J Biomol NMR (2015) 61:35–45
DOI 10.1007/s10858-014-9879-2

ARTICLE

Correlation of chemical shifts predicted by molecular dynamics simulations for partially disordered proteins

Jerome M. Karp · Ertan Erylimaz ·
David Cowburn

Received: 19 June 2014 / Accepted: 17 November 2014 / Published online: 22 November 2014

Abstract There has been a longstanding interest in being able to accurately predict NMR chemical shifts from structural data. Recent studies have focused on using molecular dynamics (MD) simulation data as input for improved prediction. Here we examine the accuracy of chemical shift prediction for intein systems, which have regions of intrinsic disorder. We find that using MD simulation data as input for chemical shift prediction does not consistently improve prediction accuracy over use of a static X-ray crystal structure. This appears to result from the complex conformational ensemble of the disordered protein segments. We show that using accelerated molecular dynamics (aMD) simulations improves chemical shift prediction, suggesting that methods which better sample the conformational ensemble like aMD are more appropriate tools for use in chemical shift prediction for proteins with disordered regions. Moreover, our study suggests that data accurately reflecting protein dynamics must be used as input for chemical shift prediction in order to correctly predict chemical shifts in systems with disorder.

Keywords Chemical shift prediction · Molecular dynamics simulation · Accelerated MD · Intein · Partially disordered proteins

Introduction

NMR chemical shifts are sensitive reporters of the local electronic environments of molecules, and there is a longstanding interest in using them to derive structural information in biopolymers, especially proteins. They are used to probe protein secondary structure (Spera and Bax 1991; Wishart 2011), overall fold (Cavalli et al. 2007; Menon et al. 2013; Shen et al. 2008), protein–protein interactions (Dominguez et al. 2003; Montalvao et al. 2008), protein–ligand interactions (Medek et al. 2000), and protein dynamics (Camilloni et al. 2012; Eisenmesser et al. 2005; Mittermaier and Kay 2006). The long-term interest in how well these experimental quantities can be predicted has been enhanced by approaches including empirical machine-learning, using structure and/or sequence data (Han et al. 2011; Kohlhoff et al. 2009; Shen and Bax 2010), as well as quantum chemical methods (Tang and Case 2011).

In addition to producing algorithms which can best predict chemical shift values from a given set of coordinates, there has also been an increased focus on generating input for these programs which will allow for the best chemical shift prediction. Most chemical shift prediction software uses PDB files as the input from which chemical shifts are predicted (Han et al. 2011; Kohlhoff et al. 2009; Neal et al. 2003; Shen and Bax 2007, 2010; Xu and Case 2001, 2002). In addition, many machine-learning algorithms are trained and tested on a set of representative PDB structures (Han et al. 2011; Shen and Bax 2007, 2010). However, most PDB structures, including those generally used to train machine-learning algorithms, are generated from static X-ray crystal structures, since it is believed that NMR structures are less accurate and less precise than crystal structures (Han et al. 2011). This may degrade chemical shift prediction since proteins are not static structures but dynamic entities, and NMR chemical shifts reflect the dynamic quality of the protein (Cavalli et al. 2007). Indeed, it has long been recognized that a

J. M. Karp · E. Erylimaz · D. Cowburn (✉)
Department of Biochemistry, Albert Einstein College of Medicine of Yeshiva University, 1300 Morris Park Avenue, Bronx, NY 10461, USA
e-mail: cowburn@cowburnlab.org;
David.cowburn@einstein.yu.edu

conformational ensemble may be necessary to accurately predict NMR chemical shifts (Han et al. 2011; Lehtivarjo et al. 2009), though clarification is necessary regarding how large an ensemble is needed to produce sufficiently accurate results. Several recent studies have examined the effect of using molecular dynamics (MD) simulation-derived ensembles as input for chemical shift prediction programs, noting that these ensembles lead to better predictions of chemical shifts than predictions based on a static crystal structure (Lehtivarjo et al. 2012; Markwick et al. 2010; Robustelli et al. 2012). An explanation for this finding is that the MD simulation ensemble reflects a dynamic picture of the protein, which better fits the NMR chemical shift values than a static X-ray crystal structure.

In addition to improving the accuracy of chemical shift prediction, MD simulation input to chemical shift prediction software might be useful in analysis of changes in shift patterns over the course of a simulation. This could allow for discrimination between populations of conformers and structures, possibly of biological significance, e.g., in catalysis. Examples of systems that are sufficiently well-studied by NMR and by MD simulation and which could benefit from such analysis include ribonucleases (Camilloni et al. 2013; Robustelli et al. 2012), protein kinases (Chen et al. 2013; Shan et al. 2012), proteases (Bondar et al. 2009; Kipp et al. 2012), and GPCRs (Deupi and Kobilka 2010; Kim et al. 2013). We sought to investigate the possible use of this approach to characterize the structural ensembles for the single turnover reaction involved in intein protein splicing, where significant dynamic structures are involved in the splicing reaction (Shah et al. 2013a) and in association of the split intein segments (Eryilmaz et al. 2014; Shah et al. 2013b) for the intein Npu (Iwai et al. 2006; Ramirez et al. 2013; Zettler et al. 2009). The rate of splicing in the Npu intein is dependent on the extein sequences, and it is suggested, based on ~500 ns MD simulations, NMR characterization, and mutations, that selection among a small number of conformers may take place (Shah et al. 2013a). For the self-association of the fragments of Npu, one split intein fragment is partly folded, while the other is completely disordered. These polypeptides capture each other through their disordered regions and form an ordered intermediate with native-like structure at their interface. This intermediate then collapses into the canonical intein fold (Shah et al. 2013b). The potential predictive power of combining shift information and MD simulations for equilibria characterization shown by others, e.g., (Camilloni et al. 2012; Robustelli et al. 2012, 2013), encouraged attempts to investigate Npu, a system with relatively complex dynamic properties (Eryilmaz et al. 2014).

In this paper, we test the effect of using ensemble input to chemical shift prediction programs in the case of this partially intrinsically disordered protein (IDP), which differs from previous studies regarding systems with predominantly high degrees of structural order. IDPs have large regions without well-defined secondary structure, and chemical shifts are widely used to characterize the degree of disorder and/or propensity to secondary structure (Ball et al. 2014; Jensen et al. 2009; Tamiola et al. 2010). Thus, we expect that a substantial time-dependent ensemble input would be necessary to best account for their dynamic features when predicting chemical shift values, assuming adequate sampling (Wereszczynski and McCammon 2012). We examined two related proteins: Npu, the DnaE intein from *Nostoc punctiforme* (Oeemig et al. 2009) described above; and the RadA intein from *Pyrococcus horikoshii* (Oeemig et al. 2012). These inteins are a protein class predominantly from *Eubacteria* and *Archea* (but also found in *Eukarya*) (Perler 2002) which are able to excise themselves from a larger precursor protein following translation (Paulus 2000), producing also a new combination protein from their N- and C-terminal extensions. We use these intein systems, which have several long unstructured regions, to study chemical shift prediction in a system with high disorder.

Methods

NMR spectroscopy

Synthesis/expression and purification of Npu DnaE split intein constructs, NpuC and NpuN, were described in detail elsewhere (Shah et al. 2013a). Uniform isotope labeling of intein fragments was achieved by growing expression cultures in minimal media (M9) supplemented with [U-^{13}C]-glucose and $^{15}NH_4Cl$.

For all NMR experiments the sample concentrations were 250 μM in NMR buffer (25 mM sodium phosphates, 100 mM NaCl, 1 mM DTT, pH 6.5). The experiments were carried out at 25 °C using Bruker (800 or 900 MHz) spectrometers equipped with cryogenic probes and capable of applying pulse field gradients along the z-axis. The initial processing was done using NMRPipe (Delaglio et al. 1995) and NMRViewJ (Johnson 2004) was used for the analysis and resonance assignments.

The backbone resonances of NpuC and NpuN constructs were assigned as described elsewhere (Shah et al. 2013a). Briefly, the constructs were assigned using ^{15}N,^{13}C-labeled fully protonated samples with standard triple resonance experiment pairs.

Molecular dynamics simulations

Molecular dynamics (MD) simulations were performed using AMBER (Case et al. 2005). The starting structure was the first structure of the NMR ensemble for the fused DnaE intein from *Nostoc punctiforme* (PDB ID: 2keq) (Oeemig et al. 2009). Prior to simulations, the structure was modified in silico to mimic the two-fragmented split intein complex with the canonical extein residues (AEY-NpuN; NpuC-CFN). The requisite number of counterions was added to neutralize the protein charges, and approximately 8,000 molecules of TIP3P water were added. The structure was minimized for 500 steps of steepest descent minimization and 500 steps of conjugate gradient minimization, holding the intein complex fixed. The structure was heated to 300 K using a Langevin thermostat over 20 ps, and then equilibrated at constant pressure (1.0 atm) for another 20 ps. Equilibration was subsequently performed for 100 ps. The equilibrated structure was used as the starting structure for MD simulation. MD was run with the AMBER99sb force field for over 200 ns using a 1 fs time step. Temperature was controlled via a Langevin thermostat with a collision rate of 5 ps^{-1}, and pressure scaling was used with a relaxation time of 2 ps to maintain the pressure at 1.0 atm. Non-bonded interactions were calculated using a cutoff of 8 Å, and long-range interactions were calculated using the Particle mesh Ewald method (Essmann et al. 1995). Hydrogen bonds were constrained using the SHAKE algorithm (Ryckaert et al. 1977). The coordinates of the intein were extracted every 5 ps, for a total of 40,270 frames.

Accelerated molecular dynamics

Accelerated molecular dynamics (aMD) (Hamelberg et al. 2004) was run using the same starting structure used for standard molecular dynamics runs. The protocol used to prepare the structure for aMD was identical to that used for MD simulation. At each step in aMD, the potential energy due to dihedral angles as well as the total potential energy is calculated, and a boost is added to the dihedral angle potential, with a second boost added based on the total potential energy. The boost to the dihedral angle potential energy is given by

$$\Delta V_D = \frac{(E_D - V)^2}{\alpha_D + (E_D - V)} \qquad (1)$$

where V is the potential energy and α_D and E_D are constants. The boost to the entire potential is then given by

$$\Delta V_P = \frac{(E_P - (V + \Delta V_D))^2}{\alpha_P + (E_P - (V + \Delta V_D))} \qquad (2)$$

The equilibration run was used to calculate reasonable constants for the aMD run, as suggested in the AMBER 12 manual (Case et al. 2012). We used $E_D = 1{,}967$ kcal/mol, $E_P = -78{,}656$ kcal/mol, $\alpha_D = 100$, and $\alpha_P = 5{,}372$. The aMD simulation was run for 200 million steps, each 1 fs, for a total of 200 ns. Coordinates and energies were extracted every 5 ps, yielding 40,000 frames.

To analyze the resulting data, principal component analysis (PCA) was performed on the Cα atom coordinates of the N-intein in each frame. We then selected frames for which $\frac{\Delta V}{k_B T}$ was >110. This produced 415 frames, which were then clustered using a complete-linkage clustering method which measured distance between two points based on their distance on the plane containing the first two PCA axes. Once 12 clusters were formed, the frame in each cluster with the lowest potential energy was selected and used as the starting point of a 1 ns standard MD simulation, performed using the parameters described above. The resulting frames were used as input for chemical shift prediction, and then these shifts were averaged to produce a final prediction, weighted based on how many frames were in the given cluster. We did not directly calculate chemical shift predictions from the aMD simulation using weights based on the potential boost, since this weighting scheme causes significant overrepresentation of low-energy states (Markwick et al. 2010; Shen and Hamelberg 2008).

The same process was repeated for the RadA intein of *Pyrococcus horikoshii* (Oeemig et al. 2012), for which there is a known X-ray structure (4e2t) and NMR ensemble (2lqm) with chemical shifts deposited in the Biological Magnetic Resonance Bank. Missing residues were added to the structure using MODELLER (Fiser et al. 2000). For this system, a shorter MD simulation was run, with data extracted every 5 ps, producing 4,000 frames. An aMD simulation was also run for 20 ns, yielding 4,000 frames. For this simulation, we used the parameters $E_D = 2{,}425$ - kcal/mol, $E_P = -72{,}180$ kcal/mol, $\alpha_D = 122$, and $\alpha_P = 5{,}026$. We selected frames for which $\frac{\Delta V}{k_B T}$ was >90, using a lower threshold than for the DnaE intein due to the low number of frames which would be selected by the higher threshold. The resulting frames were clustered into 20 clusters by the method above to run short MD simulations as described above. The results of these MD simulations were used to produce a final prediction.

Principal component analysis (PCA) was used to investigate the relative increase in conformational space visited by aMD simulation. For this analysis, we combined the trajectories of MD and aMD simulations together and performed PCA on the Cα coordinates of the N-intein.

Chemical shift prediction

Chemical shifts were predicted using SHIFTX2 version 1.07 (Han et al. 2011) and SPARTA+ version 2.80 (Shen and Bax 2010). In the SHIFTX2 program, sequence information is not used in the predictions, such that the predictions are identical to those of the SHIFTX+ program.

Results

We ran a long MD simulation (40,270 frames, >200 ns) of the NpuN intein solvated in water. The initial coordinates of the protein were extracted from the first member of the experimentally-derived NMR ensemble. We ran SHIFTX2 and SPARTA+ on each frame extracted from the simulation. The chemical shift predictions for these 40,270 sets of coordinates were then linearly averaged to make a final prediction for the ^{13}Cα, ^{13}Cβ, ^{13}C′, ^{1}H′ and ^{15}N chemical shifts. The timescale of motions in biological macromolecules spans picoseconds to seconds; hence to overcome the short timescale limitation of MD simulations and to study a more "complete" conformational space we also employed accelerated MD (aMD) simulations. We ran a long aMD simulation (40,000 frames, 200 ns) of the NpuN intein and used the resulting coordinate data for a chemical shift prediction using SHIFTX2 and SPARTA+ (see "Methods"). In addition, we ran SHIFTX2 and SPARTA+ on the coordinates from the experimentally-derived NMR ensemble, linearly averaging the predictions to produce a final prediction. We then compared these predictions with the experimentally obtained chemical shifts. Figure 1 shows a representative group of histograms of chemical shift predictions using MD and aMD. Figure 2 summarizes the simulation results in the MD and aMD simulations.

Table 1 shows the RMSDs from observed chemical shifts of the chemical shift predictions derived from standard MD simulations, accelerated MD simulations, and from the experimentally-derived NMR ensemble. It is apparent that accelerated MD simulations allow for a modest improvement in chemical shift prediction compared to standard MD simulations. Moreover, the accelerated MD produces chemical shift predictions similar in accuracy to those produced from the NMR ensemble.

To analyze the differences between the simulation techniques, we ran principal component analysis on the Cα coordinates of NpuN in the MD and aMD simulations (Fig. 3a). The plots illustrate the expectation that the aMD simulation allows the protein to visit a larger region of conformational space than the standard MD simulation. The PCA1 and PCA2 dimensions both correspond to the movement of the C-terminus of the N-intein, as the five coordinates with the largest contributions to each of these PCA dimensions (comprising more than half of the energy of the eigenvector) all belong to the terminal six residues of the N-intein. The increased variation in this C-terminus is also illustrated clearly in Fig. 3b. Thus, the principal movements of the protein correspond to motions of the C-terminus of the N-intein and its interactions with other segments of the protein.

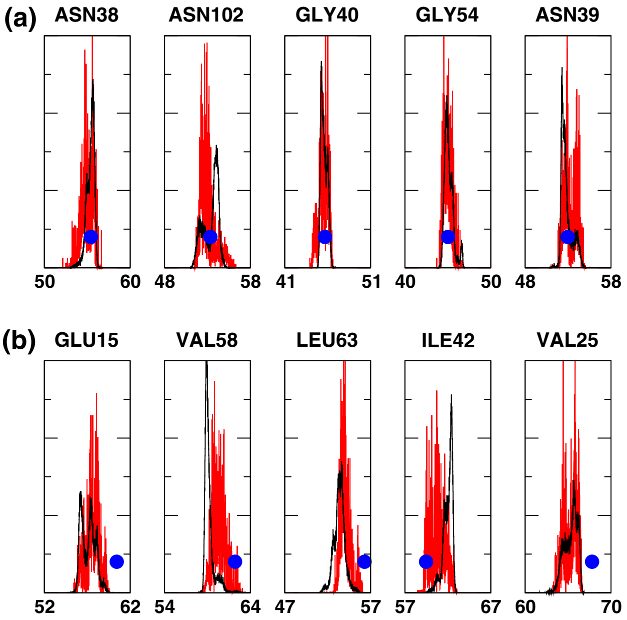


Fig. 1 Histograms of ^{13}Cα chemical shift predictions from frames of MD and aMD simulations. In each histogram, the *black line* is the distribution of chemical shift predictions based on the MD simulation, while the *red line* is the distribution of chemical shift predictions based on the aMD simulation. In each panel the abscissa scale is arbitary to accommodate the maxima of the distribution curves. The *blue marker* indicates the experimentally observed chemical shift. **a** shows the five best predictions based on the MD simulation (those with the lowest RMSD), while **b** shows the five worst predictions based on the MD simulation

Since aMD simulation samples a larger region of conformational space than standard MD simulation, permitting large-scale conformational changes to occur, we anticipated that aMD simulation would allow for better chemical shift prediction than standard MD simulation because it samples more global conformations of the protein which contribute to the experimentally observed chemical shift. However, we found that increased accuracy in chemical shift prediction based on aMD simulation does not derive from increased sampling of the protein's global conformation. Instead, the improvement stems from more accurate sampling of the conformational space of rigid segments of the protein. Figure 4a shows a plot of the difference in the errors of MD simulation ^{13}Cα shift predictions and aMD simulation ^{13}Cα shift predictions by residue, and Fig. 4b shows how these predictions correlate with spatial locations of these residues. The figure indicates that residues in which aMD performs better than MD in chemical shift prediction are clustered together sequentially and spatially. This suggests that chemical shift prediction accuracy may indeed correlate with accuracy of conformational sampling, which is expected to be a

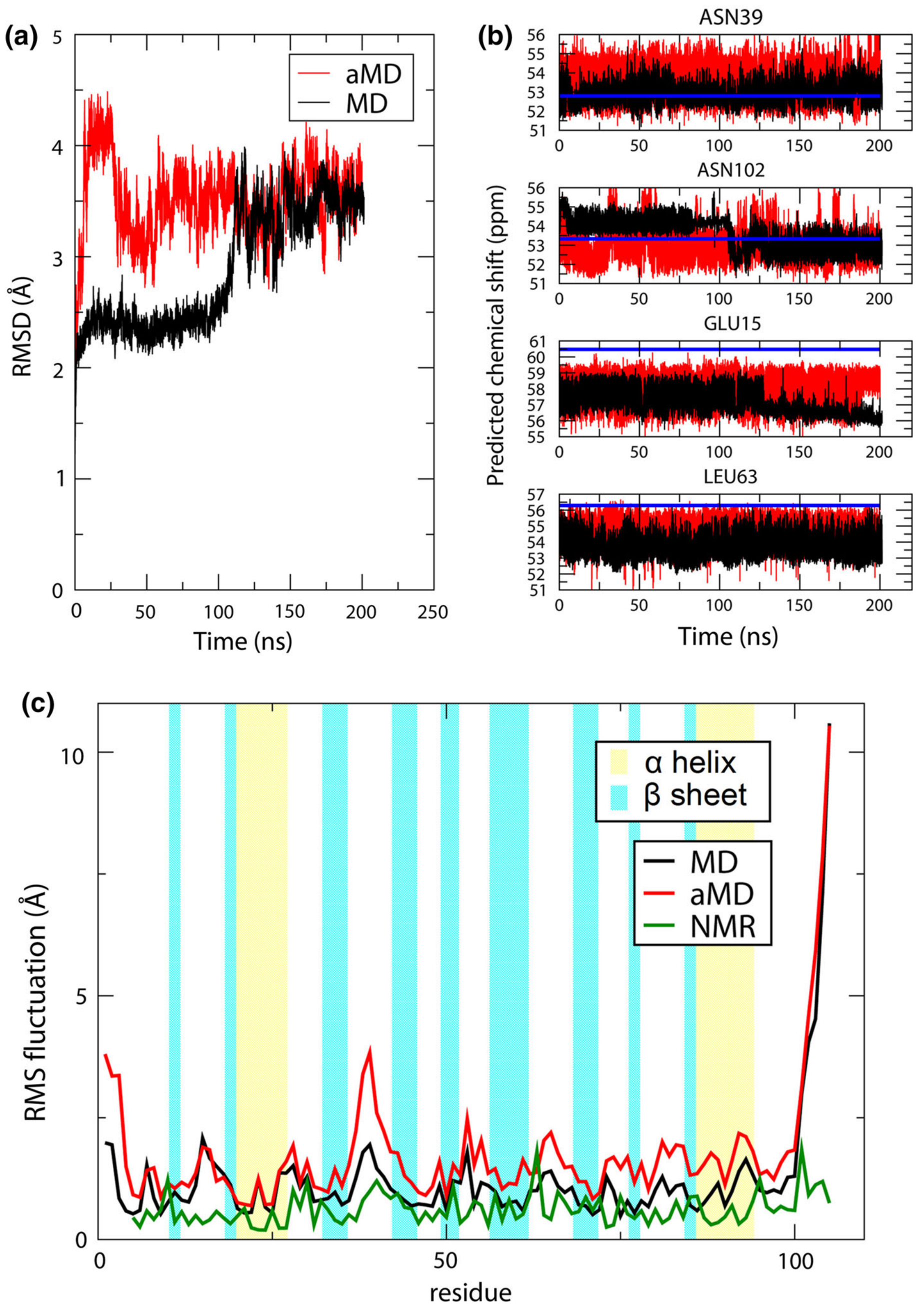


Fig. 2 **a** Plots of the RMSD of the protein structure as a function of simulation time for the MD and aMD simulations. **b** Plots of the predicted chemical shift as a function of simulation time for the MD (*black*) and aMD (*red*) simulations for four of the residues whose chemical shift prediction distribution are shown in Fig. 1. The experimentally observed chemical shift is indicated by a *blue horizontal line*. **c** The RMS fluctuation of each residue in the N-intein for the MD and aMD simulations and the NMR ensemble, with shading of the plot background to indicate secondary structure

function of the local environment of a particular segment; thus contiguous segments of the protein, which sample the local conformational space to a similar degree of accuracy, may have similar degrees of chemical shift prediction accuracy. An example of improvement due to more extensive and accurate sampling is the loop comprised of residues 79–85, seen in Fig. 5. The loop is seen to be more extensively sampled in aMD simulation, and Cα shift prediction distributions from MD and aMD simulations indicate that sampling in the aMD simulation favors conformations which have predicted chemical shifts which are closer to the experimental values.

On the other hand, it is noteworthy that the chemical shifts of the four most C-terminal residues, which undergo large conformational changes, are all predicted more accurately with standard MD than with aMD. One possible explanation for this might be that the boost potential of aMD may cause small potential wells in configurational space to be flattened. The boost potential, as discussed in "Methods", is applied in such a way as to "fill in"

Table 1 RMSD values of ensemble-derived chemical shift predictions (using SHIFTX2 and SPARTA+) from experimentally-derived values for the DnaE intein (pdb: 2KEQ)

	^{13}Cα	^{13}Cβ	^{13}C′	^{15}N	^{1}H′
SHIFTX2					
Normal MD	1.30	1.38	3.29	3.18	0.60
Accelerated MD	1.18	1.34	3.23	2.78	0.54
NMR ensemble	1.18	1.23	3.28	2.80	0.54
SPARTA+					
Normal MD	1.24	1.28	3.37	3.39	0.58
Accelerated MD	1.08	1.28	3.31	3.08	0.54
NMR ensemble	1.10	1.18	3.34	3.14	0.55

potential wells in order to increase the chances of escaping the well. However, when the well is shallow, such as in the case of very flexible terminal side chains, the boost potential may nearly flatten the potential well, leading to a nearly random walk in configurational space. This leads to a false distribution of configurations, which may lead to incorrect chemical shift prediction using these distributions. Figure 6 shows the chemical shift prediction distributions for residues 100 through 103 of the C-terminus. While the experimentally observed value for the chemical shift is contained inside the aMD prediction distribution, unlike those of other residues in which the prediction software does not predict a distribution close to the experimental value (see Fig. 1b), the aMD prediction distribution does not center at the value of the experimentally observed shift. Using aMD may be more helpful in sampling configurations which lie in deep potential wells separated by large barriers, such that a potential boost would preserve the potential well while simultaneously allowing for escape from the well and sampling of alternative configurations which may contribute to the experimental chemical shift value.

To investigate whether aMD simulation would allow for more accurate prediction of chemical shifts than a static

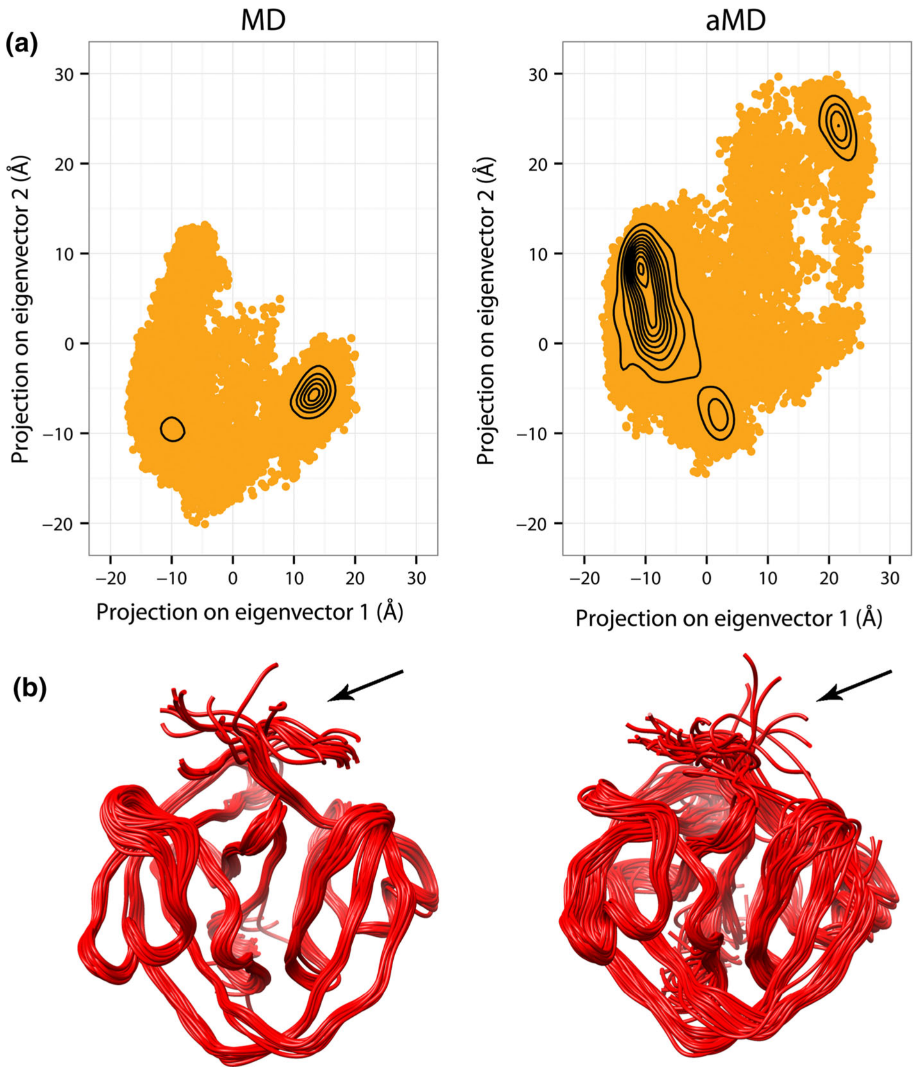


Fig. 3 **a** Plots of the MD (*left*) and aMD (*right*) frames projected onto the two largest principal component axes. Contours are shown indicating areas of high density. **b** An ensemble of 40 conformations seen in MD (*left*) and aMD (*right*) simulations generated by taking every 1,000th frame from the trajectory and aligning these to the first frame. The *drawing* indicates that aMD simulation allows for greater conformational heterogeneity throughout the protein but especially in the C-terminal tail, indicated by the *arrow*, which accounts for the majority of the variation reflected in the first two PCA axes

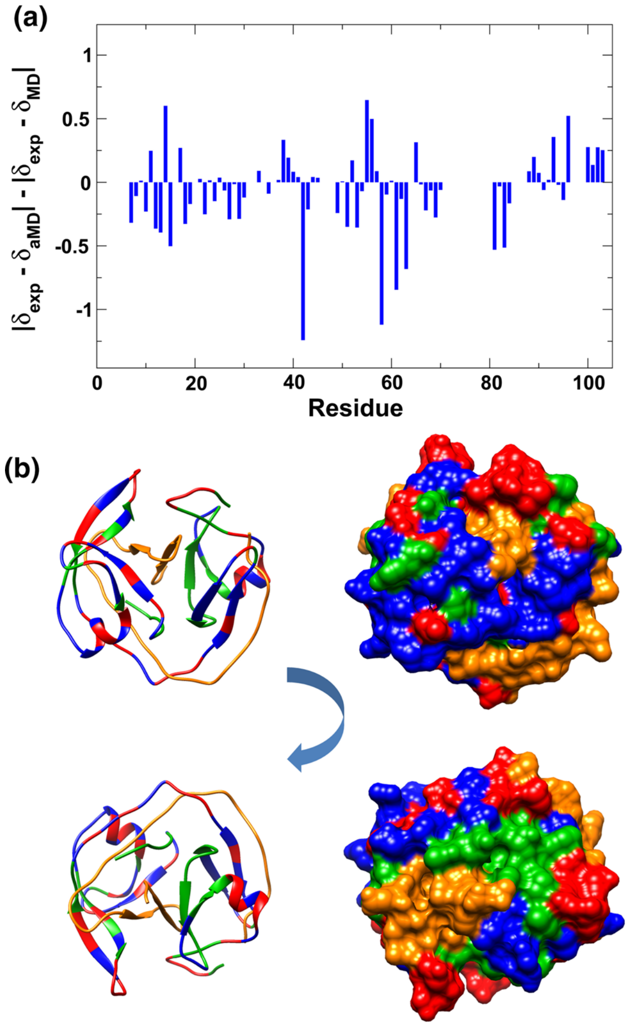


Fig. 4 **a** Plot of the difference in magnitudes of the RMSD discrepancy between the ^{13}Cα chemical shift prediction derived from MD simulations and from aMD simulations from the observed experimental value. **b** Front and rear views of the intein system, *colored* to indicate the accuracy of chemical shift prediction using MD or aMD. *Red* indicates that MD simulations predicted the chemical shift with a lower RMSD than aMD; *blue* indicates that aMD simulations predicted the chemical shift with a lower RMSD than MD. *Green* indicates residues for which no experimental shift was available. The C-intein is colored *orange*

X-ray crystal structure, we ran MD and aMD simulations for another intein structure, that of the RadA intein of *Pyrococcus horikoshii*. For this intein, both a crystal structure and an NMR structure have been deposited in the Protein Data Bank (4e2t and 2lqm, respectively), and NMR chemical shifts have been deposited in the Biological Magnetic Resonance Bank. Details of these simulations are discussed in "Methods". Results of chemical shift predictions based on the X-ray crystal structure, the NMR ensemble, MD and aMD simulations are in Table 2.

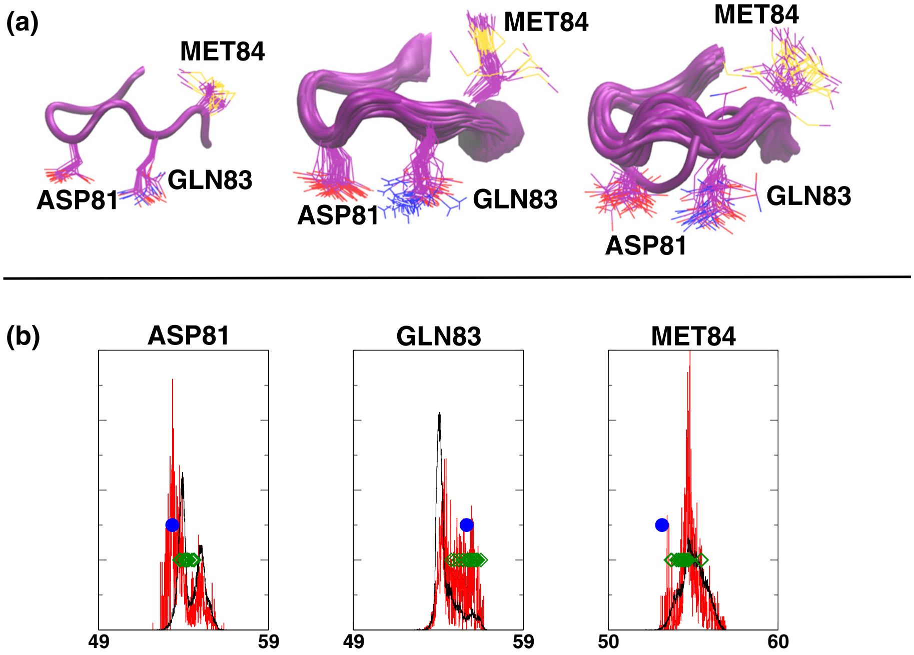
We were surprised to find that for this structure, chemical shift predictions based on the static X-ray crystal structure are slightly better than those based on standard MD simulations. This contrasts with reports that MD simulations allow for improved chemical shift prediction based on increased conformational sampling (Lehtivarjo et al. 2009, 2012; Robustelli et al. 2012). On the other hand, we found that chemical shift predictions based on accelerated MD improve upon those based on standard MD simulations, and were comparable to predictions based on the crystal structure or the NMR ensemble.

Discussion

Though the RMSDs of chemical shift prediction for both intein systems are higher than RMSDs reported elsewhere for other systems (Han et al. 2011; Robustelli et al. 2012), we attribute this to unique factors of these systems. Split inteins are partially intrinsically disordered proteins, with significant segments lacking well-defined secondary structure. The dynamic nature of these segments is likely to complicate accurate chemical shift prediction throughout the protein due to their interactions with better-defined secondary structural elements, and the lack of representation of similar dynamic structures in the databases from which prediction methods have been obtained. This is also evident in the poor RMSD values for chemical shift prediction in the DnaE intein based on the experimentally-derived NMR structural ensemble. Similarly, in the RadA intein, the NMR ensemble chemical shift predictions are not significantly more accurate than the predictions based on the X-ray crystal structure. This contrasts with an earlier study (Lehtivarjo et al. 2012) which showed more accurate chemical shift predictions with NMR structural ensembles (without MD simulations) than with X-ray crystal structures.

Our results show that use of ensembles as input for chemical shift prediction software does not always improve upon use of static X-ray structures. For the RadA intein, we found that all ensembles used—the NMR structural ensemble, MD and aMD simulation—do not improve significantly upon a static X-ray crystal structure with regard to chemical shift prediction, whereas studies on other systems have shown that all these ensembles do lead to improved shift prediction (Lehtivarjo et al. 2012; Robustelli et al. 2012). Thus, whether an ensemble input better predicts chemical shifts is likely system-dependent. However, what determines whether an ensemble will better predict chemical shifts in a given system is still unclear. More work is needed to analyze how structural features of a given protein correlate with chemical shift prediction accuracy and what type of input improves prediction. This

Fig. 5 **a** An ensemble of conformations of the loop containing residues 79 through 85 obtained from the NMR ensemble (20 conformations), the MD ensemble (40 conformations, extracted every 5 ns) and the aMD ensemble (40 conformations, extracted every 5 ns). **b** Histograms of ^{13}Cα chemical shift prediction based on MD and aMD simulations for residues 81, 83 and 84. The experimentally observed value is indicated by a *blue marker*, while the *green markers* indicate the predictions of the NMR ensemble structures



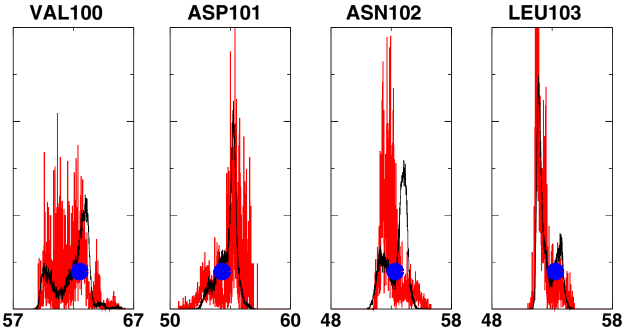


Fig. 6 Histograms of ^{13}Cα chemical shift prediction based on MD and aMD for residues 100–103. In each histogram, the *black plot* is the distribution of chemical shift predictions based on the MD simulation, and the *red plot* is the distribution of chemical shift predictions based on the aMD simulation. The *blue marker* indicates the experimentally observed chemical shift

study of the RadA intein suggests that in a case where an NMR structural ensemble does not lead to improved shift prediction, MD and aMD simulation ensembles also do not improve prediction over predictions based on a static crystal structure.

The disparity between the predicted chemical shifts and the experimental values in the intein system was noted despite the long-timescale simulation performed which was likely to explore a very large region of conformational space and produce an ensemble similar to that which exists in solution. For some residues (e.g., those in Fig. 1b), the experimental chemical shift value was not even in the distribution of predicted values, whereas for other residues, the experimental value was inside the distribution but the distribution was skewed such that the average predicted value was different from the experimental value. Indeed, one previous study (Robustelli et al. 2013) noted that multiple MD simulation runs produced different predicted shift value distributions which could be compared to the experimental value in order to evaluate the quality of the simulation run. Our results indicate that enhanced simulation methods may be an alternative method of producing ensembles which more accurately match the experimental chemical shift values.

Table 2 RMSD values of ensemble-derived chemical shift predictions from experimentally-derived values for the RadA intein. Of the 174 residues in the protein, residues 1–172 are included in the BMRB entry. We excluded residue 172 in calculating the RMSD values since that residue is completely disordered and significantly increases the RMSD values of the chemical shift predictions

	^{13}Cα	^{13}Cβ	^{13}C′	^{15}N	^{1}H′
SHIFTX2					
X-ray structure	0.98	1.08	0.94	2.89	0.51
Normal MD	1.02	1.14	1.01	3.29	0.52
Accelerated MD	0.94	1.10	1.00	2.97	0.51
NMR ensemble	0.99	1.09	0.98	3.04	0.46
SPARTA+					
X-ray structure	0.96	1.04	1.01	3.09	0.50
Normal MD	0.99	1.17	1.06	3.24	0.50
Accelerated MD	0.91	1.10	1.07	3.01	0.49
NMR ensemble	0.91	1.06	1.06	3.03	0.49

In addition to using ensemble input for shift prediction, others have suggested the incorporation of some degree of ensemble averaging in the training phase of development of an improved "neural network" chemical shift prediction approach (Han et al. 2011). Our results strongly support this approach. Indeed, it has already been shown that including dynamic information in a machine-learning algorithm significantly improves chemical shift prediction (Lehtivarjo et al. 2009). At present, the prediction of chemical shifts is not uniformly improved by combining MD simulation and current high-quality prediction methods, particularly with disordered systems like those studied here. This raises the question of how well deviations of predicted chemical shifts from the observed values can be correlated with simulations to identify populations of conformational states observed in the simulation. No doubt favorable cases will provide useful correlation which can be tested by appropriate and detailed mutational perturbation (Stafford et al. 2013). MD simulation input has been shown to improve prediction accuracy over static X-ray structures in several cases (Baskaran et al. 2010; Lehtivarjo et al. 2012; Robustelli et al. 2012).

With regards to our original hypothesis that chemical shifts might identify states associated with the context control of splicing rate, or the self-association of split inteins, the lack of agreement between results of simulation and observation indicates that in this case either the simulation does not address the range of ensemble structures necessary for the representation of structures and dynamics, or that the database used for knowledge-based shift prediction is insufficiently representative of actual dynamic structures.

With regard to continued development of shift prediction software, our results suggest that more work is needed in tuning shift prediction for use in systems with disorder. Though it is well-established that shift prediction for the so-called intrinsically disordered class of proteins or denatured proteins can be accurately obtained solely from considering the primary sequence (De Simone et al. 2009; Tamiola et al. 2010), the partially disordered segments of the split inteins discussed here are insufficiently highly disordered for their shifts to be calculated in that fashion, and additionally, their transient interactions with the more ordered segments likely produce additional shift perturbations of both classes of structure.

In terms of producing the comprehensive ensemble fully representing structure and dynamics of a protein, using accelerated MD trajectories appears to improve chemical shift prediction compared to using standard MD. Accelerated MD is expected to be especially useful in the area of chemical shift prediction because the timescale of NMR experiments is in the millisecond range, and standard MD usually probes only the nanosecond timescale, at most, whereas aMD is capable of probing very long timescales (Hamelberg et al. 2004). Our data suggest a modest improvement in chemical shift prediction accuracy for aMD trajectories over MD trajectories, likely owing to the improved conformational ensemble sampled for this partially disordered protein. Though there is still much improvement needed in NMR chemical shift prediction, particularly for proteins with disordered segments, our findings show that molecular simulation with enhanced sampling may be a key tool in chemical shift prediction for proteins with a high degree of flexibility.

Conclusion

Molecular dynamics simulations represent a potentially useful source of input data for NMR chemical shift prediction algorithms. However, more work is needed to determine when MD simulation data is likely to yield more accurate shift predictions. This study indicates that in the case of systems with partial disorder, like that studied here, use of MD simulation input for chemical shift prediction may not lead to improved prediction. We attribute this to the inadequate conformational sampling in standard MD simulation, which does not suffice for proteins with disordered regions whose conformational ensemble is likely to be more complex than that of a more rigid protein. For such systems, use of aMD simulation data for chemical shift prediction offers a way of counteracting this problem, allowing for more accurate shift prediction.

Acknowledgments Supported in part by NIH R01 GM086868-16 (Tom Muir, PI), and T32 GM007288-39. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. We are grateful to Sam Sparks for discussion.

References

Ball KA, Wemmer DE, Head-Gordon T (2014) Comparison of structure determination methods for intrinsically disordered amyloid-β peptides. J Phys Chem B 118:6405–6416

Baskaran K, Brunner K, Munte CE, Kalbitzer HR (2010) Mapping of protein structural ensembles by chemical shifts. J Biomol NMR 48:71–83

Bondar AN, del Val C, White SH (2009) Rhomboid protease dynamics and lipid interactions. Structure 17:395–405

Camilloni C, Robustelli P, De Simone A, Cavalli A, Vendruscolo M (2012) Characterization of the conformational equilibrium between the two major substates of RNase A using NMR chemical shifts. J Am Chem Soc 134:3968–3971

Camilloni C, Cavalli A, Vendruscolo M (2013) Assessment of the use of NMR chemical shifts as replica-averaged structural restraints in molecular dynamics simulations to characterize the dynamics of proteins. J Phys Chem B 117:1838–1843

Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668–1688

Case D, Darden T, Cheatham T III, Simmerling C, Wang J, Duke R, Luo R, Walker R, Zhang W, Merz K (2012) AMBER 12. University of California, San Francisco

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620

Chen H, Huang Z, Dutta K, Blais S, Neubert TA, Li X, Cowburn D, Traaseth NJ, Mohammadi M (2013) Cracking the molecular origin of intrinsic tyrosine kinase activity through analysis of pathogenic gain-of-function mutations. Cell Rep 4:376–384

De Simone A, Cavalli A, Hsu ST, Vranken W, Vendruscolo M (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. J Am Chem Soc 131: 16332–16333

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Deupi X, Kobilka BK (2010) Energy landscapes as a tool to integrate GPCR structure, dynamics, and function. Physiology 25:293–303

Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737

Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D (2005) Intrinsic dynamics of an enzyme underlies catalysis. Nature 438:117–121

Eryilmaz E, Shah NH, Muir TW, Cowburn D (2014) Structural and dynamical features of inteins and implications on protein splicing. J Biol Chem 289:14506–14511

Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593

Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein sci 9:1753–1773

Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120:11919–11929

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57

Iwai H, Züger S, Jin J, Tam P-H (2006) Highly efficient protein *trans* splicing by a naturally split DnaE intein from *Nostoc punctiforms*. FEBS Lett 580:1853–1858

Jensen MR, Markwick PR, Meier S, Griesinger C, Zweckstetter M, Grzesiek S, Bernado P, Blackledge M (2009) Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. Structure 17:1169–1185

Johnson B (2004) Using NMRView to Visualize and Analyze the NMR Spectra of Macromolecules. In: Downing AK (ed) In Protein NMR Techniques. Humana Press, pp 313–352

Kim TH, Chung KY, Manglik A, Hansen AL, Dror RO, Mildorf TJ, Shaw DE, Kobilka BK, Prosser RS (2013) The role of ligands on the equilibria between functional states of a G protein-coupled receptor. J Am Chem Soc 135:9465–9474

Kipp DR, Hirschi JS, Wakata A, Goldstein H, Schramm VL (2012) Transition states of native and drug-resistant HIV-1 protease are the same. Proc Natl Acad Sci USA 109:6543–6548

Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131:13894–13895

Lehtivarjo J, Hassinen T, Korhonen SP, Perakyla M, Laatikainen R (2009) 4D prediction of protein ^{1}H chemical shifts. J Biomol NMR 45:413–426

Lehtivarjo J, Tuppurainen K, Hassinen T, Laatikainen R, Perakyla M (2012) Combining NMR ensembles and molecular dynamics simulations provides more realistic models of protein structures in solution and leads to better chemical shift prediction. J Biomol NMR 52:257–267

Markwick PR, Cervantes CF, Abel BL, Komives EA, Blackledge M, McCammon JA (2010) Enhanced conformational space sampling improves the prediction of chemical shifts in proteins. J Am Chem Soc 132:1220–1221

Medek A, Hajduk PJ, Mack J, Fesik SW (2000) The use of differential chemical shifts for determining the binding site location and orientation of protein-bound ligands. J Am Chem Soc 122:1241–1242

Menon V, Vallat BK, Dybas JM, Fiser A (2013) Modeling proteins using a super-secondary structure library and NMR chemical shift information. Structure 21:891–899

Mittermaier A, Kay LE (2006) New tools provide new insights in NMR studies of protein dynamics. Science 312:224–228

Montalvao RW, Cavalli A, Salvatella X, Blundell TL, Vendruscolo M (2008) Structure determination of protein-protein complexes using NMR chemical shifts: case of an endonuclease colicin-immunity protein complex. J Am Chem Soc 130:15990–15996

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. J Biomol NMR 26:215–240

Oeemig JS, Aranko AS, Djupsjobacka J, Heinamaki K, Iwai H (2009) Solution structure of DnaE intein from *Nostoc punctiforme*: structural basis for the design of a new split intein suitable for site-specific chemical modification. FEBS Lett 583:1451–1456

Oeemig JS, Zhou D, Kajander T, Wlodawer A, Iwai H (2012) NMR and crystal structures of the *Pyrococcus horikoshii* RadA intein guide a strategy for engineering a highly efficient and promiscuous intein. J Mol Biol 421:85–99

Paulus H (2000) Protein splicing and related forms of protein autoprocessing. Annu Rev Biochem 69:447–496

Perler FB (2002) InBase: the intein database. Nucleic Acids Res 30:383–384

Ramirez M, Valdes N, Guan D, Chen Z (2013) Engineering split intein DnaE from *Nostoc punctiforme* for rapid protein purification. Protein Eng Des Sel 26:215–223

Robustelli P, Stafford KA, Palmer AG 3rd (2012) Interpreting protein structural dynamics from NMR chemical shifts. J Am Chem Soc 134:6365–6374

Robustelli P, Trbovic N, Friesner RA, Palmer AG (2013) Conformational dynamics of the partially disordered yeast transcription factor GCN4. J Chem Theory Comput 9(11):5190–5200

Ryckaert J-P, Ciccotti G, Berendsen HJ (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. J Comput Phys 23:327–341

Shah NH, Eryilmaz E, Cowburn D, Muir TW (2013a) Extein residues play an intimate role in the rate-limiting step of protein trans-splicing. J Am Chem Soc 135:5839–5847

Shah NH, Eryilmaz E, Cowburn D, Muir TW (2013b) Naturally split inteins assemble through a “capture and collapse” mechanism. J Am Chem Soc 135:18673–18681

Shan Y, Eastwood MP, Zhang X, Kim ET, Arkhipov A, Dror RO, Jumper J, Kuriyan J, Shaw DE (2012) Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. Cell 149:860–870

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

Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 48:13–22

Shen T, Hamelberg D (2008) A statistical analysis of the precision of reweighting-based simulations. J Chem Phys 129:034103

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and C. alpha and C. beta. ^{13}C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113: 5490–5492

Stafford KA, Robustelli P, Palmer AG 3rd (2013) Thermal adaptation of conformational dynamics in ribonuclease H. PLoS Comput Biol 9:e1003218

Tamiola K, Acar B, Mulder FA (2010) Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J Am Chem Soc 132:18000–18003

Tang SS, Case DA (2011) Calculation of chemical shift anisotropy in proteins. J Biomol NMR 51:303–312

Wereszczynski J, McCammon JA (2012) Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. Q Rev Biophys 45:1–25

Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58:62–87

Xu XP, Case DA (2001) Automated prediction of ^{15}N, ^{13}Cα, ^{13}Cβ and ^{13}C$'$ chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333

Xu XP, Case DA (2002) Probing multiple effects on ^{15}N, ^{13}Cα, ^{13}Cβ, and ^{13}C$'$ chemical shifts in peptides using density functional theory. Biopolymers 65:408–423

Zettler J, Schütz V, Mootz HD (2009) The naturally split *Npu* DnaE intein exhibits an extraordinarily high rate in the protein *trans*-splicing reaction. FEBS Lett 583:909–914