Protein structure prediction: assembly of secondary structure elements by basin-hopping

Falk Hoffmann*[1] Ioan Vancea†[1] Sanjay G. Kamat* Birgit Strodel*‡§

May 25, 2014

1 Abstract

The prediction of protein tertiary structure from primary structure remains a challenging task. A possible approach to this problem is the application of basin-hopping global optimization combined with an all-atom force field. In this work, we further improve the efficiency of basin-hopping by introducing an approach that derives tertiary structures from the secondary structure assignments of individual residues. We term this approach secondary-to-tertiary basin-hopping and benchmark it for three miniproteins, trpzip, trp-cage and ER-10. For each of the the three miniproteins the secondary-to-tertiary basin-hopping approach successfully and reliably predicts the three-dimensional structure. When it is applied to larger proteins we also obtain correctly folded structures. We thus conclude that the assembly of secondary structure elements using basin-hopping is a promising tool for *de novo* protein structure prediction.

[1]These authors contributed equally to this work.
*Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, 52425 Jülich, Germany
†European Molecular Biology Laboratory c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany
‡Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
§corresponding author: b.strodel@fz-juelich.de

2 Introduction

The prediction of protein structure from their amino acid sequence is one of the most important computational problems in bioinformatics and one of the great challenges in structural biology. Knowledge of the three-dimensional structure of proteins will give invaluable insights into the molecular basis of protein functions, and will therefore facilitate finding treatments and cures for many diseases. It is generally assumed that a protein folds to a native conformation or ensemble of conformations which is at or near the global free-energy minimum.[1] Thus, protein structure prediction can be understood as the search for an energy minimum in the conformational space of the protein. From a computational point of view, the problem of finding native-like conformations for a given primary structure, which is referred to as *de novo* protein structure prediction, can be decomposed into two subproblems: (a) developing an accurate energy function for which native protein fold and energy minimum coincide; (b) developing an efficient protocol for searching the energy landscape.

The focus of the current study is on the latter task. The extensive exploration of the whole conformational space of a protein is generally not possible as it would be a time prohibitive endeavor. Approaches based on the Metropolis Monte Carlo (MC) method offer the possibility to efficiently explore the conformational space or at least specific regions of it. Searching for the conformational space by MC methods usually involves a two-step process, a trial conformation move followed by an energy evaluation. In this work we use the basin-hopping (BH) global optimization algorithm[2,3] which is analogous in principle to the Monte Carlo-minimization approach.[4] Global optimization can be defined as the procedure of finding the lowest value of a given function. The BH algorithm is a stochastic global optimization method, which employs MC moves on a transformed potential energy surface, where a structural perturbation is followed by energy minimization. Basin-hopping has been employed successfully to find the global minimum of peptides and proteins,[5–12] including peptide complexes.[13–15]

Several possibilities to improve the efficiency of MC sampling exist, including the optimization of trial moves for proteins[16] and applying experimental restraints during an MC simulation.[17,18] The topic of the current work is the improvement of the trial moves. A typical protein MC move consists of randomly moving residues in a single MC step where these residues are often contiguous. The efficiency of the trial moves can be increased by incorporating residue-specific structural preferences derived from experimental structures.[19,20] It is well known that the Φ and Ψ angles of the protein backbone are more densely centered around some regions with the distribution of the (Φ, Ψ) densities depending on the amino acid identity.[21] Likewise, protein side chains tend to exist in a limited number of low energy conformations called rotamers.[22] Instead of considering the full geometrically possible conformational space, only populated (Φ, Ψ) regions and a small number of rotamers can be used for designing MC moves to describe the most frequently occurring amino acid conformations.

Another way to incorporate database driven information into an MC scheme can be realized by basing the protein structure prediction on secondary structure assignments of the residues.[23,24] The secondary structure is the three-dimensional form of local segments of proteins, which consists of local inter-residue interactions mediated by hydrogen bonds. Amino acids vary in their ability to form the various secondary structure elements. The dominating secondary structures are α-helices (henceforth denoted H) and sheets consisting of β-strands (henceforth denoted E). These regular secondary structure elements are linked by tight turns or loose, flexible loops. Furthermore, other types of helices, such as the 3_{10}-helix and π-helix exist. These structural elements will be collectively denoted C for 'coil' in the following sections. It should be noted that random coil is not a true secondary structure, but is the class of conformations that indicate an absence of regular secondary structure. Thus, the secondary structure of a protein is characterized by a sequence of letters over the alphabet {E,H,C}, with one letter per amino acid of the primary protein structure. Most secondary structure prediction methods are evolution-based methods (aka, homology-based methods), which either exploit neural network-based approaches (e.g., Porter[25] and Psipred[26]), hidden Markov models (e.g., SAM[27]), or the frequency analysis of amino acid conformational states (e.g., Gor IV[28]). In this work we use Porter for the prediction of secondary structure as it was identified as the best performing secondary structure prediction methods.[29] Miceli et al. compared the performance of nine secondary structure prediction tools applied to two protein data sets.[29] In the current work we confirm that Porter is superior to the other methods based on another performance criterion than those used by Miceli et al.

The secondary structure assignment is followed by the actual folding simulation using basin-hopping. Here, we apply MC moves only to the intervening amino acids in the C conformation and connecting the H or E secondary structure elements, allowing them to establish their tertiary contacts. We term this approach secondary-to-tertiary basin-hopping. It is similar in idea to fragment assembly approaches, which are applied in the *de novo* methods Rosetta[30–32] and Chunk-Tasser.[33] We show for the three peptides trpzip, trp-cage and ER-10 with PDB[34] codes 1LE0,[35] 1L2Y[36] and 1ERP,[37] respectively, that this secondary-to-tertiary BH implementation allows the reliable prediction of correctly folded structures within 2,500 BH steps. Furthermore, we demonstrate for larger proteins with up to 79 residues that our approach is able to predict correctly folded protein structures. These developments make the BH approach to global optimization a promising tool for *de novo* protein structure prediction, which is computationally less demanding compared to other prediction methods and which will be ensued in future applications.

Set	Predictor	H↔E		H↔C		E↔C	
all	Porter	9	(0.1%)	511	(2.9%)	484	(2.7%)
	Psipred	186	(1.1%)	1565	(8.8%)	1765	(10.0%)
	SAM	210	(1.2%)	1645	(9.3%)	1765	(10.0%)
≤ 100	Porter	2	(0.1%)	42	(2.8%)	54	(3.5%)
	Psipred	12	(0.8%)	126	(8.2%)	184	(12.0%)
	SAM	19	(1.2%)	174	(11.4%)	178	(11.6%)
> 100	Porter	6	(0.0%)	469	(2.9%)	430	(2.7%)
	Psipred	174	(1.1%)	1439	(8.9%)	1581	(9.8%)
	SAM	191	(1.2%)	1471	(9.1%)	1587	(9.8%)
α	Porter	0		44		3	
	Psipred	8		160		29	
	SAM	12		139		14	
β	Porter	0		3		63	
	Psipred	0		2		108	
	SAM	6		18		117	

Table 1: Performance analysis of secondary structure prediction methods. The numbers of secondary structure mix-ups are provided for Porter, Psipred and SAM for all proteins of the PDB25Select database, proteins with ≤ 100 and > 100 amino acids, and α and β proteins. For the first three protein sets, the percentage of mix-ups relative to the total number of amino acids in the set in question is given in parantheses.

3 Results and Discussion

3.1 Testing of secondary structure predictors

We evaluate the precision of the three secondary structure predictors Porter,[25] Psipred[26] and SAM[27] by counting the number of H↔E, H↔C and E↔C mix-ups for the PDB25Select database. Table 1 shows the results for all proteins of the database, which are further split into small proteins with less than or equal to 100 residues and large proteins of greater than 100 residues, and into α and β proteins that contain only α-helices and β-sheets, respectively.

The most striking result is that in most cases Porter performs much better than Psipred and SAM. The number of mix-ups is by a factor of 3 to $\gtrsim 30$ larger for Psipred and SAM in comparison to Porter for almost all cases (i.e., type of mix-up and protein set). The only exception are the Psipred predictions for β proteins. For this set, neither Porter nor Psipred wrongly predicted H instead of E (SAM has 6 such mix-ups), while there are only 3 and 2 H assignments instead of C for Porter and Psipred, respectively. With regard to E↔C, Porter again performs significantly better than Psipred. For all three prediction methods, the number of H↔E mix-ups is by at least one order of magnitude lower than the number of H↔C and E↔C mix-ups, independent of protein length and type of fold. This indicates that helices and β-sheets can be distinguished from one another by

Peptide	Secondary structure	
trpzip	assignment:	CEEECCCCEEEC
	target:	CEEEECCEEEEC
trp-cage	Porter:	CHHHHHHHCCCCCCCCCCCC
	target:	CHHHHHHHHCHHHCCCCCC
ER-10	Porter:	CHHHHHHHCCCHHHHHHCCCCCHHHHHHHHHCCCCCCC
	target:	CHHHHHHHHCCHHHHHHHCCCHHHHHHHHHHCCCCCC

Table 2: Secondary structure assignments along with the target secondary structure. For trp-cage and ER-10 the secondary structure assignments were obtained from Porter, while they were manually assigned for trpzip. The letter 'H' for residues 11-13 in the trp-cage target denotes a 3_{10} helix.

Porter, Psipred and SAM, which is important since the basis for the current BH approach is the accurate assignment of secondary structure for the subsequent assembly of the tertiary structure. This assumption is especially justified for Porter, which only has 0.1% H↔E mix-ups for the total database, that only affect α/β proteins (i.e., proteins that contain both α-helices and β-sheets) as there are no H↔E mix-ups for α and β proteins.

These findings led us to use Porter for the prediction of secondary structures as the starting point for our BH simulations. Compared to H↔E, the numbers of H↔C and E↔C mix-ups are somewhat higher but generally below 3% for Porter. For small proteins the correct prediction of β-sheets seems to be slightly more difficult with 3.5% E↔C mix-ups demonstrated by Porter. For both α and β proteins, E was predicted instead of C and H was predicted instead of C for only 3 residues in each case. This again shows that Porter is highly capable of distinguishing α- and β-folds. Based on the Porter prediction, the main task of the subsequent BH simulations is to identify the correct tertiary contacts and to correct wrongly assigned secondary structures, which mainly involve those in which C was wrongly assigned instead of H or E.

3.2 First BH round: from secondary to tertiary structure

Porter was used to determine the secondary structure of the residues of trp-cage and ER-10, while they were manually assigned in the case of trpzip based on its target structure as this peptide is too short to be treated by Porter. In Table 2 we present the assignments together with the secondary structure of the targets. The Porter predictions for the helix lengths in trp-cage and ER-10 are often short by one residue, while all other predictions are correct. The 3_{10}-helix in trp-cage (indicated by the letter 'H' for residues 11–13 in the target) is by default not considered by Porter as only α and β structures are assigned. Thus, the 3_{10}-helix has to be found by the BH approach. For trpzip we assigned residues 5 and 8 to be in the coil state in order to evaluate if the BH methodology is able to identify the full β-sheet.

The BH runs in this round employed the information from Table 2 as described in Section 6.4. For each peptide, high-temperature molecular dynamics simulations were used to generate 20 different unfolded structures, which were taken as starting structures of the BH runs. We considered three different maximum dihedral twisting angles of 30°, 60° and 90° per starting structure. Furthermore, 10 independent BH runs were performed for each starting structure and twisting angle, using different seeds for the random number generation. This amounts to $20 \times 10 \times 3 = 600$ BH runs per peptide. BH runs were conducted for 1,000, 2,000 and 5,000 Monte Carlo steps (aka BH steps) for trpzip, trp-cage and ER-10, respectively. As an example, the BH input file for trpzip with step size 60° is provided in the Supplementary Information.

Energy versus RMSD plots. The performance of each BH run was measured in terms of the energy and C_α root mean square deviation (RMSD) from the target structure, and the three best structures per run as determined by both energy and RMSD were considered for analysis. In the following, these sets of structures are denoted as low-energy and low-RMSD structures, respectively. In the ideal case the energy function ranks the native structure in first place with respect to energy (lowest energy), i.e., the sets of low-energy and low-RMSD structures are identical or at least overlap to a large extent. Figure 1 shows the energy versus RMSD plots for low-energy (blue) and low-RMSD structures (red) for the 600 BH runs per peptide, along with the structures of overall lowest energy and lowest RMSD. In this figure, the results for the maximum twisting angles of 30°, 60° and 90° are displayed together. Detailed results for the individual step sizes are provided in Figures S1 to S3 of the Supplementary Information.

The results for the energy-minimized target structures (i.e., the energy-minimized PDB structures) using the CHARMM22/FACTS energy function are displayed as yellow dots in Figure 1. The changes to the RMSD as a result of the minimization procedure are small (< 0.5 Å). In the following we will use the structure of the energy-minimized target as a reference for the RMSD calculations since within the BH procedure one cannot expect to get closer to the PDB structure than the minimized target structure. Thus, the yellow dots in Figure 1 occur at RMSD zero. The energy of the energy-minimized target structures is higher than the energy of many of the low-energy structures. The target structures are NMR solution structures, which were determined by minimizing the distance or dihedral angle violations resulting from experimental constraints. It is important to note that the ensemble of structures obtained is an 'experimental model', which is not necessarily the best solution when modeled with an empirical force field, such as CHARMM22/FACTS. We therefore subjected the three target structures to further optimization by performing BH runs of 1,000 steps with maximum dihedral angle changes of 20°, which were applied to both backbone and side chains of 3–5 randomly selected contiguous residues. This procedure generated energy-optimized structures at the cost of the RMSD, which increases. The energy and RMSD values of these structures, which

we call optimized target structures, are -307.2 kcal mol^{-1} and 1.33 Å for trpzip, -510.7 kcal mol^{-1} and 1.82 Å for trp-cage, and -907.6 kcal mol^{-1} and 2.58 Å for ER-10, respectively. These results are provided as orange dots in the energy versus RMSD plots in Figure 1.

Figure 1 shows the success of the secondary-to-tertiary BH procedure, i.e., the application of Monte Carlo moves at residues between secondary structure elements H and E to obtain tertiary structure from the secondary structure data. For all three peptides native-like structures are found, where a threshold for the C_α RMSD of 2.0 Å from the target for defining native-like conformations is used. For trpzip, trp-cage and ER-10 we identified 198, 512 and 2 native-like conformations, respectively. The lowest-RMSD structures shown in Figure 1 have RMSD values of 1.22 Å (trpzip), 0.55 Å (trp-cage), and 1.63 Å (ER-10), while the lowest-energy structures have RMSD values of 3.15 Å, 0.70 Å and 8.05 Å, respectively. The results for trp-cage indicate that for this peptide the CHARMM22/FACTS potential can distinguish the native structure from unfolded structures. This conclusion is supported by the funnel shape of the energy versus RMSD plot for trp-cage. Furthermore, the secondary-to-tertiary BH approach samples native-like structures of lower RMSD and lower energy than obtained for the optimized target structure (orange dot in Figure 1).

For trpzip the best structure obtained so far is the optimized target structure. The lowest-RMSD structure has an energy which is 35 kcal mol^{-1} higher than the energy of the optimized target structure, and the RMSD of the latter structure is almost 2 Å below the RMSD of the lowest-energy structure. The lowest-RMSD structure exhibits the hairpin structure yet no β-sheet is formed due to the missing H-bonds between the two strands. Figure S4 in the Supplementary Information shows the various structures for trpzip obtained in this work, with the H-bonds indicated in these structure plots and an analysis of the interaction energies between the residues in these structures. The higher energy of the lowest-RMSD structure compared to that of the optimized target is due to the missing H-bonds and the electrostatic stabilization between Glu5 and Lys8. Moreover, the tryptophan residues Trp2, Trp4, Trp9 and Trp11 in the lowest-RMSD structure are not oriented as they are in the target, where they are stacked and T-shaped with respect to each other, which further destabilizes the lowest-RMSD structure. On the other hand, in the lowest-energy structure the β-sheet is partially formed but the turn region deviates from the target structure. The turn folds towards the β-sheet and is stabilized by a H-bond between the side chains of Asn7 and Thr10. However, the largest energetic stabilization of this structure compared to the target results from electrostatic attraction between the N- and C-terminal residues despite C-terminal amidation. The tryptophan residues, that are in a stacked orientation, also stabilize this structure, though the Trp2–Trp4 and Trp2–Trp11 interactions are not as strong as in the target (see Figure S4). This analysis reveals how subtle the interplay between atomic positions and overall energy in all-atom energy functions is, making protein structure prediction with all-atom models a challenge. Furthermore, it has been demonstrated that implicit solvent models tend to overweight nonnative states which are

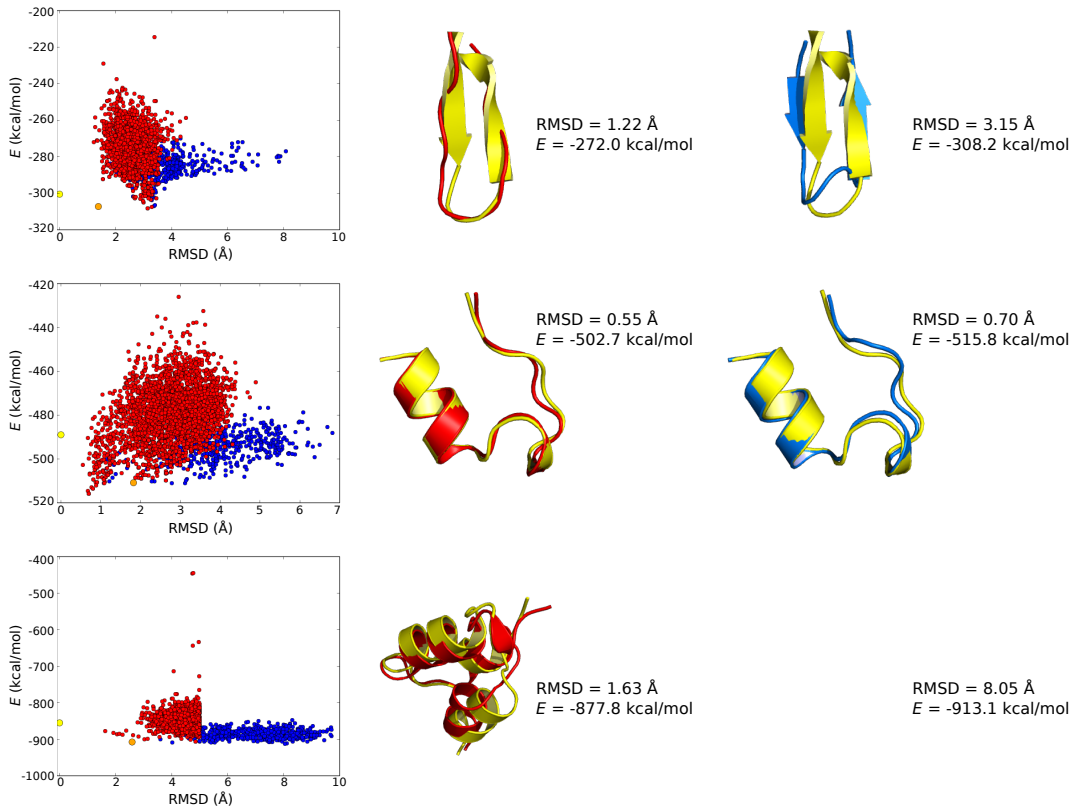


Figure 1: Results from the first BH round are shown for trpzip (top), trp-cage (middle) and ER-10 (bottom). (Left) Energy versus RMSD plots for the low-energy (blue) and low-RMSD (red) structures obtained from 600 BH runs for each peptide. For the low-RMSD structures only conformations with an RMSD < 5 Å are included, explaining the sharp cut at RMSD ≈ 5 Å for the red dots for ER-10. The minimized and the optimized target structures are represented by a yellow and an orange dot, respectively. (Middle) Lowest-RMSD structure (red) and (Right) lowest-energy (blue) structure along with the target structure (yellow). The RMSD and energy values of these structures are provided.

stabilized by nonnative electrostatic attractions.[38]

Nonetheless, for both trpzip and trp-cage we find an overlap between the sets of structures of low RMSD and low energy (i.e., between the red and blue dots in Figure 1), which indicates that the CHARMM22/FACTS energy function is able to predict native-like structures as low-energy structures for both peptides. Among the low-energy structures, there are 14 native-like structures for trpzip and 40 for trp-cage. However, for ER-10 we observe a clear separation between the red and blue dots in Figure 1 and find no native-like structure among the low-energy structures. We identify two native-like structures at energies of ≈ -880 kcal mol^{-1}, which is about 30 kcal mol^{-1} higher than the energy of the low-energy structures at rather high RMSD. The comparison between the RMSD and energy values for the lowest-RMSD and lowest-energy structures shown in Figure 1 highlights this observation. Nevertheless, there are some ER-10 conformations with RMSD values below 4 Å among the low-energy structures. This result indicates that the secondary-to-tertiary BH approach is also able to identify native-like structures for ER-10. The question rather is whether the considered energy function can distinguish between near-native and nonnative conformations for ER-10, which will be addressed in section 3.3.

The dependence of prediction efficiency on step size. Apart from reliably identifying near-native structures as discussed above, the aim is also to find them quickly. To this end, we determined for each step size and peptide the average RMSD and energy values of the low-RMSD and low-energy structures, respectively. In addition, we analyzed how many BH steps were needed to locate the lowest-RMSD and lowest-energy structures in each BH run. These quantities allow us to deduce which of the maximum twisting angles of 30°, 60° or 90° yields the best and fastest predictions. The results of this analysis are presented in Figure 2. Panels A and B of Figure 2 allow us to conclude that the step size has no large influence on the identification of low-RMSD and -energy structures. For all three peptides, the averaged RMSD and energy values are very similar for the considered step sizes and none of the step sizes consistently outperforms the others. The energy versus RMSD plots for the different step sizes (Supplementary Information) demonstrate that the identification of similar structures is independent of the maximum twisting angle. The supplementary figures also reveal that the final RMSD and energy values do not depend on the RMSD and energy values of the starting structures. That is, near-native structures are identified not only when the BH run is initiated from rather folded but also from completely unfolded conformations.

Panel C of Figure 2 shows that the use of a maximum step size of 60° or 90° enables near-native structures to be located more quickly than when the maximum twisting angle is only 30°. Low-RMSD conformations are generally produced faster than low-energy structures, implying that the RMSD did not further approve upon improving the energy. This is due to the above mentioned problem of assuming the native structure as global energy minimum and the fact that atom-based

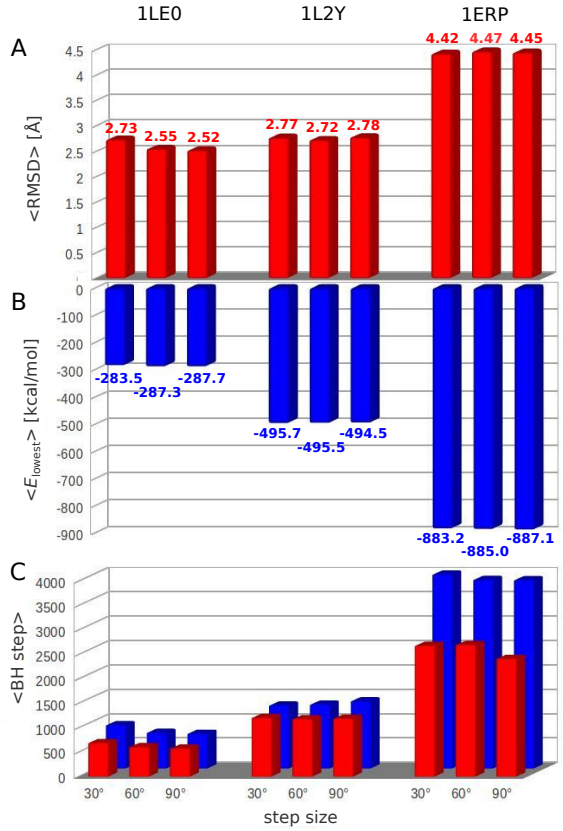


Figure 2: Results from the first BH round are shown for the maximal step size of 30°, 60° or 90°. (A) The mean of the RMSDs of the low-RMSD structures and (B) the mean of the energies of the low-energy structures, averaged over the 200 BH runs per peptide and step size, are shown. (C) The average numbers of BH steps needed to locate the structures of lowest RMSD (red) and lowest energy (blue) in each of the BH runs are provided.

potentials are particularly sensitive to the precise position of the interacting atoms, hampering the detection of native-like geometries. As already pointed out for ER-10, the CHARMM22/FACTS potential does not identify the native structure as the global minimum on the potential energy surface. Therefore, for this peptide on average 1,000 BH steps more are needed to find the lowest-energy structure than are required for the location of the lowest-RMSD structure. This problem is less aggravated for trpzip and trp-cage, for which native-like structures correspond to low-energy structures.

In summary, less than 1,000, 2,000 and 5,000 BH steps are generally sufficient to detect near-native (or low-energy) structures for trpzip, trp-cage and ER-10, respectively. The average computational time required for each BH run was 1.7 h for trpzip, 7.4 h for trp-cage and 54.1 h for ER-10 on a single 2.93 GHz Intel Xeon Processor X5570. While a smaller twisting angle does not prevent the identification of near-native structures, a larger step size of 60° or 90° helps to find them faster. Thus, we conclude that the secondary-to-tertiary BH approach works. The aim of the following BH round is to test whether the low-RMSD and low-energy conformations identified so far can be further optimized by unconstrained BH simulations.

3.3 Second BH round: refinement of tertiary contacts

From the structures obtained in the previous BH round we randomly selected 31 conformations for trpzip, 38 for trp-cage, and 54 for ER-10 with low RMSD, and 56 conformations for trpzip, 77 for trp-cage, and 70 for ER-10 with low energy. Here, we applied following upper cutoffs for the selection of structures based on either RMSD or energy: 3.0 Å and -295 kcal mol^{-1} for trpzip, 2.5 Å and -505 kcal mol^{-1} for trp-cage, 5.0 Å and -900 kcal mol^{-1} for ER-10. For each starting structure we performed three independent BH runs of 5,000 steps for trpzip and trp-cage, and 7,000 steps for ER-10. In this round we released all constraints and applied dihedral angle changes to three, four or five randomly selected contiguous residues. Here, all residues were considered independent from the initial secondary structure prediction, thereby enabling wrongly predicted secondary structures to be corrected during the BH optimization procedure. We tested different ratios of dihedral angle changes for the backbone (BB) and side chains (SC): i) alternating BB and SC moves; ii) a SC move every 5th BH step, else BB moves; iii) a BB move every 5th BH step, else SC moves. The different BB:SS frequency schemes are subsequently denoted as 1:1, 4:1 and 1:4. Hence, the number of BH runs is $(31+56)\times3\times3 = 783$ for trpzip, $(38+77)\times3\times3 = 1,035$ for trp-cage and $(70+54)\times3\times3 = 1,116$ for ER-10. The performance of each BH run was measured in terms of energy and RMSD considering the three best structures for both quantities. In addition, we monitored whether a BH run was started from a low-RMSD or a low-energy structure from the previous BH round. The maximum dihedral angle change in each run was 30° with group rotation moves[39] applied to the side chains.

The small step size was chosen since in this BH round the aim is to further optimize near-native (or low-energy) structures, and not to generate completely different structures. An example input file for such a BH run for trpzip is provided in the Supplementary Information.

The simulations of this round were analyzed in the same manner as the BH simulations of the first round. We produced energy versus RMSD plots, which are shown in Figure 3 together with the structures of lowest RMSD and lowest energy detected for each peptide. We calculated the average RMSD and energy of all low-RMSD and -energy structures, respectively, taking into account whether a BH run was started from a low-RMSD or a low-energy conformation from the first BH round. In order to be able to decide which of the BB:SC perturbation ratios works best, we monitored the average number of BH steps needed before the best structure with respect to RMSD or energy was detected. Below we present the combined results for all BB:SC move ratios while in Figures S5 to S7 results are shown separately for the 1:1, 4:1 and 1:4 move combinations.

Energy versus RMSD plots. The first, very obvious result is that unconstrained remodeling of the structures identified in the first secondary-to-tertiary BH round leads to a considerable decrease in both energy and RMSD. As before, the energy-minimized target structure is used as reference for the calculation of the RMSD. The average energy decreased by ≈ 30 kcal mol^{-1} for trpzip, by ≈ 25 kcal mol^{-1} for trp-cage, and by even ≈ 35 kcal mol^{-1} for ER-10. For all three peptides the optimized target structure (orange dot in the energy versus RMSD plots) no longer belongs to the best structures, neither in terms of RMSD nor energy. For trpzip and trp-cage many near-native structures were detected: of all saved structures, 25.6% and 44.9% have an RMSD ≤ 2Å for trpzip and trp-cage, respectively. Especially for trp-cage, it is almost unimportant whether the successful runs were initiated from low-RMSD or low-energy structures from the previous BH round. The information about the starting structure is provided in the energy versus RMSD plots in Figure 3 by using light colors for low-RMSD starting structures and dark colors for low-energy starting structures. The ratio of light and dark colored dots below 2 Å is 5.6:1 for trpzip and 1:1.2 for trp-cage.

For trpzip, many of the low-energy structures have an RMSD < 2 Å, which means that the CHARMM22/FACTS potential can distinguish between native-like and nonnative structures for the β-hairpin. However, it has to be noted that the structure of lowest RMSD (RMSD $= 0.44$ Å) has an energy of more than 20 kcal mol^{-1} above the value for the lowest-energy conformation with an RMSD of 1.68 Å. In the latter, the hairpin is properly formed yet it lacks the β-sheet. It has fewer backbone H-bonds compared to the target structure since the two strands are not perfectly aligned for β-sheet formation. Instead, this structure is mainly stabilized by a H-bond between the N- and C-terminal residues, leading to an energy decrease of more than 60 kcal mol^{-1} compared to the same inter-residue interaction in the target (see Figure S4 in the Supplementary Information). Another appreciable stabilization of ≈ 30 kcal mol^{-1} originates from another H-bond between the side chains

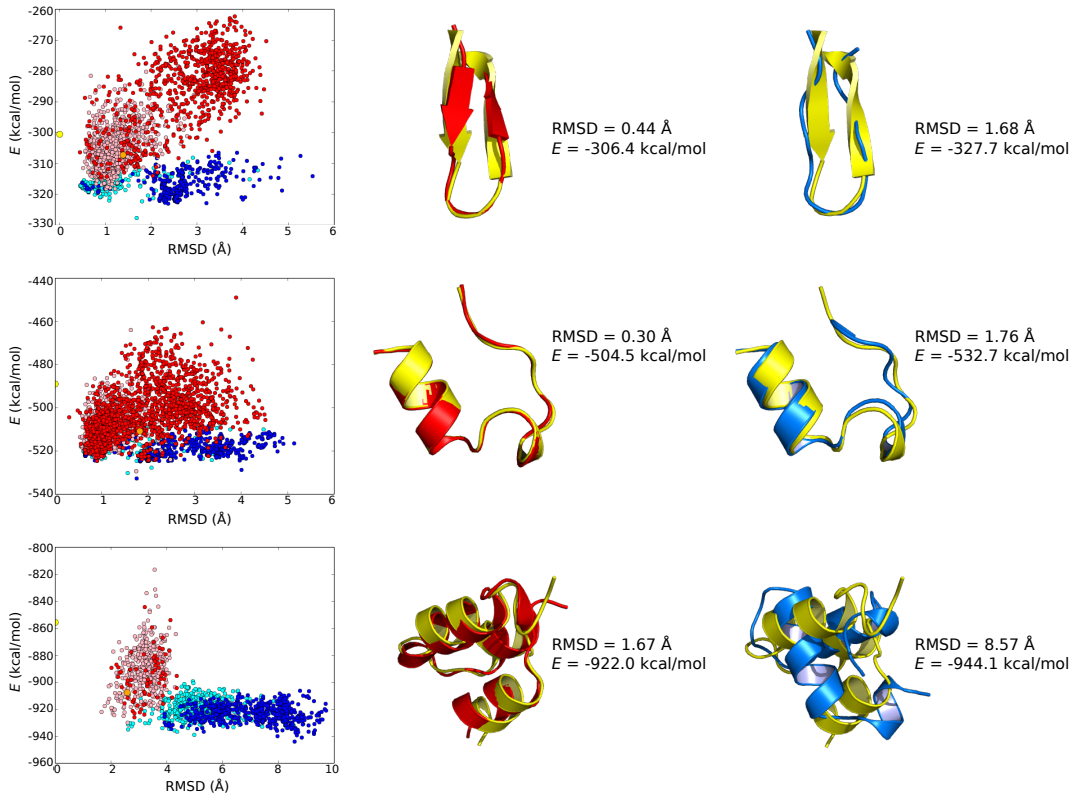


Figure 3: Results from the second BH round are shown for trpzip (top), trp-cage (middle) and ER-10 (bottom). (Left) Energy versus RMSD plots for the low-energy (blue) and low-RMSD (red) structures obtained from 783 BH runs for trpzip, 1,035 BH runs for trp-cage and 1,116 BH runs for ER-10. The darkness of the colors indicates whether a BH run was started with a structure of low RMSD (light red or blue) or of low energy (dark red or blue) obtained in the first BH round. The minimized and the optimized target structures are represented by a yellow and an orange dot, respectively. (Middle) Lowest-RMSD structure (red) and (Right) lowest-energy (blue) structure along with the target structure (yellow). The RMSD and energy values of these structures are provided.

of Glu5 and Asn7. The tryptophan residues are not perfectly oriented with respect to each other, leading to higher interaction energies compared to the target. In the lowest-RMSD structure the β-sheet is partially formed. The largest deviation from the target structure occurs around Glu5 and Lys8, which are not in the β state and have their side chains oriented differently than in the target.

For trp-cage the findings are similar to those of trpzip: a structure of rather low RMSD (0.30 Å) was detected, which has an energy of ≈ 28 kcal mol^{-1} above that of the lowest-energy conformation. Yet the latter is also a near-native structure with an RMSD of 1.76 Å, which has the α-helix and 3_{10}-helix correctly formed. Only the C-terminal residues in coil conformation are slightly different arranged than in the target structure. This can be explained by the formation of H-bonds involving the side chains of the last five residues, creating a turn that is not present in the target structure (Figure S8 in the Supplementary Information). In conclusion, the CHARMM22/FACTS potential leads to a funnel shape of the energy versus RMSD plots for both trpzip and trp-cage, enabling the prediction of native-like structures for both peptides based on energy ranking.

The situation is different for ER-10. As was seen in the first BH round, we observe a separation between low-RMSD and low-energy structures. There is almost no overlap between these two sets of conformations, i.e., between the red and blue dots in the energy versus RMSD plot for ER-10 in Figure 3. The low-RMSD set contains only six native-like structures (RMSD $\leq$ 2 Å), while the majority of the low-energy structures has an RMSD $>$ 5 Å. The reason for this discrepancy is that the three helices in ER-10 are held together by three disulfide bridges, which are between the oxidized forms of Cys3 and Cys19, Cys10 and Cys37, and Cys15 and Cys27.[37] These disulfide bridges are not present in the lowest-energy structure, where the S–S distances are 16.7 Å for Cys3–Cys19, 4.9 Å for Cys10–Cys39, and 7.5 Å for Cys15-Cys27, while the disulfide bond length is 2.0 ± 0.2 Å. Instead, a salt bridge between Asp23 and Lys24 is formed in the lowest-energy structure giving rise to an interaction energy of -84.2 kcal mol^{-1} between these two residues. Thus, this salt bridge is very stable and prevents the formation of the correct turn between the second and third helix of this structure. In the CHARMM force field disulfide bonds between cystein residues have to be defined by the user during the setup of the protein model, i.e., there is currently no possibility for a disulfide bond to form during a simulation. This shortcoming could be addressed as in the sOPEP coarse-grained force field, which permits the formulation of S–S bonds based on the distance between the cystein side chain centroids.[40,41]

The dependence of prediction efficiency on move set. The statistical analysis of the simulation results in Figure 4 highlights that for trpzip and trp-cage the low-RMSD structures have a considerably lower RMSD when the BH runs were started from low-RMSD instead of the low-energy structures obtained in the first BH round (panel A in Figure 4). For ER-10, the differences between the average RMSD values of structures obtained when starting from low-energy or low-RMSD con-

formations are rather small (< 0.2 Å). Interestingly, the energy of the low-energy structures is not affected by the choice of the starting structures for any of the peptides (panel B in Figure 4). This allows us to conclude that BH remodeling of structures obtained from the initial secondary-to-tertiary approach is robust with respect to energy minimization, while the improvement of the RMSD can depend on the starting configuration. A marginal influence of the BB:SC move ratio is observed for the number of BH steps needed before the lowest-RMSD and lowest-energy structure in each BH run is detected. On average, the 1:4 move set needs fewer BH steps than the 1:1 and 4:1 move sets yet the improvement is only minor. Thus, it seems to be of some advantage to have more side chain moves (in contiguous residues) compared to backbone moves for the efficient lowering of energy and RMSD. This observation underpins the importance of side chain packing for the native protein structure, since both side-chain–side-chain and side-chain–backbone interactions play an important role to the stabilization of folded protein structures.[42] Nonetheless, we can conclude that the BH approach is robust with respect to the step taking scheme and step size, considering also the results from the previous BH round.

As in the first BH round, we observe that often fewer BH steps are needed to find low-RMSD structures compared to low-energy conformations (i.e., compare red versus blue bars in panel C of Figure 4). Though this result is not as clear as in the previous BH round and also not universal. In case of trpzip and ER-10, the BH steps needed to locate low-RMSD and low-energy structures are smaller when started from low-RMSD instead of low-energy structures from the first BH round (i.e., compare light colored versus dark colored bars in panel C of Figure 4). However, for trp-cage the average number of BH steps needed to encounter the structures of lowest energy and RMSD is independent of the starting structure. Furthermore, it is also independent of whether a lowest-RMSD or a lowest-energy structure was identified. This finding for trp-cage can be explained with the overlap between the two sets of low-RMSD and low-energy structures, in other words, low-RMSD and low-energy structures are often identical.

In summary, this statistical analysis confirms the above conclusion that refinement of structures in this BH round is most successful when started from structures, that are already close to the target structure. This finding justifies our two-step approach with a first secondary-to-tertiary BH round, followed by structure refinement of the best candidates in a second BH round. The average computational time required for each BH run in this round was 8.3 h for trpzip, 13.1 h for trp-cage and 63.6 h for ER-10, again on a single 2.93 GHz Intel Xeon Processor X5570. The longer simulation times compared to those of the first BH round can be explained with the larger number of BH steps applied in this round. The computational time could be reduced by reducing the number of BH steps, which is justified by the results presented in Figure 4C. It shows that only 2,000–3,000 BH steps were necessary for trpzip and trp-cage and less than 5,000 steps for ER-10, i.e., about 2,000 more BH steps were performed than actually needed. Another possibility to reduce the wall-clock

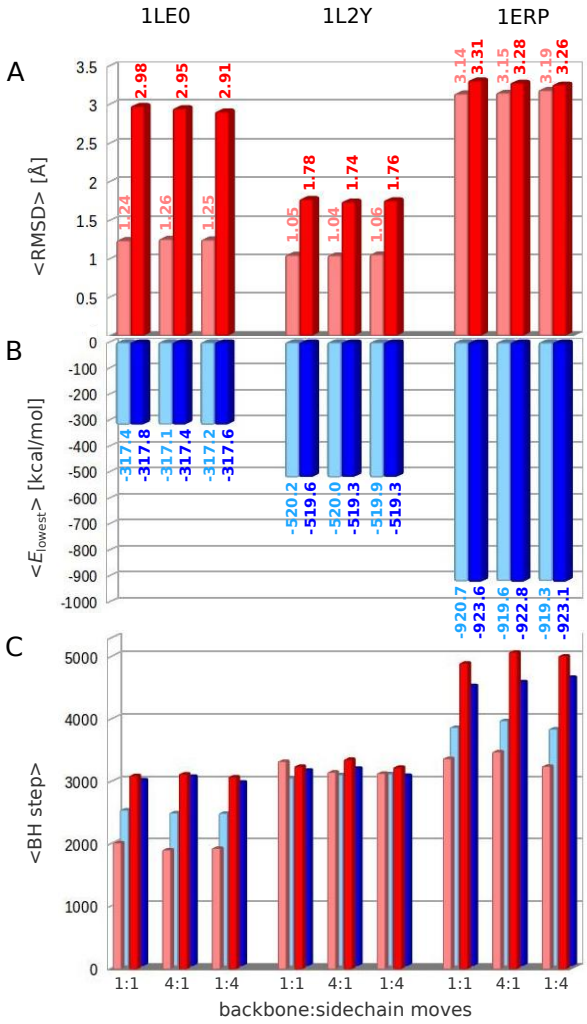


Figure 4: Results from the second BH round are shown for the BB:SC move combinations of 1:1, 4:1 and 1:4. (A) The mean of the RMSDs of the low-RMSD structures and (B) the mean of the energies of the low-energy structures, averaged over the BH runs per peptide and move combination. The results shown in light colors were obtained from initial structures taken from the low-RMSD set from the first BH round, while the darker colors are for the BH runs started from low-energy structures. (C) The average numbers of BH steps needed to locate the structures of lowest RMSD (red) and lowest energy (blue) in each of the BH runs are provided.

Peptide	Secondary structure	
villin	Porter:	CCHHHHHHHHCCCHHHHHCCCHHHHHHHHHHCCCC
	target:	CCHHHHHHHHCCCHHHHHHCCHHHHHHHHHHCCC
protein B	Porter:	CCHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHCCCHHHHHHHHHHHHHHHCC
	target:	CCCCCCCCCCHHHHHHHHHHCCCCCCCCHHHHHHHHHHCCCHHHHHHHHHHHHHHHCC
PyJ	Porter:	CCCCCCCHHHHHHHHHHHCCCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCC
	target:	CCCCCCCHHHHHHHHHHHHCCCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCC

Table 3: Secondary structure assignments along with the target secondary structure.

time is to implement a parallel version of the LBFGS minimization method used in this work.[43]

4 Protein structure prediction of larger proteins

The next step is to test whether the secondary-to-tertiary BH approach also works for larger proteins. To this end, we considered three proteins: the 35-residue villin, protein B with 53 residues, and a 79-residue DnaJ-like protein denoted PyJ. The folding of two of these proteins has recently been studied using molecular dynamics (MD)[44] and MC simulated annealing simulations.[45] The secondary structure predictions produced by Porter along with the secondary structure of the targets are shown in Table 3. It can be seen that Porter yields good predictions for the secondary structure. In some cases the predicted helices are by one residue too short. Only for protein B, Porter overestimates the helicity of the N-terminal residues next to the first helix.

We took unfolded conformations produced by high-temperature MD runs of these three proteins as starting structures for the BH simulations. The current BH simulations are only for the sake of demonstrating that our approach also works for larger proteins. Thus, we considered only two starting structures per protein with initial RMSD values from the target ranging from 8 Åfor villin to more than 30 Åfor protein B and PyJ. For each starting structure we performed 8 independent secondary-to-tertiary BH simulations using a maximum twisting angle of 60°. After 5,000 BH steps for the smaller proteins and 10,000 BH steps for PyJ the following structures of lowest RMSD were obtained: 2.9 Åfor villin, 2.4 Åfor protein B, and 6.4 Åfor PyJ. Visual inspection of these structures revealed that the helical structures adopted the correct tertiary topology. In the second BH round we aimed at refining the best structures from the first BH round by performing 56 BH runs for the lowest-RMSD structure for each protein. Here, we considered combinations of different maximal step sizes (30° or 40°) and different BB:SC move ratios (1:1, 5:1 and 1:5) and applied 3,000 BH steps in each run. We were able to further improve the structure predictions and obtained following RMSDs: 1.9 Åfor villin, 2.1 Åfor protein B, and 4.9 Åfor PyJ. The corresponding native-like structures are shown in figure 5. Given the larger size of PyJ, we assume a C_α-RMSD below 5 Åto be sufficient for the definition of native-like structures. We also tested our secondary-to-tertiary BH approach for a 36-residue WW domain (PDB code 2KCF). We obtained a reasonable RMSD value of 4.6Åwithin

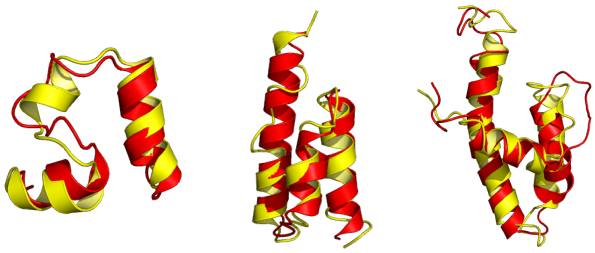


Figure 5: Native-like structures produced by basin-hopping for larger proteins. For each protein the lowest-RMSD structure (red) and target structure (yellow) are shown. The C_α-RMSDs between the two structures is given, along with the PDB entry of the target structure.

3,000 BH steps. However, the β-sheet structure was not fully established and further refinement for another 3,000 BH steps did not considerably improve this structure. Longer simulations and further methodological developments are needed for improving the prediction of long-range β-contacts with basin-hopping.

For villin and protein B our prediction results compare well to the results obtained by Lindorff-Larsen et al.[44] and Adhikari et al.[45] The RMSDs for the best structures for these two proteins are lowest for the MD predictions,[44] followed by our predictions and then those from the MC simulated annealing runs.[45] Though it should be noted that while the RMSD for our villin structure is lower than that obtained by Adhikari et al.,[45] in their structure the middle helix is better predicted than by our BH approach. For protein B, we mainly overpredict the helicity of the N-terminal helix which originates from the Porter prediction. In the protein B structure obtained by Adhikari et al. the helicity is also overestimated.[45] While the microsecond-long MD simulations involving explicitly represented solvent molecules produce the best native-like structures,[44] they are at the same time computationally most demanding.[45] The calculations by Adhikari et al. took around 600 CPU hours on an Intel 2.6 GHz Sandy Bridge Xeon E5-2670 processor for each protein. Using the same processor running NAMD, a single 10 μs MD trajectory would take around 3,000,000 CPU hours/protein.[45] The BH simulations presented here cumulated to less than 35 hours for villin, about 40 hours for protein B and 50 hours for PyJ on a single 2.93 GHz Intel Xeon Processor X5570 and counting the first and second BH round together. Thus, given the reduced computational demand of our BH approach and the good results for helical proteins, it may become a promising alternative to existing methods for protein structure prediction.

5 Conclusions

In this study, we have used the Monte Carlo (MC) based basin-hopping (BH) approach to global optimization as previous studies have shown that BH is very effective at predicting global minima of peptides[5–12] and peptide assemblies.[13–15] In order to further improve the efficiency of the BH approach to protein structure prediction, we have implemented knowledge-based MC moves by incorporating secondary structure information from secondary structure prediction. We refer to this approach as secondary-to-tertiary basin-hopping. We have evaluated the performance of the secondary-to-tertiary BH scheme for three peptides with PDB codes 1LE0 (trpzip), 1L2Y (trp-cage) and 1ERP (pheromone ER-10). To perturb the conformation of selected residues, we applied dihedral angle moves since simple Cartesian moves usually perform poorly because they tend to disrupt the bonded structure of molecules. To change the dihedral angles of the side chains we use group rotation moves, which were recently introduced to the BH scheme and shown to be very effective.[39]

Based on the primary structure, each residue of the sequence is assigned a local secondary structure, which can be either helix, extended or coil. We have compared the performance of three secondary structure predictors: Porter, Psipred and SAM. We found that Porter clearly provides the best prediction independent of protein fold and length, supporting the findings of an earlier study.[29] Thus, we use Porter for secondary structure assignment as starting point for subsequent BH simulations where only the conformation of the residues predicted to be coil are perturbed. In doing this, we enable the secondary structure elements to be assembled into their tertiary structure. In case Porter wrongly predicts coil instead of helix or strand, this can be corrected by random trial moves applied to the residues assumed to be coil. This secondary-to-tertiary BH approach is successful for the three peptides under study, as native-like structures with an RMSD of less than 2 Å from the target are found within 1,000 steps for trpzip and trp-cage, and within 2,500 steps for ER-10. We have benchmarked random dihedral angle moves applied to the coil residues with a maximum change of 30°, 60° or 90° and found that larger step sizes of 60° or 90° fold the proteins more efficiently.

To refine the structures predicted by the secondary-to-tertiary BH approach, we have performed further BH simulations of the structures of low energy and RMSD found thus far. In order to account for the possibility that Porter wrongly assigns helix or strand instead of coil, trial moves are applied to all residues in the refinement BH runs. Here, we have used dihedral angle moves for the backbone affecting Φ and Ψ and group rotation moves for the side chains, perturbing the conformation of three to five randomly chosen yet contiguous residues. We have benchmarked alternative backbone and side chain moves with different relations (1:1, 1:4 and 4:1) using a maximal dihedral angle change of 30° for both backbone and side chains. This rather small perturbation was chosen since the goal of

these BH simulations is to refine the already folded structures. This approach is successful as both energy and RMSD were considerably improved for all three peptides, leading to the identification of more native-like structures than in the initial secondary-to-tertiary BH approach. Here, we have not observed a strong dependence on the ratio of backbone and side chain moves, underpinning the importance of both backbone and side chains and their interrelation for the protein structure.

In conclusion, we have introduced secondary-to-tertiary BH optimization and benchmarked this approach for three peptides. We have demonstrated that this approach reliably and effectively identifies native-like structures, which can be further refined in subsequent BH runs without restraints placed on the trial moves. Our test runs for larger proteins have produced promising results, especially for helical proteins. In future, we will apply the secondary-to-tertiary BH approach to more proteins with more than 50 amino acid residues and aim to provide a benchmark for larger proteins as we did in this study for three miniproteins. Prior to this, further methodological developments are necessary for improving the prediction of long-range residue contacts in β-sheets. Moreover, we will validate our methodology for larger proteins of mixed secondary structure in a blind test, such as CASP. The current study has demonstrated that the basin-hopping approach to global optimization with improved Monte Carlo moves is on the route to become a promising and computationally low-demanding tool for *ab initio* protein structure prediction.

6 Computational Details

6.1 Secondary structure prediction.

Miceli. et al. compared different secondary structure predictors and found that the neuronal-network based predictors Porter[25] and Psipred[26] and the hidden Markov chain-based predictor SAM are the three most reliable prediction methods with Porter being by far the best.[29] As quality parameters they used the average performance accuracy ($Q3$)[47] and the segment overlap (SOV),[48] where $Q3$ measures the percentage of correctly guessed secondary structures of single amino acids, while SOV is obtained by computing per-segment overlaps. Both $Q3$ and SOV do not test which of the secondary structures (i.e., H, E and C) are mistaken for one another in case of misprediction. However, for the current work, which aims to predict tertiary protein structure by assembling segments of defined secondary structure, it is significant whether H and E are interchanged, or whether H or E is interchanged with C. While the latter type of false prediction can be easily corrected in the assembly process, the mix-up of H and E would hamper the tertiary structure prediction. Therefore, we compare the performance of Porter,[25] Psipred[26] and SAM[27] in terms of secondary structure mix-ups considering the cases of H$\leftrightarrow$E, H$\leftrightarrow$C and E$\leftrightarrow$C. We collect the mix-up statistics for the PDB25Select database,[49] which was used by Miceli et al.[29]

6.2 Protein models

The structures for trpzip, trp-cage and ER-10 were downloaded from the RCSB protein data bank[34] and used as target structures. Trpzip (PDB code 1LE0) is a 12 residue β-hairpin known as tryptophan zipper (trpzip);[35] trp-cage (PDB code 1L2Y) a 20 residue peptide with a short α-helix, a 3_{10}-helix, and a polyproline II helix at the C-terminus, which is known as tryptophan-cage miniprotein;[36] and ER-10 (PDB code 1ERP) a 38-residue pheromone ER-10 from the ciliated protozoan Euplotes raikovi consisting of three α-helices.[37] These miniproteins have been used as test cases in previous folding studies.[41,44,50–63] We use the CHARMM22 force field[64,65] to model the peptides, and the the generalized Born model FACTS[66] to describe the aqueous solvent. For the calculation of the nonbonded interactions, the cutoff scheme suggested in the FACTS documentation is employed, i.e., truncation of both long-range electrostatics at 12 Å using a shift function and the van der Waals energy with a polynomial switching function applied between 10 and 12 Å. We performed 20 ns MD simulations at an elevated temperature of $T = 500$ K using a Langevin thermostat with frictional coefficient 5 ps^{-1} to produce 20 unfolded starting structures per peptide for the subsequent folding simulations. The root mean square deviations (RMSDs) of the C_α atoms between the starting structures and the corresponding target structure are 5.6–10.9 Å for trpzip, 5.7–9.7 Å for trp-cage, and 8.6–14.1 Å for ER-10. In Figure 6 the target structure and one representative starting structure are shown for each peptide.

For the testing of larger proteins, the structures of a 35-residue subdomain of the chicken villin headpiece (villin, PDB code 2F4K), the 53-residue GA module of an albumin binding domain (protein B, PDB code 1PRB), and the N-terminal, DnaJ-like domain with 79 residues of murine polyomavirus tumor antigens (PyJ, PDB code 1FAF) were downloaded from the RCSB protein data bank. For the BH simulations, the structures were prepared and modeled in the same way as the miniproteins. More simulation details for these proteins are provided in section 4.

6.3 Basin-hopping

In the basin-hopping (BH) approach to global optimization[2–4] moves are proposed by perturbing the current geometry, and are accepted or rejected based upon the energy difference between the local minimum obtained by minimization from the current configuration and the previous minimum in the chain. In effect the potential energy surface is transformed into the basins of attraction[67,68] of all the local minima, so that the energy for configuration $\mathbf{r}$ is

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

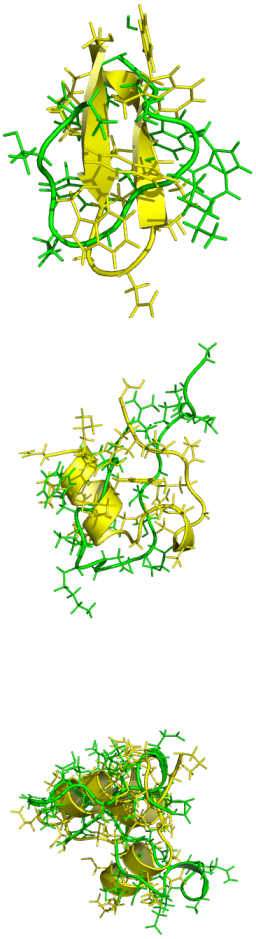


Figure 6: Representative initial structure (green) and target structure (yellow) for trpzip (top), trp-cage (middle) and ER-10 (bottom).

where min denotes 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, since the BH approach attempts to locate the global potential energy minimum and is not intended to sample thermodynamic properties. The BH algorithm has been implemented in the GMIN program[69] and has already been employed to find the global minimum of peptides and peptide complexes in previous work.[5,6,8–15,18] In GMIN, local minimization is facilitated by using a modified version of the LBFGS procedure described by Liu and Nocedal.[43]

To perturb the current geometry we have the option of taking steps in dihedral angle space for the backbones and side chains of the peptides,[8] where we consider dihedral angles defining planar structures, such as rings, as rigid in order to maintain the planar geometry.[70] In earlier work, we selected a certain number of the rotatable dihedral angles for the backbone and side chains with different twisting probabilities depending on the position of the residue along the peptide chain[8] and twisted them up to a maximum angle, which can be initially set by the user and is normally in the range of 20° to 50°. In this study we employ different approaches for dihedral trial moves. First, we develop a secondary-to-tertiary methodology, which uses the information from secondary structure prediction to determine the tertiary structure of the proteins. This approach is described in the next paragraph. Second, we introduce the possibility of applying trial moves to contiguous residues along the chain. Third, we apply generalized rotation moves to sample the rotameric states of protein side chains.[39] This scheme allows arbitrary groups of atoms to be rotated about an axis defined by a bond vector, maintaining maximum flexibility without introducing reliance on standard topologies. For instance, for a Lys side chain three such rotatable groups are defined, where atoms are rotated about the $C_\alpha-C_\beta$, $C_\beta-C_\gamma$ and $C_\gamma-C_\delta$ bonds.

6.4 Combination of basin-hopping with secondary structure predictions

Here we use the information from secondary structure prediction for the determination of the tertiary structure of a protein within the BH approach. Based on the initial secondary structure assignments we set the Ramachandran angles (Φ, Ψ) to $(-57°, -47°)$ and $(-135°, 135°)$ for α-helices (H) and β-strands (E), respectively. In the subsequent BH run we keep these angles fixed by (i) not allowing backbone dihedral angle moves for the amino acids, for which H or E is being predicted, and (ii) imposing constraints with a force constant of 1,000 kcal/(mol Å^2) on these dihedral angles during the energy minimization procedure. The constraints are necessary as otherwise the secondary structure elements would be lost during the energy minimization before the tertiary fold has been determined. This is especially true for amino acids in the E state, as β-strands are often only stable as part of a β-sheet. In this phase of a BH simulation we also conserve the side chain rotamers for the H and E amino acids. Instead, we concentrate on the amino acids predicted to be in a coil as the correct

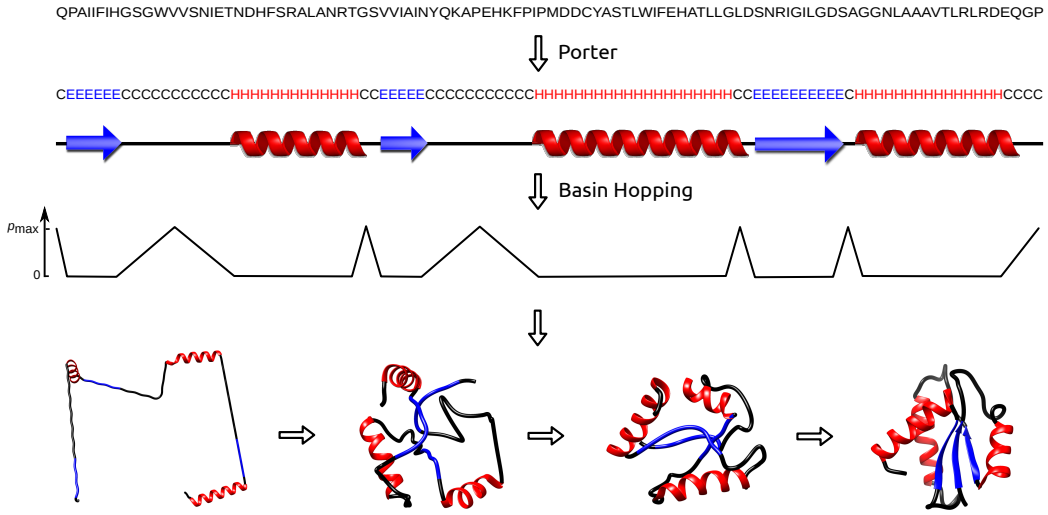


Figure 7: Schematic diagram of the combination of secondary structure prediction combined with basin-hopping. Based on the primary structure of the protein, the secondary structure is predicted using Porter. The starting structure for the BH run is modified by setting the (Φ, Ψ) angles to $(-57°, -47°)$ and $(-135°, 135°)$ for residues predicted to belong to an α-helix (H) and β-strand (E), respectively. In the BH run, trial moves are only applied to residues that are not in the H or E state, with the twisting probability being highest ($p_{\max}$) in the center of a segment of successive amino acids in the C state and decreasing linearly to zero at the ends of such a segment. During a BH run the tertiary contacts between secondary structure sequences become established, as illustrated at the bottom of this figure.

structure of the protein regions in the C state will lead to the tertiary contacts between the H and E segments. Thus, dihedral angle moves are applied only to the residues in the C state with twisting rotatable backbone and side chain dihedral angles. Here, the twisting probability is set highest in the center of a segment of successive C assignments and decreases linearly until the ends of such a segment is reached, becoming zero for the H or E residues connected to the C segment. For the N and C terminal residues the coil state is very likely. For the N terminus, the twisting probability is highest for the first residue and decreases linearly to zero until the first residue in the H or E state is reached. For the C terminus, the twisting probability increases linearly from zero for the last residues in the H or E state to its maximum for the very last residue in the sequence. The approach used to include secondary structure information in the BH methodology is also depicted in Figure 7.

For the prediction of secondary structure we use Porter[25] because Miceli et al.[29] and also our tests (see section 3.1) have found that Porter provides the most reliable secondary structure prediction.

6.5 Simulation outline

The aim of this study is to evaluate the secondary-to-tertiary BH scheme for the prediction of the tertiary structure of proteins. To this end we limit our protein test set to three rather small but well tested miniproteins with either α- or β-only structures. We start each folding simulation from 20 different initial structures per peptide. The BH simulations are divided into two rounds. First, BH runs are performed with constraints on the amino acids in the H and E conformational state according to the secondary structure prediction. From this round, the low-energy and low-RMSD (RMSD with respect to the target) structures are identified. In the ideal case, where the force field produces the lowest energy for the native structure, these two structure sets would be identical. Unfortunately, very often these two sets are different from each other as the physical yet empirical force fields are not perfect. For the three peptides under consideration the performance of the CHARMM22/FACTS potential to identify the native structure as lowest energy structure is discussed in this study. A second round of BH runs is then performed for the low-energy and low-RMSD structures but this time without any constraints. Dihedral angle moves are applied to all amino acids in the chain. For the side chains we employ group rotation moves as described in [39]. Unlike in previous work,[8–11,13–15,18] where the dihedral angles of randomly chosen residues along the chain were perturbed, we now apply dihedral angle changes to three to five contiguous residues. Furthermore, in the first BH round we test different step sizes in the intervals $(-30°, +30°)$, $(-60°, +60°)$ and $(-90°, +90°)$, while in the second BH round we test whether alternating backbone (BB) and side chain (SC) moves, SC moves only at every fifth BH step, or BB moves only at every fifth BH step perform best. To benchmark each move set, we repeat each simulation ten times in the first and three times in the second BH round, using different seeds for random number generation.

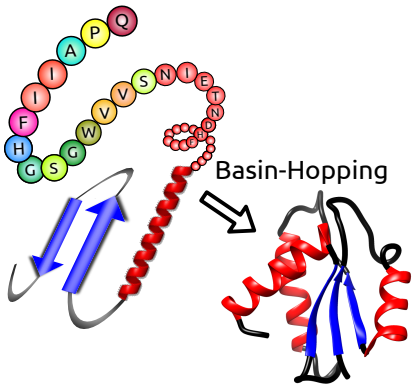
7 Acknowledgment

F.H. and B.S. gratefully acknowledge the computing time granted by the JARA-HPC Vergabegremium and provided on the JARA-HPC Partition part of the RWTH Compute Cluster in Aachen (grant number JARA0018). We thank Dr. Michael Owen for critical reading of the manuscript.

8 Keywords

protein structure prediction, basin hopping, Monte Carlo moves, secondary structure assembly

9 TOC



From secondary to tertiary structure: The basin-hopping approach to global optimization is employed for protein structure prediction. The efficiency of basin-hopping is improved by introducing a methodology that derives tertiary structures from the secondary structure assignments of individual residues. It is demonstrated that this secondary-to-tertiary basin-hopping approach successfully and reliably predicts three-dimensional protein structures.

References

[1] C. B. Anfinsen, *Science* **1973**, *181*, 223–230.

[2] D. J. Wales, J. P. K. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

[3] D. J. Wales, H. A. Scheraga, *Science* **1999**, *285*, 1368–1372.

[4] Z. Li, H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6611–6615.

[5] P. Derreumaux, *J. Chem. Phys.* **1997**, *106*, 5260–5270.

[6] P. Derreumaux, *J. Chem. Phys.* **1997**, *107*, 1941–1947.

[7] M. A. Miller, D. J. Wales, *J. Chem. Phys.* **1999**, *111*, 6610–6616.

[8] P. N. Mortenson, D. J. Wales, *J. Chem. Phys.* **2001**, *114*, 6443–6454.

[9] P. N. Mortenson, D. A. Evans, D. J. Wales, *J. Chem. Phys.* **2002**, *117*, 1363–1376.

[10] J. M. Carr, D. J. Wales, *J. Chem. Phys.* **2005**, *123*, 234901.

[11] A. Verma, A. Schug, K. H. Lee, W. Wenzel, *J. Chem. Phys.* **2006**, *124*, 044515.

[12] M. T. Oakley, R. L. Johnston, *J. Chem. Theor. Comput.* **2013**, *9*, 650–657.

[13] B. Strodel, D. J. Wales, *J. Chem. Theor. Comput.* **2008**, *4*, 657–672.

[14] B. Strodel, J. W. L. Lee, C. S. Whittleston, D. J. Wales, *J. Am. Chem. Soc.* **2010**, *132*, 13300–13312.

[15] O. O. Olubiyi, B. Strodel, *J. Phys. Chem. B* **2012**, *116*, 3280–3291.

[16] M. R. Betancourt, *J. Chem. Phys.* **2011**, *134*, 014104.

[17] P. Robustelli, A. Cavalli, C. M. Dobson, M. Vendruscolo, X. Salvatella, *J. Phys. Chem. B* **2009**, *113*, 7890–7896.

[18] F. Hoffmann, B. Strodel, *J. Chem. Phys.* **2013**, *138*, 025102.

[19] W. W. Chen, J. S. Yang, E. I. Shakhnovich, *Proteins: Struct., Func. and Bioinf.* **2007**, *66*, 682–688.

[20] S. Liang, N. V. Grishin, *Protein Sci.* **2002**, *11*, 322–331.

[21] G. Ramachandran, V. Sasisekharan, *Adv. Protein Chem.* **1968**, *23*, 283–438.

[22] J. W. Ponder, F. M. Richards, *J. Mol. Biol.* **1987**, *193*, 775–791.

[23] J. Meiler, D. Baker, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12105–12110.

[24] M. Karakas, N. Woetzel, R. Staritzbichler, N. Alexander, B. E. Weiner, J. Meiler, *PLoS ONE* **2012**, *7*, e49240.

[25] G. Pollastri, A. McLysaght, *Bioinformatics* **2005**, *21*, 1719–1720.

[26] L. J. McGuffin, K. Bryson, D. T. Jones, *Bioinformatics* **2000**, *16*, 404–405.

[27] R. Hughey, A. Krogh, *SAM: Sequence alignment and modeling software system.*, Technical Report UCSC-CRL-95-7, University of California, Santa Cruz, CA, 1995.

[28] J. Garnier, J. F. Gibrat, B. Robson, *Methods Enzymol.* **1996**, *266*, 540–553.

[29] L. Miceli, L. Palopoli, S. E. Rombo, G. Terracina, G. Tradigo, P. Veltri, *9th International Conference Baton Rouge, LA, USA, May 25-27, 2009 Proceedings, Part I.*, (Eds.:G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra, P. M. A. Sloot), Springer-Verlag, Berlin, Heidelberg, **2009**, pp. 848-857.

[30] K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J. Mol. Biol.* **1997**, *268*, 209–225.

[31] P. Bradley, D. Chivian, J. Meiler, K. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. E. M. Strauss, D. Baker, *Proteins: Struct., Func. Gen.* **2003**, *53*, 457–468.

[32] P. Bradley, L. Malmström, B. Qian, J. Schonbrun, D. Chivian, D. E. Kim, J. Meiler, K. M. S. Misura, D. Baker, *Proteins: Struct., Func. and Bioinf.* **2005**, *61*, 128–134.

[33] H. Zhou, S. B. Pandit, J. Skolnick, *Proteins: Struct., Func. and Bioinf.* **2009**, *77*, 123–127.

[34] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *J. Mol. Biol.* **1977**, *112*, 535–542.

[35] A. G. Cochran, N. J. Skelton, M. A. Starovasnik, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 5578–5583.

[36] J. W. Neidigh, R. M. Fesinmeyer, N. H. Andersen, *Nature Struct. Biol.* **2002**, *9*, 425–430.

[37] L. R. Brown, S. Mronga, R. A. Bradshaw, C. Ortenzi, P. Luporini, K. Wüthrich, *J. Mol. Biol.* **1993**, *231*, 800–816.

[38] R. Zhou, B. J. Berne, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12777–12782.

[39] K. Mochizuki, C. S. Whittleston, S. Somani, H. Kusumaatmaja, D. J. Wales, *Phys. Chem. Chem. Phys.* **2014**, *16*, 2842–2853.

[40] J. Maupetit, P. Tuffery, P. Derreumaux, *Proteins: Struct., Func. and Bioinf.* **2007**, *69*, 394–408.

[41] P. Thévenet, Y. Shen, J. Maupetit, F. Guyon, P. Derreumaux, P. Tufféry, *Nucleic Acids Res.* **2012**, *40*, W288–W293.

[42] V. Z. Spassov, L. Yan, P. K. Flook, *Protein Sci.* **2007**, *16*, 494–506.

[43] D. Liu, J. Nocedal, *Math. Prog.* **1989**, *45*, 503–528.

[44] K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, *Science* **2011**, *334*, 517–520.

[45] A. N. Adhikari, K. F. Freed, T. R. Sosnick, *Phys. Rev. Lett.* **2013**, *111*, 028103.

[46] M. V. Berjanskii, M. I. Riley, A. Xie, V. Semenchenko, W. R. Folk, S. R. Van Doren, *J. Biol. Chem.* **2000**, *275*, 36094–36103.

[47] B. Rost, C. Sander, R. Schneider, *J. Mol. Biol.* **1994**, *235*, 13–26.

[48] K. Fidelis, B. Rost, A. Zemla, *Proteins: Struct., Func. and Bioinf.* **1999**, *223*, 220–223.

[49] U. Hobohm, C. Sander, *Protein Sci.* **1994**, *3*, 522–524.

[50] C. Simmerling, B. Strockbine, A. E. Roitberg, *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.

[51] S. Chowdhury, M. C. Lee, G. Xiong, Y. Duan, *J. Mol. Biol.* **2003**, *327*, 711–717.

[52] A. Schug, T. Herges, W. Wenzel, *Phys. Rev. Lett.* **2003**, *91*, 1–4.

[53] A. Schug, T. Herges, A. Verma, K. H. Lee, W. Wenzel, *ChemPhysChem* **2005**, *6*, 2640–2646.

[54] A. Schug, W. Wenzel, U. Hansmann, *J. Chem. Phys.* **2005**, *122*, 194711.

[55] S. Piana, K. Lindorff-Larsen, D. E. Shaw, *Biophys. J.* **2011**, *100*, L47–L49.

[56] J. Maupetit, P. Derreumaux, P. Tufféry, *Nucleic Acids Res.* **2009**, *37*, W498–W503.

[57] J. W. Pitera, W. Swope, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 7587–7592.

[58] J. Juraszek, P. G. Bolhuis, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 15859–15864.

[59] D. Paschek, H. Nymeyer, A. E. García, *J. Struct. Biol.* **2007**, *157*, 524–533.

[60] K. Klenin, W. Wenzel, *International Journal of Computers and Communications* **2007**, *1*, 1–3.

[61] I. H. Radford, A. R. Fersht, G. Settanni, *J. Phys. Chem. B* **2011**, *115*, 7459–7471.

[62] T. Cellmer, M. Buscaglia, E. R. Henry, J. Hofrichter, W. A. Eaton, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 6103–6108.

[63] Y. Tang, M. J. Grey, J. McKnight, A. G. Palmer III, D. P. Raleigh, *J. Mol. Biol.* **2006**, *355*, 1066–1077.

[64] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comp. Chem.* **1983**, *4*, 187–217.

[65] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[66] U. Haberthür, A. Caflisch, *J. Comp. Chem.* **2008**, *29*, 701–715.

[67] P. G. Mezey, *Potential Energy Hypersurfaces.*, Elsevier, Amsterdam, **1987**.

[68] D. J. Wales, *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 653–657.

[69] D. J. Wales, *GMIN: A program for basin-hopping global optimisation*, http://www-wales.ch.cam.ac.uk/software.html.

[70] M. S. Bauer, B. Strodel, S. N. Fejer, E. F. Koslover, D. J. Wales, *J. Chem. Phys.* **2010**, *132*, 054101.