UC Irvine
UC Irvine Previously Published Works

Title
ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins

Permalink
https://escholarship.org/uc/item/6dj4v2xb

Journal
Chemical Biology & Drug Design, 89(1)

ISSN
1747-0277

Authors
Song, Dong
Wang, Wei
Ye, Wei
et al.

Publication Date
2017

DOI
10.1111/cbdd.12832

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Chem Biol Drug Des. Author manuscript; available in PMC 2017 June 22.

Published in final edited form as:

Chem Biol Drug Des. 2017 January ; 89(1): 5–15. doi:10.1111/cbdd.12832.

ff14IDPs Force Field Improving the Conformation Sampling of Intrinsically Disordered Proteins

Dong Song[1,#], **Wei Wang**[1,#], **Wei Ye**[1,#], **Dingjue Ji**[1,#], **Ray Luo**[2,*], and **Hai-Feng Chen**[1,3,*]

[1]State Key Laboratory of Microbial metabolism, Department of Bioinformatics and Biostatistics, College of Life Sciences and Biotechnology, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai, 200240, China

[2]Departments of Molecular Biology and Biochemistry, Chemical Engineering and Materials Science, and Biomedical Engineering, University of California, Irvine, CA 92697-3900

[3]Shanghai Center for Bioinformation Technology, 1275 Keyuan Road, Shanghai, 200235, China

Abstract

Intrinsically disordered proteins (IDPs) are proteins which lack of specific tertiary structure and unable to fold spontaneously without the partner binding. These IDPs are found to associate with various diseases, such as diabetes, cancer, and neurodegenerative diseases. However, current widely used force fields, such as *ff99SB*, *ff14SB*, *OPLS/AA*, and *Charmm27* are insufficient in sampling the conformational characters of IDPs. In this study, the CMAP method was used to correct the φ/ψ distributions of disorder-promoting amino acids. The simulation results show that the force filed parameters (*ff14IDPs)* can improve the φ/ψ distributions of the disorder-promoting amino acids, with RMSD less than 0.10% relative to the benchmark data of IDPs. Further test suggests that the calculated secondary chemical shifts under *ff14IDPs* force field are in quantitative agreement with the data of NMR experiment for five tested systems. In addition, the simulation results show that *ff14IDPs* can still be used to model structural proteins, such as tested lysozyme and ubiquitin, with better performance in coil regions than the original general Amber force field *ff14SB*. These findings confirm that the newly developed Amber *ff14IDPs* force field is a robust model for improving the conformation sampling of IDPs.

Keywords

IDPs; CMAP correction; ff14IDPs; force field

Introduction

Intrinsically disordered proteins (IDPs) or intrinsically disordered protein regions are characterized by lack specific tertiary structure and unable to fold spontaneously into globular three-dimensional structures without partner binding.[1] The results from human

*Corresponding authors: haifengchen@sjtu.edu.cn; ray.luo@uci.edu, Tel: 86-21-34204348, Fax: 86-21-34204348.
#These authors contributed equally to this work.

There is no conflict of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

proteomes suggest that there are 35-50% of proteins with more than 40 consecutive disordered residues.[2, 3] Furthermore, IDPs have been found to be included in many biological processes, such as regulation, recognition, cell cycle control, and signaling.[4] For example, the N terminal and C terminal domains of tumor suppressor p53 are intrinsically disordered and p53 is found in multiple signaling pathways.[5-8] Disease-associated proteins are also rich in predicted disordered regions.[9] Thus it is essential to research the structure-function of IDPs.

Since IDPs cannot spontaneously fold into stable tertiary structure without binding to their partners, the key experimental method for exploring the dynamics conformation of IDPs is NMR spectroscopy.[1] In addition, molecular dynamics simulation can be used to reveal the structural continuum of IDPs, from tightly folded single domains and multidomain proteins with flexible or disordered regions, to disordered molten globules, highly extended, and heterogeneous unstructured states. Force field plays a key role in applications of MD simulations. However our previous studies show that most protein force fields could not reproduce the flexible conformers of intrinsically disordered proteins due to their original intended applications of folded proteins.[10] Other researchers also reached similar conclusions regarding the use of existing force fields for sampling IDPs.[11-13] To overcome these limitations, optimized water model and specific modification of protein–water Lennard-Jones parameters were used to simulate IDPs.[11-14] These studies also suggest that redesigning and reparameterizing the protein force field would be required.[13]

Previous experimental study shows that the X-ray-, NMR-, and CD-characterized disordered segments have similar amino acid compositions and are significant different from ordered segments.[15] These residues of G, A, S, P, R, Q, E, and K are enriched in disorder regions and named disorder-promoting amino acids. To improve the performance of existing force fields, an Amber force field (*ff99IDPs)* was developed to correct the main chain torsional distributions for these disorder promoting residues. Our tests on the multiple representative IDPs show that *ff99IDPs* better reproduces the conformers of IDPs than its generic counterpart, *ff99SBildn*.[16]

In order to improve the accuracy of conformational sampling for structural proteins, the Amber *ff14SB* force field was recently developed to refine the protein side chain and backbone torsion terms[17] and improved secondary structure content in small peptides and reproduced NMR measurements for proteins in solution. In this study, we intend to refresh our IDPs-specific force field to bring it up-to-date with the generic protein force field *ff14SB*. Indeed, our tests of the generic force fields with multiple IDPs, such as measles virus nucleoprotein (MeV N_{TAIL}) in apo state;[18-23] N-terminal domain of free *p53*,[24, 25] arginine-rich HIV-1 Rev (HIVRev),[26] aspartic proteinase inhibitor IA_3,[27] and α-synuclein,[28] show that *ff99SB*, *ff14SB*, *OPLS/AA*, and *CHARMM27* are all insufficient in sampling the conformational characteristics of IDPs. Based on these tests, we updated our IDPs force field (termed *ff14IDPs*) to be consistent with the latest generic force field. Our tests show that *ff14IDPs* leads to diverse disordered conformers for the five tested IDPs. Furthermore, predicted secondary chemical shifts of *ff14IDPs* are in agreement with those of experimental measures. In addition, both *ff14IDPs* and *ff14SB* force fields can sample the conformers of structural proteins.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Methods

The hypothesis in developing a special-purpose force field for IDPs

Previous work suggests that disordered segments have similar amino acid compositions and are different from ordered segments.[15] We hypothesized that the torsional terms of the eight disorder-promoting amino acids could be altered to improve the sampling of IDPs by the generic protein force fields. An issue in the revision of a generic force field is the lack of training data to improve the dihedral angle terms. To overcome this initial difficulty, we relied on a working hypothesis to use coiled regions from crystal structures to model the conformations of disordered amino acids in the construction of an IDP-specific force field. Here the coil regions are these residues without in any secondary structure as defined by DSSP.[29] Obviously there is a difference between the coiled amino acids and disordered amino acids. However, there are reports that intrinsically disordered proteins can be divided into intrinsically pre-molten globules and intrinsically random coils based on their conformational properties.[30, 31] The final validation of the adopted strategy is the reproduction of experimental properties as to be shown below.

Data collection of disordered protein

17,540 structures with sequence identity less than 30% and R factor less than 0.25 were extracted from PDB database. The secondary structures and dihedral angles of these structures were calculated with DSSP[32, 33]. The definition of IDPs benchmark is consecutive five or more residues classified as “coil structure”, i.e. without any secondary structure assignment in DSSP.[29, 34, 35] Our previous work confirms the reasonability for five or more consecutive residues with coil structure as training data for IDPs.[10] The statistical results are shown in Figure 1. In summary, 54,838 coil fragments containing 346,335 pairs of backbone dihedrals for eight disorder-promoting amino acids were collected.

CMAP method

In order to evaluate the hypothesis, CMAP energy correction term [36-38] was utilized to minimize the difference of the dihedral distributions between disorder promoting residues and IDPs benchmark. This method is previously integrated in the CHARMM software package and was ported to the Amber simulation package.[39, 40] To correct the dihedral energy of disorder promoting residues, E_{CMAP} energy term was introduced into the potential energy function of AMBER, as shown in Eq. 1.

$$E_{\mathrm{MM}} = E_{\mathrm{ff14SB}} + E_{\mathrm{CMAP}} \quad (1)$$

We corrected the dihedral distribution with E_{CMAP} for the eight disorder-promoting amino acids and other amino acids remain the same to that of the previous force field *ff14SB*. The di-peptide models (Nme-X-Ace, Nme means aminomethyl, X means a certain amino acid, Ace represents acetyl) were used in this study. Similar models were also applied in the previous Amber force field *ff14SB* developments.[41, 42]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The method of CMAP correction for IDPs is described in our previous work,[10] and will be briefly introduced here. The CMAP is a matrix of corrections between grid points with a two-dimensional bicubic interpolation method.[38] Root mean square deviations of dihedral population between MD and benchmark populations were used to quantitatively evaluate the distribution difference. In order to obtain the convergence of RMSp, an optimization was conducted for every disorder-promoting amino acid. Up to seven iteration steps were used to optimize the parameters for the eight disorder-promoting amino acids. At each iteration step, the dipeptides were simulated 100ns to collect conformations.

Integrating with Amber *ff14SB* force field

The CMAP parameters of the eight disorder-promoting amino acids were created and integrated into a text file. Therefore, *ff14IDPs* force field can easily be used to process these disorder-promoting amino acids. Firstly, LEaP module was used to build standard topology and coordinate files under *ff14SB*. Secondly, the parameters of CMAP term were added to the standard topology file. Finally, the standard simulation can be performed with Amber package.

Molecular dynamics simulations

All IDPs were simulated and processed with AMBER12.[40] LEaP module was used to create the di-peptide models for eight disorder-promoting amino acids. Counter-ions were introduced to maintain system neutrality. A truncated octahedron box of TIP3P waters was used to solvate all systems with a buffer of 10 Å. Long-range electrostatic interactions were calculated with Particle Mesh Ewald (PME)[43] algorithm.[40] CUDA version of PMEMD[44] was used to accelerate the MD simulations. The bonds involving hydrogen atoms were constrained with the SHAKE algorithm.[45] All systems were minimized for 20,000-step with steepest descent method, then with heating for 20 ps and equilibrating for 20 ps in the NVT ensemble at room temperature.

To compare the performances between *ff14IDPs* and *ff14SB* to sample IDPs, ten individual trajectories for each system were created for bound and free MeV N_{TAIL}, free *p53, free* HIVRev, free IA_3, and free-αSyn, under both force fields. 140 ns simulations were needed for the convergence of apo-HIVRev, and 100 ns for the other systems. Lysozyme and ubiquitin, as structural proteins, are usually used to test the previous force fields. Therefore they were also applied to evaluate the compatibility of *ff14IDPs* on structural proteins. Ten trajectories and 100ns are sufficient for the sample and equilibration under both force fields. In order to compare the performance of *OPLS/AA* force field for sampling the conformer of IDPs, GROMACS 5.0.4 package was used to simulate five trajectories for apo-HIVRev.[46] At the same time, *Charmm27* force field was also used to sample the conformers of apo-HIVRev.[47] Total 9.8 μs trajectories for seven test systems were simulated at 298K with about 4,600 GPU hours. Table 1 is gathered the simulation conditions for these systems.

Data Analyses

PTRAJ module was used to process the routine analysis of conformational sampling.[40] Kclust program with the mode of phi and radius of 30 degree in the MMTSB tool set was performed the structural clustering.[48] DSSP was used to assign the secondary

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

structures.[32, 33] The experimental values of secondary chemical shift data for seven tested systems were extracted from the BRMB database[49] and the access numbers are shown in Table 1. SPARTA version 1.01 was used to calculate secondary chemical shift data for test systems.[50] N-H order parameters (S^2) were calculated with the Karplus equation.[51] NOE parameters were calculated using the method in the previous work.[52] The literature method was used to calculate the helicity content.[10]

Results and Discussion

Database statistics

Previous study shows that there are eight disorder-promoting residues and are enriched in disorder regions. We hypothesized that correction of the main-chain dihedral terms of these residues could improve the sampling of IDPs. To evaluate the strategy, we processed the following statistics of the main-chain dihedral angle distributions for these residues. The differences of dihedral distributions between benchmark of IDPs and CMAP correction for Glu, Lys and Pro are shown in Figure 2. The RMSp between the dihedral distributions of benchmark of IDPs and *ff14SB* is about 0.6135%, 0.3834%, and 0.3857%, for Glu, Lys, and Pro, respectively. The *P* values of two-sample KS test are less than 0.001, which suggests the significant differences for the distributions. In order to compare the difference of dihedral distribution for Glu, Lys, and Pro from coiled and structured regions of PDB, the Φ/Ψ distributions are shown in Figure S1 (supplementary). This figure indicates that the Φ/Ψ distributions of coiled regions are significantly different from structured regions. For example, the most popular distribution of Glu mainly adopts the PPII conformer in coil regions, while it adopts the other conformer in the structured regions.

CMAP energy term optimization

The RMS*p* gradually decreases between CMAP optimization and benchmark of IDPs for the eight disorder-promoting amino acids and the value remains stable after seven iteration steps. The supplementary Figure S2 shows the PMF differences for eight disorder-promoting amino acids between *ff14SB* and final CMAP. The results suggest that the PMF differences are significant. The iterative optimization processes are shown in Figure 3. This suggests that the optimized CMAP reproduces the distribution of disorder-promoting amino acids comparing with that of IDPs benchmark.

Force Field Validation

Five representative intrinsically disordered proteins (MeV N_{TAIL}, *p53*, HIVRev, IA_3, and αSyn) were used to evaluate the performance of *ff14IDPs*. MeV N_{TAIL} could fold into highly ordered α-helices upon binding to XD domain of measles virus phosphoprotein.[53] HIVRev includes 17 disorder-promoting amino acids over 21 residues. On the contrary, IA_3 has 14 order-promoting amino acids over 31-mer polypeptide. αSyn is an another type of IDPs combined with a long structured region. These five IDPs were extensively simulated in *ff14IDPs and ff14SB*. In summary, 5 independent trajectories were simulated for each protein with *ff14IDPs and ff14SB* force fields. The secondary chemical shift and other NMR parameters were calculated and compared with experimental observations. The

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

performances of *ff14IDPs* and *ff14SB* were also compared for structural bound MeV which experimental data are available.

In order to evaluate the convergence of sampled conformers, the conformer clusters over increasing simulation time (0-50 ns, 0-55 ns, 0-60 ns, etc) was analyzed (Figure 4). The conformer cluster number did not increase after 140 ns, indicating that apo-HIVRev becomes dynamics equilibration. While the numbers of conformer cluster remain stable after 100 ns. This indicates that 100 ns simulation samples enough diverse conformations of other IDPs systems.

HIV Rev ARM (HIVRev)—RMSD, RMSF, secondary chemical shift, conformer clustering, and the helicity under *ff14IDPs* and *ff14SB* force fields are shown in Figure 5. Top 10 conformer clusters under *ff14IDPs* occupy 31.18% of conformation ensemble (top 89 for 70%). Most conformers include high ratio of disordered regions. However, top 10 conformer clusters under *ff14SB* occupy up to 99.94% of the conformation ensemble (top 2 for 70%). This suggests that the structural clusters are significant different under two force fields. The potential of mean force (PMF) free energy landscapes with the reaction coordinates of the radius of gyration (RG) and RMSD (Figure 5C) show that the distribution of conformer from *ff14IDPs* is between RMSD at 1~10 Å and RG at 9 Å ~18 Å, and between RMSD at 1~8 Å and RG at 9 Å ~12 Å for *ff14SB*. This suggests that *ff14IDPs* samples more diverse flexible disordered conformers than *ff14SB*, which is consistent with the conformer clustering analysis. The helicity of free HIVRev is shown in Figure 5E. This figure shows that *ff14IDPs* reproduces significantly lower helical secondary structures than *ff14SB*. The experimental and computational J-coupling constants for HIVRev are shown in supplementary Figure S3. This figure suggests that the prediction from *ff14IDPs* is more similar to experiment than that from *ff14SB*. The representative structures from the clusters occupying no less than 70% conformations were used to calculate the secondary Cα chemical shifts. The full length RMSD between calculated and experimental chemical shifts (shown in Figure 5F) was 0.738 ppm for *ff14IDPs*, 1.499 ppm for *ff14SB*, 1.054 ppm for the *OPLS* force field, and 1.543 ppm for the *Charmm27* force field, respectively. This indicates that the performance of *ff14IDPs* for reproducing the chemical shifts is better than that of *ff14SB, OPLS*, and *Charmm27*. Comparison with *ff99IDPs* (0.676 ppm) shows that the full length RMSD of secondary chemical shift for *ff14IDPs* is slightly larger,[16] highlighting the effect of the refinement in the side chain terms in the newer Amber force fields.

Apo-p53 TAD—To further compare the performance between *ff14IDPs* and ff14SB on sampling IDPs, multiple trajectories were simulated on the free *p53* (residues 17-29), respectively. The analysis results are shown in Figure 6. The RMSF from *ff14IDPs* demonstrates slightly larger than that from *ff14SB*.

Conformer clusters were also applied to explore the heterogeneity of *p53* conformers. Figure 6 shows the representative conformers of top 10 clusters under *ff14SB* and *ff14IDPs*. The top 10 conformers from *ff14SB* and *ff14IDPs* occupy 79.05% and 52.30% of the total conformers, respectively. The results suggest that *ff14IDPs* samples more heterogeneous conformations than *ff14SB*.[54] The helicity content of p53 is shown in Figure 6E. The helicity content for L22-L25 under *ff14IDPs* is 31.61%, 56.07% under *ff14SB*, and about

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

30% from NH^N RDCs experiment[24]. This shows that the *ff14IDPs* data is in quantitative agree with that of experiment. However, the helicity content under *ff14SB* is clearly over-estimated. The representative conformers and their occupancies were used to calculate the secondary chemical shifts. The full length RMSD was 0.966 ppm for *ff14IDPs* and 1.146 ppm for *ff14SB*, respectively. This indicates that the secondary chemical shifts *from ff14IDPs* are more approach to experiment data than those from *ff14SB*. It is also interested to note that the full length RMSD for *ff14IDPs* is slightly smaller than that for *ff99IDPs* (1.032 ppm). However, the significant difference is focused on the N-terminal domain between experiment and prediction from ff14IDPs. After the analysis of residue composition, we found that most residues (T, F, S, and D) are not disordered promoting residues and have not the parameters of CMAP correction. Therefore, specific CMAP corrections based on each type of residues should be necessary in the next version of force field.

α-Synuclein (αSyn)—The NMR structures of αSyn include a disordered and two long α-helices region.[55] To test the performance of *ff14IDPs*, we just analyzed the character of disordered region (Res. 95-140). The analyzed results are shown in Figure 7. The top 10 conformer clusters occupy 14.18% (*ff14IDPs*) and 33.66% (*ff14SB*), respectively. As shown in Figure 7G, *ff14IDPs* samples extensively disordered conformers; however, *ff14SB* creates partially helical conformers. The Rg value of alpha-synuclein under *ff14IDPs* is 18.33Å and 17.54Å under *ff14SB*. This suggests that the *ff14IDPs* simulation is closer to NMR[56]. Similar to HIVRev and p53, representative structures and their occupancy were used to predict the secondary Cα chemical shifts. At the same time, these data were directly compared with the values of NMR experiment. The full length RMSD is 0.455 ppm for *ff14IDPs* and 0.566 ppm for *ff14SB*, respectively. This further suggests that the secondary chemical shifts *from ff14IDPs* are more approach to experiment data than those from *ff14SB*. Finally the full length RMSD for *ff14IDPs* is slightly smaller than that for *ff99IDPs* (0.472 ppm).[16]

Aspartic Proteinase Inhibitor (IA_3)—The RMSD, RMSF, PMF landscapes, secondary chemical shift, average helicity, and conformer clustering of apo-IA3 under *ff14IDPs* and *ff14SB* are shown in supplementary Figure S4. The ratio of top 10 conformer clusters under *ff14IDPs* is 53.10% of the total conformations (top 24 for 70%). The representative conformers of 8 clusters include high ratio of coiled regions. However, top 10 clusters under *ff14SB* occupy up to 87.90% of the conformation ensemble (top 6 for 70%) and the ratio of disordered structures is much lower than those under *ff14IDPs*. The PMF landscapes between the radius of gyration (Rg) and RMSD (supplementary Figure S4C) show that the conformer distribution from *ff14IDPs* is between Rg at 8 Å ~21 Å and RMSD at 1~13 Å, and between RG at 8 Å ~18 Å and RMSD at 1~13 Å from *ff14SB*. This suggests that *ff14IDPs* improves the conformer sampling of IDPs, which is in accord to the conformer clustering. Representative conformers and their occupancies were used to predict the secondary Cα chemical shifts and shown in supplementary Figure S4F. The RMSD between calculated and experimental chemical shifts was 1.175 ppm for *ff14IDPs* and 2.061 ppm for *ff14SB*, respectively. Thus the conformer sampling from *ff14IDPs* is more approach to NMR data than that from *ff14SB*. Furthermore, the full length RMSD of chemical shifts for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ff14IDPs is similar to that for *ff99IDPs* (1.121 ppm).[16] The helicity of apo-IA_3 is shown in supplementary Figure S4E. This figure indicates that *ff14IDPs* samples significant lower ratio of helical structures than *ff14SB*.

MeV N_{TAIL}—RMSD, RMSF, PMF landscapes, average helicity, secondary Cα chemical shift, and conformer clustering for MeV N_{TAIL} under *ff14IDPs* and *ff14SB* are shown in supplementary Figure S5. The PMF landscapes show that the conformer space from *ff14IDPs* is between RMSD at 1~10 Å and RG at 7 Å ~15 Å, and between RMSD at 1~9 Å and RG at 7 Å ~12 Å for *ff14SB*. This indicates that *ff14IDPs* could sample more disordered conformers than *ff14SB*. Cα RMSFs in *ff14IDPs* are larger than those in *ff14SB*, which suggests *ff14IDPs* introducing a larger conformational adjustment (supplementary Figure S5D). The helicity of MeV is shown in supplementary Figure S5E. This figure indicates that *ff14SB* samples more helical structures than *ff14IDPs*. The helicity content is 28.56% under *ff14IDPs*, 62.89% under *ff14SB*, and 26.36% from NMR experiment.[53] This suggests that the helicity content under *ff14IDPs* is in quantitative agree with that of experiment. However, the helicity content under *ff14SB* is also over-estimated. Furthermore, more flexible conformers for *ff14IDPs* can also be found in structural clustering. Top 10 clusters under *ff14IDPs* and *ff14SB* occupy 31.50% and 65.36% of the whole conformations, respectively. 70% conformers under *ff14IDPs* include high ratio disordered regions. However, only one conformer under *ff14SB* includes partially disordered region. The full length RMSD (shown in supplementary Figure S5F) was 0.727 ppm for *ff14IDPs* and 1.033 ppm for *ff14SB* between experimental chemical shifts and predicted data, respectively. This indicates that the difference between predicted chemical shifts and experimental data under *ff14IDPs* is significant lower than that under *ff14SB*. Finally the full length RMSD under *ff14IDPs* is slightly larger than that under *ff99IDPs* (0.699 ppm).

To further evaluate the performance of *ff14IDPs* on IDPs complex, bound MeV N_{TAIL} was also used in this study. The results are shown in supplementary Figure S6. RMSD and RMSF indicate that bound MeV N_{TAIL} is less flexible under two force fields. Furthermore, the secondary structure under *ff14IDPs* is similar to that under *ff14SB* with highly ordered structures (shown in supplementary Figure S6G). The full length RMSD was 0.470 ppm for *ff14IDPs* and 0.341 ppm for *ff14SB*, respectively. This indicates that *ff14IDPs* can also sample the conformers of IDPs complex.[20, 21]

Lysozyme and Ubiquitin—Lysozyme and ubiquitin, as structural proteins, have often been used to test the performance of force field.[42, 57] In order to evaluate the performance of *ff14IDPs* on structured proteins, these two widely used proteins were also employed in this study. Figure 8 shows the order parameter (S^2), the secondary Cα chemical shift, NOE, and conformer cluster derived by *ff14IDPs* and *ff14SB* for ubiquitin. These parameters from *ff14IDPs* are similar to those from *ff14SB* and are in quantitative accord to those of experimental values. Especially, the performance for *ff14IDPs* is better in loop regions than *ff14SB*. These parameters for lysozyme are shown in supplementary Figure S7. Similar results are found for lysozyme. The results from ubiquitin and lysozyme indicate that *ff14IDPs* can also model the structured proteins.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Conclusion

IDPs play important biological function in cell cycle control, regulation, recognition, and signaling. These IDPs are also associated with many diseases. However, latest developed force field (*ff14SB*) could not reproduce the diverse conformers of IDPs. In this study, *ff14IDPs* force field was developed to address the sampling of IDPs. Previous work shows that there are eight disorder-promoting amino acids that are enriched in disordered regions. We hypothesized that the dihedral angle terms of these residues could be altered to improve the sampling of IDPs. Following this idea, we revised the φ/ψ dihedral terms using a customized CMAP energy term. The CMAP term was iteratively optimized to reproduce the φ/ψ dihedral distribution of the eight disordered-promoting residues as observed in a database compiled from the coil regions of well-resolved crystal structures. To validate our strategy used to develop the special purpose *ff14IDPs* force field, five typical IDPs were simulated extensively. The results confirm that *ff14IDPs* samples the diverse conformers of IDPs. The predicted secondary chemical shift data are in quantitative agreement with experimental data. These results are significantly better than those from *ff14SB*. Finally, *ff14IDPs* can also be used for modeling the stable complex conformations of IDPs and structured proteins. The performance of *ff14IDPs* was also compared with that of *ff99IDPs* and we found similar performance between the two. As we mentioned, the newer Amber force field was also improved the side chain parameters. Apparently these changes in principle would change the quality of the sampling of IDPs, and we recommend the newer generation Amber force field for more accurate protein structure sampling.

Acknowledgments

This work was supported by Center for HPC at Shanghai Jiao Tong University, by the National High-tech R&D Program of China (863 Program) (2014AA021502), by grants from the Ministry of Science and Technology of China (2012CB721003), the National Natural Science Foundation of China (J1210047), by Medical Engineering Cross Fund of Shanghai Jiaotong University (YG2013MS68, YG2014MS47, and YG2015MS56), and by National Institute of General Medicines (GM093040 & GM079383).

References

1. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nature reviews Molecular cell biology. 2005; 6:197–208. [PubMed: 15738986]
2. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Informatics Series. 2000:161–171. [PubMed: 11700597]
3. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry. 2002; 41:6573–6582. [PubMed: 12022860]
4. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nature reviews Molecular cell biology. 2015; 16:18–29. [PubMed: 25531225]
5. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. Journal of molecular biology. 2002; 323:573–584. [PubMed: 12381310]
6. Tsai CJ, Ma B, Nussinov R. Protein–protein interaction networks: how can a hub protein bind so many different partners? Trends in biochemical sciences. 2009; 34:594–600. [PubMed: 19837592]
7. Collavin L, Lunardi A, Del Sal G. p53-family proteins and their regulators: hubs and spokes in tumor suppression. Cell Death & Differentiation. 2010; 17:901–911. [PubMed: 20379196]
8. Chouard T. Structural biology: Breaking the protein rules. Nature. 2011; 471:151–3. [PubMed: 21390105]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

9. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annual review of biochemistry. 2014; 83:553–84.
10. Wang W, Ye W, Jiang C, Luo R, Chen HF. New force field on modeling intrinsically disordered proteins. Chemical biology & drug design. 2014; 84:253–69. [PubMed: 24589355]
11. Best RB, Zheng W, Mittal J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. J Chem Theory Comput. 2014; 10:5113–5124. [PubMed: 25400522]
12. Best RB, Zheng W, Mittal J. Correction to Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. J Chem Theory Comput. 2015; 11:1978. [PubMed: 26574399]
13. Piana S, Donchev AG, Robustelli P, Shaw DE. Water dispersion interactions strongly influence simulated structural properties of disordered protein States. The journal of physical chemistry B. 2015; 119:5113–23. [PubMed: 25764013]
14. Palazzesi F, Prakash MK, Bonomi M, Barducci A. Accuracy of current all-atom force-fields in modeling protein disordered states. J Chem Theory Comput. 2015; 11:2–7. [PubMed: 26574197]
15. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins: Structure, Function, and Bioinformatics. 2001; 42:38–48.
16. Ye W, Ji D, Wang W, Luo R, Chen HF. Test and Evaluation of ff99IDPs Force Field for Intrinsically Disordered Proteins. Journal of chemical information and modeling. 2015; 55:1021–9. [PubMed: 25919886]
17. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015; 11
18. Wright PE, Dyson HJ. Linking folding and binding. Curr Opin Struct Biol. 2009; 19:31–8. [PubMed: 19157855]
19. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Curr Opin Struct Biol. 2011; 21:441–6. [PubMed: 21482101]
20. Longhi S, Receveur-Bréchot V, Karlin D, Johansson K, Darbon H, Bhella D, Yeo R, Finet S, Canard B. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. Journal of Biological Chemistry. 2003; 278:18638–18648. [PubMed: 12621042]
21. Bourhis JM, Johansson K, Receveur-Bréchot V, Oldfield CJ, Dunker KA, Canard B, Longhi S. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. Virus research. 2004; 99:157–167. [PubMed: 14749181]
22. Bourhis JM, Receveur-Bréchot V, Oglesbee M, Zhang X, Buccellato M, Darbon H, Canard B, Finet S, Longhi S. The intrinsically disordered C-terminal domain of the measles virus nucleoprotein interacts with the C-terminal domain of the phosphoprotein via two distinct sites and remains predominantly unfolded. Protein science. 2005; 14:1975–1992. [PubMed: 16046624]
23. Wang Y, Chu X, Longhi S, Roche P, Han W, Wang E, Wang J. Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. Proceedings of the National Academy of Sciences. 2013; 110:E3743–E3752.
24. Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI, Blackledge M, Fersht AR. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc Natl Acad Sci U S A. 2008; 105:5762–7. [PubMed: 18391200]
25. Chen HF, Luo R. Binding induced folding in p53-MDM2 complex. J Am Chem Soc. 2007; 129:2930–7. [PubMed: 17302414]
26. Casu F, Duggan BM, Hennig M. The arginine-rich RNA-binding motif of HIV-1 Rev is intrinsically disordered and folds upon RRE binding. Biophys J. 2013; 105:1004–17. [PubMed: 23972852]
27. Li M, Phylip LH, Lees WE, Winther JR, Dunn BM, Wlodawer A, Kay J, Gustchina A. The aspartic proteinase from Saccharomyces cerevisiae folds its own inhibitor into a helix. Nature structural biology. 2000; 7:113–7. [PubMed: 10655612]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

28. Rao JN, Jao CC, Hegde BG, Langen R, Ulmer TS. A combinatorial NMR and EPR approach for evaluating the structural ensemble of partially folded proteins. J Am Chem Soc. 2010; 132:8657–68. [PubMed: 20524659]
29. Jiang F, Han W, Wu YD. The intrinsic conformational features of amino acids from a protein coil library and their applications in force field development. Physical chemistry chemical physics: PCCP. 2013; 15:3413–28. [PubMed: 23385383]
30. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein science: a publication of the Protein Society. 2002; 11:739–56. [PubMed: 11910019]
31. Tamiola K, Acar B, Mulder FA. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J Am Chem Soc. 2010; 132:18000–3. [PubMed: 21128621]
32. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]
33. Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G. A series of PDB related databases for everyday needs. Nucleic acids research. 2011; 39:D411–D419. [PubMed: 21071423]
34. Jiang F, Zhou CY, Wu YD. Residue-specific force field based on the protein coil library. RSFF1: modification of OPLS-AA/L. The journal of physical chemistry B. 2014; 118:6983–98. [PubMed: 24815738]
35. Zhou CY, Jiang F, Wu YD. Residue-specific force field based on protein coil library. RSFF2: modification of AMBER ff99SB. The journal of physical chemistry B. 2015; 119:1035–47. [PubMed: 25358113]
36. MacKerell AD, Bashford D, Bellott M, Dunbrack R, Evanseck J, Field MJ, Fischer S, Gao J, Guo H, Ha Sa. All-atom empirical potential for molecular modeling and dynamics studies of proteins. The Journal of Physical Chemistry B. 1998; 102:3586–3616. [PubMed: 24889800]
37. MacKerell AD, Feig M, Brooks CL. Improved treatment of the protein backbone in empirical force fields. Journal of the American Chemical Society. 2004; 126:698–699. [PubMed: 14733527]
38. MacKerell AD, Feig M, Brooks CL. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. Journal of computational chemistry. 2004; 25:1400–1415. [PubMed: 15185334]
39. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. Journal of computational chemistry. 2005; 26:1668–1688. [PubMed: 16200636]
40. Case, D., Darden, T., Cheatham, T., III, Simmerling, C., Wang, J., Duke, R., Luo, R., Walker, R., Zhang, W., Merz, K. AMBER 12. University of California; San Francisco: 2012.
41. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society. 1995; 117:5179–5197.
42. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins: Structure, Function, and Bioinformatics. 2006; 65:712–725.
43. Darden T, York D, Pedersen L. Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. J Chem Phys. 1993; 98:10089–92.
44. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born. Journal of Chemical Theory and Computation. 2012; 8:1542. [PubMed: 22582031]
45. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys. 1977; 23:327–41.
46. Hess B, Kutzner C, Spoel Dvd, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput. 2008; 4:435–447. [PubMed: 26620784]
47. Feller SE, M AD. An Improved Empirical Potential Energy Function for Molecular Simulations of Phospholipids. J Phys Chem B. 2000; 104

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

48. Feig M, Karanicolas J, Brooks CL III. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. Journal of Molecular Graphics and Modelling. 2004; 22:377–395. [PubMed: 15099834]

49. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z. Bio Mag Res Bank. Nucleic acids research. 2008; 36:D402–D408. [PubMed: 17984079]

50. Shen Y, Bax A. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. Journal of biomolecular NMR. 2007; 38:289–302. [PubMed: 17610132]

51. Pardi A, Billeter M, Wuthrich K. Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein. Use of 3JHN alpha for identification of helical secondary structure. J Mol Biol. 1984; 180:741–51. [PubMed: 6084720]

52. Prompers JJ, Bruschweiler R. General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation. J Am Chem Soc. 2002; 124:4522–34. [PubMed: 11960483]

53. Jensen MR, Communie G, Ribeiro EA Jr, Martinez N, Desfosses A, Salmon L, Mollica L, Gabel F, Jamin M, Longhi S, Ruigrok RW, Blackledge M. Intrinsic disorder in measles virus nucleocapsids. Proc Natl Acad Sci U S A. 2011; 108:9839–44. [PubMed: 21613569]

54. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci U S A. 2010; 107:8183–8. [PubMed: 20404210]

55. Rao JN, Jao CC, Hegde BG, Langen R, Ulmer TS. A Combinatorial NMR and EPR Approach for Evaluating the Structural Ensemble of Partially Folded Proteins. J Am Chem Soc. 2010; 132:8657–8668. [PubMed: 20524659]

56. Morar AS, Olteanu A, Young GB, Pielak GJ. Solvent-induced collapse of alpha-synuclein and acid-denatured cytochrome c. Protein science: a publication of the Protein Society. 2001; 10:2195–9. [PubMed: 11604526]

57. Buck M, Bouguet-Bonnet S, Pastor RW, Mac Kerell AD Jr. Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme. Biophys J. 2006; 90:L36–L38. [PubMed: 16361340]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

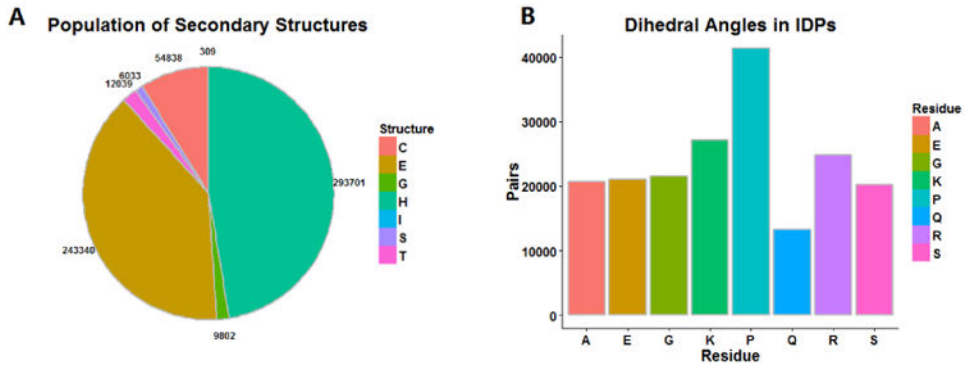

Figure 1.

Distribution of secondary structure and dihedral of IDRs in PDB. A: Distribution of secondary structure and coil in PDB. E for β-sheet, H for α-helix, G for 3_{10} helix, I for π helix, T for turn, S for bend, C for coil. B: Distribution of dihedral for eight disorder-promoting amino acids.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

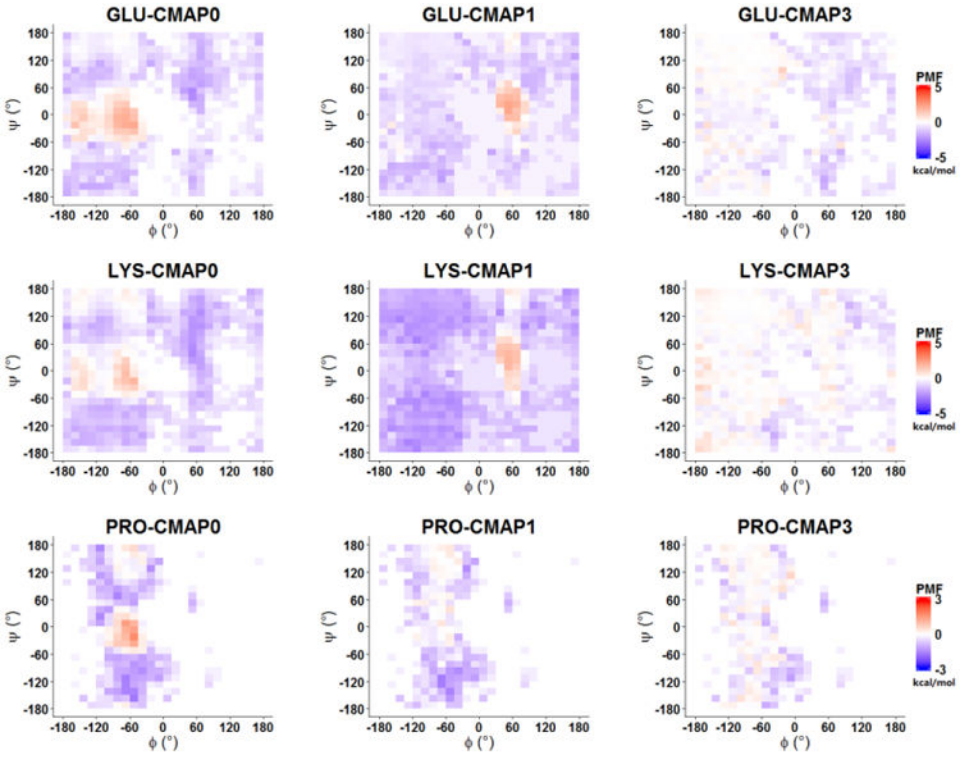

Figure 2.

The difference of dihedral distribution between benchmark of IDPs and CMAP correction for Glu, Lys and Pro.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

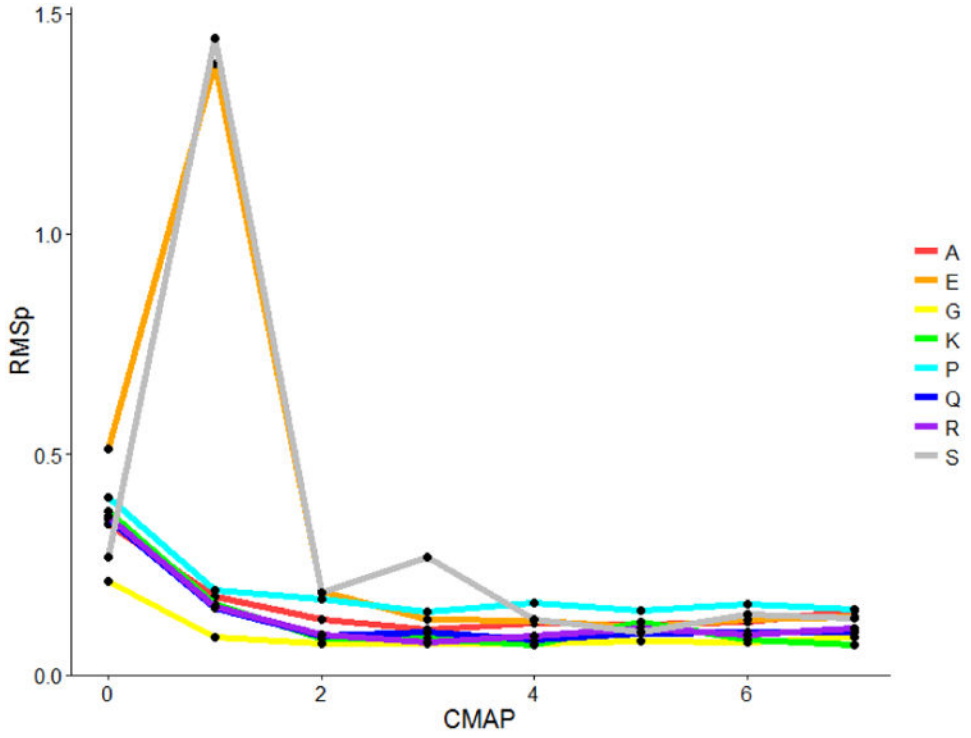

Figure 3.
RMSp of CMAP optimization for eight disorder-promoting amino acids.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

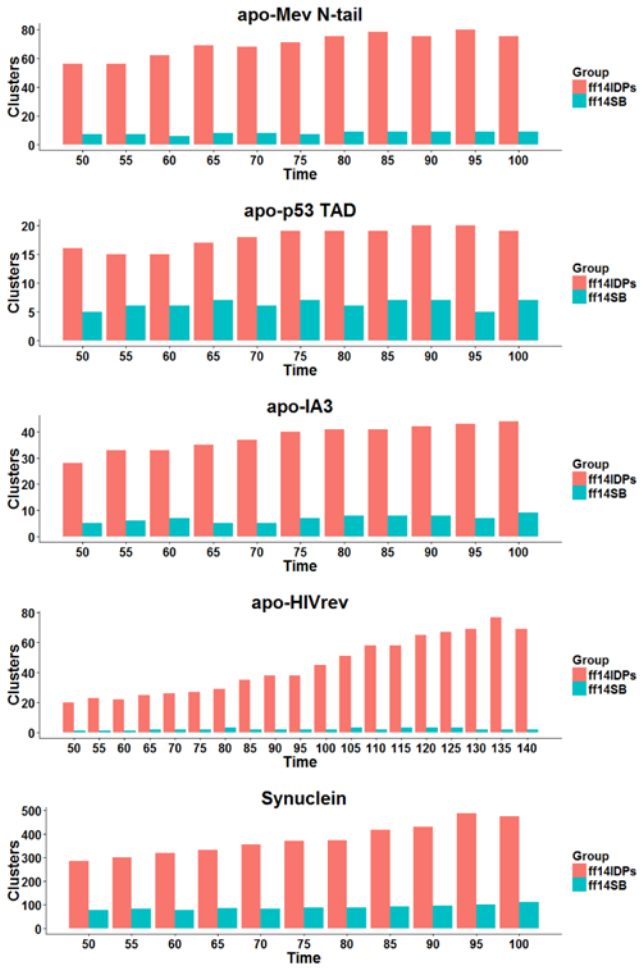

Figure 4.

Number of clusters occupying 70% or more conformations over increasing simulation time (checked every 5ns) with both tested force fields.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

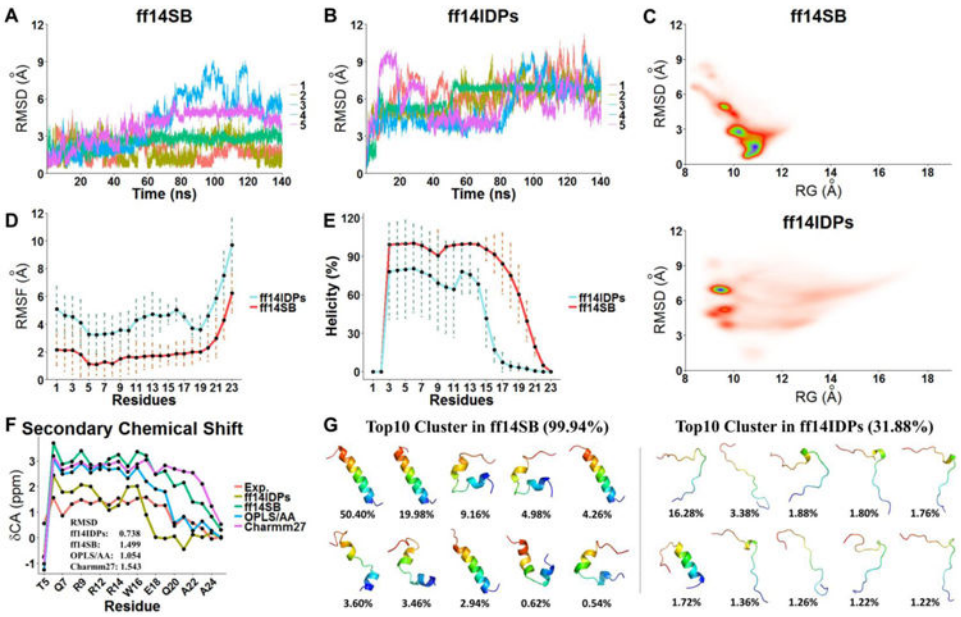

Figure 5.

Simulation and thermodynamic data derived from *ff14IDPs* and *ff14SB* for HIVRev. (A)(B) Cα RMSD for five trajectories. (C) PMF free energy landscape on 2D space of radius gyration (RG) and RMSD, showing *ff14IDPs* could sample wider and more flexible conformation space. (D) RMSF. (E) Comparison of the average helicity under both force fields. (F) Comparison of the secondary chemical shift data. (G) Representative structures of top 10 clusters and their occupancies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

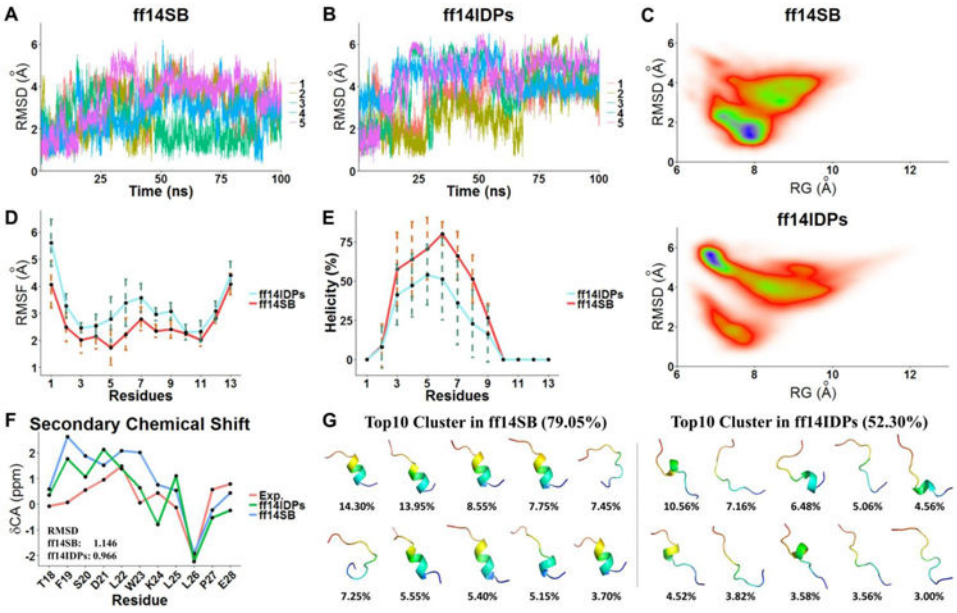

Figure 6.

Simulation and thermodynamic data derived from *ff14IDPs* and *ff14SB* for apo-p53. (A)(B) Cα RMSD for five trajectories. (C) PMF free energy landscape on 2D space of radius gyration (RG) and RMSD, showing *ff14IDPs* could sample wider and more flexible conformation space. (D) RMSF. (E) Comparison of the average helicity under both force fields. (F) Comparison of the secondary chemical shift data. (G) Representative structures of top 10 clusters and their occupancies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

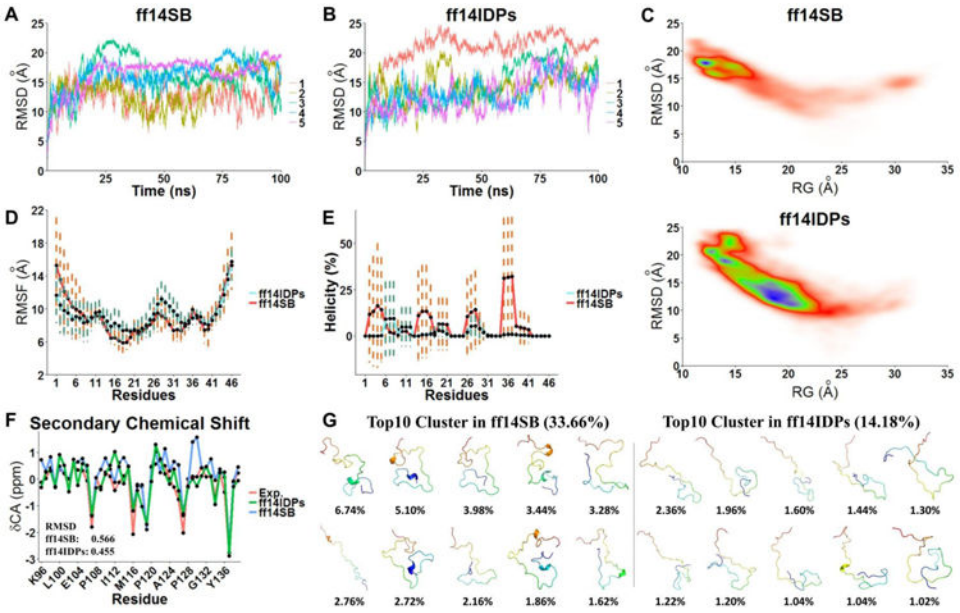

Figure 7.

Simulation and thermodynamic data derived from *ff14IDPs* and *ff14SB* for α-synuclein. (A) (B) Cα RMSD for five trajectories. (C) PMF free energy landscape on 2D space of radius gyration (RG) and RMSD, showing *ff14IDPs* could sample wider and more flexible conformation space. (D) RMSF. (E) Comparison of the average helicity under both force fields. (F) Comparison of the secondary chemical shift data. (G) Representative structures of top 10 clusters and their occupancies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

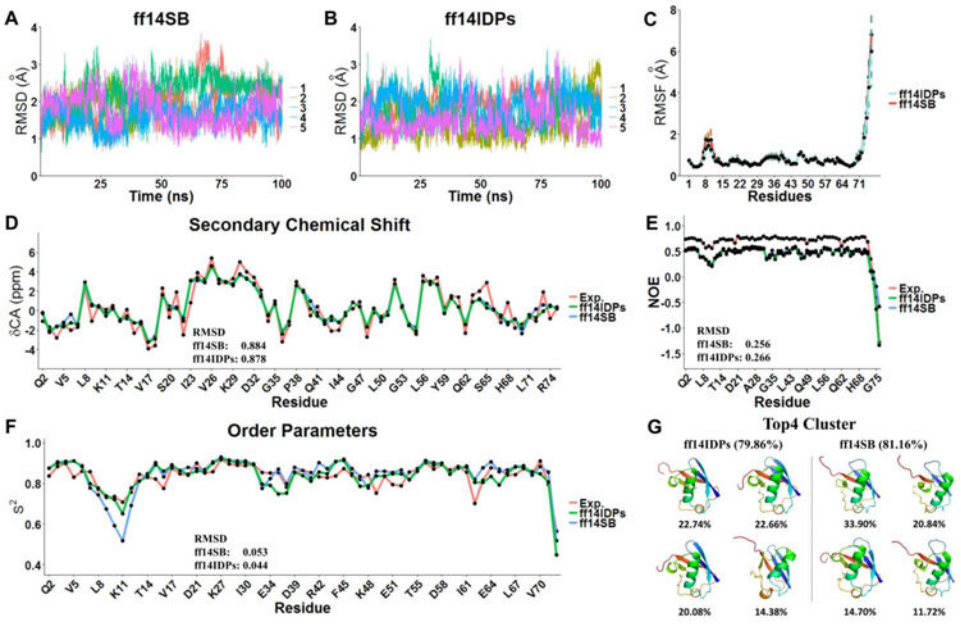

Figure 8.

Simulation and thermodynamic data derived from *ff14IDPs* and *ff14SB* for ubiquitin. (A)(B) Cα RMSD for five trajectories. (C) RMSF. (D) Comparison of the secondary chemical shift data. (E) Comparison of NOE data. (F) Comparison of order parameter (S^2). (G) Representative structures of top 4 clusters and their occupancies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Simulation conditions for all models.

System	Force field	Trajectories	Time (ns)	BMRB Index
Apo-HIVRev	*ff14SB*	5	140	18851
	ff14IDPs	5	140	
	OPLS	5	140	
	Charmm27	5	140	
Apo-p53 TAD	*ff14SB*	5	100	17760
	ff14IDPs	5	100	
Apo-MeV N-tail	*ff14SB*	5	100	6566
	ff14IDPs	5	100	
Bound-MeV N-tail	*ff14SB*	5	100	6567
	ff14IDPs	5	100	
Apo-IA_3	*ff14SB*	5	100	6078
	ff14IDPs	5	100	
Apo-αSyn	*ff14SB*	5	100	19337
	ff14IDPs	5	100	
Lysozyme	*ff14SB*	5	100	4562, 18304(S^2)
	ff14IDPs	5	100	
Ubiquitin	*ff14SB*	5	100	5387, 6470(NOE,S^2)
	ff14IDPs	5	100	