Protein structure prediction using global optimization by basin-hopping with NMR shift restraints

Falk Hoffmann and Birgit Strodel

Citation: The Journal of Chemical Physics **138**, 025102 (2013); doi: 10.1063/1.4773406
View online: http://dx.doi.org/10.1063/1.4773406
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/138/2?ver=pdfcov
Published by the AIP Publishing

Protein structure prediction using global optimization by basin-hopping with NMR shift restraints

Falk Hoffmann[1] and Birgit Strodel[1,2,a)]
[1]*Institute of Complex Systems: Structural Biochemistry, Research Centre Jülich, 52425 Jülich, Germany*
[2]*Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany*

(Received 30 September 2012; accepted 12 December 2012; published online 11 January 2013)

Computational methods that utilize chemical shifts to produce protein structures at atomic resolution have recently been introduced. In the current work, we exploit chemical shifts by combining the basin-hopping approach to global optimization with chemical shift restraints using a penalty function. For three peptides, we demonstrate that this approach allows us to find near-native structures from fully extended structures within 10 000 basin-hopping steps. The effect of adding chemical shift restraints is that the α and β secondary structure elements form within 1000 basin-hopping steps, after which the orientation of the secondary structure elements, which produces the tertiary contacts, is driven by the underlying protein force field. We further show that our chemical shift-restraint BH approach also works for incomplete chemical shift assignments, where the information from only one chemical shift type is considered. For the proper implementation of chemical shift restraints in the basin-hopping approach, we determined the optimal weight of the chemical shift penalty energy with respect to the CHARMM force field in conjunction with the FACTS solvation model employed in this study. In order to speed up the local energy minimization procedure, we developed a function, which continuously decreases the width of the chemical shift penalty function as the minimization progresses. We conclude that the basin-hopping approach with chemical shift restraints is a promising method for protein structure prediction. © *2013 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4773406]

I. INTRODUCTION

The determination of protein structures is one of the most important challenges in biochemistry. Computational techniques can help find the three-dimensional arrangement of atoms. However, the exact determination of native structures from denatured or unfolded proteins is still a challenge. The usage of structural restraints obtained from experiments such as nuclear magnetic resonance (NMR) measurements shows significant improvement in this field of research.[1–11] About 12% of the structures saved in the RCSB protein data bank[12] are produced from NMR data. Chemical shifts are the most readily and accurately measurable NMR observables in solution and in the solid state,[5] and can be used to predict the molecular structure,[4–9,13–18] including the structure of a low-populated, on-pathway folding intermediate.[19]

Many of the simulations for NMR based structure determination use sequence homology information.[4,5,8,9,16,20] In such approaches structural motifs are selected from databases of existing protein structures based on NMR data, such as chemical shifts, residual dipolar couplings (RDCs), *J*-couplings, or nuclear Overhauser effect (NOE) data.[21] However, the usage of molecular fragment replacement approaches with chemical shift information depends on the structural model and cannot be easily used to calculate conformational changes or combined with RDC, *J*-couplings, or NOE data. Applying chemical shift restraints using a penalty function avoids these problems. Here, the Hamiltonian is applied such that it reduces the conformational search to structures with small shift restraints. This approach was used successfully to perform structural refinements of proteins.[6,7]

In the works by Vendruscolo and co-workers[6,7] the CamShift method[22] was used for the incorporation of chemical shift restraints. CamShift is a tool recently introduced for the rapid prediction of NMR chemical shifts from protein structures based on an approximation of the chemical shifts as polynomial functions of interatomic distances.[22] This chemical shift predictor is combined with a tunable soft-square harmonic well as a penalty function to compute the differences between calculated and experimental chemical shifts.[6,7] Furthermore, the chemical shifts are differentiable functions of the atomic coordinates, which enables the calculation of forces. Vendruscolo and co-workers were able to find the structures of a set of proteins with 56–108 residues with a resolution of 0.8–2.2 Å using CamShift molecular dynamics (MD) simulations of previously partially folded proteins.[7] The determination of peptide structures from unfolded conformations using a Monte Carlo (MC) approach was also possible with a simulated annealing (SA) protocol.[6]

In this study, we combine the basin-hopping (BH)[23,24] approach to global optimization with NMR chemical shift restraints using CamShift. The BH method, which is a generalization of the Monte Carlo-minimization approach,[25] has

a)Author to whom correspondence should be addressed. Electronic mail: b.strodel@fz-juelich.de.

been successfully used to identify the global minimum of peptides and proteins,[26–31] including structures of peptide complexes.[32–34] The availability of forces in CamShift enables us to combine it with the BH method. In this work we demonstrate that this approach allows us to find near-target structures from fully extended peptide conformations. We present the results from chemical shift-restrained BH simulations of three peptides with the PDB[12] codes 1LE0,[35] 1L2Y,[36] and 1YRF.[37] We show that we are able to find the folded structures within 10 000 BH steps, while the unrestrained BH simulations of same run length fail to locate near-native structures.

II. METHODS

A. Structural models

The structures for 1LE0, 1L2Y, and 1YRF were downloaded from the RCSB protein data bank[12] and used as target structures for the BH simulations. 1LE0 is a 12 amino acid β-hairpin;[35] 1L2Y is a 20 amino acid peptide with a short α-helix, a 3_{10}-helix, and a polyproline II helix at the C-terminus;[36] and 1YRF is a 35-residue subdomain of the villin headpiece consisting of three α-helices.[37] These minipeptides have been used as test cases in previous folding studies.[38–56] We employed CamShift[22] to calculate ^{1}Hα, amide ^{1}H, ^{13}Cα, ^{13}Cβ, carbonyl ^{13}C, and amide ^{15}N chemical shifts from the target structures and used them as target chemical shifts for the definition of the restraint function. These are denoted δ_{exp} in the following. Fully extended structures of the peptides were generated from their structural sequence using VMD[57] and employed as starting structures for the BH simulations (Figure S1 of the supplementary material[58]).

We used the CHARMM22 force field parameters[59,60] to model the peptides. To model the aqueous solvent we employed the generalized Born model FACTS.[61] For the calculation of the nonbonded interactions, the cutoff scheme suggested in the FACTS documentation was employed, i.e., truncation of both long-range electrostatics at 14 Å using a shift function and the van der Waals energy with a polynomial switching function applied between 10 and 12 Å.

B. Basin-hopping

The BH approach to global optimization[23,24] is analogous in principle to the Monte Carlo-minimization approach.[25] Moves are proposed by perturbing the current geometry and are accepted or rejected on the basis of the Metropolis criterion,[62] which uses the energy difference between the local minimum obtained by minimization from the instantaneous configuration and the previous minimum in the Markov chain. In effect, the potential energy surface is transformed into the basins of attraction of all the local minima so that the energy for configuration $\mathbf{r}$ is

$$\widetilde{E}(\mathbf{r}) = \min\{E(\mathbf{r})\}, \tag{1}$$

where "min" denotes local minimization. Large steps can be taken to sample this transformed landscape, since the objective is to step between local minima. Furthermore, there is no need to maintain detailed balance when taking steps because the BH approach attempts to locate the global potential energy minimum and is not intended to sample thermodynamic properties. The BH algorithm is implemented in the GMIN program.[63]

Basin-hopping has been employed successfully to find the global minimum of peptides and proteins,[26–31,64] including structures of peptide complexes.[32–34] In our study, we performed BH simulations using between 100 and 10 000 BH steps. The moves for perturbing the current geometry of the peptides were taken in backbone and sidechain dihedral angle space.[28] At each BH step, on average 30% of these dihedrals were randomly chosen and then twisted by an angle of maximally 60°. Dihedral angles which define planar structures, such as rings, were considered non-twistable to keep their planarity.[65] In all BH runs the temperature was set to 300 K. We use a limited-memory variation of the Broyden-Fletcher-Goldfarb-Shanno update by Nocedal[66] (LBFGS) for energy minimization.

C. Chemical shift restraints

We implemented chemical shift restraints into the GMIN program with a modified version of the program CamShift.[22] CamShift calculates chemical shifts using distance dependent functions of the atomic coordinates for the influence of backbone atoms, sidechain atoms, and nonbonded atoms. Furthermore, it also includes a dipole approach for the influence of aromatic rings and a parametrized function for dihedral angles. CamShift enables us to calculate chemical shifts quickly and accurately.[22] Furthermore, it allows to calculate forces from chemical shifts.

We use a soft-square harmonic potential as introduced by Vendruscolo and co-workers[6] to define the chemical shift penalty function E_{CS}, which restrains the structures to conformations in agreement with the chemical shifts of the target structure. Figure 1 shows that E_{CS} is split into three regions: a flat-bottomed region that takes into account inaccuracies in the chemical shift predictions, a harmonic region that penalizes statistically significant deviations between the computed and experimental shifts, and a linear region that prevents large deviations of individual chemical shifts from dominating the magnitude of E_{CS} and thus frustrating the conformational search.[6] The width of the potential well E_{CS} is governed by the parameter n.

The CamShift energy E_{CS} and force $\mathbf{F}_{\text{CS}}$ are added to the CHARMM22 energy E_{FF} and force $\mathbf{F}_{\text{FF}}$, respectively,

$$E = E_{\text{FF}} + \alpha E_{\text{CS}}, \tag{2}$$

$$\mathbf{F} = \mathbf{F}_{\text{FF}} + \alpha \mathbf{F}_{\text{CS}}. \tag{3}$$

Here, the adjustable parameter α defines the contribution of the chemical shift restraints to the total energy E. If the value of α is too high, the forces resulting from CamShift are too large, creating instabilities during the energy minimization process. If the value of α is too low or the tolerance parameter n is too large, the influence of CamShift is too weak to provide an improvement over unrestrained simulations. If the

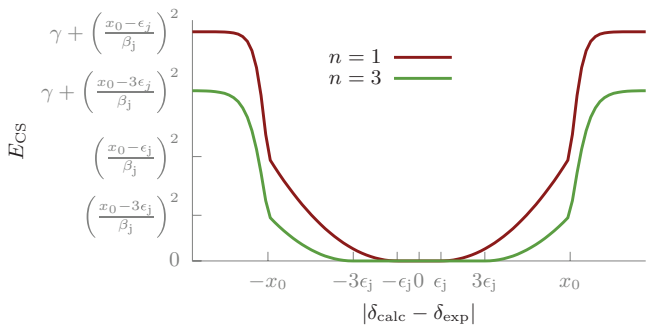

FIG. 1. Chemical shift penalty energy E_{CS} as a function of the difference between the chemical shifts of the simulated (δ_{calc}) and the target structure (δ_{exp}) for $n = 1$ (red) and $n = 3$ (green). The width of the flat-bottomed part is $2n\epsilon_j$, which is adjustable by modifying n with ϵ_j being the accuracy of the chemical shift predictions for atom type j. x_0 is the cutoff of the harmonic part of the energy function, β_j is a scaling parameter determining the magnitude of the energy penalty, and γ influences how large the energy penalty can grow beyond x_0.

value of n is too small, small deviations from the target structure will generate chemical shifts that result in large penalties, thus creating a rugged energy landscape. It will therefore be more difficult to locate the global minimum as the system can easily become trapped in deep local minima.[6] In the first part of our study we identified the optimal values of α and n for the combination of GMIN and CamShift as described in Sec. III A. It should be noted that the unit of E_{FF} is kcal/mol in the CHARMM force field,[59] while E_{CS} is a dimensionless quantity.[6]

III. RESULTS AND DISCUSSION

A. Optimization of the parameters α and n

We prepared the systems as described in Sec. II A. First, we performed two types of short chemical shift-restrained simulations with 100 BH steps for 1LE0 and 1L2Y, one using value pairs with $\alpha = 1$ and varying n from $n = 0.5$ to $n = 4$, and the second with constant $n = 1$ but varying α from $\alpha = 0$ to $\alpha = 3$. For the latter we chose $n = 1$ because from the runs with varying n only the one with $n = 1$ could find parts of the β-hairpin for 1LE0, as the structural results in Figure S2 of the supplementary material[58] show. In the runs with varying α, the unrestrained simulation ($\alpha = 0$) was not able to produce a structure resembling the β-hairpin, while the other values for α were more successful. Figure S2 of the supplementary material[58] shows that the simulations with $(\alpha, n) = (1, 1)$ and $(\alpha, n) = (2, 1)$ find structures fitting best to the β-hairpin within the short 100 step-BH runs. To test if this choice of values for α and n is universal or peptide specific, we performed 100 step-BH simulations of 1L2Y using the various (α, n) pairs. Figure S3 of the supplementary material[58] shows that only the simulations with $(\alpha, n) = (1, 0.5)$ and $(1, 1)$ could find parts of the α-helix. In a previous chemical shift-restrained MC study using a SA protocol, Robustelli *et al.* also chose $n = 1$ yet in connection with higher values for α.[6] The ideal value of α depends on the absolute value of the force field energy E_{FF}: the larger this value, the larger α has to be chosen.

During our systematic test of the interplay between α and n, we further observed that n had to be optimized for the LBFGS minimizer, while keeping $\alpha = 1$ constant throughout each BH simulation. For n we found that the local minimization at a given BH step is more successful in terms of robustness and speed if n is decreased while the minimization progresses. We use the root mean square force (RMSF) during the minimization as progression variable to determine n:

$$n = \begin{cases} 3 & \text{RMSF} > 1, \\ 3 + \frac{2}{3}\log(\text{RMSF}) & 10^{-3} < \text{RMSF} \leq 1, \\ 1 & \text{RMSF} \leq 10^{-3}, \end{cases} \quad (4)$$

where the unit of RMSF is computed for the change of the total energy E. We start with the relatively large value $n = 3$ to make sure that the first few minimization steps after changing the dihedral angles are mainly force-field driven. Figure 1 shows that for large values of n, the calculated chemical shifts of a wide range of conformations fall near the flat-bottomed region of E_{CS} and thus generate relatively small energetic penalties. Once the minimization has sufficiently progressed, the conformation is increasingly forced towards the target structure by decreasing n, i.e., by increasing the penalties for the calculated shifts of atoms j, which deviate by more than $n\epsilon_j$ from the experimental chemical shifts. We reduce the value of n continuously to the previously determined $n = 1$.

B. Results for 1LE0, 1L2Y, and 1YRF

We performed chemical shift restrained and unrestrained BH simulations using 1000 and 10 000 BH steps for the peptides 1LE0 and 1L2Y, and 1000 and 5000 BH steps for 1YRF. We did not continue the BH run for 1YRF up to 10 000 BH steps because we did not observe a significant improvement during the last 3000 BH steps, and the result after 2000 BH steps is already very convincing. The best structures, which we obtained for the three peptides within 1000 steps and full-length BH simulations, are shown in Figure 2. Here, the definition of the best structures is with respect to the total energy $E = E_{FF} + E_{CS}$, which is lowest for these structures. This allows us to test, if by using the total energy as criterion, structures with low E correspond to structures with low backbone root mean square displacement (RMSD) from the target structure. This can only be the case when the force field correctly predicts the target structure as the global minimum. Thus, we included a β-sheet structure and helical structures in our test set in order to check if both structural elements are correctly supported by the CHARMM22 potential in connection with the FACTS solvation model.

1. 1LE0

The structures for 1LE0 in Figure 2 show that the β-hairpin can be determined with very high accuracy. Within 1000 BH steps the β-sheet is already correctly identified, while the turn region still needs improvement. After 10 000 steps this deficiency was resolved, so that the RMSD of the best structure is only 0.86 Å from the target structure. The

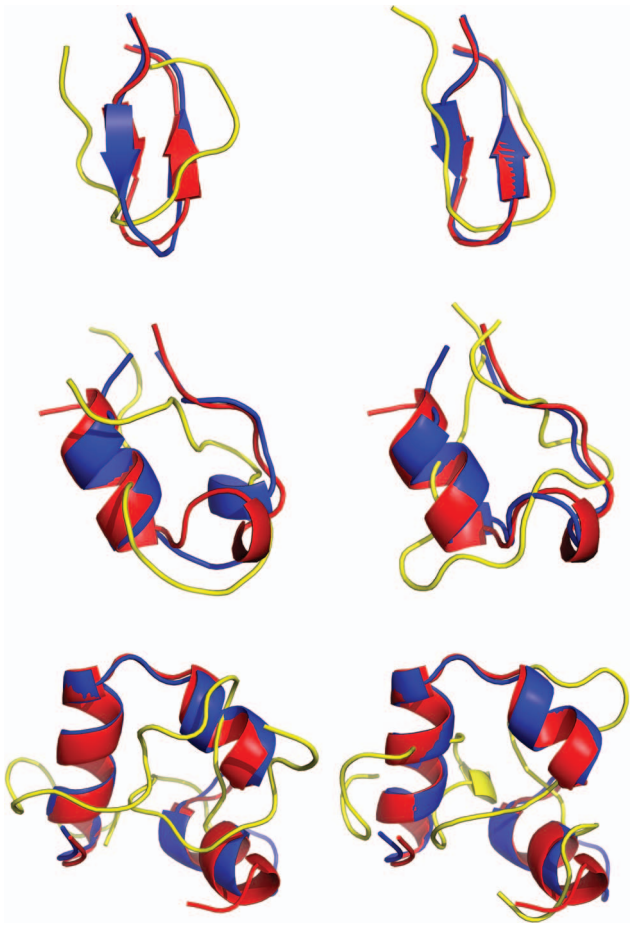

FIG. 2. Target structures (red), structures of unrestrained (yellow), and chemical shift restrained (blue) simulations after 1000 (left) and 10 000 (right) BH steps for 1LE0 (top) and 1L2Y (center), and after 1000 (left) and 5000 (right) BH steps for 1YRF (bottom).

unrestrained BH run was not able to produce the β-sheet within 10 000 steps.

2. 1L2Y

The correct structure of 1L2Y was also identified within 10 000 BH steps using chemical shift restraints, while the unrestrained BH simulation did not even find the α-helix for the first nine residues. Imposing chemical shift restraints, the α-helix was found quickly (within 1000 BH steps) and accurately. The biggest deviations from the target structure are seen for the termini and for the transition between the α-helix and the 3_{10} helix (residues 10 and 11). The middle part of the peptide needed longest before its correct structure was located. The RMSD for the best structure after 10 000 BH steps is 2.18 Å. In order to pinpoint the origin of the deviation between the predicted and the target structure, we plotted the deviation between computed and target chemical shifts, $\delta_{\text{calc}} - \delta_{\text{exp}}$, for each C$\alpha$ atom of 1L2Y (Figure 3). This analysis reveals that for residues 3–9 and 12–17 the predicted shifts are restrained to their target shifts since $|\delta_{\text{calc}} - \delta_{\text{exp}}|/\epsilon < 1$, which for $n = 1$ corresponds to the flat-bottomed region of the chemical shift penalty function (Figure 1) leading to $E_{\text{CS}} = 0$ for these atoms.

FIG. 3. Deviation between the Cα chemical shifts of the predicted and the target structures for 1L2Y. $\delta_{\text{calc}} - \delta_{\text{exp}}$ is shown for each residue apart from residues 1 and 20, because CamShift does not provide chemical shifts for the first and last residue. The chemical shift deviation is given in units of the CamShift accuracy $\epsilon_{\text{C}\alpha}$ for the prediction of Cα chemical shifts.

Figure 3 shows that residues 2, 18, and 19 produce the largest deviations from the target structure. This is because CamShift does not calculate the chemical shifts for the first and last amino acids in the chain, because the CamShift prediction for a given atom relies on the distances to atoms in the two neighboring residues. Therefore, the structures for the first and last residues have to be predicted without chemical shift restraints. In general, this means that the largest structural deviations come from the terminal residues, as the predicted structure for 1L2Y in Figure 2 supports. The wrong structures for the first and last residues give rise to wrong interatomic distances, which are needed for the chemical shift calculations for residues 2 and $N_{\text{res}} - 1$, with N_{res} being the total number of residues in the chain. In turn, this leads to inaccurate chemical shifts δ_{calc} for residues 2 and $N_{\text{res}} - 1$, hampering the structure prediction for these residues. This effect propagates to residues 3 and $N_{\text{res}} - 2$ before eventually leveling off. For small peptides such as 1L2Y the deviation of only a few residues leads to an appreciably increased RMSD from the target structure. This effect will decrease for larger peptides.

3. 1YRF

The 5000 step-BH run with chemical shift restraints produced a structure for 1YRF with a RMSD of 3.81 Å. The best structure, which was found within 1000 BH steps looked already very good and could only slightly improved during the subsequent 4000 BH steps. From the structures in Figure 2 it is visible that, as discussed above for 1L2Y, the largest deviations originate from the terminal residues. If we exclude residues 1 and 36 from the RMSD calculation we obtain a RMSD of 2.44 Å, which further decreases to 1.88 Å by excluding residues 1, 2, 35, and 36, and to 1.39 Å for the RMSD between residues 4 and 33. Excluding more residues does not further improve the RMSD. As for the other two peptides, the unrestrained BH run did not produce a structure resembling the target structure. None of the helices were found during this run.

In order to better understand the interplay between the force field energy and the chemical shift penalty, and their influence on folding the helical peptide 1YRF, we plotted the total energy $E = E_{\text{FF}} + E_{\text{CS}}$, the CamShift penalty energy

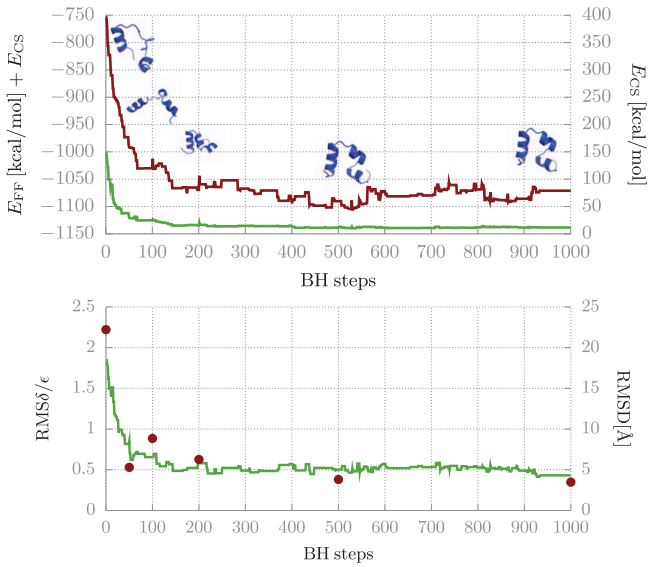

FIG. 4. Folding of 1YRF during the first 1000 BH steps. (Top) Energy for each state of the Markov chain. The red line represents the total energy (E_{FF} + E_{CS}) and the green line represents the CamShift energy (E_{CS}). Structures after 50, 100, 200, 500, and 1000 BH steps are shown in blue. (Bottom) RMSδ is shown for Cα atoms for each state of the Markov chain (green line). It was normalized with regard to the atom type specific CamShift accuracy $\epsilon_{C\alpha}$. For the structures after 0, 50, 100, 200, 500, and 1000 BH steps the RMSD is provided as well (red dots).

E_{CS}, the structural RMSD, and the root mean square chemical shift deviation RMSδ (for Cα atoms) from the target chemical shifts during the first 1000 BH steps (Figure 4). All four quantities reach a plateau in less than 200 BH steps, which are sufficient for the structure to find the secondary structure elements, i.e., the three α-helices (see blue structure after 200 steps). Identifying the α-helices is accompanied by a marked decrease of E_{CS}. During the following 800 BH steps the structure improved by finding the correct arrangement of the α-helices with respect to each other and local refinements. These improvements are mainly force field driven as E_{FF} decreases more strongly than E_{CS} for the near-target structures. The penalty energy plateaus at $E_{CS} \approx 11.5$, implying that within 1000 BH steps not all predicted chemical shifts fall into the flat-bottomed region of the chemical shift penalty function (Figure 1). As discussed above, the largest deviations originate from the amino acids in neighborhood to the terminal residues. The improvement of the structure is confirmed by the RMSD. It decreases from ≈7 Å at BH step 200 to ≈4 Å at BH step 1000, which is already close to the final RMSD of 3.81 Å for the best structure after 5000 BH steps.

4. Incomplete chemical shift assignments

It is often not possible to measure and assign all chemical shifts in a NMR experiment. To test the robustness of our approach with respect to incomplete chemical shift assignments, we performed BH simulations of 1YRF where only one of the six chemical shift types, ^{1}Hα, amide ^{1}H, ^{13}Cα, ^{13}Cβ, carbonyl ^{13}C, or amide ^{15}N chemical shifts were used in the restraining function. The number of chemical shift restraints applied is given by $N_{shift} \times (N_{res} - 2)$, with N_{shift} being the number of chemical shift types considered (for the first and last residue no chemical shifts are calculated). In the simulations above we set $N_{shift} = 6$, while in the simulations with only one chemical shift type $N_{shift} = 1$. Additionally, we performed one simulation with $N_{shift} = 3$, where restraints for ^{1}H, ^{13}Cα, and ^{15}N chemical shifts were included. We chose these three shift types as these are the most frequently measured chemical shifts for proteins as the statistics derived from a total of about 5.6×10^6 chemical shifts in the Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu/) reveals (see Figure S4 of the supplementary material[58]). Figure 5 shows the structures obtained after 1000 BH steps with reduced chemical shift restraints. Apart from the simulation with only ^{13}Cβ chemical shift restraints, the other simulations with $N_{shift} = 3$ and $N_{shift} = 1$ are able to fold parts of the peptide into α-helices. The predicted structures from these simulations are much closer to the target structure than the structure from the unrestrained 1000 step BH simulation (Figure 2).

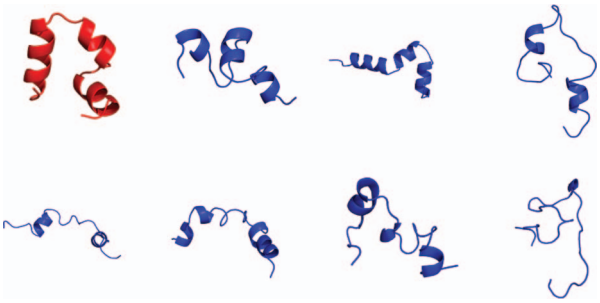

FIG. 5. Structures for 1YRF after 1000 BH steps with reduced chemical shift restraints. Top (from left to right): Target structure (red) and predicted structures (blue) with chemical shift assignments for (^{1}H, ^{13}Cα, ^{15}N), ^{13}C, and ^{15}N. Bottom: Predicted structures with chemical shift assignments for ^{1}Hα, ^{1}H, ^{13}Cα, and ^{13}Cβ. The results are sorted according to decreasing prediction quality.

For a more detailed analysis of the performance of the BH simulations with reduced chemical shift restraints, we determined the secondary structure of each residue in the structures given in Figure 5 using STRIDE[67] (Figure 6). The simulation with only carbonyl ^{13}C chemical shift restraints succeeded to predict all three α-helices at almost identical positions to the target structure. This can be explained by

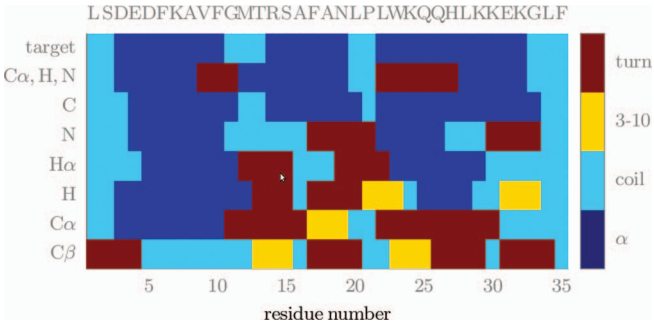

FIG. 6. Secondary structure per residue in the structures shown in Figure 5. On the top line the one letter code of each residue is given, on the bottom line the residue number. The left column designates the chemical shift restraints applied in the simulations.

considering that the backbone torsional angles Φ and Ψ are strong determinants of the ^{13}C chemical shifts. Their influence on this chemical shift is about 50%, while on $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts their effect is only 25% and 10%, respectively.[1] Therefore, $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts are less effective as restraints for secondary structure prediction. With $^{13}C\alpha$ chemical shift restraints one of the three α-helices can still be predicted within 1000 BH steps, while $^{13}C\beta$ chemical shift restraints fail to fold any of the helices. $^{13}C\beta$ chemical shifts are shifted downfield by about 2.5 ppm in β-sheets, but have nearly random coil values in helices.[1] Thus, it is not surprising that the information from merely $^{13}C\beta$ chemical shifts is not sufficient for the identification of the helices of 1YRF. Individual ^{15}N, $^{1}H\alpha$, and ^{1}H chemical shift restraints are successful in the prediction of two of the three helices within 1000 BH steps. While both ^{15}N and ^{1}H chemical shifts are not very good predictors of dihedral angles or indicators of secondary structure, they are very sensitive to hydrogen bonding[1] and are therefore helpful as restraints in protein folding simulations. $^{1}H\alpha$ chemical shifts are known as a reliable indicator of secondary structure, and backbone dihedral torsional effects are the most important contribution to $^{1}H\alpha$ chemical shift deviations. This explains the good performance of the BH run with only $^{1}H\alpha$ chemical shift restraints. The combined application of ^{1}H, $^{13}C\alpha$, and ^{15}N chemical shift restraints leads to the prediction of all three helices in less than 1000 BH steps. Here, even the length of the coil sequences between the second and third helix (residues 21 and 22), and at the N- and C-termini (residues 1–2 and 33–35, respectively) are correctly predicted.

IV. CONCLUSION

Computational methods that utilize chemical shifts to produce protein structures at atomic resolution have recently been introduced. These methods use the information contained in experimental chemical shifts together with structural homology of proteins in structural databases such as the RCSB protein data bank to generate new structures,[4,5,8,9,16,20] or directly incorporate chemical shifts as restraints in molecular simulations with an energetic penalty function analogous to those used in standard NMR structure calculations.[6,7] In the current work, we applied the latter idea and combined the basin-hopping (BH) approach to global optimization[23,24] with chemical shift restraints by using the chemical shift penalty function introduced by Vendruscolo and co-workers.[6,7] For the calculation of NMR chemical shifts from protein structures we used the CamShift method, which approximates chemical shifts as polynomial functions of interatomic distances.[22]

For the proper implementation of chemical shift restraints into the BH approach we determined the optimal weight of the chemical shift penalty energy with respect to the CHARMM22 force field[59,60] employed in conjunction with the solvation model FACTS.[61] Furthermore, we developed a function, which continuously decreases the width of the chemical shift penalty function during each local energy minimization procedure, which thereby becomes more robust. We demonstrated for three peptides that the BH approach with chemical shift restraints is able to find near-native structures from fully extended structures within 10 000 BH steps. The conformational searches were able to fold α and β secondary structure elements in less than 1000 BH steps, and correctly orient their tertiary contacts in subsequent BH steps. The unrestrained BH runs, on the other hand, failed to fold any of the secondary structure elements within 10 000 BH steps. Much longer unrestrained BH runs would be needed for the conformational searches to succeed without guidance from chemical shift restraints. In another study we tested whether or not the CHARMM22/FACTS potential supports the target structures of 1LE0, 1L2Y, and 1YRF as global minima. We found that the RMSD values of the global minima from the respective targets are between 1.5 and 3 Å. Our conclusion therefore is that it is rather inefficient sampling and not the CHARMM22/FACTS potential that precluded the generation of near-native structures in the unrestrained BH simulations of the current study.

We tested our approach for incomplete chemical shift assignments, where the information from only one chemical shift type was included in each of the chemical shift-restraint BH simulations. Apart from the simulation with $^{13}C\beta$ chemical shift restraints, these simulations succeeded to predict secondary structure elements within 1000 BH steps. For each of the chemical shift types, the success (and failure) can be explained based on the relation between structure and chemical shifts in proteins.[1] The usage of fewer chemical shifts speeds up the restrained BH simulations as the computational overhead compared to unrestrained BH simulations scales linearly with N_{shift}. However, in order to obtain as good prediction results as from the runs with more chemical shift restraints, more BH steps have to be conducted. The full-length BH simulations with $N_{\text{shift}} = 6$, for which the results are shown in Figure 2, required 10 CPU days for 1LE0, 12 CPU days for 1L2Y, and 16 CPU days for 1YRF. All BH simulations were run on a single 2.93 GHz Intel Xeon Processor X5570. For the folding of proteins of comparable length using chemical shift restrained Monte Carlo simulations with a simulated annealing protocol Robustelli *et al.*[6] needed between 380 and 473 CPU days. This comparison reveals that it is more effective to apply chemical shift restraints via both energy and energy gradients, as it is realized in BH and molecular dynamics,[7] than considering only the energy as in simulated annealing based on Monte Carlo simulations.[6] Like Robustelli *et al.*,[6] we found that the major bottleneck of the chemical shift restrained simulations is the computation of chemical shifts with each call to the energy function. The unrestrained BH simulations of the same length required less than a CPU day for 1LE0 and 1L2Y, and 2.5 CPU days for 1YRF. We currently work on a relief of this computational cost.

We conclude that the BH approach with chemical shift restraints is a promising method for protein structure prediction. The approach is an addition to existing methods based on chemical shift restrained Monte Carlo simulations using a simulated annealing protocol,[6] molecular dynamics simulations with chemical shift restraints,[7] and various molecular fragment replacement approaches with chemical shift information.[4,5,8,9,16,20] The three proteins that we considered as test cases contain fewer than 50 amino acids, and

have relatively simple topologies. It is expected that the amount of computational time required to achieve convergence will significantly increase for larger proteins with more complex topologies, which will probably limit the application of the current implementation of the BH approach with chemical shift restraints to proteins not much larger than 50 to 60 residues. Therefore, we are currently implementing knowledge-based Monte Carlo moves into the GMIN program, which should speed up the folding of secondary structure elements for BH runs with and without chemical shifts restraints. Additionally, the BH approach could easily be combined with restraints traditionally used in NMR structure calculations such as NOEs, *J*-couplings, and RDCs, which, in connection with chemical shift restraints, will open the possibility for the BH approach to become a valuable tool in structural biology.

ACKNOWLEDGMENTS

The authors thank Dr. Andrea Cavalli for helpful discussions and Professor David Wales for proofreading the manuscript.

[1]D. S. Wishart and D. A. Case, Method Enzymol. **338**, 3 (2002).
[2]C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore, J. Magn. Reson. **160**, 65 (2003).
[3]A. T. Brunger, Nat. Protoc. **2**, 2728 (2007).
[4]A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, Proc. Natl. Acad. Sci. U.S.A. **104**, 9615 (2007).
[5]P. Robustelli, A. Cavalli, and M. Vendruscolo, Structure (London) **16**, 1764 (2008).
[6]P. Robustelli, A. Cavalli, C. M. Dobson, M. Vendruscolo, and X. Salvatella, J. Phys. Chem. B **113**, 7890 (2009).
[7]P. Robustelli, K. Kohlhoff, A. Cavalli, and M. Vendruscolo, Structure (London) **18**, 923 (2010).
[8]Y. Shen *et al.*, Proc. Natl. Acad. Sci. U.S.A. **105**, 4685 (2008).
[9]D. S. Wishart *et al.*, Nucleic Acids Res. **36**, W496 (2008).
[10]M. Berjanskii *et al.*, Nucleic Acids Res. **37**, W670 (2009).
[11]R. Das *et al.*, Proc. Natl. Acad. Sci. U.S.A. **106**, 18978 (2009).
[12]F. C. Bernstein *et al.*, J. Mol. Biol. **112**, 535 (1977).
[13]J. Kuszewski, A. M. Gronenborn, and G. M. Clore, J. Magn. Reson. Ser. B **107**, 293 (1995).
[14]J. G. Pearson, J.-F. Wang, J. L. Markley, H.-B. Le, and E. Oldfield, J. Am. Chem. Soc. **117**, 8823 (1995).
[15]P. Luginbühl, T. Szyperski, and K. Wüthrich, J. Magn. Reson. Ser. B **109**, 229 (1995).
[16]H. Gong, Y. Shen, and G. D. Rose, Protein Sci. **16**, 1515 (2007).
[17]R. Montalvao, A. Cavalli, and X. Salvatella, J. Am. Chem. Soc. **130**, 15990 (2008).
[18]Y. Shen, R. Vernon, D. Baker, and A. Bax, J. Biomol. NMR **43**, 63 (2009).
[19]P. Neudecker *et al.*, Science **336**, 362 (2012).
[20]F. Delaglio, G. Kontaxis, and A. Bax, J. Am. Chem. Soc. **122**, 2142 (2000).
[21]G. M. Clore and A. M. Gronenborn, Proc. Natl. Acad. Sci. U.S.A. **95**, 5891 (1998).
[22]K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, and M. Vendruscolo, J. Am. Chem. Soc. **131**, 13894 (2009).
[23]D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111 (1997).
[24]D. J. Wales and H. A. Scheraga, Science **285**, 1368 (1999).
[25]Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
[26]P. Derreumaux, J. Chem. Phys. **106**, 5260 (1997).
[27]P. Derreumaux, J. Chem. Phys. **107**, 1941 (1997).
[28]P. N. Mortenson and D. J. Wales, J. Chem. Phys. **114**, 6443 (2001).
[29]P. N. Mortenson, D. A. Evans, and D. J. Wales, J. Chem. Phys. **117**, 1363 (2002).
[30]J. M. Carr and D. J. Wales, J. Chem. Phys. **123**, 234901 (2005).
[31]A. Verma, A. Schug, K. H. Lee, and W. Wenzel, J. Chem. Phys. **124**, 044515 (2006).
[32]B. Strodel and D. J. Wales, J. Chem. Theor. Comput. **4**, 657 (2008).
[33]B. Strodel, J. Lee, C. Whittleston, and D. Wales, J. Am. Chem. Soc. **132**, 13300 (2010).
[34]O. O. Olubiyi and B. Strodel, J. Phys. Chem. B **116**, 3280 (2012).
[35]A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, Proc. Natl. Acad. Sci. U.S.A. **98**, 5578 (2001).
[36]J. Neidigh, R. Fesinmeyer, and N. Andersen, Nat. Struct. Biol. **9**, 425 (2002).
[37]T. K. Chiu *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 7517 (2005).
[38]C. Simmerling, B. Strockbine, and A. E. Roitberg, J. Am. Chem. Soc. **124**, 11258 (2002).
[39]S. Chowdhury, M. C. Lee, G. Xiong, and Y. Duan, J. Mol. Biol. **327**, 711 (2003).
[40]A. Schug, T. Herges, and W. Wenzel, Phys. Rev. Lett. **91**, 158102 (2003).
[41]A. Schug, T. Herges, A. Verma, K. H. Lee, and W. Wenzel, ChemPhysChem **6**, 2640 (2005).
[42]A. Schug, W. Wenzel, and U. Hansmann, J. Chem. Phys. **122**, 194711 (2005).
[43]K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[44]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Biophys. J. **100**, L47 (2011).
[45]J. Maupetit, P. Derreumaux, and P. Tufféry, Nucleic Acids Res. **37**, W498 (2009).
[46]P. Thévenet *et al.*, Nucleic Acids Res. **40**, W288 (2012).
[47]J. W. Pitera and W. Swope, Proc. Natl. Acad. Sci. U.S.A. **100**, 7587 (2003).
[48]J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U.S.A. **103**, 15859 (2006).
[49]D. Paschek, H. Nymeyer, and A. E. García, J. Struct. Biol. **157**, 524 (2007).
[50]K. Klenin and W. Wenzel, Int. J. Comput. Commun. **1**, 1 (2007).
[51]I. H. Radford, A. R. Fersht, and G. Settanni, J. Phys. Chem. B **115**, 7459 (2011).
[52]J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **359**, 546 (2006).
[53]T. Cellmer, M. Buscaglia, E. R. Henry, J. Hofrichter, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **108**, 6103 (2011).
[54]Y. Tang, M. J. Grey, J. McKnight, A. G. Palmer III, and D. P. Raleigh, J. Mol. Biol. **355**, 1066 (2006).
[55]P. L. Freddolino and K. Schulten, Biophys. J. **97**, 2338 (2009).
[56]D. R. Ripoll, J. A. Vila, and H. A. Scheraga, J. Mol. Biol. **339**, 915 (2004).
[57]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[58]See supplementary material at http://dx.doi.org/10.1063/1.4773406 for a graphical presentation of the starting structures, the performance of chemical shift restrained BH runs for various (α, n) pairs, and a graphical presentation showing the statistics of how frequently the different chemical shifts are measured in proteins as derived from a total of about 5.6×10^6 chemical shifts in the Biological Magnetic Resonance Data Bank.
[59]B. R. Brooks *et al.*, J. Comput. Chem. **4**, 187 (1983).
[60]A. D. MacKerell, Jr. *et al.*, J. Phys. Chem. B **102**, 3586 (1998).
[61]U. Haberthür and A. Caflisch, J. Comput. Chem. **29**, 701 (2008).
[62]N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[63]D. J. Wales, *GMIN: A Program for Basin-Hopping Global Optimisation*, see http://www-wales.ch.cam.ac.uk/software.html.
[64]M. A. Miller and D. J. Wales, J. Chem. Phys. **111**, 6610 (1999).
[65]M. Bauer, B. Strodel, S. Fejer, E. Koslover, and D. Wales, J. Chem. Phys. **132**, 054101 (2010).
[66]R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, SIAM J. Sci. Stat. Comput. **16**, 1190 (1995).
[67]D. Frishman and P. Argos, Proteins: Struct., Funct., Bioinf. **23**, 566 (1995).