Biochemical and Biophysical Research Communications 458 (2015) 352–355

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

At least 10% shorter C–H bonds in cryogenic protein crystal structures than in current AMBER forcefields

Yuan-Ping Pang

Computer-Aided Molecular Design Laboratory, Mayo Clinic, Stabile 12-26, 200 First Street SW, Rochester, MN 55905, USA

ARTICLE INFO

Article history:
Received 21 January 2015
Available online 4 February 2015

Keywords:
Cryogenic crystal structures
Bond parameters
Force field
Chignolin
β-Hairpin
Protein folding

ABSTRACT

High resolution protein crystal structures resolved with X-ray diffraction data at cryogenic temperature are commonly used as experimental data to refine forcefields and evaluate protein folding simulations. However, it has been unclear hitherto whether the C–H bond lengths in cryogenic protein structures are significantly different from those defined in forcefields to affect protein folding simulations. This article reports the finding that the C–H bonds in high resolution cryogenic protein structures are 10–14% shorter than those defined in current AMBER forcefields, according to 3709 C–H bonds in the cryogenic protein structures with resolutions of 0.62–0.79 Å. Also, 20 all-atom, isothermal–isobaric, 0.5-μs molecular dynamics simulations showed that chignolin folded from a fully-extended backbone formation to the native β-hairpin conformation in the simulations using AMBER forcefield FF12SB at 300 K with an aggregated native state population including standard error of 10 ± 4%. However, the aggregated native state population with standard error reduced to 3 ± 2% in the same simulations except that C–H bonds were shortened by 10–14%. Furthermore, the aggregated native state populations with standard errors increased to 35 ± 3% and 26 ± 3% when using FF12MC, which is based on AMBER forcefield FF99, with and without the shortened C–H bonds, respectively. These results show that the 10–14% bond length differences can significantly affect protein folding simulations and suggest that re-parameterization of C–H bonds according to the cryogenic structures could improve the ability of a forcefield to fold proteins in molecular dynamics simulations.

1. Introduction

Since exposed protein side chains and surrounding water molecules are more ordered at cryogenic temperatures than at higher temperatures, cryogenic temperatures are now routinely used to obtain X-ray diffraction data [1]. The resulting cryogenic protein crystal structures are commonly used as experimental data to develop forcefield parameters and evaluate protein folding simulations [2–4]. Of 318 crystal structures [5] used to improve side-chain torsion parameters of the AMBER FF99SB forcefield [2], 117 have data collection temperatures described at the Protein Data Bank. Of the 117 structures, 47 (40%) are resolved with X-ray diffraction data collected at 90–150 K (Table 1). According to a report in 2011 [3], four (80%) out of five crystal structures that were used to evaluate protein folding simulations are low-temperature crystal structures (Table 1). In a report in 2014 [4], five (71%) out of seven crystal structures used for simulation evaluation are low-temperature structures as well (Table 1).

C–H bonds are critical to protein dynamics simulations as they directly affect the volume of a protein. In addition, as C–H bonds are typically constrained in molecular dynamics (MD) simulations, they are more prone than the bonds that are not constrained to the exaggeration of short-range repulsion caused by the 6–12 Lennard-Jones potential and a nonpolarizable charge model [6]. Interestingly, the C–H bonds in high-resolution, cryogenic protein crystal structures are visibly shorter than those defined in current AMBER forcefields (Fig. 1). However, until now it has been unclear whether this difference can significantly affect the ability of a forcefield to fold proteins in MD simulations.

This article reports the comparison of 3709 C–H bonds in the cryogenic protein structures with resolutions of 0.62–0.79 Å to those currently defined in AMBER forcefields in order to quantify the C–H bond length difference. It also reports 80 unique, independent, all-atom, isothermal–isobaric, and 0.5-μs MD simulations

Abbreviations: MD, molecular dynamics; CαβRMSD, Cα and Cβ root mean square deviation; NMR, nuclear magnetic resonance.

E-mail address: pang@mayo.edu.

http://dx.doi.org/10.1016/j.bbrc.2015.01.115

Table 1
Diffraction data collection temperatures of X-ray protein crystal structures used for forcefield development and evaluation of protein folding simulations.

Code	Res (Å)	Tem (K)	Code	Res (Å)	Tem (K)	Code	Res (Å)	Tem (K)	Code	Res (Å)	Tem (K)
Group 1			1gai	1.70	293	1yal	1.70	293	1cgh	1.80	293
1a2p	1.50	287	1gvp	1.60	285	1yge	1.40	100	1der	2.40	100
1a62	1.55	100	1hyt	1.70	298	1yna	1.55	298	1ema	1.90	295
1a8u	1.60	90	1ido	1.70	100	2arc	1.50	100	1exf	2.10	293
1aay	1.60	295	1irn	1.20	298	2hft	1.69	100	1ft1	2.25	96
1ab1	0.89	150	1ixh	0.98	100	2ilk	1.60	100	1fur	1.95	298
1ab9	1.60	292	1jcv	1.55	93	2wea	1.25	110	1hav	2.00	295
1agj	1.70	273	1jer	1.60	300	3bto	1.66	100	1kel	1.90	100
1ah7	1.50	300	1jev	1.30	120	3cyr	1.60	298	1kzu	2.50	300
1aho	0.96	287	1jhg	1.30	287	3lck	1.70	110	1lmb	1.80	258
1aie	1.50	293	1jsf	1.15	285	3nul	1.60	100	1ppr	2.00	283
1ajj	1.70	93	1llp	1.70	277	3vub	1.40	300	1slu	1.80	298
1ako	1.70	292	1lt5	1.70	295	4mbp	1.70	287	1tdt	2.20	293
1akz	1.57	275	1mh1	1.38	101	4pga	1.70	293	1wba	1.80	286
1am3	1.70	100	1mrp	1.60	292	5nul	1.60	150	1wej	1.80	90
1amm	1.20	150	1msi	1.25	277	6cel	1.70	120	2 fha	1.90	293
1aoe	1.60	295	1mty	1.70	100	6fd1	1.35	100	3bct	2.10	100
1aop	1.60	277	1myr	1.64	100	8ruc	1.60	283	5cro	2.30	290
1aqz	1.70	298	1nls	0.94	110	1a17	2.45	100	7ahl	1.90	287
1ar5	1.60	295	1not	1.20	287	1a28	1.80	100	Group 2		
1awd	1.40	103	1nox	1.59	283	1a31	2.10	100	2f21	1.50	100
1bfd	1.60	277	1nwp	1.60	298	1a4y	2.00	289	2hba	1.25	100
1bkr	1.10	100	1pdo	1.70	277	1a8m	2.30	290	1lmb	1.80	258
1bpi	1.10	125	1pen	1.10	289	1acc	2.10	100	2f4k	1.05	95
1brt	1.50	283	1qoa	1.70	277	1af0	1.80	290	1mi0	1.85	298
1btk	1.60	100	1ra9	1.55	298	1aik	2.00	100	Group 3		
1ctj	1.10	293	1rhs	1.36	100	1aim	2.00	288	cln025	1.11	290
1dos	1.67	293	1rie	1.50	100	1air	2.20	291	2f21	1.50	100
1edm	1.50	293	1tal	1.50	120	1aly	2.00	123	2hba	1.25	100
1fdr	1.70	277	1tfe	1.70	113	1aol	2.00	100	1mi0	1.85	298
1fds	1.70	290	1vie	1.70	277	1ba7	2.50	291	1whz	1.52	100
1fvk	1.70	289	1wer	1.60	100	1baz	1.90	293	1lmb	1.80	258
1g3p	1.46	100	1whi	1.50	295	1beo	2.20	277	1qys	2.50	100

Group 1: the 117 structures that have data collection temperatures disclosed at the Protein Data Bank and were used to develop improved side-chain torsion parameters of the AMBER FF99SB-ILDN forcefield [2]. Group 2: the five structures used to evaluate the protein folding simulations described in Ref. [3]. Group 3: the seven structures used to evaluate the protein folding simulations described in Ref. [4].

to determine the effect of the quantified bond length difference on folding chignolin, one of the smallest, fast-folding proteins [7].

2. Methods

2.1. Molecular dynamics simulations to fold chignolin

Chignolin in the anti-parallel β-strand conformation was surrounded by two sodium ions and three sodium chloride molecules and solvated with 1281 TIP3P water molecules [8] to keep the closest distance between any atom of chignolin and the edge of the periodic solvent box at 8.2 Å using LEAP of AmberTools 1.5 (University of California, San Francisco). The anti-parallel β-strand conformation was generated by MacPyMOL Version 1.5.0 (Schrödinger LLC, Portland, OR). Because pH 5.5 was used for the nuclear magnetic resonance (NMR) structure determination for chignolin [7], two sodium ions were added for neutrality of the protein. Three sodium chloride molecules were also added to keep the ionic strength of the system close to physiological ionic strength at 150 mM NaCl. The neutralized, slightly brined, and solvated chignolin was then energy-minimized for 100 cycles of steepest-descent minimization followed by 900 cycles of conjugate-gradient minimization to remove close van der Waals contacts using SANDER of AMBER 11 (University of California, San Francisco) with AMBER forcefield FF12SB or FF12MC (see Section 3.2 for information of the two forcefields), heated from 0 to 300 K at a rate of 10 K/ps under constant temperature and constant volume, and finally simulated in 20 unique, independent, and 0.5-μs MD simulations using PMEMD of AMBER 11 with a periodic boundary condition at a constant temperature of 300 K and a constant pressure of 1 atm with isotropic molecule-based scaling. The 20 unique seed numbers for initial velocities of Simulations 1–20 are 1804289383, 846930886, 1681692777, 1714636915, 1957747793, 424238335, 719885386, 1649760492, 596516649, 1189641421, 1025202362, 1350490027, 783368690, 1102520059, 2044897763,

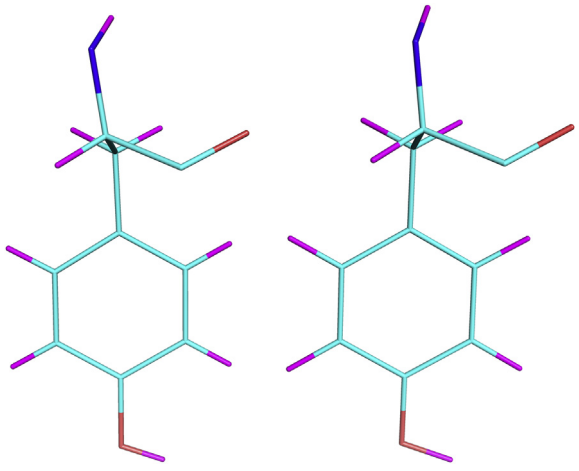

Fig. 1. Shorter C–H bonds in a cryogenic protein crystal structure than in AMBER forcefield FF14SB. Left: Residue 73 of a high resolution crystal structure of type-III antifreeze protein RD1 (Protein Data Bank ID: 1UCS). Right: The corresponding structure defined in FF14SB.

1967513926, 1365180540, 1540383426, 304089172, and 1303455736, respectively. These all-atom, isothermal–isobaric MD simulations used (1) a dielectric constant of 1.0, (2) the Berendsen coupling algorithm [9], (3) the Particle Mesh Ewald method to calculate long-range electrostatic interactions [10], (4) a time step of 1.0 fs, (5) SHAKE-bond-length constraints applied to all the bonds involving the H atom, (6) a protocol to save the image closest to the middle of the "primary box" to the restart and trajectory files, (7) a formatted restart file, and (8) default values of all other inputs of PMEMD. Each simulation was performed on either an Apple Xserve with two G5 processors (2.2/2.4 GHz) or a 12-core Apple Mac Pro with Intel Westmere (2.40/2.93 GHz).

2.2. Aggregated native state population calculation

The native β-hairpin conformation of chignolin in the NMR structure has Tyr2 and Trp9 on the same side of the hairpin [7]. Analysis of the trajectories obtained from MD simulations using FF12SB at 300 K showed that chignolin could fold to native-like β-hairpins with Tyr2 on one side of the hairpin and Trp9 on the other side. The lowest Cα and Cβ root mean square deviation (CαβRMSD) between one of the native-like β-hairpins and the NMR structure is 1.99 Å, whereas the corresponding Cα root mean square deviation is 1.58 Å. To distinguish the native β-hairpin from the native-like ones, conformations with CαβRMSDs of ≤1.96 Å relative to the NMR structure are considered to be at the native or folded state. The native state population of chignolin in an MD simulation was calculated as the number of the native state conformation divided by the number of all chignolin conformations saved at 100-ps intervals. Averaging the native state populations of a set of 20 unique and independent MD simulations gave rise to the aggregated native state population for the set. The standard deviation and error of the aggregated native state population were calculated according to Eqs. (1) and (2), respectively, wherein N is the number of all simulations, P_i is the native state population of the ith simulation, and $\overline{P}$ is the aggregated native state population. The aggregated native state population with its standard deviation and error is introduced here to assess the convergence of a set of simulations and for comparison of the relative abilities of different forcefields to fold a protein in multiple MD simulations.

$$\text{Standard Deviation} = \sqrt{\sum \left(P_i - \overline{P}\right)^2/(N-1)} \quad (1)$$

$$\text{Standard Error} = \text{Standard Deviation}\Big/\sqrt{N} \quad (2)$$

2.3. Root mean square deviation calculation

CαβRMSDs and Cα root mean square deviations were calculated automatically using PTRAJ of AmberTools 1.5 or manually using ProFit V2.6 (http://www.bioinf.org.uk/software/profit/).

3. Results and discussion

3.1. Shorter C–H bonds in high resolution cryogenic protein crystal structures

There are two types of C–H bonds in a protein. One type includes aromatic C–H bonds that are hydrogen-containing bonds involving carbon with sp^2 hybridization (denoted as C^{sp^2}–H). The other covers aliphatic C–H bonds that comprise hydrogen and carbon with sp^3 hybridization and has two subtypes. One subtype has aliphatic C–H bonds at the main chain of a protein (denoted as $C\alpha^{sp^3}$–H), and the other has those at the side chain (denoted as C^{sp^3}–H).

Analysis of 3709 C–H bonds in nine cryogenic protein crystal structures with resolutions of 0.62–0.79 Å showed that the average $C\alpha^{sp^3}$–H and C^{sp^3}–H bonds in these structures are 10% and 12% shorter than those defined in AMBER forcefields FF99 [11], FF12SB (AmberTools 13 reference manual, 27–29), and FF14SB (AMBER 14 reference manual, 29–31), respectively (Table 2). For C^{sp^2}–H bonds that are slightly shorter than the aliphatic ones as defined in current AMBER forcefields (1.08 Å versus 1.09 Å), the average bond length in the cryogenic structures is 14% shorter than the one presently defined (Table 2).

Given the 1% difference between the aliphatic and aromatic C–H bonds defined in the AMBER forcefields (1.09 Å versus 1.08 Å), it is reasonable to assume that the 10–14% bond length difference can significantly affect MD simulations of protein folding.

3.2. The effect of the C–H bond length difference on protein folding simulations

To test the above assumption, a set of 20 unique, independent, all-atom, isothermal–isobaric, and 0.5-μs MD simulations of chignolin in the anti-parallel conformation was carried out at 300 K and 1 atm using FF12SB to autonomously fold the extended backbone conformation to the native β-hairpin conformation. As a control, the aggregated native state population of this set with its standard error was found to be 10 ± 4% (Table 3), which demonstrates that FF12SB with the standard aliphatic and aromatic C–H bond lengths (1.09 Å and 1.08 Å, respectively) has the ability to fold chignolin at 300 K and 1 atm in MD simulation within the aggregated timescale of 10 μs. When this set of simulations was repeated under the same simulation conditions except that the aliphatic (including both $C\alpha^{sp3}$–H and C^{sp3}–H bonds) and aromatic C–H bonds were shortened to 0.98 Å and 0.93 Å, respectively, the aggregated native state population with its standard error reduced to 3 ± 2% (Table 3), which is significantly smaller than the population of the simulations using standard C–H bonds (10 ± 4%).

To further support the assumption, the two sets of simulations described above were then repeated using AMBER forcefield FF12MC with and without the shortened C–H bonds. Developed by this author, FF12MC is based on AMBER forcefield FF99 with changes of (i) reducing atomic masses systemically by tenfold to

Table 2
C–H bond lengths defined in AMBER forcefields and observed in high-resolution cryogenic protein crystal structures.

Source	Aliphatic bond				Aromatic bond	
	$C\alpha^{sp^3}$–H		C^{sp^3}–H		C^{sp^2}–H	
	(Å)	Δ (%)[a]	(Å)	Δ (%)[a]	(Å)	Δ (%)[a]
Nine crystal structures[b]	0.98[c]	–	0.96[d]	–	0.93[e]	–
FF99, FF12SB, FF14SB	1.09	10	1.09	12	1.08	14
FF12MC	0.98	0	0.98	2	0.93	0

[a] $\Delta = (\text{bond}_{\text{forcefield}} - \text{bond}_{\text{crystal\_structure}})/\text{bond}_{\text{forcefield}}$.
[b] The Protein Data Bank IDs along with resolutions and X-ray diffraction data collection temperatures in parentheses of nine selected protein crystal structures are 1UCS (0.62 Å; 110 K), 2VB1 (0.65 Å; 100 K), 1YK4 (0.69 Å; 100 K), 3A38 (0.70 Å; 93 K), 2B97 (0.75 Å; 100 K), 1HJE (0.75 Å; 150 K), 1GCI (0.78 Å; 100 K), 1X6Z (0.78 Å; 100 K), and 2PVE (0.79 Å; 100 K).
[c] Average of 1140 $C\alpha^{sp^3}$–H bonds calculated from all residues in the nine structures.
[d] Average of 2230 C^{sp^3}–H bonds calculated from side chains of Ala, Val, Leu, Ile, and Thr in the nine structures.
[e] Average of 339 C^{sp^2}–H bonds calculated from His, Phe, Tyr, and Trp in the nine structures.

Table 3
Folding of chignolin in 20 0.5-μs isothermal–isobaric molecular dynamics simulations at 300 K and 1 atm using different C–H bond lengths.

Forcefield	C–H bond length	Aggregated native state population (%)		
		Mean	Standard deviation	Standard error
FF12SB	Standard	10	19	4
FF12SB	Short	3	10	2
FF12MC	Standard	26	11	3
FF12MC	Short	35	12	3

Aggregated native state population: The number of chignolin conformations with CαβRMSDs of ≤1.96 Å divided by the number of all chignolin conformations from the 20 simulations. Standard: 1.09 Å for $C\alpha^{sp^3}$–H and C^{sp^3}–H bonds; 1.08 Å for C^{sp^2}–H bond. Short: 0.98 Å for $C\alpha^{sp^3}$–H and C^{sp^3}–H bonds; 0.93 Å for C^{sp^2}–H bond.

improve configurational sampling, according to the report that low-mass MD simulation enhances configurational sampling [12]; (ii) shortening C–H bond lengths (1.09 Å to 0.98 Å for the aliphatic; 1.08 Å to 0.93 Å for the aromatic) to reduce the exaggeration of short-range repulsion caused by the 6–12 Lennard-Jones potential and a nonpolarizable charge model [6] without resorting to the 6–exponential potential and a polarizable charge model according to the results above; (iii) zeroing torsion potentials involving a nonperipheral sp^3 atom to eliminate effects of shortening C–H bonds on related torsion potentials; (iv) reducing the 1–4 interaction scaling factors of protein backbone torsions ϕ and ψ (from 2.00 to 1.00 for the van der Waals interaction; from 1.20 to 1.18 for the electrostatic interaction) to balance the forcefield propensities for adopting secondary structure elements, according to the report that 1–4 interaction scaling factors control the conformational equilibrium between α-helix and β-strand [13]. The aggregated native state population with its standard error was 26 ± 3% when using FF12MC with standard C–H bonds, whereas the population with its standard error increased to 35 ± 3% when using FF12MC with the shortened C–H bonds (Table 3). These results confirm that the 10–14% bond length difference can significantly affect protein folding simulations.

3.3. The benefit of shortening C–H bonds for protein folding simulations

The present work shows that FF12SB with either standard or shortened C–H bonds is able to fold chignolin in 20 all-atom, isothermal–isobaric, and 0.5-μs MD simulations at 300 K with an aggregated native state population including its standard deviation of 10 ± 19% or 3 ± 10%, respectively (Table 3). As indicated by the large standard deviations relative to the means, the aggregated native state populations of the two sets of simulations using FF12SB are not well converged at the aggregated timescale of 10 μs. Under the same simulation conditions, FF12MC, with either the standard or the shortened bonds, is able to fold chignolin with 26 ± 11% or 35 ± 12%, respectively (Table 3). Here, the small standard deviations indicate the convergences of the two sets of simulations using FF12MC at the aggregated timescale of 10 μs.

Taken together, the results show that FF12MC can fold chignolin more efficiently than FF12SB and suggest that FF12MC can reduce the simulation timescale required to capture folding events. More importantly, these results show that shortening C–H bonds by up to 14% can increase the aggregated native state population of the simulations using FF12MC with standard C–H bonds by ~35% (Table 3). The ~35% increase is presumably due to the constraint applied to hydrogen-containing bonds in MD simulations that makes these bonds inflexible and more prone than the bonds that are not constrained to the exaggeration of short-range repulsion caused by the 6–12 Lennard-Jones potential and a nonpolarizable charge model [6]. It is conceivable that shortening C–H bonds can reduce the repulsion exaggeration and hence shorten the simulation timescale for capturing protein folding events. Therefore, these findings suggest that re-parameterization of C–H bonds according to cryogenic protein crystal structures could improve the ability of a forcefield to fold proteins in MD simulations.

Conflict of interest

The author declares no conflict of interest.

Acknowledgments

Yuan-Ping Pang acknowledges the support of this work from the US Defense Advanced Research Projects Agency (DAAD19-01-1-0322), the US Army Medical Research Material Command (W81XWH-04-2-0001), the US Army Research Office (DAAD19-03-1-0318 and W911NF-09-1-0095), the US Department of Defense High Performance Computing Modernization Office, and the Mayo Foundation for Medical Education and Research. The contents of this article are the sole responsibility of the author and do not necessarily represent the official views of the funders. The author also appreciates the comments from an anonymous reviewer.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.bbrc.2015.01.115.

References

[1] J.S. Richardson, The anatomy and taxonomy of protein structure, Adv. Protein Chem. 34 (1981) 167–339.
[2] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J.L. Klepeis, R.O. Dror, D.E. Shaw, Improved side-chain torsion potentials for the AMBER ff99SB protein force field, Proteins 78 (2010) 1950–1958.
[3] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science 334 (2011) 517–520.
[4] H. Nguyen, J. Maier, H. Huang, V. Perrone, C. Simmerling, Folding simulations for proteins with diverse topologies are accessible in days with physics-based force field and implicit solvent, J. Am. Chem. Soc. 136 (2014) 13959–13962.
[5] S.C. Lovell, J.M. Word, J.S. Richardson, D.C. Richardson, The penultimate rotamer library, Proteins 40 (2000) 389–408.
[6] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz Jr., D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, J. Am. Chem. Soc. 117 (1995) 5179–5197.
[7] S. Honda, K. Yamasaki, Y. Sawada, H. Morii, 10 residue folded peptide designed by segment statistics, Structure 12 (2004) 1507–1518.
[8] W.L. Jorgensen, J. Chandreskhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926–935.
[9] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. Di Nola, J.R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys. 81 (1984) 3684–3690.
[10] T.A. Darden, D.M. York, L.G. Pedersen, Particle mesh Ewald: an N log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089–10092.
[11] J.M. Wang, P. Cieplak, P.A. Kollman, How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comput. Chem. 21 (2000) 1049–1074.
[12] Y.-P. Pang, Low-mass molecular dynamics simulation: a simple and generic technique to enhance configurational sampling, Biochem. Biophys. Res. Commun. 452 (2014) 588–592.
[13] Y.-P. Pang, Use of 1–4 interaction scaling factors to control the conformational equilibrium between α-helix and β-strand, Biochem. Biophys. Res. Commun. 457 (2015) 183–186.