Energy Landscapes of a Hairpin Peptide Including NMR Chemical Shift Restraints

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID:	CP-ART-03-2015-001259.R2
Article Type:	Paper
Date Submitted by the Author:	03-Jul-2015

Complete List of Authors:	Carr, Joanne; University of Cambridge, Chemistry Whittleston, Chris; University of Cambridge, Department of Chemistry Wade, David; University of Cambridge, Chemistry Wales, David; Chemistry,

Physical Chemistry Chemical Physics
CREATED USING THE RSC LaTeX PCCP ARTICLE TEMPLATE – SEE www.rsc.org/electronicfiles FOR DETAILS
ARTICLE TYPE
www. rsc.org/xxxxxx | XXXXXXXX

Energy Landscapes of a Hairpin Peptide Including NMR Chemical Shift Restraints

Joanne M. Carr,[a] Chris S. Whittleston,[a] David C. Wade,[a] and David J. Wales[*a]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

Methods recently introduced to improve the efficiency of protein structure prediction simulations by adding a restraint potential to a molecular mechanics force field introduce additional input parameters that can affect the performance. Here we investigate the changes in the energy landscape as the relative weight of the two contributions, force field and restraint potential, is systematically altered, for restraint functions constructed from calculated nuclear magnetic resonance chemical shifts. Benchmarking calculations were performed on a 12-residue peptide, tryptophan zipper 1, which features both secondary structure (a β-hairpin) and specific packing of tryptophan sidechains. Basin-hopping global optimization was performed to assess the efficiency with which lowest-energy structures are located, and the discrete path sampling approach was employed to survey the energy landscapes between unfolded and folded structures. We find that inclusion of the chemical shift restraints improves the efficiency of structure prediction because the energy landscape becomes more funnelled and the proportion of local minima classified as native increases. However, the funnelling nature of the landscape is reduced as the relative contribution of the chemical shift restraint potential is increased past an optimal value.

1 Introduction

Significant progress has been made on improving the computational efficiency of protein structure prediction simulations by making direct use of nuclear magnetic resonance (NMR) observables. One approach is based on molecular fragment replacement using sequence homology, combined with databases of structures and experimental NMR observables, and methods for predicting the observables given a structure.[1–4] An alternative approach that does not require sequence homology involves a conformational search of the energy landscape obtained by combining a biomolecular force field with a restraint potential that biases the search towards structures consistent with some reference observables.[5,6] Here we consider only restraint potentials that introduce an energy penalty as a function of the difference between reference and calculated NMR chemical shifts; such terms were introduced for the refinement of structures determined by NMR,[7,8] in order to make good use of these precise and readily available spectroscopic observables. More recently, chemical shift restraint potentials together with molecular mechanics force fields and more extensive conformational searches have been employed to predict the native structures of various proteins in studies using Monte Carlo,[9] molecular dynamics[10] and basin-hopping global optimization[11] simulations. In each study, significant improvements in the quality of the predictions were obtained over unrestrained simulations. The overall approach depends upon the link between biomolecular chemical shifts and three-dimensional structure (see Ref 12 and references therein).

In the current work, we analyse the changes in the energy landscape as the relative weight of the two contributions, force field and restraint potential, is systematically altered. We employ order-parameter-free visualizations of the landscape via disconnectivity graphs.[13,14] The aim is to gain insight into how the efficiency of structure prediction varies using such an approach and, hence, might be optimized in future applications. Due to the bias introduced by the restraint potential, we do not consider thermodynamics or dynamics. Our test system is tryptophan zipper 1 (PDB[15] code 1LE0[16]), a 12-residue peptide that features both secondary structure (a β-hairpin) and specific packing of tryptophan sidechains, yet remains computationally tractable in terms of the total amount of sampling required. We consider the region of the landscape relevant for folding from extended (rather than partially unfolded[10]) structures, and use basin-hopping global optimization[17–19] and discrete path sampling[20–22] as the search methods. The calculated chemical shifts and restraint energies are obtained using the CamShift methodology,[23] which also provides analytical gradients with respect to the atomic coordinates and is therefore amenable to these search methods, which are based on efficient geometry optimization.

[a] *University Chemical Laboratories, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. E-mail: dw34@cam.ac.uk*

2 Methods

2.1 The potential

The energies and gradients were calculated using a molecular mechanics force field in combination with a restraint potential based on NMR chemical shifts. The CHARMM22/CMAP dihedral-potential-corrected all atom force field[24–26] with the FACTS implicit solvation model[27] were used to ensure protein-like behaviour of the polypeptide chain. The restraint potential was obtained via a Fortran implementation[11] of the CamShift program and methodology.[23] CamShift predicts the ^{1}H$_{\alpha}$, amide ^{1}H, ^{13}C$_{\alpha}$, ^{13}C$_{\beta}$, carbonyl ^{13}C, and amide ^{15}N chemical shifts of a given protein structure using calculations based on polynomial functions of the interatomic distances (and therefore allows analytic gradients to be obtained straightforwardly). The terms in the CamShift function that we included here account for the influence of backbone, sidechain and nonbonded atoms, aromatic rings via point-dipole terms, and an improved description of backbone dihedral angles.

The overall CamShift penalty function is a sum of soft-square harmonic wells applied atom by atom to the difference between the chemical shift predicted for that atom in the current structure and a corresponding reference shift representing the target conformation.[10] The form of this function penalises statistically significant deviations in chemical shift (harmonic region), whilst also allowing a margin of error in the shifts (flat bottom region) and not allowing large deviations to dominate the overall potential (hyperbolic tangent region).

A parameter α determines the relative weight of the two contributions:

$$E_{\mathrm{tot}} = \alpha E_{\mathrm{CS}} + (1-\alpha) E_{\mathrm{FF}}, \tag{1}$$

with $0 \leq \alpha \leq 1$, E_{CS} the CamShift restraint energy, and E_{FF} the energy from CHARMM22/CMAP and FACTS. An equivalent expression exists for the gradients. We note that E_{CS} is a dimensionless quantity, whereas E_{FF} has units of kcal mol^{-1}. This form for the total energy differs from previous work,[9–11] in which $E_{\mathrm{tot}} = \alpha E_{\mathrm{CS}} + E_{\mathrm{FF}}$ for $\alpha \geq 0$.

The CHARMM22 potential was symmetrized with respect to feasible permutations of identical atoms,[28] as was CamShift for the relevant atoms in ARG, GLU, ASP, TYR and PHE residues. The two hydrogens in GLY residues are not permutable here because CamShift treats them slightly differently. To avoid the unphysical complication of pairs of structures with similar but non-identical energies for exchange of these two hydrogens, only the conformations with the spatial order matching the native structure were retained.

2.2 Basin-hopping global optimization

Since the main aim of including "experimental" restraints is to improve the computational efficiency of protein structure prediction, we performed global optimization simulations using the basin-hopping approach,[17–19] as implemented in the `GMIN` program,[29] in order to identify putative lowest-energy minima.

To obtain statistics, 10 independent simulations were performed for each landscape. Each run was started from a different conformation with no native contacts, taken from a preliminary high-temperature basin-hopping simulation. For the production runs, $k_{\mathrm{B}}T$ in the Metropolis criterion[30] was fixed at 2.5 energy units, and 100 000 basin-hopping steps were performed. Each step involved perturbing a randomly chosen set of backbone and sidechain dihedral angles by an angle selected at random up to the maximum step size, either clockwise or anticlockwise. The maximum step size, initially 40°, was dynamically adjusted to maintain a Metropolis acceptance ratio of approximately 0.3. Local minima were converged to a root-mean-square gradient of 10^{-3} Å^{-1} after each basin-hopping step, and 10^{-6} Å^{-1} during the refinement of the 50 lowest-energy structures, using a slightly modified version of the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm.[31,32]

2.3 Discrete path sampling

To explore the energy landscapes more widely, from unfolded to folded conformations, we employed the discrete path sampling framework[20–22] to generate kinetic transition networks. Each network was set up with an initial discrete path[20] (connected sequence of minima and the intervening transition states) between an unfolded and a folded conformation, and then expanded to improve the ensemble of folding paths and refine the overall kinetics. Discrete paths were identified using an iterative missing connection procedure[33] based on Dijkstra's shortest path algorithm,[34] as implemented in the `OPTIM` program.[35] Transition state candidates were obtained using the doubly-nudged[36] elastic band method,[37–39] and then converged tightly to a root-mean-square gradient of 10^{-6} Å^{-1} using hybrid eigenvector-following.[37,40,41] Paths were refined iteratively using various procedures implemented in the `PATHSAMPLE` program,[42] to reduce the overall number of transition states in a path,[43,44] to find alternative routes that avoid high energy barriers,[44] and to remove artificial frustration from under-sampling.[44] The resulting landscapes were visualized using disconnectivity graphs.[13,14]

3 Results and Discussion

We investigate the energy landscapes for tryptophan zipper 1 (PDB code 1LE0[16]), a 12-residue de novo peptide that readily forms a β-hairpin in water, stabilized by two cross-chain TRP–TRP interactions. The sequence is SER–TRP–THR–TRP–GLU–GLY–ASN–LYS–TRP–THR–TRP–LYS, and in our simulations we employed standard, zwitterionic capping groups at the termini, following the work of Hoffmann and Strodel.[11]

We consider the landscapes defined by four values of the parameter α: 0, 0.3, 0.5 and 0.7. Higher values were found in preliminary work to have insufficient contribution from the force field to distinguish protein-like structures from unphysical ones. For each value of α, we performed basin-hopping global optimization and also assembled a kinetic transition network using discrete path sampling, as described in Section 2. The general input parameters in the basin-hopping runs were held constant across the different landscapes, at values chosen using shorter, preliminary simulations. In each case, the initial unfolded conformation used as the "reactant" endpoint in the discrete path sampling was obtained by locally minimizing a fully extended structure with the prevailing overall potential. For non-zero values of α, the folded conformation (the "product" endpoint) was the appropriate putative global minimum from preliminary global optimization runs at each value of α. For $\alpha = 0$, the product endpoint was initially taken as the locally minimized PDB structure (with standard capping groups), since the true global minimum in this case is a more difficult target, as discussed below. The set of reference chemical shifts used throughout to represent the target conformation was calculated using CamShift for the unoptimized PDB structure with standard capping groups. This conformation is illustrated in Fig. 1, using PyMOL.[45]

For the basin-hopping results on each landscape (defined by the value of α; all other CamShift parameters[10,23] held constant), we consider the lowest-energy structure found in each of the 10 independent runs and analyse these in terms of energies and structural order parameters that characterise the folded state. The order parameters we consider are the number of native backbone hydrogen bonds (denoted O1, maximum 4), using the default geometrical definition of a hydrogen bond from the CHARMM program;[26,46] and the number of distances between centres of mass of neighbouring pairs of TRP sidechains (in terms of structure not sequence) that match the PDB structure to within a tolerance of ± 0.5 Å for the two closest pairs and ± 1.0 Å for the other (denoted O2, maximum 3). These order parameters are also used, one at a time, to add information to the disconnectivity graphs by colouring the branches according to the values for the minima. The tolerances were selected by considering the changes in the relevant distances and angles on minimization of the PDB conformation using the CHARMM-only potential, and also by observing consistent plateaux in the number of structures defined as matching the reference as the values were increased, for each kinetic transition network. Reasonable changes in the tolerances do not affect the conclusions.

We found that all the potentials (including $\alpha = 0$) supported unphysical structures as stationary points with reasonably low energies, but separated from corresponding physically reasonable structures by high barriers. Examples include highly non-tetrahedral –CH_2– and –NH_3^+ groups in sidechains. Such structures are kinetic traps and cause unrealistic frustration in the network,[47,48] both at the sampling stage and in the analysis. We removed such structures from our networks using a criterion based on the bond angle component of the molecular mechanics energy for transition states, as this was found to clearly distinguish the unphysical conformations, and is more general than individual geometric criteria.

3.1 $\alpha = 0$

Energies and order parameters for the lowest minima found in the 10 independent basin-hopping runs are presented in Table 1. The runs did not all produce the same lowest minimum, either in terms of energy or the structural order parameters, indicating that this system is quite challenging for global optimization. The overall lowest-energy structure has the hairpin and turn not quite as well-formed as in the reference structure, and a non-native packing of the TRP sidechains (all-atom root-mean-square deviation of 2.8 Å from the PDB reference structure).

To highlight the presence of non-native structures lower in energy than native, the putative global minimum (run 9) and the lowest minima from basin-hopping runs 2, 5, 7 and 10 were connected in to the landscape sampled between unfolded and native conformations, using subsequent applications of the discrete path sampling approach. These five structures can be described as hairpin-like with some but not all of the native hydrogen bonds present and either one or zero native TRP–TRP contacts, and are the five lowest in energy of the set of 10 from the global optimization. The lowest-energy minima from the remaining runs are more than five energy units above the overall lowest and are structurally distinct from hairpins (some possess helical turn sections); we chose not to connect them to the main kinetic transition network here to simplify the presentation, though it should be noted that the force field supports these structures at lower energies than native conformations.

Disconnectivity graphs showing the resulting landscape are presented in Fig. 2. The kinetic transition network contains 74 007 connected minima and 88 838 transition states. The colouring by order parameter shows that native structures lie above the putative global minimum for this potential as discussed above, are not prevalent (they comprise fewer than 3% of the minima), and can be separated by high downhill barriers of up to 24 energy units. These results are all consistent with the global optimization runs.

Short, constant-temperature molecular dynamics simulations were run using CHARMM[26,46] for various minima from the lowest-energy part of the landscape shown in Fig. 2, after an initial heating phase that did not form part of the subsequent analysis. These structures, which include the putative global minimum, were found to be reasonably stable for at least 3 ns in terms of the structural order parameters for backbone hydrogen bonding and TRP–TRP interactions. Whilst there are fluctuations away from the original order parameter values, and the trajectories leave the original basins of attraction, there are no substantial periods of time throughout which the order parameter values of the starting minima are lost. Furthermore, none of the snapshots from these trajectories are classified as native via the order

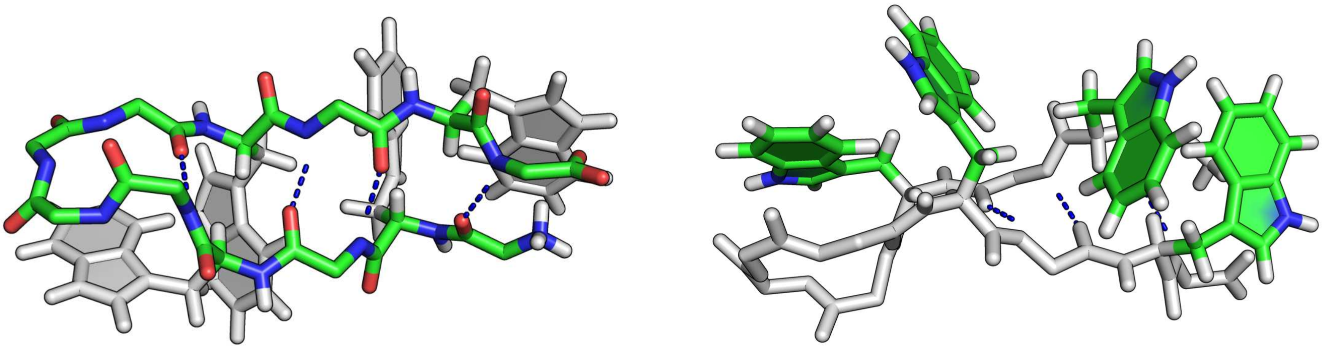

Fig. 1 The reference structure of tryptophan zipper 1. Left: view from below, highlighting the backbone hydrogen bonds (tryptophan sidechain atoms are grey). Right: side view, highlighting the interactions of the tryptophan sidechains (backbone atoms are grey). Other sidechains are not shown

Table 1 Analysis of the lowest-energy minimum found in each of the 10 independent basin-hopping global optimization runs for four values of α. The total energies given are relative to the lowest found for each landscape, and the CamShift energies (E_{CS}) are the values of the restraint potential, not including the factor of α. The structural order parameters are the number of native backbone hydrogen bonds (denoted O1, maximum four), and the number of distances between centres of mass of neighbouring pairs of TRP sidechains that match the PDB structure to within a tolerance of ± 0.5 Å for the two closest pairs and ± 1.0 Å for the other (denoted O2, maximum three)

Run	1	2	3	4	5	6	7	8	9	10
$\alpha = 0$										
E_{tot} (relative)	6.70	1.58	5.90	9.47	3.04	7.20	1.83	8.58	0.00	0.998
O1 (out of 4)	0	3	0	0	3	0	2	0	2	3
O2 (out of 3)	1	1	0	1	0	0	1	0	1	1
$\alpha = 0.3$										
E_{tot} (relative)	0.0865	1.64	0.00	1.10	0.787	0.321	1.44	1.83	1.66	0.547
E_{CS}	3.71	4.88	2.54	2.24	2.97	3.80	9.31	2.03	4.13	2.99
O1 (out of 4)	4	4	4	4	4	4	4	4	4	4
O2 (out of 3)	3	3	3	2	3	3	2	2	3	2
$\alpha = 0.5$										
E_{tot} (relative)	1.21	1.39	0.239	0.684	2.28	0.892	1.75	0.831	0.604	0.00
E_{CS}	1.59	2.80	1.51	2.16	3.89	1.72	1.14	1.31	1.46	1.79
O1 (out of 4)	4	4	4	4	4	4	4	4	4	4
O2 (out of 3)	2	1	3	3	2	2	2	2	3	2
$\alpha = 0.7$										
E_{tot} (relative)	2.91	0.00	2.96	4.77	3.28	2.10	1.51	5.32	4.95	1.97
E_{CS}	0.693	0.401	0.851	1.55	0.917	0.842	0.598	2.86	0.947	0.654
O1 (out of 4)	4	4	4	3	4	4	4	2	4	4
O2 (out of 3)	0	2	1	1	1	3	3	1	2	1

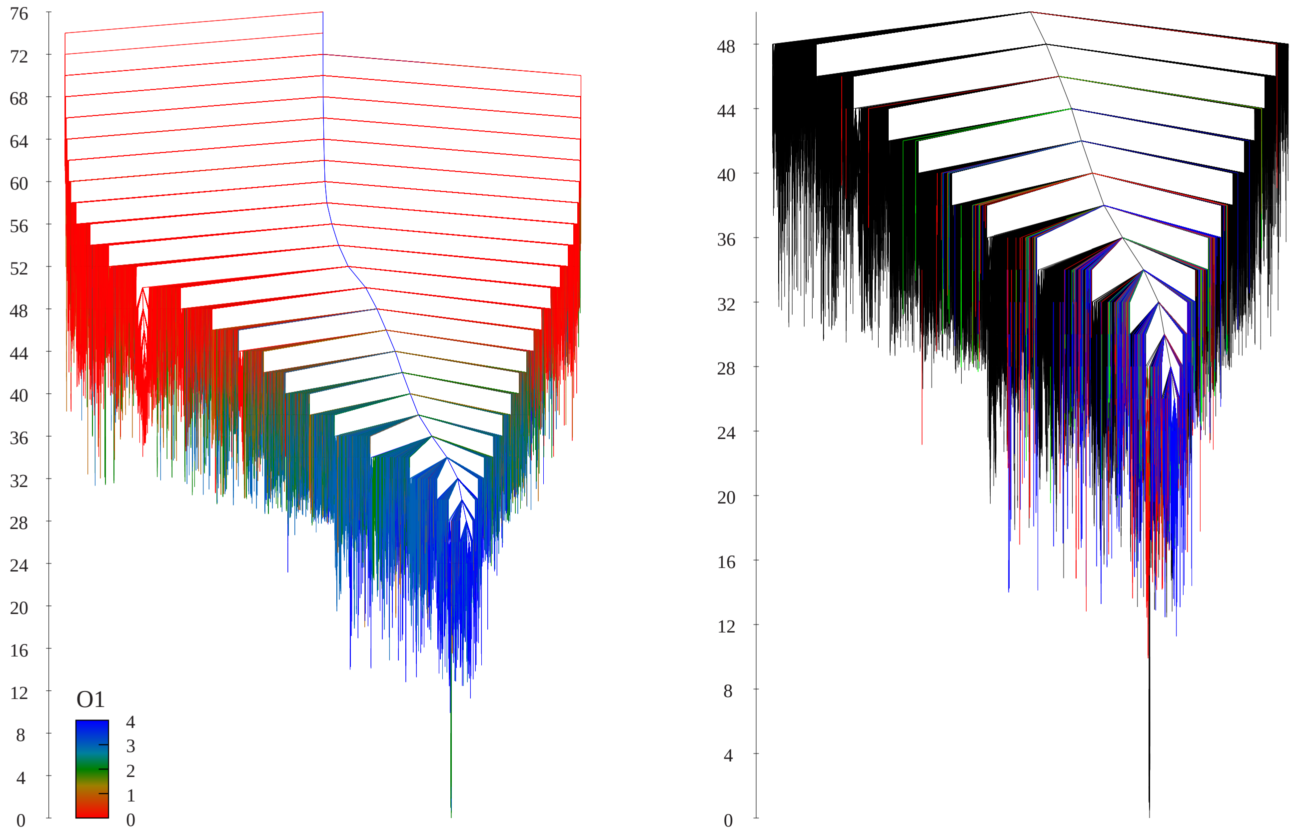

Fig. 2 Disconnectivity graphs for $\alpha = 0$. The vertical axes are the total energy, relative to the lowest-energy minimum. Left: full graph, with the branches coloured according to the number of native backbone hydrogen bonds in the corresponding local minimum, with blue representing the maximum (four). Right: magnification of the low-energy region. Here, only the minima with four native backbone hydrogen bonds are coloured, and the colour scheme now displays the number of distances between centres of mass of neighbouring pairs of TRP sidechains that match the PDB structure (tolerances given in the text). Red: one. Green: two. Blue: three

parameters.

3.2 $\alpha = 0.3$

Energies and order parameters for the lowest minima found in the 10 independent basin-hopping runs are presented in Table 1. Inclusion of the restraint potential significantly improved the success rate in locating native-like structures to 60%, according to the two order parameters. Consensus is again not reached at the level of an individual minimum; this is the case for all the values of α considered here. All 10 runs located native hairpin structures, but four of these did not manage to fully pack the TRP sidechains in a native manner. This result may be due to limitations of the geometrical move set employed here, or because of the particular sensitivity of $^1H_\alpha$ and ^{13}C chemical shifts to backbone dihedral angles.[49]

Disconnectivity graphs[13,14] showing the landscape from the discrete path sampling simulations are presented in Fig. 3. There are 45 426 connected minima and 61 208 transition states. Native-like structures are lowest in energy, occupying a significant fraction of the landscape interspersed with near-native minima. The downhill barriers to native structures are also lower than for $\alpha = 0$.

3.3 $\alpha = 0.5$

The success rate for $\alpha = 0.5$ was lower than for $\alpha = 0.3$, at 30 %. Again, all 10 runs located the correct hairpin structure, but there are more runs that correctly predicted only one or two out of the three pairs of TRP–TRP distances in the order parameter (Table 1).

Disconnectivity graphs showing the landscape are presented in Fig. 4. There are 44 152 connected minima and 59 159 transition states. Although native hairpin structures are numerous and low in energy, only 22 % of them also possess the fully correct packing of the TRP sidechains.

3.4 $\alpha = 0.7$

Although the values of the CamShift restraint potential (E_{CS}) are now on average the smallest, these runs did not perform as well as $\alpha = 0.3$ and 0.5 in terms of locating native-like structures, though they are better than $\alpha = 0$ (Table 1). The success rate is 20 %, but now two out of 10 runs did not locate the full set of native backbone hydrogen bonds within the fixed number of basin-hopping steps. The prediction of the TRP sidechain packing is also the poorest of the non-zero values of α.

Disconnectivity graphs are presented in Fig. 5. There are 35 689 connected minima and 51 182 transition states. Many of the minima are native hairpin structures of comparable, low energy. Among these hairpins, only 36 % also have correctly packed TRP sidechains, and they are not well separated in energy from partially folded conformations, thus hindering the search for the global minimum.

3.5 Overall Trends

Given the composition of the total energy [$E_{\mathrm{tot}} = \alpha E_{\mathrm{CS}} + (1-\alpha)E_{\mathrm{FF}}$], and the relative magnitudes of the two parts in this case ($|E_{\mathrm{FF}}| \sim 300\,\mathrm{kcal\,mol^{-1}}$ and $|E_{\mathrm{CS}}| \sim 30$), the range of energies decreases as α increases. By the time α reaches 0.7, the contribution of neither the force field nor CamShift is sufficient to distinguish clearly between native and partially folded structures. The form of the energy landscape changes from frustrated, with high barriers between native-like structures ($\alpha = 0$), to funnelling ($\alpha = 0.3$ and 0.5), to frustrated again, with many competing structures of similar energy ($\alpha = 0.7$). The number of stationary points also decreases as α increases. More importantly, the proportion of minima classified as native according to structural order parameters for backbone hydrogen bonding and TRP–TRP sidechain packing is significantly higher when CamShift is included.

The values of the restraint potential, E_{CS}, for the lowest-energy minima are not negligible, even for structures classified as native, though they decrease as α increases. The non-negligible values are therefore probably due to competition between CamShift and the force field. In general, this issue will be affected by the relative orders of magnitude of the two sets of energies and gradients, and some extra adjustment to the weighting may be necessary.[11] However, for this system at least, it was not necessary to obtain conformations with restraint energies very close to zero. Whilst it would be possible to reduce the value of the restraint potential by increasing the margin of error allowed between the calculated and reference shifts (Section 2.1), this change may also have the undesirable effect of reducing the driving force towards native structures;[9] the relevant CamShift tolerance parameter must therefore also be chosen with care[9,11] ($n = 0.5$ was used throughout the current work). It has also been noted that predicted chemical shifts can be very sensitive to small changes in protein structure.[10]

4 Conclusions

We have systematically explored the effects on the energy landscape of adding a restraint potential term based on calculated and reference NMR chemical shifts to the energy of a protein from a molecular mechanics force field. Our test molecule is the tryptophan zipper 1 peptide (1LE0), the force field is CHARMM22/CMAP[24–26] with the FACTS implicit model of solvation,[27] and CamShift[23] (recoded[11] in Fortran) provided the restraint potential and calculated chemical shifts. The general aim of including such restraint terms is to improve the efficiency and accuracy of structure prediction simulations, by incorporating experimental information about the target conformation. This approach is most useful when the force field supports an unphysically large number of local minima, and/or does not have the native state as the global minimum. We therefore performed basin-hopping global optimization simulations and assembled kinetic transition networks using discrete path sampling, for a series of total energy functions with increasing contributions from the restraint potential, controlled by a mixing parameter α.

The results show that locating a native-like structure for this system is relatively difficult without any restraint terms but, as expected, this situation can be improved significantly by incorporating restraints from CamShift. We postulate that this improvement arises because the proportion of minima classified as native, according to structural order parameters for backbone hydrogen bonding and TRP–TRP sidechain packing, is significantly higher when CamShift is included, and also because the organization of the energy landscape changes. Of the values tested, $\alpha = 0.3$ was found to give optimal per-

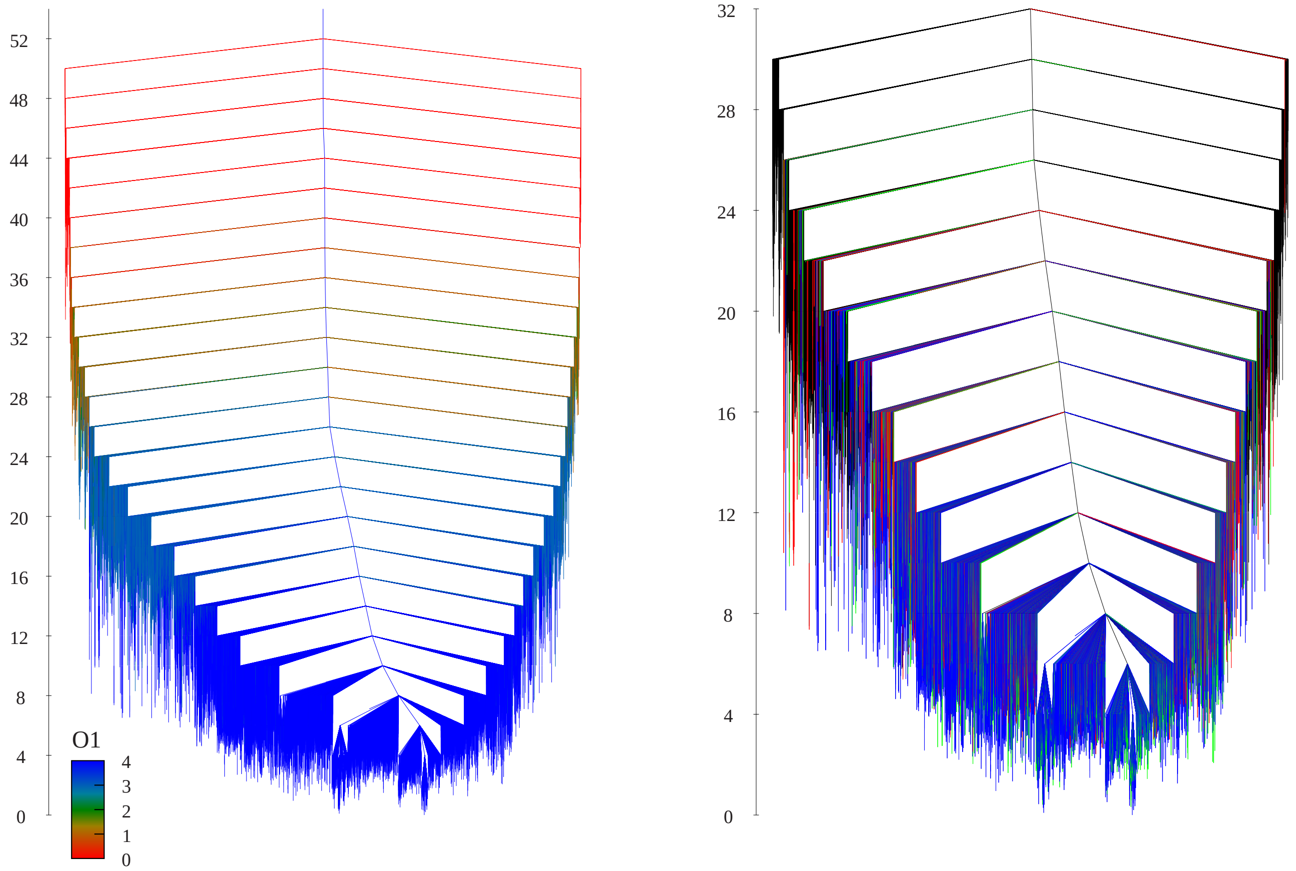

Fig. 3 Disconnectivity graphs for $\alpha = 0.3$. The vertical axes are the total energy, relative to the lowest-energy minimum. Left: full graph, with the branches coloured according to the number of native backbone hydrogen bonds in the corresponding local minimum, with blue representing the maximum (four). Right: magnification of the low-energy region. Here, only the minima with four native backbone hydrogen bonds are coloured, and the colour scheme now displays the number of distances between centres of mass of neighbouring pairs of TRP sidechains that match the PDB structure (tolerances given in the text). Red: one. Green: two. Blue: three

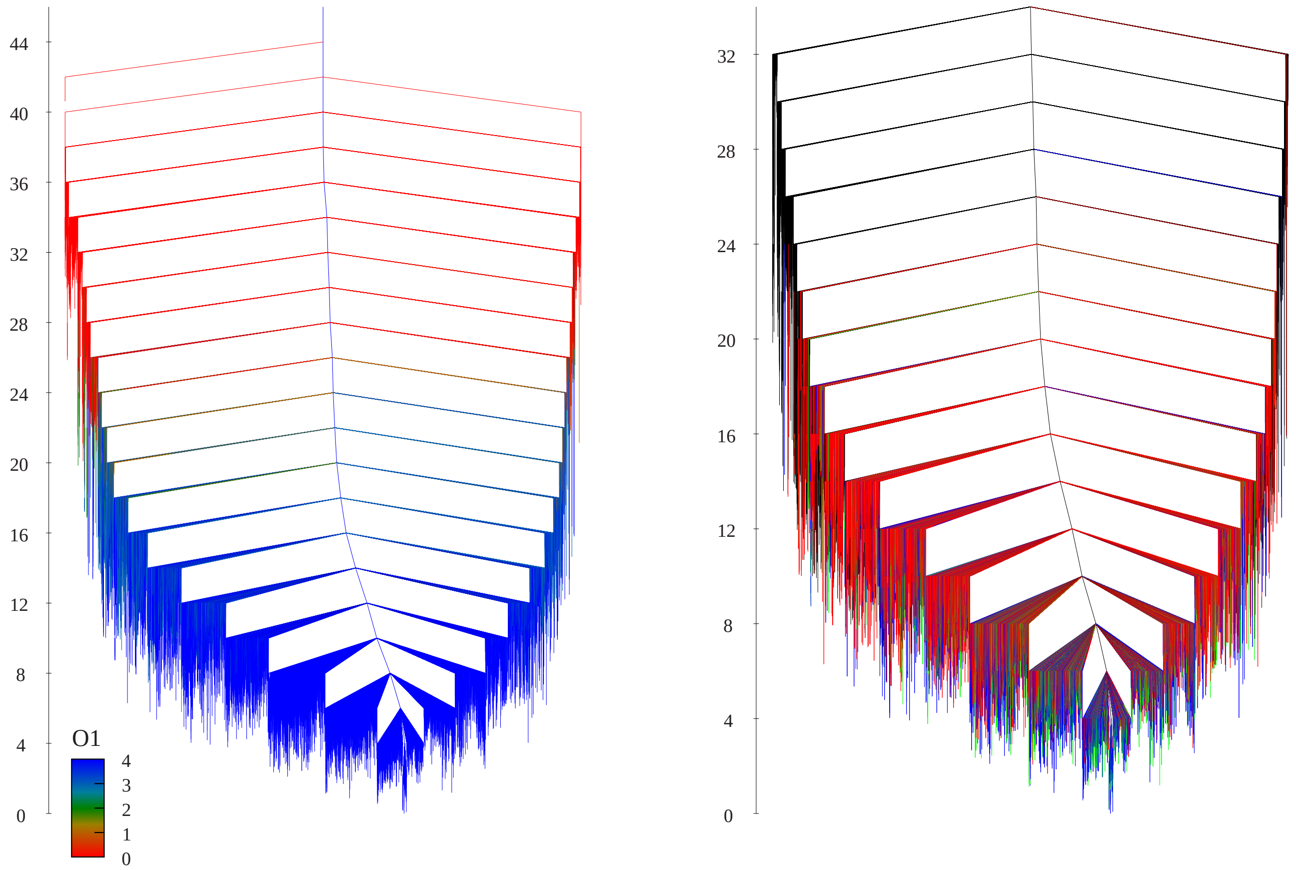

Fig. 4 Disconnectivity graphs for $\alpha = 0.5$. The vertical axes are the total energy, relative to the lowest-energy minimum. Left: full graph, with the branches coloured according to the number of native backbone hydrogen bonds in the corresponding local minimum, with blue representing the maximum (four). Right: magnification of the low-energy region. Here, only the minima with four native backbone hydrogen bonds are coloured, and the colour scheme now displays the number of distances between centres of mass of neighbouring pairs of TRP sidechains that match the PDB structure (tolerances given in the text). Red: one. Green: two. Blue: three

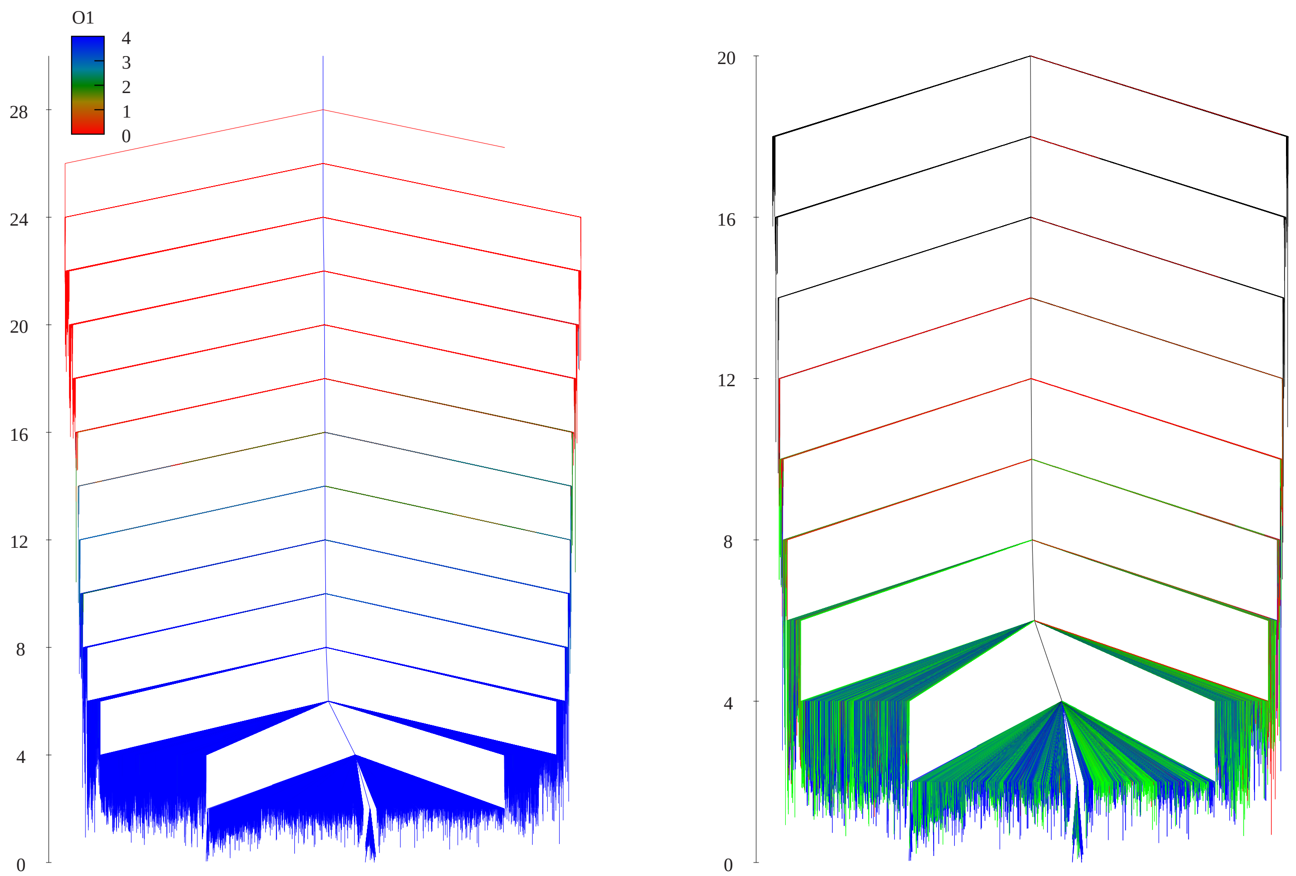

Fig. 5 Disconnectivity graphs for $\alpha = 0.7$. The vertical axes are the total energy, relative to the lowest-energy minimum. Left: full graph, with the branches coloured according to the number of native backbone hydrogen bonds in the corresponding local minimum, with blue representing the maximum (four). Right: magnification of the low-energy region. Here, only the minima with four native backbone hydrogen bonds are coloured, and the colour scheme now displays the number of distances between centres of mass of neighbouring pairs of TRP sidechains that match the PDB structure (tolerances given in the text). Red: one. Green: two. Blue: three

formance in terms of both the basin-hopping statistics and the structure of the observed landscape, which is the most funnelling. Different systems may have different optimal values of α; however, an advantage of the form of the total energy function employed here where the force field component is weighted by $(1-\alpha)$ is that the range of α values is bounded ($0 \leq \alpha \leq 1$, compared with $\alpha \geq 0$ in previous work[9–11]). There is a computational overhead associated with the CamShift potential that must also be considered: the average CPU time required per basin-hopping step is longer by a factor of around 2.5 when CamShift is included compared with CHARMM22/CMAP only, increasing slightly with α.

It was also found that including CamShift improves the efficiency of locating the native secondary structure (backbone hydrogen bonding pattern) to a greater extent than the native sidechain packing. It therefore seems likely that, for difficult cases, structure prediction could be achieved efficiently via a hierarchical procedure with alternating phases. The secondary structure could first be optimized using a potential including restraint terms, followed by a period of refinement of the tertiary structure and local sidechain packing with the force field only and an appropriate geometrical move set, such as group rotations,[50,51] in which sets of atoms are rotated as rigid bodies about chosen axes.

5 Acknowledgements

The authors thank Dr Carlo Camilloni and Dr Falk Hoffmann for helpful discussions about the CamShift program. This research was supported by the EPSRC and the ERC.

6 Data availability

Additional data related to this publication is available at the xxxxxxx data repository.

References

1 H. Gong, Y. Shen and G. D. Rose, *Prot. Sci.* , 2007, **16**, 1515–1521.
2 A. Cavalli, X. Salvatella, C. M. Dobson and M. Vendruscolo, *Proc. Natl. Acad. Sci. USA* , 2007, **104**, 9615–9620.
3 D. S. Wishart, D. Arndt, M. Berjanskii, P. Tang, J. Zhou and G. Lin, *Nucleic Acids Res.*, 2008, **36**, W496–W502.
4 Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker and A. Bax, *Proc. Natl. Acad. Sci. USA* , 2008, **105**, 4685–4690.
5 M. Vendruscolo, E. Paci, C. M. Dobson and M. Karplus, *Nature*, 2001, **409**, 641–645.
6 M. Vendruscolo and C. M. Dobson, *Philos. Trans. Roy. Soc. Lond. A* , 2005, **363**, 433–452.
7 J. Kuszewski, A. Gronenborn and G. Clore, *J. Magn. Reson. Ser. B*, 1995, **107**, 293–297.
8 J. Kuszewski, J. Qin, A. Gronenborn and G. Clore, *J. Magn. Reson. Ser. B*, 1995, **106**, 92–96.
9 P. Robustelli, A. Cavalli, C. M. Dobson, M. Vendruscolo and X. Salvatella, *J. Phys. Chem. B* , 2009, **113**, 7890–7896.
10 P. Robustelli, K. Kohlhoff, A. Cavalli and M. Vendruscolo, *Structure*, 2010, **18**, 923–933.
11 F. Hoffmann and B. Strodel, *J. Chem. Phys.* , 2013, **138**, 025102 1–7.
12 D. S. Wishart, *Prog. Nucl. Mag. Res. Sp.*, 2011, **58**, 62–87.
13 O. M. Becker and M. Karplus, *J. Chem. Phys.*, 1997, **106**, 1495–1517.
14 D. J. Wales, M. A. Miller and T. R. Walsh, *Nature*, 1998, **394**, 758–760.
15 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
16 A. G. Cochran, N. J. Skelton and M. A. Starovasnik, *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 5578–5583.
17 Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 6611–6615.
18 D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111–5116.
19 D. J. Wales and H. A. Scheraga, *Science*, 1999, **285**, 1368–1372.
20 D. J. Wales, *Mol. Phys.*, 2002, **100**, 3285–3306.
21 D. J. Wales, *Mol. Phys.*, 2004, **102**, 891–908.
22 D. J. Wales, *Int. Rev. Phys. Chem.*, 2006, **25**, 237–282.
23 K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella and M. Vendruscolo, *J. Am. Chem. Soc.* , 2009, **131**, 13894–13895.
24 A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, 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. WiorkiewiczKuczera, D. Yin and M. Karplus, *J. Phys. Chem. B* , 1998, **102**, 3586–3616.
25 A. D. Mackerell, M. Feig and C. L. Brooks, *J. Comp. Chem.* , 2004, **25**, 1400–1415.
26 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comp. Chem.*, 1983, **4**, 187–217.
27 U. Haberthuer and A. Caflisch, *J. Comp. Chem.* , 2008, **29**, 701–715.
28 E. Małolepsza, B. Strodel, M. Khalili, S. Trygubenko, S. Fejer and D. J. Wales, *J. Comp. Chem.*, 2010, **31**, 1402–1409.
29 D. J. Wales, *GMIN: A program for basin-hopping global optimisation, basin-sampling, and parallel tempering*.
30 N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.*, 1953, **21**, 1087–1092.
31 J. Nocedal, *Mathematics of Computation*, 1980, **35**, 773–782.
32 D. Liu and J. Nocedal, *Math. Prog.*, 1989, **45**, 503–528.
33 J. M. Carr, S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.*, 2005, **122**, 234903.
34 E. W. Dijkstra, *Numerische Math.*, 1959, **1**, 269–271.
35 D. J. Wales, *OPTIM: A program for optimising geometries and calculating pathways*.
36 S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.*, 2004, **120**, 2082–2094.
37 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 1999, **111**, 7010–7022.
38 G. Henkelman, B. P. Uberuaga and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9901–9904.
39 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978–9985.
40 L. J. Munro and D. J. Wales, *Phys. Rev. B*, 1999, **59**, 3969–3980.
41 Y. Kumeda, L. J. Munro and D. J. Wales, *Chem. Phys. Lett.*, 2001, **341**, 185–194.
42 D. J. Wales, *PATHSAMPLE: A program for generating connected stationary point databases and extracting global kinetics*.
43 J. M. Carr and D. J. Wales, *J. Chem. Phys.*, 2005, **123**, 234901.
44 B. Strodel, C. S. Whittleston and D. J. Wales, *J. Amer. Chem. Soc.*, 2007, **129**, 16005–16014.
45 The PyMOL Molecular Graphics System, Version 1.4.1 Schrödinger, LLC.
46 B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable,

H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *Journal of Computational Chemistry*, 2009, **30**, 1545–1614.

47 J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 7524–6915.

48 J. N. Onuchic and P. G. Wolynes, *Curr. Op. Struct. Biol.* , 2004, **14**, 70–75.

49 D. S. Wishart and D. A. Case, *Nuclear Magnetic Resonance of Biological Macromolecules Part A*, Academic Press, 2002, vol. 338, pp. 3–34.

50 C. Whittleston, *PhD Thesis*, University of Cambridge, 2011.

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