Folding of Small Proteins by Monte Carlo Simulations with Chemical Shift Restraints without the Use of Molecular Fragment Replacement or Structural Homology

Folding of Small Proteins by Monte Carlo Simulations with Chemical Shift Restraints without the Use of Molecular Fragment Replacement or Structural Homology

Paul Robustelli,[†] Andrea Cavalli,[†] Christopher M. Dobson,[†] Michele Vendruscolo,*[,†] and Xavier Salvatella*[,†,‡]

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K., and ICREA and Institute for Research in Biomedicine Barcelona, Baldiri Reixac 10-12, 08028 Barcelona, Spain

Received: January 27, 2009; Revised Manuscript Received: March 30, 2009

It has recently been shown that protein structures can be determined from nuclear magnetic resonance (NMR) chemical shifts using a molecular fragment replacement strategy. In these approaches, structural motifs are selected from existing protein structures on the basis of chemical shift and sequence homology and assembled to generate new structures. Here, we demonstrate that it is also possible to determine structures of proteins by directly incorporating experimental NMR chemical shifts as structural restraints in conformational searches, without the use of structural homology and molecular fragment replacement. In this approach, a protein is folded from an extended conformation to its native state using a simulated annealing procedure that minimizes an energy function that combines a standard force field with a term that penalizes the differences between experimental and calculated chemical shifts. We provide an initial demonstration of this procedure by determining the structure of two small proteins, with α and β folds, respectively.

Introduction

Chemical shifts are highly sensitive probes of molecular structure and are the most readily accessible and precisely measurable nuclear magnetic resonance (NMR) observables in solution and in the solid state.[1,2] The chemical shift of a nuclear spin is the net result of many factors that influence its environment, and is dependent on through-bond effects such as bond hybridization and geometry and on through-space effects such as those resulting from electric fields and magnetic anisotropies. The complex dependency of chemical shifts on a large number of factors makes them a rich source of structural information, although one that is challenging to interpret in terms of conformational properties in large molecules such as proteins. Recent advances in semiempirical chemical shift calculations, however, enable the rapid and accurate prediction of backbone chemical shifts for proteins based on main chain and side chain torsion angles, electric field and ring current effects, and the presence of hydrogen bonding interactions.[3−6] The availability of these relatively fast and reliable methods for the computation of chemical shifts from structural data is stimulating their use in methods of protein structure determination and refinement; recent examples have shown that specific conformations can be determined to a resolution of 2 Å or better using chemical shifts as the only experimental restraints.[7−13]

In these techniques,[7−10] backbone chemical shifts and an analysis of known structures are used together to determine the local conformational preferences of protein fragments, which are then assembled into structures that are further refined in order to provide a final conformation. A key component of these methods is therefore an inference of the structures of local fragments from databases of protein structures in combination with experimental restraints derived from chemical shifts. By focusing the search on regions of conformational space that are preferentially sampled by folded proteins, the search for the correct average structure is rendered very efficient. These developments show that the structural information provided by chemical shifts is sufficient, when combined with state-of-the-art force fields and structural databases, to determine accurately and reliably the average structure of proteins[7−10,13] and protein complexes[11] in solution and in the solid state.[12,13] So far, these approaches have made use of chemical shifts to identify commonly occurring protein motifs and folds in order to rapidly suggest possible structures for the complete protein prior to the refinement procedure; a major objective of these studies has been to reduce the amount of data acquisition required for the determination of the native structure of globular proteins.

Chemical shifts have in addition a great potential for the structural characterization of many important classes of proteins and protein states that have thus far proved to be very difficult to characterize with conventional means of structure determination, including non-native states,[14,15] folding intermediates,[16,17] intrinsically disordered proteins,[18] and protein aggregates.[19,20] For such species, NMR spectroscopy is the only technique likely to be able to provide extensive residue-specific structural information, and chemical shifts and residual dipolar couplings (RDCs) are expected to be the only experimental NMR parameters that can be measured with a good degree of completeness. In order to extend further the use of NMR spectroscopy in structural biology and to enable the structural characterization of such states of proteins, a powerful strategy would be to incorporate chemical shift restraints in methods analogous to those used for NMR structure determination based on nuclear Overhauser effects (NOEs) and RDCs.[21−23] In such methods, a conformational search is used to identify protein structures that minimize an energy function given by the sum of a standard molecular mechanics force field and a penalty function derived from experimental restraints. Minimization of the penalty function biases the conformational search to structures consistent with the experimental restraints, while the

* To whom correspondence should be addressed. E-mail: mv245@cam.ac.uk (M.V.); xavier.salvatella@irbbarcelona.org (X.S.).

[†] University of Cambridge.

[‡] ICREA and Institute for Research in Biomedicine Barcelona.

10.1021/jp900780b CCC: $40.75
Published on Web 05/08/2009

inclusion of the force field ensures that the search is limited to energetically favorable structures. This approach has been extensively utilized with penalty functions derived from NOEs, scalar couplings, and RDCs,[21,22] but its use for structure determination using chemical shifts as restraints has been limited to the refinement of predetermined structures.[24]

In the present study, we demonstrate the use of a penalty function for chemical shifts as restraints in molecular simulations and provide initial evidence of the ability of conformational searches based on chemical shifts to determine structures of proteins starting from random conformations without any additional experimental restraints. These results demonstrate that NMR chemical shift restraints can be directly incorporated into structure calculation protocols to determine the structures of proteins without utilizing homology information or requiring any other experimental data. In addition, they make it possible to envisage the use of this most fundamental NMR observable for the determination of the detailed structures and, possibly, dynamics of protein states that are not easily amenable to X-ray crystallography and to more conventional methods of NMR spectroscopy.

Results

We first present the development of a penalty function to restrain molecular simulations to those regions of conformational space that are compatible with a given set of experimentally measured chemical shifts. The penalty function converts the differences between experimentally measured and calculated $^{1}H_{\alpha}$,$^{13}C_{\alpha}$,$^{13}C_{\beta}$, and ^{15}N backbone chemical shifts of a specific protein conformation into a bias that drives the simulation toward the specific conformation of the protein that matches the data more closely, i.e., to the solution structure of the protein. We then demonstrate the ability of a Monte Carlo protocol that incorporates this penalty function to determine the solution structures of two proteins, SDA, a 46-residue two-helix bundle, and the Itch E3 ubiquitin ligase third WW domain, a 37-residue all β protein, and compare these structures with those determined previously by conventional methods.

Chemical Shift Penalty Function. The chemical shift restraint function, E_{CS}, developed in this work is a tunable soft-square harmonic well, equivalent to the penalty functions used to restrain other NMR observables such as NOEs, RDCs, and scalar couplings, that is designed to reproduce the approximately Gaussian distribution of experimental errors. In order to render the structure determination procedure efficient, E_{CS} is split into three regions, as shown in Figure 1: a flat-bottomed region that takes into account inaccuracies in the chemical shift predictions, a harmonic region that penalizes statistically significant deviations between the computed and experimental shifts, and a linear region that prevents large deviations of individual chemical shifts from dominating the magnitude of E_{CS} and thus frustrating the conformational search (see Methods).

In order to compute the chemical shifts of candidate structures, we use SHIFTX, a semiempirical method developed by Wishart and co-workers[3] that is sufficiently fast to be used at each step in molecular simulations; quantum mechanical calculations of chemical shifts from protein structures are rapidly improving in accuracy and could provide in principle a valid alternative to semiempirical methods but are too computationally demanding for current structure determination protocols.[25] An important property of SHIFTX is that it deconvolutes the relationship between chemical shifts and protein structure by fitting combinations of structural variables, that include hydrogen bonding distances as well as main chain and side chain torsion angles, to the chemical shifts measured experimentally. To this aim, it uses a database of high-quality protein structures for which chemical shifts are available to produce a general expression for the computation of the shifts that is transferable to other proteins and protein states. As SHIFTX is an approximate method, it is important to account for its inaccuracies to ensure that only statistically relevant violations affect the outcome of the restrained simulations. E_{CS} has thus been implemented with a flat-bottomed function with a width determined by the reported accuracy of the SHIFTX calculations for each atom type[3] that is scaled by an adjustable tolerance n, where the width of the flat region is equal to the reported accuracy of the shift calculations when $n = 1$. In addition, in order to ensure that chemical shift violations are weighted according to their statistical significance, they are normalized by a quantity that accounts for the different gyromagnetic ratios of the nuclei found in proteins and for the degree of variability of each chemical shift type in protein structures (see Methods).

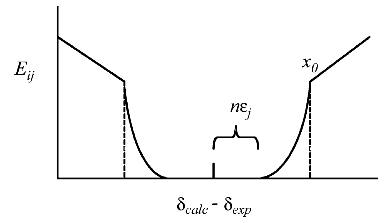

Figure 1. Chemical shift penalty function used in the Monte Carlo simulations described in this work. The energy term E_{ij} gives the contribution of each chemical shift to the total penalty E_{CS} from the difference between the calculated (δ_{calc}) and experimental shift (δ_{exp}); j is the chemical shift type with $j = \{^{1}H_{\alpha}, ^{13}C_{\alpha}, ^{13}C_{\beta}, ^{15}N\}$ and i is the residue number. The function E_{ij} is flat-bottomed with a width determined by the term $n\varepsilon_j$, where n is an adjustable parameter called the tolerance and ε_j is the accuracy of the predictions used for chemical shifts of the type j. The penalty is harmonic until the deviation reaches a cutoff value x_0. Deviations in excess of the value of x_0 contribute linearly to the size of the penalty.

Computational Experiments. Standard NMR structure calculation protocols often use a combination of molecular dynamics simulations and simulated annealing (SA) techniques[26] to minimize functions and yield configurations of the protein in agreement with experimental data. In this procedure standard NMR experimental restraints are converted into geometric restraints with penalty functions that are differentiable with respect to the positions of the atoms of the protein. Since current methods of rapid semiempirical chemical shift calculations such as SHIFTX are not fully differentiable, E_{CS} was implemented in a Monte Carlo conformational search (see Methods).

The landscape obtained by combining the energy of the force field, E_{FF}, with that of the chemical shift restraint, E_{CS}, can be very rugged primarily because of the very high sensitivity of chemical shifts to small changes in the structure of the protein, particularly in compact and relatively rigid regions of the fold that are rich in aromatic side chains.[27] In order to assess in detail the effect of varying the value of n in an SA scheme and to increase the ability of the simulations to search effectively conformational space, we carried out an initial series of computational experiments in which we calculated the structures of two protein fragments, a 12-residue α-helix (residues 23−34 of ubiquitin, PDB code 1ubq)[28] and a 16-residue β-hairpin (residues 41−56 of protein G, PDB code 2gb1).[29] In these preliminary calculations, we used chemical shifts computed from the target structures by SHIFTX as target chemical shifts in the definition of E_{CS}. For the calculations of the structure of two proteins, SDA and the E3 WW domain, we instead used the

TABLE 1: Assessment of the Influence of the Value of the Tolerance n on the Accuracy of the Average Structures and on the Ability of the SA Procedure to Minimize Simultaneously E_{FF} and E_{CS}[a]

	α-helix		β-strand	
n	⟨rms⟩	μ	⟨rms⟩	μ
0	n.a.	0	n.a.	0
1	0.4	3	0.6	1
2	0.6	9	1.1	4
3	0.9	9	2.2	7
4	2.0	10	2.8	10
∞	2.8	n.a.	4.2	n.a.

[a] We report the root mean square (⟨rms⟩) value of the average backbone pairwise distance of the ensemble of μ selected average structures obtained after 10 rounds of SA. Structures were selected if they had an average violation lower than 0.25 ppm in $^{13}C^{\alpha}$ chemical shift units and if their energy was lower than −90 kcal·mol− 1 for the α-helix and −120 kcal·mol− 1 for the β-strand.

experimental chemical shifts reported in the BMRB data bank (http://www.bmrb.wisc.edu) with reference corrections made by the program SHIFTCOR.[30]

In these experiments, the weight, α, of E_{CS} was progressively raised from 0.1 to 1.0, while the temperature was annealed from 1000 to 300 K at a constant value of n. The results of this experiment, which are summarized in Table 1, demonstrate that the probability of minimizing successfully both E_{CS} and E_{FF}, and therefore the quality of the resulting structures, is determined by the value of n. For large values of this parameter, we found that most attempts at minimizing the energy were successful but that the quality of the lowest energy structures was lower, whereas when $n = 1$ the probability of successfully minimizing the energy was low, but structures with low energies were very accurate. When no flat bottom was included, i.e., $n = 0$, the landscape was too rugged and the SA was unsuccessful at minimizing the energy.

To summarize, therefore, for large values of n, the calculated chemical shifts of a wide range of conformations fall near the flat-bottomed region of E_{CS} and thus generate relatively small energetic penalties; the energy landscape is therefore smoother and contains lower barriers and shallower minima. By contrast, for small values of n, more conformations generate chemical shifts that result in large penalties, thus creating a rugged energy landscape with barriers of larger heights and a well-defined global minimum. When smaller values of n are used in the simulations, searches are thus much more likely to become frustrated in deep local minima; the global minima of conformational space will, however, be well-defined, and only a small range of structures will successfully minimize the value of E_{CS} and E_{FF} simultaneously.

Monte Carlo Simulated Annealing Protocols. We found that, by running several cycles of SA beginning with large values of n and incrementally decreasing n after each cycle, E_{CS} and E_{FF} could be progressively minimized more efficiently and more successfully than with cycles of SA run at constant n. This process of tolerance annealing is illustrated in Figure 2. During each cycle of SA, chemical shift violations progressively decrease to generate structures that successfully minimize E_{CS} for a given value of n. By gradually decreasing n after each minimization, the magnitude of E_{CS} does not become large enough to frustrate the conformational search or to exceed the contributions of E_{FF} and result in distorted structures. Chemical shift violations are minimized by a conformational search on an energy landscape that has minimal E_{CS} penalty barriers while still providing a driving force for minimization. Running successive rounds of tolerance annealing, where n is repeatedly annealed from larger values to 1, allows the large structural adjustments required to free the system from local energy minima to take place.

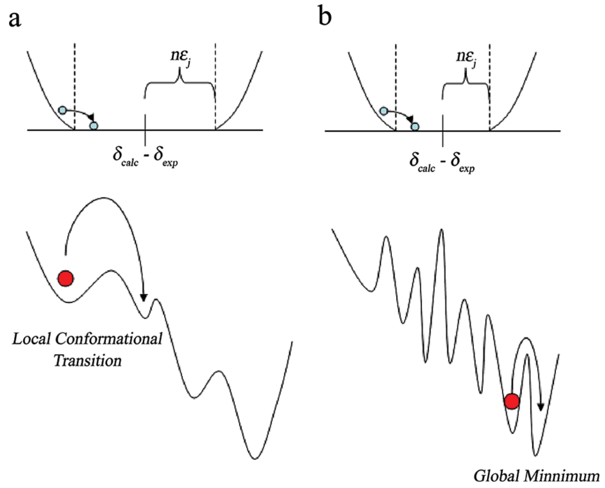
a
b
$n\varepsilon_j$
$n\varepsilon_j$
$\delta_{calc} - \delta_{exp}$
$\delta_{calc} - \delta_{exp}$
Local Conformational Transition
Global Minnimum

Figure 2. Schematic illustration of the effect of the tolerance n on the ruggedness of the energy landscape. (a) For large values of n, few conformations generate penalties, resulting in a smoother energy landscape that contains few low barriers. (b) For small values of n, several conformations generate large penalties, so that the energy landscape is rugged with numerous barriers of large heights. Progressively decreasing the tolerance over the course of several rounds of SA provides the advantage of searching a conformational landscape with E_{CS} barriers that are as small as possible while still providing a driving force for minimization.

In addition to the gradual annealing of the tolerance of E_{CS} over multiple cycles of SA, we found that the best results were obtained when the weight of E_{CS}, α, was annealed adaptively[31] in each cycle of SA. To prevent E_{CS} from becoming too large with respect to E_{FF}, and in doing so driving the search to nonphysical regions of conformational space, the maximum value of α and the increment by which it was raised were determined by the extent to which chemical shift violations had been minimized (see Methods).

Protein Structure Calculations. We applied the Monte Carlo scheme with NMR chemical shift restraints introduced in this work to determine the solution structures of two proteins with very different architectures, SDA, a 46-residue two-helix bundle, and the Itch E3 ubiquitin ligase third WW domain, a 37-residue all β protein. The simulations were initiated from extended conformations without any experimental restraints other than chemical shifts (see Methods). The minimization of the restrained energy $E = E_{FF} + E_{CS}$ over the course of successive rounds of tolerance annealing, each consisting of several cycles of SA where α was increased while the temperature was decreased at constant n, is shown for both proteins in Figure 3. The contributions of E_{FF} and E_{CS} during the simulations are illustrated in Figure 3c and d, which show that E_{FF} remains relatively constant throughout the simulations compared to E_{CS}. The energy landscapes generated in both conformational searches are presented as a function of the root mean square deviation (rmsd) from the target structures and are compared to control simulations run without chemical shift restraints in Figure 4. The latter simulations failed to find native-like structures for either protein, thus demonstrating the crucial contribution of the information provided by the chemical shifts.

The lowest energy structures of SDA and the E3 WW domain obtained are compared to previously determined structures of

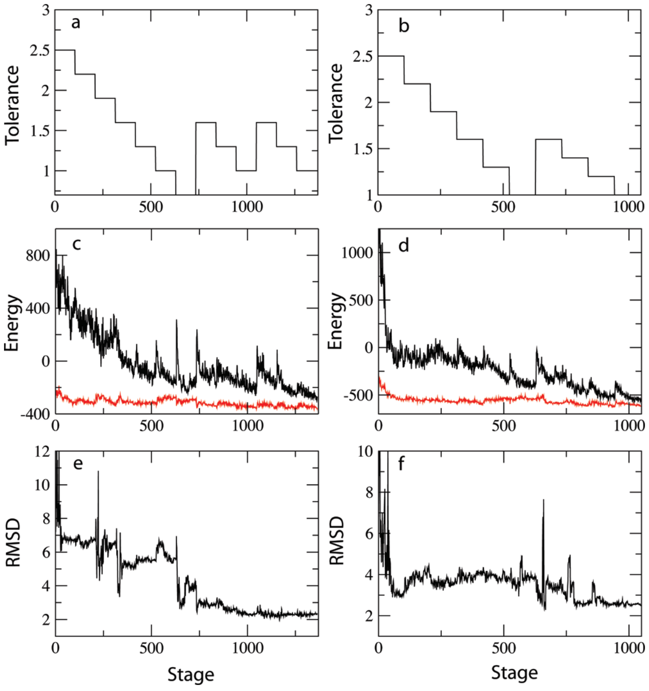

Figure 3. Representative trajectories for the folding of E3 WW (left-hand panels) and SDA (right-hand panels). The tolerance, *n*, of E_{CS} is annealed over multiple cycles of SA. Each cycle of SA consists of 21 Monte Carlo runs. The temperature is progressively decreased after each run, while the weight (α) of E_{CS} is increased adaptively (see Methods). To demonstrate the progress of the simulations, structures were selected at regular intervals from each Monte Carlo run, and the properties of the structures are reported (each Monte Carlo run is represented by five stages). (a, b) Process of tolerance annealing for E3 WW(a) and SDA (b). Rounds of tolerance annealing were repeated until no further improvement of total energy was observed. (c, d) Energy of the structure from each stage for E3 WW (c) and SDA (d). Black lines represent the total energy ($E_{FF} + E_{CS}$). Red lines represent the force field energy alone (E_{FF}). (e, f) The backbone rmsd of the structure from each stage is shown for E3 WW (e) and SDA (f).

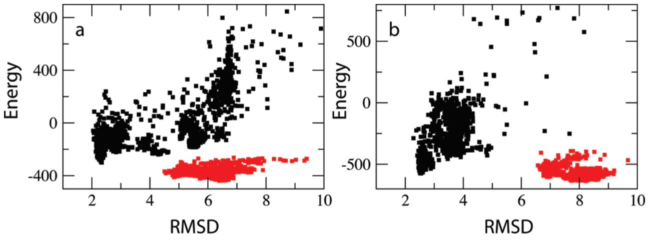

Figure 4. Energy landscapes of the structures generated in the folding of E3 WW (a) and SDA (b) as a function of the rmsd from previously determined structures. The total energy ($E_{FF} + E_{CS}$) for the restrained simulations is shown in black. The energies of control simulations, run without chemical shift restraints, are shown in red.

the two proteins in Figure 5. The lowest energy structures calculated in this study have backbone (BB) rmsd's of 2.50 and 2.51 Å from the experimental X-ray crystal structures of SDA (PDB code 1pvo) and E3 WW domain NMR structure (PDB code 1yiu), respectively, computed by taking into account all residues between the first and last elements of ordered secondary structure.

Discussion

Many current methods for the determination of protein structures from chemical shifts utilize molecular fragment replacement strategies.[7−10] In these methods, chemical shifts are first used to identify local structural propensities and to select potential structures for fragments of the protein from structural databases; these fragments are then assembled combinatorially to generate trial structures that are subsequently filtered and subjected to a high-resolution refinement again using the chemical shift data.

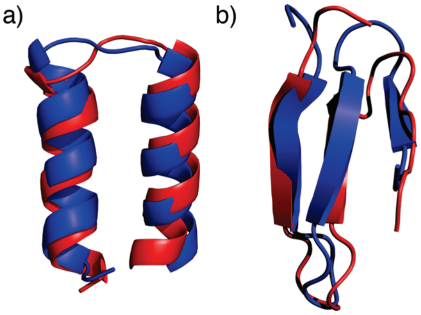

Figure 5. Comparison of the lowest energy structures calculated here (red) with the reference structures (blue), which were determined by standard methods. (a) The SDA structure has a backbone rmsd of 2.51 Å from the previously determined NMR structure (PDB code 1pvo). (b) The Itch E3 ubiquitin ligase third WW domain has a backbone rmsd of 2.50 Å from the previously determined NMR structure (PDB code 1yiu).

In the present work, we have described an approach that exploits a direct mapping from chemical shifts to structure without reference to sequence and structural homology; this method does not require the use of structural databases (see Methods). In order to validate this approach, we have shown that it can be used to determine native protein structures by incorporating experimental chemical shifts as restraints in molecular simulations. We have illustrated the procedure by determining the native states of two small proteins and by assessing the accuracy of the results. The methodology that we have presented, in its current implemenation, is not as efficient or accurate as existing techniques that utilize molecular fragment replacement and structural homology for the determination of structures of globular proteins. For proteins of this size, molecular fragment replacement techniques require roughly one tenth of the computational time, and typically produce structures with backbone rmsd's of less than 2.0 Å from structures solved by convential means.[7,9,10] The approach that we have presented here does not aim at replacing molecular fragment replacement as a strategy for native protein structure determination from chemical shifts, since the latter is a particularly suitable tool for exploiting the wealth of information about structural motifs contained in structural databases. Our goal was instead to provide an additional tool to perform structural calculations and refinements of existing structural models using the information provided by chemical shifts in a manner that is analogous with standard NMR structure calculation protocols and can easily be combined with restraints derived from traditional NMR observables; subject to further development, this method could be used to investigate systems for which structural homology could be adopted less effectively than for native states.

The current major limitation of this approach is its high computational cost, caused by the rugged nature of the energy landscape in the presence of the chemical shift restraint, which renders the Monte Carlo simulations inefficient when a high percentage of moves is rejected. Although we have shown that it is in part possible to overcome this difficulty by adjusting the tolerance of the penalty function throughout the conformational search, the simulations are however still of limited efficiency, with Monte Carlo move acceptance probablities of about 10% when chemical shift restraints are used for structure determination. We nevertheless anticipate that it will be possible to improve the efficiency and accuracy of this procedure by using optimized sampling algorithms and exploiting further increases in computational power.

The amount of time required for each chemical shift calculation is also a limiting factor. Control simulations run without chemical shift restraints were approximately six times faster then restrained simulations that required the computation of shifts at every Monte Carlo step. The length of time required for these simulations at present would make applications of systems larger than those demonstrated here feasible only with significant computational resources. The development of more rapid chemical shift predictors will, however, greatly improve the efficiency and applicability of this method.

Conclusion

Computational methods that exploit chemical shifts to yield protein structures at atomic resolution have recently been introduced, potentially allowing for very significant decreases in the amount of experimental and user time required to determine protein structures using NMR spectroscopy. These methods use the information contained in experimental chemical shifts together with structural homology of proteins in structural databases such as the protein data bank to generate new structures. Here, we introduce an approach that directly incorporates chemical shifts as restraints in molecular simulations with an energetic penalty function analogous to those used in standard NMR structure calculations and demonstrate that it is possible to use these conformational searches to fold α and β secondary structure elements and correctly orient their tertiary contacts. These results suggest that it should be possible to determine protein structures from chemical shifts without the use of structural homology information. The two proteins that we considered as test cases contain fewer than 50 amino acids, and have relatively simple topologies. The amount of computational time required to achieve convergence is expected to significantly increase for larger proteins with more complex topological elements, suggesting that the size limit for the current implementation of this computational procedure is probably not much larger than 50 residues, compared to the 130 achievable through current molecular fragment replacement procedures.[7,9,10] While folding proteins from extended conformations using only chemical shift restrained conformational searches is a computationally intensive task, this method could easily be combined with restraints traditionally used in NMR structure calculations such as NOEs, scalar couplings, and RDCs, and could be used to refine initial structural models. We thus suggest that further developments in the speed and accuracy of chemical shift calculations and the direct incorporation of chemical shift restraints in conformational searches will make chemical shifts an increasingly important tool in structural biology.

Methods

Chemical Shift Penalty Function. At each step in the Monte Carlo simulations, the chemical shift predictor SHIFTX[3] is used to compute the chemical shifts of all $^1H_\alpha$, $^{13}C_\alpha$, $^{13}C_\beta$, and ^{15}N backbone atoms of the protein. The differences between the experimentally measured, δ_{exp}, and calculated, δ_{calc}, chemical shifts for each atom are converted into an energetic penalty according to E_{CS}:

$$E_{CS} = \alpha\rho \tag{1}$$

where α defines the contribution of the chemical shift restraints to the total energy of the system ($E = E_{FF} + E_{CS}$). In this equation, ρ is a measure of chemical shift violations of a given conformation:

$$\rho = \sum_{i}^{N} \sum_{j} \beta_j E_{ij} \tag{2}$$

where N is the number of assigned residues in the protein, j designates the chemical shift type with $j = \{^1H_\alpha, {}^{13}C_\alpha, {}^{13}C_\beta, {}^{15}N\}$, and β_j is a weight associated with each type that is a function of the variability of that chemical shift in folded proteins reported in the BMRB database. The penalty E_{ij} calculated for each chemical shift, shown in Figure 1, is given by

$$E_{ij} = \begin{cases} 0 & \text{if } |\delta_{calc} - \delta_{exp}| < n\varepsilon_j \\ (|\delta_{calc} - \delta_{exp}| - n\varepsilon_j)^2 & \text{if } n\varepsilon_j < |\delta_{calc} - \delta_{exp}| < x_0 \\ b \times (|\delta_{calc} - \delta_{exp}| - x_0) + (x_0 - n\varepsilon_j)^2 & \text{if } x_0 < |\delta_{calc} - \delta_{exp}| \end{cases} \tag{3}$$

This function is flat-bottomed to ensure that chemical shifts calculated to be within a designated accuracy of the experimental value do not produce any penalty. The width of the flat region of the potential is set by the term $n\varepsilon_j$, where n is an adjustable generic "tolerance" parameter and ε_j is the accuracy of the predictions used for chemical shifts of the type j reported by the developers of SHIFTX.[3] For deviations larger than $n\varepsilon_j$, the penalty is harmonic until the deviation reaches a cutoff value x_0. Deviations in excess of the value of x_0 contribute linearly to the size of the penalty. The properties of the penalty function E_{CS} are flexible and can be adjusted through the parameters n, x_0, b, and α. We chose to use the subset of $^1H_\alpha$, $^{13}C_\alpha$, $^{13}C_\beta$, and ^{15}N chemical shifts, which was previously demonstrated to be sufficient to describe native states of proteins,[7] as the inclusion of additional shifts did not appear to increase substantially the quality of the calculated structures but required longer simulation times.

Monte Carlo Simulations. The Monte Carlo simulations were performed with the molecular simulation package almost ("all atom molecular simulation toolkit"; www.open-almost.org) using an implicit water all-atom model with a potential function adopted from the CHESHIRE protocol.[7]

$$E_{FF} = E_{CHARMM19} + E_{HB} + E_{PMF} + E_{SASA} \tag{4}$$

where $E_{CHARMM19}$ is the energy of the CHARMM19 force field,[32] E_{HB} is a potential modeling backbone hydrogen bond formation,[33] E_{PMF} is a knowledge based pairwise potential of mean force implemented following Zhou and Zhou,[34] and E_{SASA} models solvation.[35] The chemical shift penalty E_{CS} was added to E_{FF}. The weights of the terms contributing to E_{FF} were adopted from the CHESHIRE protocol,[7] where they were empirically optimized on the landscapes of several proteins with varying topologies. Our energy function should therefore be equally applicable for a range of topologies.

The Monte Carlo simulations in this work employed move sets containing backbone and side chain moves.[7] The backbone moves simultaneously rotated the ϕ and ψ dihedral angles of between one and four randomly selected residues, while the side chain moves rotated a single randomly selected χ angle. The step size of each move was drawn from a Gaussian distribution with a designated mean of 0 and a standard deviation between 2 and 15°. ω torsion angles were not adjusted by the MC move sets. The source code as well as the input files used in this work can be obtained from the Web site or directly from the authors.

Structure Calculation Protocol. All protein structure calculations began with a polypeptide chain in an extended conformation, with ω torsion angles set to 180°. Protein structure calculations consisted of multiple rounds of tolerance annealing, in which the value of the tolerance n was annealed over the course of several cycles of SA. Cycles of SA were carried out with constant values of n, b, and x_0 (eq 3). In all SA cycles, we set $b = 4$ and $x_0 = 2n\varepsilon_j$ for each shift type j. Cycles of SA began with a short run (20 ps) of unrestrained molecular dynamics to release the system from local energetic minima. Each cycle consisted of 21 Monte Carlo stages, each containing 100 000 steps. The temperature of the simulations was linearly annealed from 600 to 300 K over the course of the 21 stages. This range of temperatures, which is narrower than that used for the computational experiments carried out with peptide fragments (1000−300 K), was modified to minimize the probability of unfolding elements of structure formed in previous cycles. The weight α of E_{CS} was increased adaptively[31] throughout the stages, with the size of the increment $\Delta\alpha$ between stages and the maximum value of α, α_{max}, determined by the degree to which ρ had been minimized. $\Delta\alpha$ and α_{max} were respectively set to 0.1 and 2.0 for $\rho > 9.0$, 0.15 and 2.0 for $9.0 > \rho > 5.0$, and 0.2 and 3.0 for $\rho < 5.0$. The range of selected values of α was determined empirically; the magnitude of α was increased until it was observed that minimizations of E_{CS} only occurred at the expense of significant increases in E_{FF}.

A radius of gyration (R_g) restraint was added after the first two cycles of SA in each round of tolerance annealing to favor compact structures.[36,37] Structures with $R_g > R_{max}$, where $R_{max} = 2.83N^{0.34}$ and N is the number of amino acids in the protein, were penalized by $0.25(R_g - R_{max})^2$. In every cycle of SA, 12 Monte Carlo simulations were run with variable move sets. Each move set contained one side chain move and four backbone moves, and the number of residues and the width of the step size in each move were varied between the simulations. Move sets were selected to range from conservative, using only one and two residue moves with narrow step size distributions, to liberal, using up to four residue moves with wide step size distributions. Of the 12 move sets used, the most conservative move set contained a side chain move with a step size distribution with a standard deviation of 10°, two single residue backbone moves with step size distributions with standards deviations of 4 and 10°, and two double residue backbone moves with step size distributions with standard deviations of 2 and 5°. The most liberal move set contained a side chain move with a step size distribution with a standard deviation of 15°, a double residue backbone move with a step size distribution with a standard deviation of 15°, and a quadruple residue backbone move with a step size distribution with a standard deviation of 7.5°.

When all of the simulations were completed, the energies of all structures saved during sampling were minimized and recomputed with a standard weight of E_{CS} ($\alpha = 1$), and the lowest energy structure was selected as the start structure for the next cycle of SA. Only the Monte Carlo simulations containing the lowest energy structures were used for analysis here. Structure calculations began with a SA cycle at $n = 2.5$, and n was lowered by 0.3 after each cycle. A round of tolerance annealing was considered complete after a cycle was run at $n = 1$. Additional rounds of tolerance annealing began from the lowest energy structure from the preceding round, and began with cycles at $n = 1.6$.

Rounds of tolerance annealing were repeated until a small value of E_{CS} ($\rho < 10.0$) was observed at $n = 1$, and an additional round of tolerance annealing failed to find structures of lower energy. Previous studies have demonstrated that it is unlikely that an incorrect structure will have a very low value of a molecular mechanics force field while maintaining good agreement with experimental chemical shifts.[7−10] By gradually decreasing the size of the tolerance and monitoring the force field energy, E_{FF}, over the course of the simulation, it is possible to tell when E_{CS} is being satisfied only at the expense of sacrifices to E_{FF}. In a successfully converging simulation, force field energies should remain relatively stable as E_{CS} is reduced; a large increase in E_{FF} indicates that a structure is unlikely to be correct. As chemical shift based structure calculations are recent developments and are not yet routine, it is important to bear in mind that there could be regions of conformational space that are not well described by chemical shift predictors and incorrect structures could be identified as the best model.

For the two proteins studied in this investigation, one round of tolerance annealing was run from $n = 2.5$ to $n = 1$, consisting of six cycles of SA, and additional rounds of tolerance annealing from $n = 1.6$ to $n = 1$, each consisting of three cycles of SA, were run three times for E3 WW and two times for SDA before convergence was achieved. This resulted in a total of 15 and 12 cycles of SA for E3 WW and SDA, respectively. Each cycle of SA requires 3 h of CPU time on a 2.4 GHz Dual Core AMD Opteron(TM) Processor 280 and was run with different move sets on 12 processors. The folding of the proteins required approximately 473 CPU days for E3 WW and 380 CPU days for SDA. In comparison, each cycle of SA in the control simulations run without computing chemical shifts at each Monte Carlo step with SHIFTX required only 0.5 h of CPU time on a 2.4 GHz Dual Core AMD Opteron(TM) Processor 280. Control simulations of the same length required 79 CPU days for E3 WW and 63 CPU days for SDA, indicating that the major bottleneck in these simulations is the computation of chemical shifts at each step. The determination of the structure of these two fragments using the CHESHIRE approach,[7] that focuses the search on native-like regions of the configurational space available to the protein sequence using molecular fragment replacement, takes about 25 CPU days.

The Monte Carlo protocol described here could be further optimized to increase its efficiency and accuracy. Significant gains in computational time could be achieved by parallelizing the annealing of both the temperature and the tolerance with multidimensional replica exchange. Rather than adjusting the tolerance of the penalty over multiple cycles of SA, simulations could simultaneously be run at several temperatures, with several replicas running at a range of values of the tolerance for each temperature. Replica exchanges could then be attempted between replicas running at the same temperature with different tolerances and between replicas running at the same tolerance and different temperatures.

Acknowledgment. The authors thank Prof. Wishart (University of Alberta) for sharing the source code of SHIFTX.[3] This work was supported by the Gates Foundation (P.R.), the EU (A.C.), EMBO (M.V.), the Royal Society (M.V.), the Wellcome Trust (X.S. and C.M.D.) and the Leverhulme Trust (A.C., C.M.D., M.V., and X.S.).

References and Notes

(1) Wishart, D. S.; Case, D. A. *Methods Enzymol.* **2001**, *338*, 3–34.
(2) Cornilescu, G.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1999**, *13*, 289–302.
(3) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215–240.

(4) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321–333.

(5) Meiler, J. *J. Biomol. NMR* **2003**, *26*, 25–37.

(6) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289–302.

(7) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615–20.

(8) Gong, H.; Shen, Y.; Rose, G. D. *Protein Sci.* **2007**, *16*, 1515–21.

(9) Shen, Y.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685–90.

(10) Wishart, D. S.; Arndt, D.; Berjanskii, M.; Tang, P.; Zhou, J.; Lin, G. *Nucleic Acids Res.* **2008**, *36*, W496–502.

(11) Montalvo, R. W.; Cavalli, A.; Salvatella, X.; Blundell, T. L.; Vendruscolo, M. *J. Am. Chem. Soc.* **2008**, *130*, 15990–15996.

(12) Robustelli, P.; Cavalli, A.; Vendruscolo, M. *Structure* **2008**, *16*, 1–6.

(13) Shen, Y.; Vernon, R.; Baker, D.; Bax, A. *J. Biomol. NMR* **2009**, *43*, 63–78.

(14) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–22.

(15) Fuentes, G.; Nederveen, A. J.; Kaptein, R.; Boelens, R.; Bonvin, A. M. J. J. *J. Biomol. NMR* **2005**, *33*, 175–86.

(16) Whittaker, S. B.-M.; Spence, G. R.; Grossmann, J. G.; Radford, S. E.; Moore, G. R. *J. Mol. Biol.* **2007**, *366*, 1001–15.

(17) Korzhnev, D. M.; Salvatella, X.; Vendruscolo, M.; Di Nardo, A. A.; Davidson, A. R.; Dobson, C. M.; Kay, L. E. *Nature* **2004**, *430*, 586–590.

(18) Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430–1435.

(19) Wasmer, C.; Lange, A.; Melckebeke, H. V.; Siemer, A. B.; Riek, R.; Meier, B. H. *Science* **2008**, *319*, 1523–6.

(20) van der Wel, P. C. A.; Lewandowski, J. R.; Griffin, R. G. *J. Am. Chem. Soc.* **2007**, *129*, 5117–30.

(21) Braun, W.; Bösch, C.; Brown, L. R.; Go, N.; Wüthrich, K. *Biochim. Biophys. Acta* **1981**, *667*, 377–96.

(22) Schwieters, C. D.; Kuszewski, J. J.; Clore, G. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48*, 47–62.

(23) Wüthrich, K. *J. Biomol. NMR* **2003**, *27*, 13–39.

(24) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson,. Ser. B* **1995**, *107*, 293–7.

(25) de Dios, A. C.; Pearson, J. G.; Oldfield, E. *Science* **1993**, *260*, 1491–1496.

(26) S. Kirkpatrick, C. D. G.; Vecchi, M. P. *Science* **1983**, *220*, 671–680.

(27) Vila, J. A.; Villegas, M. E.; Baldoni, H.; Scheraga, H. A. *J. Biomol. NMR* **2007**, *38*, 221–235.

(28) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–44.

(29) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657–661.

(30) Zhang, H.; Neal, S.; WIshart, D. *J. Biomol. NMR* **2003**, *25*, 173–195.

(31) Triki, E.; Collette, Y.; Siarry, P. *Eur. J. Oper. Res.* **2005**, *166*, 77–92.

(32) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(33) Kortemme, T.; Morozov, A. V.; Baker, D. *J. Mol. Biol.* **2003**, *326*, 1239–1259.

(34) Zhou, H.; Zhou, Y. *Protein Sci.* **2002**, *11*, 2714–26.

(35) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2002**, *46*, 24–33.

(36) Gong, H.; Fleming, P. J.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16227–32.

(37) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 1337–1338.

JP900780B