures independently and simultaneously by conventional Monte Carlo (MC) or molecular dynamics (MD) methods. Pairs of replicas at neighboring temperatures attempt to exchange simulation temperatures according to a Metropolis type algorithm after a number of steps of MC or MD simulation. Replicas with lower potential energy tend to occupy the lower temperature conditions, while exchanging to higher temperature is highly probable even for replicas with lower energies compared with their higher temperature neighbors. In the course of an REX simulation, replicas can travel up and down the temperature space automatically in a self-regulated fashion, which, in turn, induces a non-trivial walk in temperature space. It has been demonstrated that initial structures that are poorly converged due to a lack of experimental data can be rapidly moved toward the native basin through the REX/GB refinement (Chen et al., 2004). In addition, an ensemble of most native-like structures can be automatically selected from a potentially diverse initial ensemble generated by conventional software such as CNS (Brunger et al., 1998) or DYANA (Güntert et al., 1997).

In this communication, we apply the REX/GB refinement protocol to structure calculations for the zinc-bound C-terminal domain of *E. coli* chaperone Hsp33 (residues 227–287). We demonstrate that our REX/GB refinement procedure can be effectively used to generate high-quality native-like models with limited experimental data at very early stages of NMR structure determination. Hsp33 belongs to a novel class of redox-regulated proteins and contains a C-terminal

Table 1. Summary of restraints for calculation of initial and final structures[f]

Distance[c]	Initial set			Final set		
	$d_{N*}(i,j)$[a]	$d_{CC}(i,j)$[b]	Total	$d_{N*}(i,j)$[a]	$d_{CC}(i,j)$[b]	Total
Intra	143	0	143	185	48	233
Seq.	202	0	200	275	27	302
Med.	140	0	140	181	59	240
Long	27	0	27	57	60	117
ZN[d]	–	–	6	–	–	6
Total	512	0	518	698	194	898
Dihedral[e]	ϕ	ψ	Total	ϕ	ψ	Total
	0	0	0	31	31	62

[a]Distance restraints from amide protons obtained from the ^{15}N-edited NOESY spectrum.
[b] Distance restraints between carbon-attached protons obtained from the ^{13}C-edited NOESY spectrum.
[c] Intra: $i = j$; seq.: $|i - j| = 1$; med.: $2 \leq |i - j| \leq 4$; long: $|i - j| > 4$.
[d]Distance restraints imposed to enforce the bound zinc coordination (Lee et al., 1989).
[e] Backbone dihedral angle restraints in helical and β-strand regions were derived from an HNHA experiment and from the TALOS program (Cornilescu et al., 1999).
[f] The restraint data sets are available upon request from the authors.

zinc-binding domain that modulates activity by a so-called 'redox switch' (Jakob et al., 1999). This domain becomes unfolded to activate the protein under conditions of oxidative stress, while in normal reducing environment it adopts a well folded structure and blocks the substrate-binding site (Graf et al., 2004). The solution structure of the redox-switch domain of *E. coli* Hsp33 in the reduced, zinc-bound state has been recently solved by NMR (Won et al., 2004). The structure has a novel fold that has not been previously documented. The conventional method was applied to determine the structure. Distance restraints were obtained from the 3D ^{15}N-edited and ^{13}C-edited NOESY spectra. About 1400 distance restraints were built up through many iterations of recursive processes, using combinations of manual and automatic NOE assignment. Several distance restraints were imposed to enforce the bound zinc coordination (Lee et al., 1989). All of the distance restraints were then summarized into 898 actual restraints that were responsible for the final structure calculation, by removing redundant and meaningless restraints. About 62 additional dihedral angle restraints were also used in the final structure calculation (see Table 1).

In the earliest stages of the calculation only a limited number of NOE distance restraints could be unambiguously assigned. The first round of assignments was performed on the ^{15}N-edited NOESY spectrum, as it generally provides better spectral quality than the ^{13}C-edited NOESY spectrum. Thus, the initial restraints included no NOEs between carbon-attached protons. A total of 512 NOE restraints were summarized from the 607 restraints initially assigned. Six additional distance restraints were imposed to enforce the bound zinc coordination (Lee et al., 1989). Table 1 summarizes the restraints used for the calculation and refinement of the initial structures, in comparison to the final restraints that were used to calculate the final NMR structures (Won et al., 2004). Note that the initial restraint set severely lacks long-range NOEs, which are required to define the tertiary structure. The number of long-range distance restraints in the initial set is less than a quarter of that in the final set. Moreover, as illustrated by the residue-residue NOE contact map (Figure 1b, left panel), all of the initially assigned long-range NOEs are localized within the anti-parallel β-sheet and between the zinc-coordinating residues. As a result, both CNS and DYANA calculations failed to identify a unique initial fold, generating ensembles with large structural uncertainties on the order of 6 Å while sufficiently satisfying all the distance restraints (see Table 2). Note that solvent effects are either ignored or greatly simplified in DYANA and CNS calculations: the non-bonded interactions are simply represented by repulsive soft-sphere

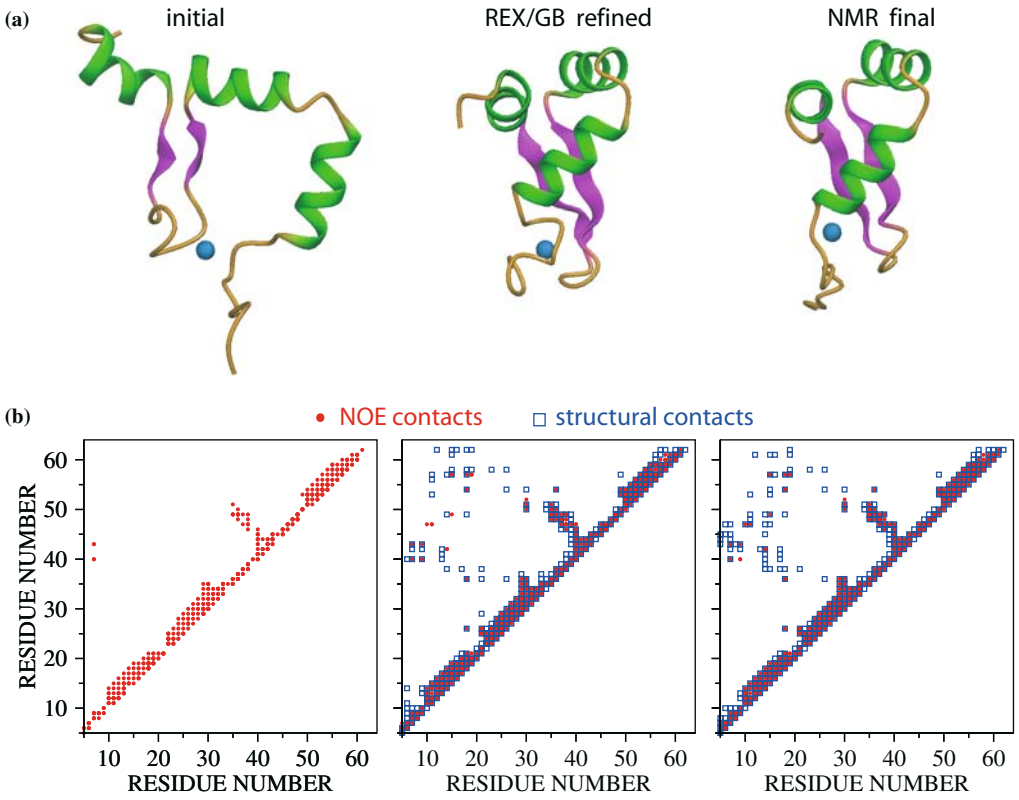


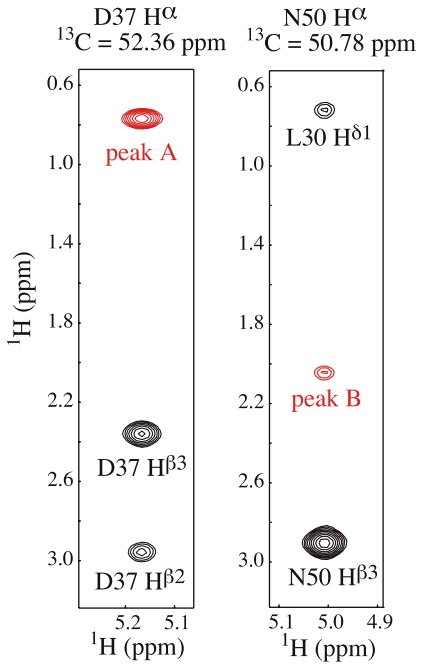
Figure 1. (a) Representative structures before and after the REX/GB refinement in comparison with the final NMR structure. The representative structures belong to a replica from Run I of Table 2 that has an occupancy 33% in the lowest temperature ensemble during the last 100 REX steps. The final NMR structure is an energy-minimized average structure of previously reported ensemble of 20 structures (Won et al., 2004). The backbone RMSD values (residues 7–60) with respect to the final structure are 10.1 Å (initial) and 2.8 Å (REX/GB refined). (b) NOE and structural contact maps. The left panel shows the NOEs that were initially assigned; the center and right panels show the final NOEs (red dots), overlaid with the structural contact maps (blue squares). The center structural contact map was computed from the average REX/GB refined structures and the right panel from the average final NMR structure. An NOE contact forms between two residues if there is one or more NOE restraint(s) defined between their atomic members. Two residues are in structural contact if the distance between their geometric centers is less than 8.5 Å.

Table 2. Comparison of statistics before and after the REX/GB refinement of the structures calculated using initial restraints

Runs	Before		After	
	RMSD[a]	NOE[b]	RMSD[a]	NOE[b]
I	8.8 ± 5.8	2.1/0.021	2.2 ± 2.7	4.4/0.020
II	9.6 ± 6.7	2.9/0.023	2.3 ± 2.9	5.6/0.024

[a] Backbone RMSD of the ensemble average from the final NMRstructure ± backbone RMS fluctuation around the average (in Å). Only structured regions (residues 7–60) were included in the RMSD calculations.
[b] Average number of NOE restraints violated by more than 0.2 Å/RMSD of NOE restraints for all structures in the ensemble (in Å).
The structures were first generated by CNS then refined by the REX/GB approach. Note that no NOE was violated by over 0.5 Å in any structure.

interactions in DYANA, and electrostatic interactions with a distance dependent dielectric constant and full Lennard–Jones potential are used for non-bonded interactions in the final cooling and energy minimization stages of CNS calculations. An example of such an initial structure is shown in Figure 1a. Low-quality initial models such as these are not very useful in further assignment of long-range NOEs. In contrast, when the same set of initial structures were refined by 400 steps of REX calculation in the GBSW implicit solvent, the quality of the models was dramatically improved, with backbone RMSD values and uncertainties reduced to about 2–3 Å. Each REX step involved 0.5 ps constant-temperature molecular dynamics, enabled by the Multiscale Modeling Tools in Structural Biology



Peak A at (5.17 ppm, 0.77 ppm)

Destination	δ (ppm)	$d_{REX/GB}$ (Å)	d_{Final} (Å)
L21 $H^{\delta1}$	0.78	8.0	7.6
L48 $H^{\delta2}$	0.77	4.3	3.2
L48 $H^{\delta1}$	0.77	6.6	3.3
I54 $H^{\gamma2}$	0.77	7.6	7.6
I57 $H^{\delta1}$	0.76	11.9	10.9

Peak B at (5.01 ppm, 2.05 ppm)

Destination	δ (ppm)	$d_{REX/GB}$ (Å)	d_{Final} (Å)
E4 H^{γ}	2.05	17.8	26.0
E25 $H^{\beta3}$	2.05	12.2	12.8
E32 $H^{\beta2}$	2.05	9.9	9.9
E35 $H^{\beta2}$	2.05	2.8	2.4
E35 H^{γ}	2.04	4.9	4.4

Figure 2. Selected strips from the ^{13}C-edited NOESY spectrum. Two cross peaks that could not be unambiguously assigned without the structural information are colored in red. The assignment candidates are shown in the tables. The most favorable assignments are highlighted in red font. $d_{REX/GB}$ was computed from average REX/GB refined structure and d_{Final} from the average final NMR structure. A $(\sum r^{-6})^{-1/6}$ summation was used for NOEs that involved more than two protons.

(MMTSB) tool set (available from http://mmtsb.scripps.edu) (Feig et al., 2001, 2004) and CHARMM (Brooks et al., 1983) with the PARAM22 all-hydrogen parameter set (MacKerell Jr. et al., 1998). The NOE restraint force constant was set to 10 kcal/(mol· Å^2) during the REX calculation and 75 kcal/(mol· Å^2) during minimization. Each refinement calculation took about 9 h on a cluster of 16 Intel XEON 2.4 GHz CPUs. Structures of replicas that contributed to the lowest temperature ensemble during the last 100 REX steps were collected and minimized to compute the statistical data shown in Table 2. A representative refined structure from one of the lowest temperature ensembles is shown in Figure 1(a), in comparison with the final NMR structure. Examination of the structural contact maps, shown in Figure 1(b), reveals good agreement between the average REX/GB refined structure and final NMR structure. It needs to be pointed that the current case represents an extremely well behaviored example of the REX/GB refinement approach, where the correct tertiary fold is almost completely predicted by the force field. However, even though the GB implicit solvent models have been shown to accurately characterize the solvent effects (Bashford and Case, 2000; Feig and Brooks III, 2004), *ab initio* structure prediction using physics based force fields is still limited to mini-proteins. Therefore, some reasonable amount of long-range distance restraints will generally be required for structural convergence and the exact amount of required data is system dependent. Nevertheless, when there is a severe lack of experimental data such that the REX/GB refinement fails, it will fail in a predictable way: multiple simulations will produce diverse structures, similar to the case of conventional NMR structure calculations.

We further examine how reliably the refined structures can be used to select the correct assignment candidates by carrying out assignment exercises. Figure 2 shows two cross peaks in the ^{13}C-edited NOESY spectrum that were initially difficult to assign due to chemical shift degeneracy and/or peak overlap during the actual assignment process. For each peak, there are five

possible candidates which have essentially degenerate chemical shift values within a spectral uncertainty of 0.01 ppm. Using the structural information from the REX/GB refined models, most likely candidates can be successfully selected and given top priority for the next round of structure calculation. The same selection would be made if the final NMR structure were used to resolve the ambiguities. Similar observations could be made when attempting to assign many other ambiguous NOESY cross peaks (data not shown), demonstrating that the REX/GB refined models could be reliably used to resolve ambiguous NOEs and speed up the whole structure determination process.

In conclusion, the present results suggest that REX/GB refinement of the initial structures generated from incomplete sets of NOE distant restraints by conventional tools such as CNS and DYANA is an effective way of obtaining highly native-like initial fold information with limited experimental data in the early stages of NMR structure determination. Improved characterization of solvent effects by GB implicit solvent is the crucial determinant in identifying the native folds. The REX method provides enhanced conformational sampling compared to conventional simulated annealing and automatic selection of an ensemble of optimal structures based on the average potential energies. The efficacy of the REX/GB approach relies on both components. By efficiently utilizing information from both experimental measurements and theoretical models, such an approach can provide better initial estimates of the protein fold, and could dramatically speed up the structure determination process, whether a manual, semi-automated or fully automated NOE assignment strategy is adopted.

Acknowledgements

CJH thanks a La Jolla Interface of Science postdoctoral fellowship for partial financial support. HSW acknowledges a grant from the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF). This work was supported by grants from the National Institutes of Health (GM48807, RR12255,CLB).

Supplemental Information Available

The supplementary information provides another control of the protocols presented here in which the force field used is a simple distance-dependent dielectric. This material is available free of charge via the Internet at http://kluweronline.com/issn/0925-2738

References

Bashfold, D. and Case, D.A. (2000) *Annu. Rev. Phys. Chem.*, **51,** 129–152.
Bowers, P.M. et al. (2000) *J. Biomol. NMR*, **18,** 311–318.
Brooks, B.R. et al. (1983) *J. Comput. Chem.*, **4,** 187–217.
Brunger, A.T. et al. (1998) *Acta Cryst.*, **D54,** 905–921.
Chen, J. et al. (2004) *J. Am. Chem. Soc.*. in press.
Cornilescu, G. et al. (1999) *J. Biomol. NMR*, **13,** 289–302.
Duggan, B.M. et al. (2001) *J. Biomol. NMR*, **19,** 321–329.
Feig, M. and Brooks III, C.L. (2004) *Curr. Opin. Struct. Biol.*, **14,** 217–224.
Feig, M. et al. (2001).
Feig, M. et al. (2004) *J. Comp. Graph. Modl.*, in press.
Graf, P.C.F. et al. (2004) *J. Biol. Chem.*, **19,** 20529–20538.
Güntert, P. et al. (1997) *J. Mol. Biol.*, **273,** 283–298.
Hansmann U.H.E. and Okamoto, Y. (1999) *Curr. Opin. Struct. Biol.*, **9,** 177–183.
Herrmann, T. et al. (2002) *J. Mol. Biol.*, **319,** 209–227.
Im, W. et al. (2003) *J. Comput. Chem.*, **24,** 1691–1702.
Jakob, U. et al. (1999) *Cell*, **96,** 341–352.
Kuszewski, J. et al. (2004) *J. Am. Chem. Soc.*, **126,** 6258–6273.
Lee, M.S. et al. (1989) *Science*, **245,** 635–637.
Lee, M.S. et al. (2002) *J. Chem. Phys.*, **116,** 10606–10614.
Li, W. et al. (2003) *Proteins*, **53,** 290–306.
MacKerell, Jr., A.D. et al. (1998) *J. Phys. Chem. B*, **102,** 3586–3616.
Mitsutake, A. et al. (2001) *Biopolymers*, **60,** 96–123.
Nilges, M. et al. (1997) *J. Mol. Biol.*, **269,** 408–422.
Roux, B. and Simonson, T. (1999) *Biophys. Chem.*, **78,** 1–20.
Skolnick, J. et al. (1997) *J. Mol. Biol.*, **265**, 217–241.
Still, W.C. et al. (1990) *J. Am. Chem. Soc.*, **112,** 6127–6129.
Sugita, Y. and Okamoto, Y. (1999) *Chem. Phys. Lett.*, **314,** 141–151.
Won, H.-S. et al. (2004) *J. Mol. Biol.*, **341,** 893–899.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.