UC San Diego
UC San Diego Previously Published Works

Title
A practical implicit solvent potential for NMR structure calculation

Permalink

Authors
Tian, Ye
Schwieters, Charles D
Opella, Stanley J
et al.

Publication Date

DOI

Peer reviewed

A Practical Implicit Solvent Potential for NMR Structure Calculation

Ye Tian[a,b], **Charles D. Schwieters**[c], **Stanley J. Opella**[b], and **Francesca M. Marassi**[a,*]

[a]Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

[b]Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0307, USA

[c]Division of Computational Bioscience, Building 12A, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624

Abstract

The benefits of protein structure refinement in water are well documented. However, performing structure refinement with explicit atomic representation of the solvent molecules is computationally expensive and impractical for NMR-restrained structure calculations that start from completely extended polypeptide templates. Here we describe a new implicit solvation potential, EEFx (Effective Energy Function for XPLOR-NIH), for NMR-restrained structure calculations of proteins in XPLOR-NIH. The key components of EEFx are an energy term for solvation energy that works together with other nonbonded energy functions, and a dedicated force field for conformational and nonbonded protein interaction parameters. The initial results obtained with EEFx show that significant improvements in structural quality can be obtained. EEFx is computationally efficient and can be used both to fold and refine structures. Overall, EEFx improves the quality of protein conformation and nonbonded atomic interactions. Moreover, such benefits are accompanied by enhanced structural precision and enhanced structural accuracy, reflected in improved agreement with the cross-validated dipolar coupling data. Finally, implementation of EEFx calculations is straightforward and computationally efficient. Overall, EEFx provides a useful method for the practical calculation of experimental protein structures in a physically realistic environment.

[*]**Corresponding Author: Francesca M. Marassi** Sanford-Burnham Medical Research Institute 10901 North Torrey Pines Road La Jolla, CA 92037, USA fmarassi@sbmri.org. Phone: 858-795-5282.

1. Introduction

The physical nature of a protein's physiological environment has long been recognized as an important factor governing molecular structure and biological function [1]. For example, the interactions of key amino acid residues with water and solvated molecules play key roles in protein activity and the highly anisotropic environment of the lipid bilayer membrane poses significant constraints on the structures and functions of membrane proteins.

Among the methods for three-dimensional molecular structure determination, the principal advantage of NMR spectroscopy is its ability to examine proteins in samples that are very close to their functional environments [2, 3]. Yet, typically, even when NMR spectra are measured for soluble proteins in aqueous solutions or for membrane proteins in lipid bilayers, structure calculations are carried out with energy functions that do not include contributions from solvation energy and, instead, represent all the nonbonded electrostatic and van der Waals interactions by a single, purely repulsive term. Such a simplified treatment is very useful because it enables fast, restrained molecular dynamics (MD) calculations of high quality structures from fully extended polypeptide templates by simulated annealing, to enhance sampling of conformational space and efficiently overcome the local-minimum problem [4, 5]. However, this approach can also lead to structures with suboptimal quality parameters, such as poor packing, unsatisfied hydrogen bond donors or acceptors and unbalanced salt bridges.

Performing structure refinement in a full MD force field, with explicit atomic representation of the solvent molecules (water and/or lipid), is one way to improve structural quality and obtain information about a protein's interactions with its surroundings, as shown for structure refinement of both soluble proteins [6–11] and membrane proteins [12–14]. However, this approach is computationally expensive due to the large amount of time that must be devoted to calculating solvent-solvent interactions and, hence, is not practical for ab-initio NMR structure calculations starting from completely extended polypeptide templates.

Methods in which solvent effects are treated implicitly have also been used in the refinement stages of NMR-restrained structure calculations. For example, refinement of initial structures using restrained MD calculations with generalized Born (GB) implicit solvent models have been shown to improve structural quality, particularly in cases where the experimental data are limited and the characterization of solvent effects is critical for identifying native fold [15–17]. However, GB methods have not been implemented as an integral parts of NMR structure calculation protocols from unfolded templates and remain too computationally intensive for routine calculations.

Various other models have been developed for the implicit treatment of solvent effects (reviewed in [18–22]). Of these, the effective energy function EEF1, developed by Lazaridis and Karplus [23], is particularly well suited to NMR applications for several reasons. EEF1 is based on the thermodynamic hypothesis that the native fold of a protein is the state of lowest free energy under physiological conditions and is determined by the amino acid sequence within the given solvent environment [1]. It contains terms for both intramolecular

energy and solvation free energy and has been shown both to provide a realistic first approximation to the effective energy hypersurface of proteins and to work well for protein folding-unfolding studies [23–25]. Importantly, it enables fast calculation of energies and energy derivatives, and its extension to an implicit membrane model (IMM) [26] enables its application to membrane proteins, thus providing access to the majority of protein families found in nature.

Here we describe the development and implementation of an implicit solvation model based on EEF1 for NMR-restrained structure calculations in the program XPLOR-NIH [27, 28]. The model is named EEFx (effective energy function for XPLOR-NIH) to highlight its origins. The principal components of EEFx are a new nonbonded energy function for solvation free energy (E_{slv}) and the related parameters that enable its implementation with the other XPLOR-NIH energy terms for NMR structure calculations. The XPLOR-NIH package is derived from XPLOR [29], which itself evolved from the CHARMM program [30, 31]. XPLOR-NIH has many completely new features, designed to facilitate its applications and continuous development for NMR structure calculations. However, it also contains all of the original XPLOR functionality, including many energy functions derived from CHARMM. This is a significant factor facilitating the implementation of EEFx in XPLOR-NIH, since its EEF1 progenitor was originally developed for CHARMM and works in conjunction with CHARMM energy functions and force fields.

We show that EEFx yields significant improvements in structural quality, accuracy and precision for NMR-restrained structure calculations from unfolded templates of several proteins with sizes ranging from 60 to 260 amino acids. Structure calculations with EEFx are computationally efficient and can be easily implemented together with standard simulated annealing protocols. We anticipate that structure calculations using EEFx with extension to an implicit membrane environment potential (in progress) will be particularly useful for membrane proteins where the highly anisotropic environment of the lipid bilayer membrane poses significant constraints on protein structure [32].

2. Description of EEFx

The XPLOR-NIH energy function (*ETOTAL*) can be grouped into three distinct classes [27–31]:

$$E_{TOTAL} = E_{EXP} + E_{KNOW} + E_{SYS} \quad (1)$$

E_{EXP} contains experimental restraining energy terms derived from the NMR data, E_{KNOW} contains knowledge-based restraining terms and E_{SYS}, describing the energy of the molecular system, contains conformational and nonbonded energy terms. Many E_{EXP} and E_{KNOW} potentials have been developed for NMR-restrained structure calculations in XPLOR-NIH, including the widely used terms for distance, dipolar coupling and dihedral angle restraints [27, 28], as well as various statistical torsion angle potentials [33, 34].

In typical NMR structure calculations the conformational energy comprises terms for covalent bonds (E_{BON}), covalent bond angles (E_{ANG}) and improper dihedral angles (E_{IMP}),

and can include a term for proper dihedral angles (E_{DIHE}), although this is usually more effectively replaced by a statistical knowledge-based term. Furthermore, the nonbonded energy is described collectively by a single repulsive potential, implemented by turning on the *repel* option of the van der Waals energy function (*EVDW-rpl*) and turning off the electrostatic energy function; this simplified term is used to prevent atomic overlap and can be scaled down to allow atoms to move through each other in the early stages of simulated annealing to accelerate the calculations [4, 5].

By contrast, the nonbonded energy function of EEFx contains terms for three types of interactions: a Lennard-Jones van der Waals term (E_{VDW}), to describe both repulsive and attractive forces; an electrostatic energy term (E_{ELEC}), computed with the atomic charges specified in the topology file; and a new term for solvation free energy (E_{slv}), introduced here to enable protein structure calculations in implicit solvent with EEFx.

E_{slv} is an empirical function for the solvation free energy of a protein in water, parametrized with experimental solvation free energy data for small model molecules. It works together with the XPLOR functions for nonbonded energy (E_{VDW} and E_{ELEC}) and conformational energy (E_{BON}, E_{ANG}, E_{IMP} and E_{DIHE}), plus a new set of protein topology and parameters to generate the EEFx force field, such that:

$$
\begin{aligned}
E_{SYS} &= F_{EEFx} \\
&= (k_{BON} E_{BON}) + (k_{ANG} E_{ANG}) + (k_{IMP} E_{IMP}) + (k_{DIHE} E_{DIHE}) \\
&\quad + (k_{VDW} E_{VDW}) + (k_{ELEC} E_{ELEC}) + (k_{SLV} E_{SLV})
\end{aligned} \quad (2)
$$

where E_{EEFx} is the effective energy of the solvated system and each energy term is scaled by its respective force constant k. The derivation of E_{slv} has been described [23] and its implementation in XPLOR-NIH is described below.

E_{slv} is defined as the sum of the solvation free energy contributions from all i atomic groups in the protein, each described as the solvation free energy of group i in its fully solvated state minus the reduction in solvation due to the presence of surrounding groups j. The functional form of E_{slv}, expressed as an empirical energy function of the protein's atomic coordinates is given by:

$$
\begin{aligned}
E_{SLV} &= \sum_i \Delta G_i^{slv} \\
&= \sum_i \Delta G_i^{ref} - \sum_i \sum_{j \neq i} SW\left(r_{ij}\right) \left(\frac{2\Delta G_i^{free}}{4\pi\sqrt{\pi}\lambda_i}\right) \exp\left[-\frac{(r_{ij}-R_i)^2}{\lambda_i^2}\right] \frac{V_j}{r_{ij}^2}
\end{aligned} \quad (3)
$$

where G_i^{ref} and G_i^{free} each represent the solvation free energy of atomic group i in its fully solvated state (G_i^{ref}) and in its isolated, fully solvated state (G_i^{free}), SW_{slv} is a switching function, r_{ij} is the distance between groups i and j, R_i is the van der Waals radius of group i, V_j is the volume of group j, and $_i$ is the correlation length of the solvation free energy density centered at group i. In equation (3), the solvation free energy density of each group is modeled as a Gaussian function, exhibiting strong distance-dependence with its maximum magnitude centered at group i and decaying to zero away from it. The switching function SW_{slv} is similar to the XPLOR switching function used to control the atomic distances at

which the nonbonded interaction terms for E_{VDW} and E_{ELEC} become effective [29]. SW_{slv} has the form:

$$SW_{slv}(r_{ij}) = \begin{array}{ll} 0 & if \quad r_{ij} > r_{off} \\ \dfrac{\left(r_{ij}^2 - r_{off}^2\right)^2 \left(r_{off}^2 + 2r_{ij}^2 - 3r_{on}^2\right)}{\left(r_{off}^2 - r_{on}^2\right)^3} & if \quad r_{off} > r_{ij} > r_{on} \\ 1 & if \quad r_{on} > r_{ij} \end{array} \quad (4)$$

Equations 3 and 4, as well as equations for evaluating the analytical derivatives of G_i^{slv} needed to generate gradients of the pairwise solvation free energies for all groups in the system, were coded in the C++ base framework of XPLOR-NIH, and new Python modules, *eefxPot* and *eefxPotTools*, were added to facilitate set up of the solvation energy term for EEFx calculations.

Evaluation of the solvation free energy using equations 3–4 requires specification of the following parameters for each group i in the protein: atom type, atom radius R_i, atom volume V_i, atom correlation length $_i$, and values of G_i^{ref} and G_i^{free}. These parameters, plus related values of heat capacity and enthalpy, were encoded in the *eefxPotTools* module and are called by *eefxPot* during structure calculations. Their values (Table 1) were selected to be the same as those of Lazaridis and Karplus [23], with the values for hydrogen atoms taken to be zero. The values of G_i^{ref} were taken from the experimental data of Privalov and Makhatadze [35] while those of G_i^{free} are derived from G_i^{ref} as described [23]. The values of R_i correspond to CHARMM19 atomic radii and the volumes V_i are derived from R_i, as described [23]. The values of $_i$ are set to the radius of the first hydration shell: 6.0 Å for NH3, NC2 and OC groups and 3.5 Å for all other groups.

Solvation free energy has a strong dependence on temperature. The CHARMM EEF1 model was found to reproduce thermodynamic parameters of protein folding-unfolding events very well and EEFx retains this thermodynamic functionality. The solvation free energy parameters encoded in *eefxPotTools* are values determined experimentally at 298.15 K [35]. For structure calculations based on NMR experiments performed at temperatures different than 298.15 K, *eefxPot* can perform temperature-dependent calibration of the free energy parameters in *eefxPotTools*, using values of heat capacity and enthalpy, also derived experimentally from model small molecules [35–38]. This is accomplished with the following functional forms of the Gibbs–Helmholtz equation, encoded in *eefxPot*:

$$\Delta G_i^{ref}(T) = \Delta G_i^{ref}(T_0) - \Delta S_i^{ref}(T_0)(T - T_0) - \Delta C p_i T \ln\left(\frac{T}{T_0}\right) + \Delta C p_i(T - T_0)$$

$$\Delta S_i^{ref}(T_0) = \frac{\Delta H_i^{ref}(T_0) - \Delta G_i^{ref}(T_0)}{T_0} \quad (5)$$

EEFx also requires dedicated protein topology and parameter files, containing the chemical information for specific residue and atom types and the various force constants for the conformational and nonbonded energy terms. The original EEF1 model was designed to work with the CHARMM19 polar hydrogen force field, in which aliphatic hydrogens are treated implicitly by representing aliphatic groups as unified atoms with increased mass and specific van der Waals properties. Since NMR structure calculations require explicit inclusion of all hydrogen atoms, we generated a new set of parameter and topology files for specific use with EEFx. The new files, proteinEEFx.par and proteinEEFx.top, were derived from the CHARMM19 [30, 31], PARALLHDG5.3 [6, 8] and OPLS [39] force fields. Since PARALLHDG5.3 is itself derived from CHARMM and since PARALLHDG5.3 and OPLS are jointly effective for explicit water refinement in XPLOR-NIH [6, 8], we reasoned that they would be good starting points for defining EEFx topology and parameters.

The proteinEEFx files were generated by making the following modifications to the amino acid parameters of PARALLHDG5.3: (i) the atom groupings were redefined to be those of CHARMM19 so as to be compatible with the solvation energy parameters in Table 1; (ii) for non-ionic residues, the partial atomic charges were replaced with those of CHARMM19 partial charges; (iii) for ionic residues (Arg, Lys, Asp, Glu and termini), the partial atomic charges were replaced with those of Lazaridis and Karplus [23], which were designed to obtain polar, albeit neutralized, residues that yield the proper stabilizing interactions for salt bridges. In addition, the force field also retains the full set of dihedral angle parameters defined in PARALLHDG5.3, thus enabling EEFx structure calculations to be performed with the XPLOR E_{DIHE} energy function instead of a statistical torsion angle potential, if desired.

Finally, to obtain effective EEFx calculations, E_{slv} is implemented together with the XPLOR van der Waals and electrostatics energy functions, E_{VDW} and E_{ELEC} [29], with their switching function effective between 7 Å and 9 Å; all nonbonded interactions beyond this value are neglected thus significantly reducing the computational cost. Dielectric screening, due to the influence of a protein's electrostatic properties on the density and distribution of the surrounding solvent molecules, is approximated by turning on the distance-dependent dielectric constant (= r) option of the E_{ELEC} function. This effect is neglected for nonpolar groups, where it is expected to be small.

3. Methods

3.1 Structure calculations

All calculations were performed with XPLOR-NIH [27, 28]. The new EEFx potential, eefxPot, is part of the XPLOR-NIH software suite (as of version 2.36), downloadable from the web (http://nmr.cit.nih.gov/xplor-nih/).

Free MD simulations were performed at 300 K, in Cartesian space, and implemented with four different models for nonbonded interactions (Table 2). The structures were downloaded from the PDB, energy minimized (500 steps of Powell minimization) and then subjected to 100 ps (and 1 ns in the case of SpAZ) of MD simulation, processed with normal atomic

masses instead of the uniform mass setup that is routinely used in NMR structure calculation protocols. The trajectories were saved every 200 steps.

NMR-restrained structure calculations were performed using two conventional simulated annealing protocols [34]: the first for folding an initially extended conformation and the second for subsequent refinement of a folded model selected from the first folding protocol. Both protocols are based on the internal variable module [40] and share the same basic scheme comprising four stages: (i) torsion angle dynamics at high-temperature (3,500 K for folding, 3,000 K for refinement) for a time of 15 ps or 15,000 timesteps; (ii) torsion angle dynamics with simulated annealing, where the temperature is reduced from the initial high temperature value to 25 K in steps of 12.5 K, for a time of 0.2 ps or 200 timesteps per temperature step (folding protocol), or a time of 0.63 ps or 630 timesteps per temperature step (refinement protocol); (iii) 500 steps of Powell torsion angle minimization; and (iv) 500 steps of Powell Cartesian minimization.

In the high temperature stage, experimental dihedral angle restraints and distance restraints were applied with respective force constants of k_{CDIH}=10 kcal mol^{-1} rad^{-2} and k_{DIST}=2 kcal mol^{-1} Å^{-2}. In the simulated annealing stage, k_{CDIH} was set to 200 kcal mol^{-1} rad^{-2} and k_{DIST} was increased geometrically from 2 to 30 kcal mol^{-1} Å^{-2}. In selected calculations, the torsionDB statistical torsion angle potential [34] was included with a force constant set to k_{tDB}=0.02 kcal mol^{-1} rad^{-2} in the high temperature stage and ramped geometrically from 0.02 to 2 kcal mol^{-1} rad^{-2} during simulated annealing. Atomic overlap was prevented by limiting allowed repulsions to those between atoms separated by three or more covalent bonds (*nbxmod*=5), except for calculations performed with torsionDB where allowed repulsions were limited to those between atoms separated by four or more covalent bonds (*nbxmod*=4).

Calculations were performed with either one of two different models for nonbonded interactions: REPEL or EEFx (Table 2). In the REPEL calculations, the simple repulsive van der Waals function [29] was used in conjunction with the default XPLOR-NIH protein topology and parameters (protein.top/par version 1.0). In the high temperature stage, only CA-CA atomic interactions were active, the van der Waals force constant was set to C_{rep}=0.004 kcal mol^{-1} Å^{-4} and the van der Waals radius scale factor was set to k_{rep}=1.2. In the simulated annealing stage, all atom-atom interactions were active, C_{rep} was ramped from 0.004 to 4 kcal mol^{-1} Å^{-4}, and k_{rep} was ramped down from 0.9 to 0.8.

The EEFx calculations were performed with the proteinEEFx topology and parameters. The van der Waals and electrostatic energy terms were implemented with a distance cutoff of 9 Å, a switching function for the Lennard-Jones potential between 7 and 9 Å, and distance-dependent dielectric. The new solvation energy term, E_{slv}, was implemented with a distance cutoff of 9 Å and the switching function for the potential between 7 and 9 Å. The force constants for van der Waals (k_{VDW}), electrostatic (k_{ELEC}) and solvation (k_{slv}) energy terms were each set to 1.

During the folding protocol of the EEFx calculations, the 15 ps high temperature stage was further divided into two equal parts, the first performed as described for the REPEL

calculations and the second performed with k_{VDW}, k_{ELEC} and k_{slv} set to 0.1. This was done to prevent fatal atomic overlap in the early stages of calculations from extended templates. In the subsequent simulated annealing stage, the force constants were ramped geometrically from 0.1 to 1 kcal mol^{-1}. The values of k_{VDW}, k_{ELEC} and k_{slv} were set to 1 throughout the refinement protocol of all EEFx calculations. For calculations of the largest protein EIN, the high temperature stage of the folding protocol was performed with REPEL and without EEFx, while EEFx was used during the annealing stage. Explicit water refinement was implemented as described previously [7–10], using the *wrefine.p*y script available in XPLOR-NIH.

3.2 Generation of partial NOE restraints

The complete data set of long-range distances (defined here as the 100% data set) included only distance restraints between atoms more than 4 residues apart in the protein sequence. Partial distance restraint data sets were generated by randomly selecting restraints from the 100% data set, to cover the percentage range from 1% to 100%. Five independent restraint sets of equal size were generated for each percentage value. Each set was used to fold 100 structures from extended templates, with the folding protocol described above, and the structure with lowest total energy was taken as input for the refinement protocol. Statistics were generated for the 10 structures with lowest energy from a total of 100 refined structures from each independent restraint set.

3.3 Structure analysis, validation and display

XPLOR-NIH was used to evaluate the precision and accuracy of the calculated structures, as well as to fit the experimental residual dipolar coupling (RDC) data to the calculated structures by singular value decomposition [41] and report the RMSD (root mean square deviation) measure of fit [42]. The backbone conformations, sidechain conformations and nonbonded atomic interactions of the calculated structures were assessed using the programs WHAT-IF [43, 44] and MolProbity [45–47]. Hydrogen bonds were computed in PyMol [48] using a script [49] with distance cutoff of 3.2 Å and angle cutoff of 55°. Structures were rendered with PyMol.

3.4 Generation of structural decoys

Blind structure predictions were performed using Rosetta, starting from the sequences of the proteins GB1 and BAF, excluding the PDB structural coordinates of the two proteins as well as their homologues from the structure prediction database. For each protein, 5,000 coarse-grained structural models were generated and then refined by all-atom relaxation, performed with the implicit aqueous environment protocol of Rosetta3.4. The refined all-atom structures were clustered according to their overall energy and their backbone CA atom RMSD to the lowest energy structure with a cutoff of 5 Å. For each protein, the most populated cluster encompassed more than 20% of the entire sampling space and contained 3554 decoys for GB1 and 1225 decoys for BAF. These decoys were subjected to two sets of Powell energy minimization (500 steps) in XPLOR-NIH, and then scored by calculating the total XPLOR-NIH energy with EEFx.

4. Results

4.1 Unrestrained molecular dynamics simulations

We first tested EEFx for its ability to model a physically realistic environment that can sustain stable, native protein structures. We performed unrestrained MD simulations of ten proteins at room temperature and examined the deviations of the resulting coordinates from those of the experimentally determined structures, taken to be representative of the native state. For each protein, we compared the results of 100 ps MD simulations performed with four different force fields (REPEL, VDW, vacuum, EEFx), each defined by a specific set of nonbonded energy function, topology and parameters for the system (Table 2). The proteins selected for analysis (Table 3) have sizes ranging from about 60 to 260 amino acids and a variety of structures, all determined by NMR spectroscopy, with coordinates and experimental restraints publicly available in the protein data bank (PDB). Three of these proteins (GB1, Eglin-c and Ubiquitin) were part of the set used for the initial development of CHARMM EEF1 [23] and, hence, provide useful benchmarks for direct comparison of EEFx performance.

A longer, 1 ns, MD simulation, performed for one of the proteins (SpAZ) demonstrates that simulations with EEFx are highly stable, as are those with VDW and vacuum (Fig. 1A). This behavior is similar to previous observations for CHARMM simulations with EEF1 [23]. Most of the changes in protein conformation occur within the first 30–40 ps of dynamics, indicating that 100 ps simulations are sufficient to compare the effects of the different force fields on protein structure. By contrast, as expected, the simple repulsive potential alone (REPEL) is unable to maintain the native structure in the absence of additional physical forces or experimental restraints, and the fold quickly and continuously unravels reaching RMSD values near 20 Å at 1 ns (Fig. 1A, black).

The data show that EEFx effectively maintains the native protein structure for the entire duration of dynamics (Fig. 1A, red). Furthermore, although simulations in vacuum (Fig. 1A, blue) or with VDW alone (Fig. 1A, gray) are also stable, EEFx yields a structure that is substantially closer to native (RMSD~1 Å; Fig. 1B), while the structures from vacuum (RMSD>2 Å; Fig. 1C) or VDW (RMSD>3 Å; Fig. 1D) simulations both differ significantly from the native state. Notably, both the vacuum and EEFx simulations were performed with distance-dependent dielectric screening (= r), as described above. By contrast comparisons in the original development of EEF1 were made to the vacuum force field with fixed unity dielectric constant (= 1) because this is the accepted standard in the field and because it corresponds to an extreme of zero dielectric screening. Thus, even though the use of = r and ionic sidechain neutralization (implemented in proteinEEFx.top) are expected to improve the results of vacuum simulations, EEFx still gives a significantly better result compared to vacuum.

The improved performance of EEFx compared to both vacuum and VDW, is also observed in the 100 ps simulations of the other test proteins (Fig. 2). Notable improvements in structural accuracy (Fig. 2A, B) are observed for all cases. Furthermore, for all proteins, simulations with EEFx yield significantly better cross validation with the experimental RDC data (Fig. 2C), providing independent evidence that EEFx produces close representations of

the native structures. For all proteins tested, EEFx relieves the notable molecular contraction that is observed in the vacuum simulations, as evidenced by the higher gyration radii of the EEFx structures compared to vacuum (Fig. 2D). Contraction is a well-known effect of MD simulations performed in vacuum, where electrostatic interactions are amplified by the lack of solvent screening [63, 64] and is readily visible in the 1 ns vacuum simulation of SpAZ (Fig. 1C).

Finally, simulations with EEFx yield a significant increase in the number of hydrogen bonds compared to the experimentally determined structures, further demonstrating that the model maintains stable secondary structural elements (Fig. 2E). For all of the test cases, the number of hydrogen bonds observed with EEFx is higher than the number observed for the experimental structures but lower than the number observed for vacuum simulations where the effects of electrostatics are not dampened by solvation screening, while VDW alone significantly decreases the number of hydrogen bonds concomitant with structural distortion. We conclude that EEFx provides a physically realistic implicit solvent environment capable of supporting stable, native protein structures.

4.2 Recognition of native fold

We next tested EEFx for its ability to discriminate among native and unfolded protein states. Protein structure prediction has become a major tool in structural biology that can be used very effectively to supplement sparse experimental restraints during protein structure determination by NMR [65–67]. Starting with the amino acid sequences of the proteins GB1 and BAF, we performed blind structure predictions using the Rosetta program, which is very successful at predicting three-dimensional structures of proteins from their amino acid sequences [68]. For each protein, we generated 5,000 coarse-grained decoys and then refined them by full-atom relaxation in Rosetta. The most populated clusters were then subjected to energy minimization in XPLOR-NIH and then scored with the EEFx energy function. Minimization produces only very minor alterations (maximum 0.2 Å) of the decoys's original structures.

Rosetta represents proteins by their backbone heavy atoms plus CB atoms for the sidechains; its full-atom energy function is a hybrid of statistical, empirical and physically realistic terms, including: PDB-derived sidechain and backbone torsion angle potentials, orientation-dependent hydrogen bonds, short-range knowledge-based electrostatic energy, reference energies for the unfolded states of the twenty amino acids, solvation effects based on the EEF1 solvation free energy function, and Lennard-Jones nonbonded interactions. By contrast, EEFx does not include any statistical terms.

Analysis of the Rosetta and EEFx energy landscapes (Fig. 3A, B, G, H) and comparison of the lowest energy structures with the NMR structures of either GB1 (Fig. 3C–F) or BAF (Fig. 3I–L), show that both the Rosetta and EEFx energy functions effectively recognize the overall, native fold of the two proteins. It is remarkable that these decoys were generated *de novo* with no other input than the protein's amino acid sequence and the PDB, which now contains a sufficient number of structures to enable fragment-based structure predictions. For each protein, the ten lowest energy decoys selected by either Rosetta or EEFx have very similar precision (Fig. 3E, F, K, L), indicating the tendency of each energy function towards

a specific structural ensemble. However, Rosetta and EEFx each select a different decoy based on lowest energy.

For GB1, the decoy with lowest Rosetta energy exhibits the overall features of the native fold, but is 3.6 Å RMSD away from the experimental structure (Fig. 3C). By contrast the decoy selected for lowest EEFx energy is very close to the experimental structure, with an RMSD of 1.2 Å (Fig. 3D). This difference in accuracy is also reflected in the shapes of the Rosetta and EEFx energy landscapes of GB1: while the EEFx energy landscape has a marked funneling shape towards the native structure (Fig. 3B), the Rosetta landscape has significantly less pronounced funneling features (Fig. 3A). In the case of BAF, the decoys with lowest Rosetta and EEFx energies also correspond to the decoys with lowest RMSD relative to the experimental structure (Fig. 3I–L) and both have Rosetta and EEFx energy landscapes with marked funneling shape towards the native fold (Fig. 3G, H).

For both proteins, the EEFx energy landscapes have significantly (six-fold) greater energy dispersion. The EEFx energy bandwidth is 300 kcal/mol, compared to the 50–60 kcal/mol observed for Rosetta and, thus, provides a greater degree of discrimination among protein folds. Analysis of the decoy EEFx energies shows that electrostatic energy makes the most significant contribution to the overall value. On average, the ten decoys of GB1 with lowest Rosetta energy have 83 hydrogen bonds, while those with lowest EEFx energy have 95 hydrogen bonds. Similarly for BAF, the ten decoys with lowest Rosetta energy have, on average, 142 hydrogen bonds, while those selected by EEFx have 163. In the lowest EEFx energy decoys, additional hydrogen bonds are formed both among backbone and sidechain atoms and contribute to lowering the total energy of the system. We conclude that the new EEFx term gives results comparable to Rosetta over a wide range of conformations and may provide a wider dynamic range for discrimination of folded states.

4.3 NMR-restrained protein structure calculations

EEFx was developed with the specific objective of providing a more physically realistic energy landscape for NMR-restrained structure calculations, without significantly sacrificing calculation speed and ease of implementation. To test its performance in this regard, we performed NMR-restrained calculations for six proteins in Table 3, using the experimental distance and dihedral angle restraints available in the PDB and retaining the RDC restraints only for cross validation. The calculations were started from extended templates, as is typically done in NMR structure determination, and performed with standard simulated annealing protocols, executed using either the simple repulsive function of the van der Waals energy term (REPEL) with the default XPLOR-NIH protein topology and parameters, or the EEFx energy function with proteinEEFx topology and parameters, each with or without inclusion of the statistical torsion angle potential torsionDB [34].

The restraints used for structure calculations necessarily reflect heterogeneity both in the way they were measured and evaluated from the experimental data and also in their number relative to protein length. For example, the interpretation of NOE signals in terms of inter-atomic distances can vary substantially among research groups, and the number of long-range NOE restraints for each protein in Table 3 varies between 1.8 and 10.6 per residue.

This situation reflects the typical range of variables associated with NMR structure calculations and hence provides a good test case for evaluating the performance of EEFx.

We first examined the ability of EEFx to produce folded protein structures with limited numbers of restraints. These tests were performed for two proteins, GB1 (Fig. 4) and ArfAB (Fig. S1), whose different sizes and distinct topologies make them excellent candidates for assessing the performance of EEFx. The NMR structure of GB1 is based on a large set of experimental restraints, including a complete set of NOEs, and is exceptionally well defined [50]. ArfAB is a larger polypeptide (131 residues) with an unusual fold whose structure was determined to very good precision [60].

To examine the dependence of the calculations on the number of distance restraints, we used data sets with decreasing numbers of distances. Each set was generated from experimental hydrogen bonds and NOEs by first, removing all distances between sites separated by less than five residues in the protein sequence, and then, randomly eliminating long-range distances from the remaining restraints. Thus, each resulting data point reflects the average over five independent structure calculations performed with five unique sets of long-range distances of equal size. For each set, 100 structures were calculated and statistics were generated for the 10 structures with lowest total energy. This approach reduces the bias associated with the inherent information content of the distance restraints, a factor that also influences structural quality [69].

Both EEFx and REPEL are capable of determining the correct global folds of GB1 and ArfAB with as few as 0.2–0.4 long-range distances per residue. However, in this case of very limited restraints EEFx produces structures that are significantly closer to the native fold and more precise than those calculated with REPEL. For GB1 (Fig. 4), calculations performed with REPEL and a partial data set containing only 4% of the long-range distance restraints (~0.2 restraints per residue) produce structures with an accuracy of 4.9 Å and a backbone precision of 4.4 Å. By contrast, structures calculated with EEFx, and the same partial data set, have an accuracy of 2.9 Å and a precision of 1.8 Å. Similarly for ArfA-B (Fig. S1), calculations with EEFx using only 20% of the long-range distance data (~0.5 restraints per residue) yield structures with better accuracy (2.9 Å) and precision (2.3 Å) while structures calculated with REPEL have both lower accuracy (3.6 Å) and precision (3.0 Å). When all available long-range distances are used (5.4 per residue in GB1, and 2.5 per residue in ArfAB), structures calculated with EEFx and REPEL have similar accuracy but the EEFx structure still have distinctly greater precision. Similar trends are observed for the precision and accuracy determined for all heavy atoms.

To the extent that structural accuracy can be assessed relative to the actual native structure, the precision of NMR structures is typically higher than their accuracy [70]. The number of restraints available for calculation is the principal factor influencing the accuracy and precision of NMR structures, but the nature of the nonbonded energy function also plays an important role [70]. The principal effect of EEFx is to direct the calculation towards the native structure even in the absence of large numbers of restraints. The ability of EEFx to fold structures with limited numbers of distance restraints correlates with its ability to bury solvent accessible protein groups, form hydrogen bonds and optimize the radius of gyration.

This is a significant advantage of EEFx, since modern methods for NMR structure determination are increasingly designed to shift the burden away from time-consuming measurements of multiple long-range distances and facilitate the determination of high-quality three-dimensional structures with very few or no distance restraints [65, 66].

We next tested the performance of EEFx for the generation of high quality structures. Introduction of a new term in the target energy function can induce deterioration in the agreement between calculated structures and the other experimental and conformational energy terms. The data in Figs. 5 and 6, obtained for six proteins with increasing sizes and an assortment of structures, demonstrate that the improvements in precision and accuracy afforded by EEFx are not accompanied by significant costs to either conformational terms or terms associated with the NMR data - on the contrary.

In all cases, the precision of both backbone and heavy atom coordinates improves significantly for structures calculated with EEFx, with the sole exception of IIBMt, where the precision decreases slightly (Fig. 5A, B). Furthermore, calculations with EEFx produce similar or improved agreement with the experimental RDC data (Fig. 5C), which were purposely excluded from structure calculations. RDCs depend on the orientation of interatomic vectors relative to the external magnetic field, and their exclusion from structure calculation provides a useful independent test of structural accuracy [42]. All of the structures calculated and refined with EEFx have better or similar agreement with the RDCs reflecting an improvement in accuracy.

Finally, calculations with EEFx produce similar levels of agreement between the structures and experimental distance and dihedral angle restraints used in the calculations (Fig. 5D, E). Although, in some cases a slight deterioration is observed when EEFx is used, the combined use of EEFx with the statistical potential torsionDB [34] produces results with similar or better agreement than those obtained with torsionDB alone. In the case of ubiquitin, EEFx actually produces a slight improvement in the agreement between structure and distance restraints.

The structures calculated with EEFx also compare very favorably with those refined in explicit water, using the *wrefine.py* refinement protocol adapted from refs. [7–10] and available in XPLOR-NIH. Correlations to the experimental data are similar for both EEFx and water-refined structures, while structural precision is somewhat better for EEFx (Fig. S2).

We further examined the quality of structures generated with EEFx with respect to WHAT-IF [43] and MolProbity [46, 47] validation metrics (Fig. 6). The results show that EEFx improves the quality of the backbone conformation in every case compared to results obtained with REPEL, regardless of whether torsionDB is included or not. Use of EEFx improves the WHAT-IF Ramachandran plot appearance (Fig. 6A). Similarly, MolProbity indicates that EEFx causes the favored regions of the Ramachandran plot to become more populated (Fig. 6D) and the percent of Ramachandran outliers to drop significantly (Fig. 6E). With regards to sidechain conformation, both WHAT-IF and MolProbity show that EEFx alone results in worse 1/ 2 rotamer normality scores (Fig. 6C) and higher numbers of

poor rotamers (Fig. 6F) for all proteins. This is expected for calculations performed without any dihedral angle potential term and is also observed for calculations performed with REPEL alone. However, these effects are readily corrected by the use of torsionDB [34], which was developed precisely for this purpose, or by inclusion of the XPLOR dihedral angle conformation energy term (E_{DIHE}) in the calculations (Fig. S3), which is enabled by the more complete force field available in the proteinEEFx.top/par files that work with EEFx.

The validation results further show that EEFx improves the quality of protein conformation and nonbonded atomic interactions. The WHAT-IF packing quality score (the atomic distributions around different molecular fragments) [71] and the MolProbity clashscore (the number of serious atomic overlaps per thousand atoms) [72] provide estimates of the quality of nonbonded atomic interactions or atomic packing. Notably, all structures generated with EEFx display marked improvements in both of these key metrics (Fig. 6B, G), even when compared with water-refined structures (Fig. S3). This is reflected in the overall MolProbity score [46, 47] (the lower the better), which improves with EEFx for every protein tested (Fig. 6H). Generally, NMR structures tend to be somewhat less well packed and expanded relative to X-ray structures [73, 74] and often the experimental NMR data are more consistent with high-resolution crystal structures than the corresponding NMR structures [75]. Indeed, the improved packing obtained with EEFx is also reflected in the improved agreement with the experimental RDC data.

Overall the best results are obtained when EEFx is used in conjunction with TorsionDB. However, the use of EEFx with the E_{DIHE} energy term also yields very favorable results (Fig. S3), thus providing a non-statistical, albeit empirical, alternative to the use of a statistical knowledge-based potential (torsionDB) for dihedral angles.

Finally, we report that calculations with EEFx are computationally efficient. For the proteins tested in this study, NMR-restrained calculations performed with EEFx were only 2.5 times longer in elapsed wall clock time than those with REPEL.

5. Conclusions

The benefits of protein structure refinement in water are well documented [6–11, 15]. However, performing structure calculations with explicit atomic representation of the solvent molecules is computationally expensive and impractical for NMR-restrained structure determination. We conclude that the new EEFx potential described in this paper provides an effective energy function for the implicit solvation of proteins during NMR-restrained calculations.

The initial results show EEFx outperforms the simple repulsive potential that is typically used in NMR structure calculations. The EEFx energy function effectively discriminates native from misfolded conformations and yields significant improvements in structural precision and accuracy, as well as conformational and nonbonded protein packing properties. Notably, EEFx can be used both to fold as well as refine NMR-restrained structures and improves the precision and accuracy of structure calculations performed with

limited numbers of experimental distance restraints. Finally, implementation of EEFx in XPLOR-NIH is straightforward and computationally efficient enabling structure calculations to be easily carried out on standard laboratory computers.

Additional studies on different proteins will be needed to fully explore the XPLOR-NIH EEFx energy landscape. However these initial results indicate that EEFx is a useful step forward towards the practical calculation of experimental protein structures in a physically realistic environment that closely resembles their native state.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

[1]. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973; 181:223–230. [PubMed: 4124164]

[2]. Banci L, Bertini I, Luchinat C, Mori M. NMR in structural proteomics and beyond. Prog. Nucl. Magn. Reson. Spectrosc. 2010; 56:247–266. [PubMed: 20633354]

[3]. Zhou HX, Cross TA. Influences of membrane mimetic environments on membrane protein structures. Annu Rev Biophys. 2013; 42:361–392. [PubMed: 23451886]

[4]. Nilges M, Gronenborn AM, Brunger AT, Clore GM. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. Protein Eng. 1988; 2:27–38. [PubMed: 2855369]

[5]. Clore GM, Gronenborn AM. Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy. Crit. Rev. Biochem. Mol. Biol. 1989; 24:479–564. [PubMed: 2676353]

[6]. Linge JP, Nilges M. Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. J. Biomol. NMR. 1999; 13:51–59. [PubMed: 10905826]

[7]. Spronk CA, Linge JP, Hilbers CW, Vuister GW. Improving the quality of protein structures derived by NMR spectroscopy. J. Biomol. NMR. 2002; 22:281–289. [PubMed: 11991356]

[8]. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. Refinement of protein structures in explicit solvent. Proteins. 2003; 50:496–506. [PubMed: 12557191]

[9]. Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AM, Vuister GW, Vriend G, Spronk CA. DRESS: a database of REfined solution NMR structures. Proteins. 2004; 55:483–486. [PubMed: 15103611]

[10]. Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CA, Nabuurs SB, Guntert P, Livny M, Markley JL, Nilges M, Ulrich EL, Kaptein R, Bonvin AM. RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins. 2005; 59:662–672. [PubMed: 15822098]

[11]. Bertini I, Case DA, Ferella L, Giachetti A, Rosato A. A Grid-enabled web portal for NMR structure refinement with AMBER. Bioinformatics. 2011; 27:2384–2390. [PubMed: 21757462]

[12]. Sharma M, Yi M, Dong H, Qin H, Peterson E, Busath DD, Zhou HX, Cross TA. Insight into the mechanism of the influenza A proton channel from a structure in a lipid bilayer. Science. 2010; 330:509–512. [PubMed: 20966252]

[13]. Cheng X, Im W. NMR observable-based structure refinement of DAP12-NKG2C activating immunoreceptor complex in explicit membranes. Biophys. J. 2012; 102:L27–29. [PubMed: 22500771]

[14]. Cheng X, Jo S, Marassi FM, Im W. NMR-based simulation studies of Pf1 coat protein in explicit membranes. Biophys. J. 2013; 105:691–698. [PubMed: 23931317]

[15]. Xia B, Tsui V, Case DA, Dyson HJ, Wright PE. Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water. J. Biomol. NMR. 2002; 22:317–331. [PubMed: 12018480]

[16]. Chen J, Im W, Brooks CL 3rd. Refinement of NMR structures using implicit solvent and advanced sampling techniques. J. Am. Chem. Soc. 2004; 126:16038–16047. [PubMed: 15584737]

[17]. Chen J, Won HS, Im W, Dyson HJ, Brooks CL 3rd. Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling. J. Biomol. NMR. 2005; 31:59–64. [PubMed: 15692739]

[18]. Roux B, Simonson T. Implicit solvent models. Biophys. Chem. 1999; 78:1–20. [PubMed: 17030302]

[19]. Feig M, Brooks CL 3rd. Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr. Opin. Struct. Biol. 2004; 14:217–224. [PubMed: 15093837]

[20]. Baker NA. Improving implicit solvent simulations: a Poisson-centric view. Curr. Opin. Struct. Biol. 2005; 15:137–143. [PubMed: 15837170]

[21]. Chen J, Brooks CL 3rd, Khandogin J. Recent advances in implicit solvent-based methods for biomolecular simulations. Curr. Opin. Struct. Biol. 2008; 18:140–148. [PubMed: 18304802]

[22]. Bashford D, Case DA. Generalized born models of macromolecular solvation effects. Annu. Rev. Phys. Chem. 2000; 51:129–152. [PubMed: 11031278]

[23]. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins. 1999; 35:133–152. [PubMed: 10223287]

[24]. Lazaridis T, Karplus M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. Science. 1997; 278:1928–1931. [PubMed: 9395391]

[25]. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J. Mol. Biol. 1999; 288:477–487. [PubMed: 10329155]

[26]. Lazaridis T. Effective energy function for proteins in lipid membranes. Proteins. 2003; 52:176–192. [PubMed: 12833542]

[27]. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. J. Magn. Reson. 2003; 160:65–73. [PubMed: 12565051]

[28]. Schwieters CD, Kuszewski JJ, Marius Clore G. Using Xplor‚ÄìNIH for NMR molecular structure determination. Prog. Nucl. Magn. Reson. Spectrosc. 2006; 48:47–62.

[29]. Brünger, AT. Version 3.1 : a system for X-ray crystallography and NMR. Yale University Press; New Haven: 1992. X-PLOR.

[30]. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 1983; 4:187–217.

[31]. Brooks BR, Brooks CL 3rd, Mackerell AD Jr. Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J. Comput. Chem. 2009; 30:1545–1614. [PubMed: 19444816]

[32]. Cross TA, Sharma M, Yi M, Zhou HX. Influence of solubilizing environments on membrane protein structures. Trends Biochem. Sci. 2011; 36:117–125. [PubMed: 20724162]

[33]. Kuszewski J, Gronenborn AM, Clore GM. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci. 1996; 5:1067–1080. [PubMed: 8762138]

[34]. Bermejo GA, Clore GM, Schwieters CD. Smooth statistical torsion angle potential derived from a large conformational database via adaptive kernel density estimation improves the quality of NMR protein structures. Protein Sci. 2012; 21:1824–1836. [PubMed: 23011872]

[35]. Privalov PL, Makhatadze GI. Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration. J. Mol. Biol. 1993; 232:660–679. [PubMed: 8393941]

[36]. Makhatadze GI, Privalov PL. Contribution of hydration to protein folding thermodynamics. I. The enthalpy of hydration. J. Mol. Biol. 1993; 232:639–659. [PubMed: 8393940]

[37]. Privalov PL, Makhatadze GI. Contribution of hydration and non-covalent interactions to the heat capacity effect on protein unfolding. J. Mol. Biol. 1992; 224:715–723. [PubMed: 1314903]

[38]. Privalov PL, Makhatadze GI. Heat capacity of proteins. II. Partial molar heat capacity of the unfolded polypeptide chain of proteins: protein unfolding effects. J. Mol. Biol. 1990; 213:385–391. [PubMed: 2160545]

[39]. Jorgensen WL, Tirado-Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J. Am. Chem. Soc. 1988; 110:1657–1666.

[40]. Schwieters CD, Clore GM. Internal coordinates for molecular dynamics and minimization in structure determination and refinement. J. Magn. Reson. 2001; 152:288–302. [PubMed: 11567582]

[41]. Losonczi JA, Andrec M, Fischer MW, Prestegard JH. Order matrix analysis of residual dipolar couplings using singular value decomposition. J. Magn. Reson. 1999; 138:334–342. [PubMed: 10341140]

[42]. Clore GM, Garrett DS. R-factor, Free R, and Complete Cross-Validation for Dipolar Coupling Refinement of NMR Structures. J. Am. Chem. Soc. 1999; 121:9008–9012.

[43]. Vriend G. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 1990; 8:52–56. [PubMed: 2268628]

[44]. Doreleijers JF, Sousa da Silva AW, Krieger E, Nabuurs SB, Spronk CA, Stevens TJ, Vranken WF, Vriend G, Vuister GW. CING: an integrated residue-based structure validation program suite. J. Biomol. NMR. 2012; 54:267–283. [PubMed: 22986687]

[45]. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003; 50:437–450. [PubMed: 12557186]

[46]. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB 3rd, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007; 35:W375–383. [PubMed: 17452350]

[47]. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D. Biol. Crystallogr. 2010; 66:12–21. [PubMed: 20057044]

[48]. DeLano, WL. PyMol. 2005. www.pymol.org

[49]. Campbell, R. Personal communication. 2013.

[50]. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science. 1991; 253:657–661. [PubMed: 1871600]

[51]. Kuszewski J, Gronenborn AM, Clore GM. Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. J. Am. Chem. Soc. 1999; 121:2337–2338.

[52]. Zheng D, Aramini JM, Montelione GT. Validation of helical tilt angles in the solution NMR structure of the Z domain of Staphylococcal protein A by combined analysis of residual dipolar coupling and NOE data. Protein Sci. 2004; 13:549–554. [PubMed: 14718654]

[53]. Kaieda S, Matsui C, Mimori-Kiyosue Y, Ikegami T. Structural basis of the recognition of the SAMP motif of adenomatous polyposis coli by the Src-homology 3 domain. Biochemistry. 2010; 49:5143–5153. [PubMed: 20509626]

[54]. Hyberts SG, Goldberg MS, Havel TF, Wagner G. The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci. 1992; 1:736–751. [PubMed: 1304915]

[55]. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. J. Am. Chem. Soc. 1998; 120:6836–6837.

[56]. Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A. Solution structure of DinI provides insight into its mode of RecA inactivation. Protein Sci. 2000; 9:2161–2169. [PubMed: 11152126]

[57]. Cai M, Huang Y, Zheng R, Wei SQ, Ghirlando R, Lee MS, Craigie R, Gronenborn AM, Clore GM. Solution structure of the cellular factor BAF responsible for protecting retroviral DNA from autointegration. Nat. Struct. Biol. 1998; 5:903–909. [PubMed: 9783751]

[58]. Baber JL, Libutti D, Levens D, Tjandra N. High precision solution structure of the C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K, a c-myc transcription factor. J. Mol. Biol. 1999; 289:949–962. [PubMed: 10369774]

[59]. Legler PM, Cai M, Peterkofsky A, Clore GM. Three-dimensional solution structure of the cytoplasmic B domain of the mannitol transporter IImannitol of the Escherichia coli phosphotransferase system. J. Biol. Chem. 2004; 279:39115–39121. [PubMed: 15258141]

[60]. Teriete P, Yao Y, Kolodzik A, Yu J, Song H, Niederweis M, Marassi FM. Mycobacterium tuberculosis Rv0899 adopts a mixed alpha/beta-structure and does not form a transmembrane beta-barrel. Biochemistry. 2010; 49:2768–2777. [PubMed: 20199110]

[61]. Garrett DS, Seok YJ, Peterkofsky A, Gronenborn AM, Clore GM. Solution structure of the 40,000 Mr phosphoryl transfer complex between the N-terminal domain of enzyme I and HPr. Nat. Struct. Biol. 1999; 6:166–173. [PubMed: 10048929]

[62]. Garrett DS, Seok YJ, Liao DI, Peterkofsky A, Gronenborn AM, Clore GM. Solution structure of the 30 kDa N-terminal domain of enzyme I of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system by multidimensional NMR. Biochemistry. 1997; 36:2517–2530. [PubMed: 9054557]

[63]. Levitt M, Sharon R. Accurate simulation of protein dynamics in solution. Proc. Natl. Acad. Sci. U. S. A. 1988; 85:7557–7561. [PubMed: 2459709]

[64]. van Gunsteren WF, Karplus M. Protein dynamics in solution and in a crystalline environment: a molecular dynamics study. Biochemistry. 1982; 21:2259–2274. [PubMed: 6178423]

[65]. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. U. S. A. 2008; 105:4685–4690. [PubMed: 18326625]

[66]. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D. NMR structure determination for larger proteins using backbone-only data. Science. 2010; 327:1014–1018. [PubMed: 20133520]

[67]. Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc. Natl. Acad. Sci. U. S. A. 2012; 109:10873–10878. [PubMed: 22733734]

[68]. Das R, Baker D. Macromolecular modeling with rosetta. Annu. Rev. Biochem. 2008; 77:363–382. [PubMed: 18410248]

[69]. Nabuurs SB, Krieger E, Spronk CA, Nederveen AJ, Vriend G, Vuister GW. Definition of a new information-based per-residue quality parameter. J. Biomol. NMR. 2005; 33:123–134. [PubMed: 16258830]

[70]. Clore GM, Robien MA, Gronenborn AM. Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. J. Mol. Biol. 1993; 231:82–102. [PubMed: 8496968]

[71]. Vriend G, Sander C. Quality control of protein models: directional atomic contact analysis. J. Appl. Crystallogr. 1993; 26:47–60.

[72]. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J. Mol. Biol. 1999; 285:1711–1733. [PubMed: 9917407]

[73]. Gronenborn AM, Clore GM. Structures of protein complexes by multidimensional heteronuclear magnetic resonance spectroscopy. Crit. Rev. Biochem. Mol. Biol. 1995; 30:351–385. [PubMed: 8575189]

[74]. Abagyan RA, Totrov MM. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. J. Mol. Biol. 1997; 268:678–685. [PubMed: 9171291]

[75]. Clore GM, Gronenborn AM. New methods of structure refinement for macromolecular structure determination by NMR. Proc. Natl. Acad. Sci. U. S. A. 1998; 95:5891–5898. [PubMed: 9600889]

HIGHLIGHTS

- EEFx is an implicit solvation potential for XPLOR-NIH.

- EEFx can be used to both fold and refine NMR structures.

- Use of EEFx improves structural precision, accuracy and quality.

- EEFx is computationally efficient.

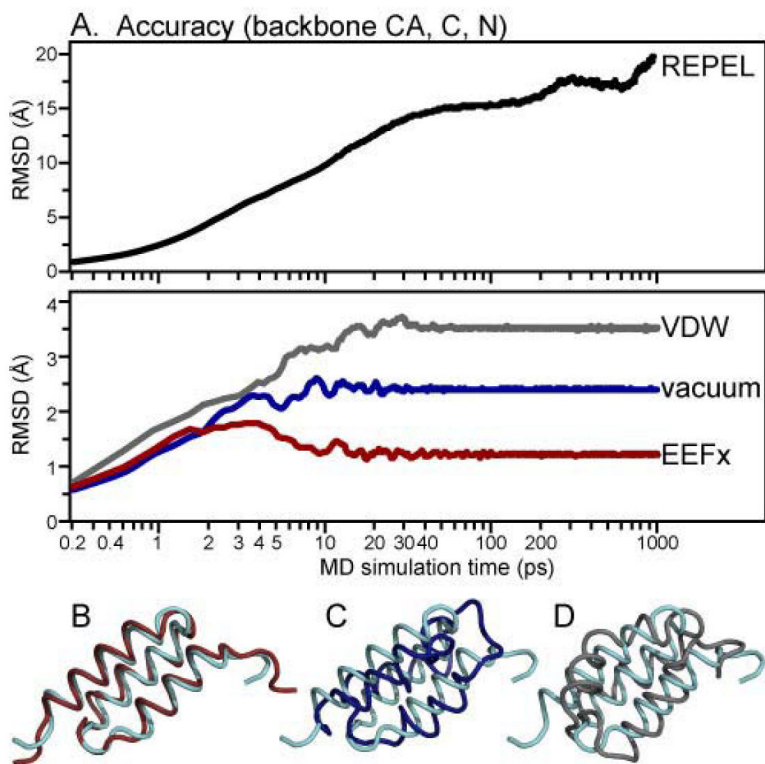- EEFx is easy to implement on standard laboratory computers.

Figure 1. Time dependence of unrestrained 1 ns MD simulations of SpAZ performed with EEFx (red), vacuum (blue), VDW (gray) or REPEL (black) force fields
(A) Structural accuracy reported as RMSD of backbone atoms (N, CA, C) to the experimental structure (PDB: 1Q2N). (B–D) Structures obtained from simulations with EEFx (C), vacuum (D) and VDW (E) superimposed on the experimentally-determined structure (cyan).
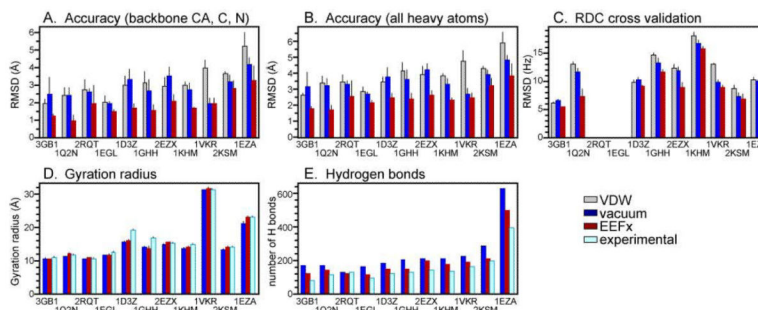
Figure 2. Free MD simulations of native protein structures
(A, B) Accuracy to native structure reported as backbone atom (CA, C, N) and heavy atom RMSD. (C) Cross correlation to experimental RDC data reported as RMSD. (D) Gyration radii of the proteins. (E) Number of hydrogen bonds observed for each simulation. Data are shown for simulations with EEFx (red), vacuum (blue) and VDW (gray) force fields. PDB codes correspond to protein names in Table 3.
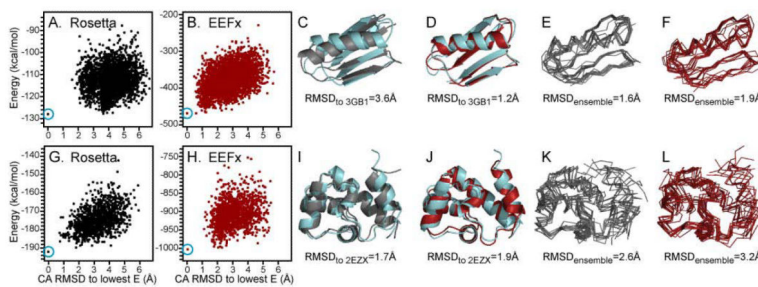
Figure 3. Recognition of native protein fold

(A, B, G, H) Rosetta (black) and EEFx (red) energy landscapes of GB1 (A, B) and BAF (G, H). RMSD values are computed for CA atoms relative to the decoy with lowest energy (blue circles). (C, D, I, J) Cartoon representations of the decoys of GB1 (C, D) and BAF (I, J) with lowest Rosetta energy (gray) or lowest EEFx energy (red) and superimposed experimental PDB structures (cyan). RMSDs represent structural accuracy, computed for CA atoms relative to the experimental PDB structures. (E, F, K, L) Ribbon representations of the 10 decoys of GB1 (E, F) and BAF (K, L) with lowest Rosetta energy (gray) or lowest EEFx energy (red). RMSDs represent precision of the structural ensembles, evaluated as average pairwise values for CA atoms.
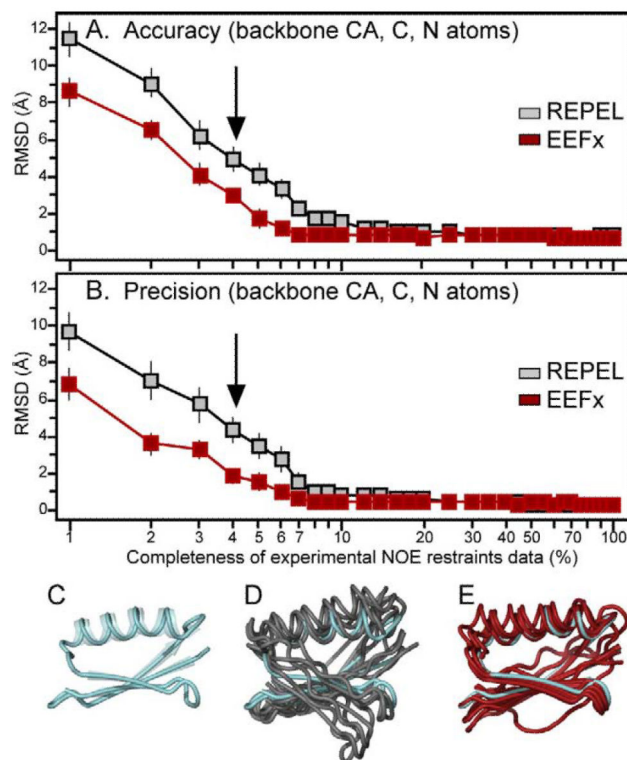
Figure 4. Effect of EEFx on NMR-restrained structure calculations of GB1 with limited distance restraints

(A, B) Effect of the number of long-range (>4 residues apart) distance restrains on structural accuracy and precision. The total number of restraints was reduced by randomly eliminating distances from the full data set. Accuracy was evaluated as pairwise RMSD of backbone CA, C, N atoms relative to the experimental structure. Precision was evaluated as average pairwise RMSD of backbone CA, C, N atoms. (C–E) Cartoon representations of the native structure (cyan) and the ensembles of five lowest energy structures obtained with 4% of the NOE data. Structures were calculated with REPEL (gray) or EEFx (red). Arrows indicate the 4% data points taken for structure illustration in panels C–E.
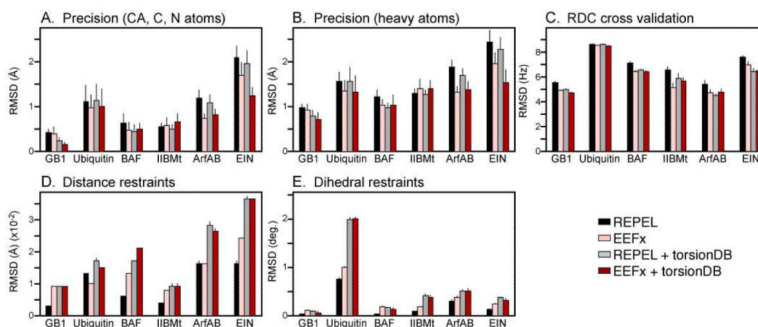
Figure 5. Structural statistics of NMR-restrained calculations performed with EEFx
(A, B) Structural precision evaluated as average pairwise RMSD of (A) backbone CA, C, N atoms and (B) all heavy atoms. (C) Agreement between structures and experimental RDC restraints excluded from structure calculations. (D, E) Agreement between structures and experimental distance and dihedral angle restraints used in the structure calculations. For each protein, the errors represent the mean ± standard deviation evaluated for ensembles of 10 lowest energy structures. Bars represent data obtained in four ways: the standard simple repulsive XPLOR potential REPEL (black); EEFx (pink); REPEL plus torsionDB (gray); and EEFx plus torsionDB (red).
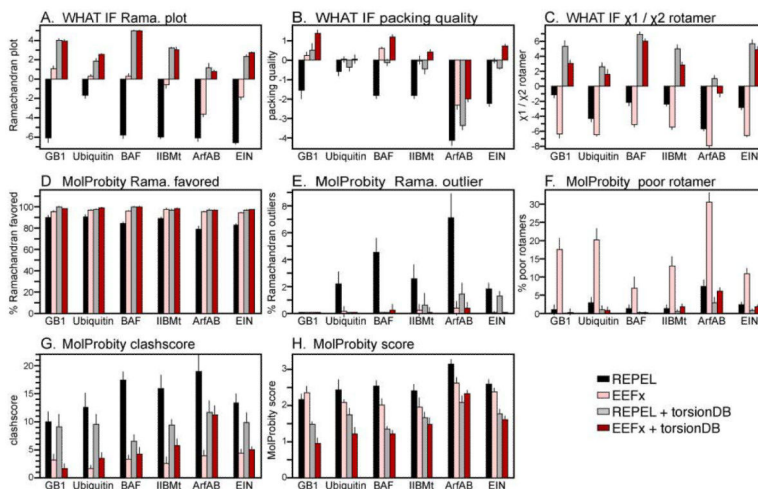
Figure 6. Structural validation analyses of NMR-restrained calculations performed with EEFx
(A–C) WHAT IF validation statistics for (A) Ramachandran plot appearance; (B) protein packing quality; and (C) 1/ 2 torsion angles. (D–H) MolProbity validation statistics for (D) percent of residues in favored regions of the Ramachandran plot; (E) percent of residues in unfavored regions of the Ramachandran plot; (F) percent of residues with poor sidechain torsion angles; (G) clashscore; and (H) overall Molprobity score. For each protein, the errors represent the mean ± standard deviation evaluated for ensembles of 10 lowest energy structures. Bars represent data obtained in four ways: the standard simple repulsive XPLOR potential REPEL (black); EEFx (pink); REPEL plus torsionDB (gray); and EEFx plus torsionDB (red). The MolProbity clashscore and MolProbity score are costs: the lower the better.

Table 1

Parameters of E_{slv} used for EEFx calculations.[a]

Atom Type	V_i	G_i^{ref}	G_i^{free}	H_i^{ref}	Cp_i^{ref}	λ_i	R_i
C	14.7	0.000	0.00	0.000	0.00	3.50	2.100
CR	8.3	−0.890	−1.40	2.220	6.90	3.50	2.100
CH1E	23.7	−0.187	−0.25	0.876	0.00	3.50	2.365
CH2E	22.4	0.372	0.52	−0.610	18.60	3.50	2.235
CH3E	30.0	1.089	1.50	−1.779	35.60	3.50	2.165
CR1E	18.4	0.057	0.08	−0.973	6.90	3.50	2.100
NH1	4.4	−5.950	−8.90	−9.059	−8.80	3.50	1.600
NR	4.4	−3.820	−4.00	−4.654	−8.80	3.50	1.600
NH2	11.2	−5.450	−7.80	−9.028	−7.00	3.50	1.600
NH3	11.2	−20.000	−20.00	−25.000	−18.00	6.00	1.600
NC2	11.2	−10.000	−10.00	−12.000	−7.00	6.00	1.600
N	0.0	−1.000	−1.55	−1.250	8.80	3.50	1.600
OH1	10.8	−5.920	−6.70	−9.264	−11.20	3.50	1.600
O	10.8	−5.330	−5.85	−5.787	−8.80	3.50	1.600
OC	10.8	−10.000	−10.00	−12.000	−9.40	6.00	1.600
S	14.7	−3.240	−4.10	−4.475	−39.90	3.50	1.890
SH1E	21.4	−2.050	−2.70	−4.475	−39.90	3.50	1.890
H	0.0	0.000	0.00	0.000	0.00	0.00	0.800

[a]Parameters are listed in *eefxPotTools*. Values of G_i^{ref}, G_i^{free}, H_i^{ref} and Cp_i^{ref} are from data of Privalov and Makhatadze [35–38]; they were determined experimentally at 298.15 K for small model molecules. Values of R_i correspond to CHARMM19 van der Waals radii. Values of V_i and λ_i were taken from the EEF1 model [23].

Table 2

Energy functions, topology and parameters of the four different force fields used in the structure calculations.[a]

Model	E_{NONB} [b]	topology/parameters	nonbonded parameters
REPEL	$E_{VDW-REPEL}$	protein.top/protein.par	k$_{rep}$ >0, C$_{rep}$ >0
VDW	E_{VDW}	proteinEEFx.top/proteinEEFx.par	k$_{rep}$=0, group, vswitch, ctonnb=7 Å, ctofnb=9Å
vacuum	$E_{VDW} + E_{ELEC}$	proteinEEFx.top/proteinEEFx.par	k$_{rep}$=0, group, vswitch, ctonnb=7Å, ctofnb=9Å, switch, rdie
EEFx	$E_{vdw} + E_{ELEC} + E_{slv}$	proteinEEFx.top/proteinEEFx.par	k$_{rep}$=0, group, vswitch, ctonnb=7Å, ctofnb=9Å, switch, rdie, r_{on}=7Å, r_{off}=9Å

[a] All calculations were performed with *nbxmod*=5, or *nbxmod*=3 for REPEL, to allow repulsions only between atoms separated by more than two covalent bonds. Calculations using the torsionDB potential were performed with *nbxmod*=4 to allow repulsions only between atoms separated by more than three covalent bonds.

[b] E_{NONB} is the XPLOR-NIH nonbonded energy function where $E_{VDW-EWPWL}$ is the simple repulsive form of the XPLOR van der Waals function, E_{VDW} is the switched Lennard-Jones form of the XPLOR van der Waals function and E_{ELEC} is the switched distance-dependent dielectric form of the XPLOR electrostatic function [29]. E_{slv} and its switching function with parameters r_{on} and r_{off} are described in equations 3–4.

Table 3

Proteins used for test structure calculations with EEFx.

Protein[a]	PDB	Length[b]	Residues[c]	Fold
GB1 [50, 51]	3GB1	56	all	α β
SpAZ [52]	1Q2N	58	6–55	α
DDEF1-SH3 [53]	2RQT	61	all	β
Eglin-c [54]	1EGL	70	8–70	α β
Ubiquitin [55]	1D3Z	76	all	α β
Din-I [56]	1GHH	81	1–71	α β
BAF [57]	2EZX	89	all	α
RNPK [58]	1KHM	89	12–84	α β
IIBMt [59]	1VKR	125	12–107	α β
ArfA-b [60]	2KSM	131	80–195	α β
EIN [61, 62]	1EZA	259	1–230	α β

[a] GBl, protein G Bl domain; SpA-Z, Staphylococcal protein A Z domain; DDEF1-SH3, human DDEF1 SH3 domain; Din, DNA-damage-inducible protein I; BAF, human barrier to autointegration factor; RNPK, nuclear ribonucleoprotein K KH domain; IIBMt, mannitol transporter enzyme II B domain; ArfA-B, *M. tuberculosis* ArfA B domain; EIN, enzyme I N-terminal domain.

[b] Full length of polypeptide.

[c] Residues used in calculations.