[6] Rapid Assessment of Protein Structural Stability and Fold Validation via NMR

By Bernd Hoffmann, Christian Eichmüller, Othmar Steinhauser, and Robert Konrat

Abstract

In structural proteomics, it is necessary to efficiently screen in a high-throughput manner for the presence of stable structures in proteins that can be subjected to subsequent structure determination by X-ray or NMR spectroscopy. Here we illustrate that the ^{1}H chemical distribution in a protein as detected by ^{1}H NMR spectroscopy can be used to probe protein structural stability (e.g., the presence of stable protein structures) of proteins in solution. Based on experimental data obtained on well-structured proteins and proteins that exist in a molten globule state or a partially folded α-helical state, a well-defined threshold exists that can be used as a quantitative benchmark for protein structural stability (e.g., foldedness) in solution. Additionally, in this chapter we describe a largely automated strategy for rapid fold validation and structure-based backbone signal assignment. Our methodology is based on a limited number of NMR experiments (e.g., HNCA and 3D NOESY-HSQC) and performs a Monte Carlo–type optimization. The novel feature of the method is the opportunity to screen for structural fragments (e.g., template scanning). The performance of this new validation tool is demonstrated with applications to a diverse set of proteins.

Introduction

The genome sequencing projects are delivering vast amounts of protein sequences encoding functionally important proteins, which are putative protein therapeutics and/or targets for the pharmaceutical industry. The concept of ''structural proteomics'' or ''structural genomics'' [e.g., the elucidation of the three-dimensional (3D) structures of the encoded proteins] is based on the empirical finding that protein function cannot always be deduced from the primary sequence but is coded in its 3D shape (Jones and Thornton, 1997; Kasuya and Thornton, 1999; Russell, 1998; Russel *et al*., 1998; Thornton *et al*., 1991). Beyond that, structural proteomics efforts will also enlarge the database of known protein structures and provide a sufficiently large basis set of structures to allow for an efficient

0076-6879/05 $35.00

determination of structure based on homology modeling techniques (Karplus *et al.*, 1999; Koppensteiner *et al.*, 2000; Ota *et al.*, 1999; Sander and Schneider, 1991; Sippl and Weitckus, 1992). To date protein structures are determined either by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. One important issue in large-scale structural proteomics is target selection or the identification of promising proteins suitable for determination of structure.

The ^{1}H NMR chemical shifts are governed by the details of the 3D solution structures of proteins. Although an enormous amount of ^{1}H experimental data exist that underscores this relationship (Seavey *et al.*, 1991), ^{1}H chemical shift information was mainly used as a prerequisite for assignment of relevant structural constraints [e.g., distance-dependent nuclear Overhauser enhancement (NOE) or dihedral angle constraints] (Wüthrich, 1986). Here we use the statistics of the ^{1}H chemical shift distribution to probe protein structural stability in solution. The method uses the autocorrelation function of the ^{1}H spectra of proteins, which are easily obtained and do not require isotope labeling of the proteins. We demonstrate that a significant correlation exists between the autocorrelation function, the topological complexity (expressed as the relative contact order), and protein structural stability of the protein. Data obtained on a diverse set of folded proteins with native structures and partially folded proteins [e.g., the molten globule state of α-lactalbumin (Kuwajima, 1996) and the partially folded oncogenic transcription factor v-Myc] (Fieber *et al.*, 2001) demonstrate that the method can be used to efficiently screen for protein structural stability in a high-throughput manner, with possible beneficial applications to large-scale structural genomics efforts (Kim, 1998) currently underway in the United States (Terwilliger, 2000), Europe (Heinemann, 2000), and Japan (Yokoyama *et al.*, 2000). Additionally, the statistical significance of the observed empirical correlation can also be used to study tertiary structural features of proteins without the need of tedious ^{1}H signal assignment. As a first example, the Ca^{2+}-induced fold stabilization in α-lactalbumin is discussed. The sensitive dependence of ^{1}H chemical shift distribution in the low contact order regime (e.g., partially folded and/or unfolded proteins) also suggests fruitful applications to protein folding studies.

NMR spectroscopy continues to make significant contributions in the challenging area of structural genomics (Prestegard *et al.*, 2001; Staunton *et al.*, 2003), and even high-throughput applications are becoming feasible. This growing impact is due to recent advances in protein preparations, spectrometer hardware, data analysis, and pulse sequence developments. One of the most time-consuming bottlenecks in the process of structure elucidation by NMR is the signal assignment of backbone and side chain

^{1}H, ^{13}C, and ^{15}N resonances, which is a prerequisite for the subsequent gathering of information about protein structure, dynamics, and intermolecular interactions from NMR spectra. Ongoing progress in the development of more powerful spectrometer equipment and pulse sequences has been accompanied by increasing efforts to partly or fully automate the signal assignment procedure. In recent years, numerous research groups reported the development of assignment programs or software packages. A detailed description of these methods is beyond the scope of this chapter. Instead, we refer to an exhaustive review by Gronwald and Kalbitzer (2004) and references therein. The signal assignment process can be subdivided into several steps: (1) grouping of resonances from one or more spectra to spin systems, (2) association of spin systems with amino acid types, (3) linking of spin systems to shorter or longer fragments, and (4) mapping of fragments to the primary sequence. Although some of the reported programs concentrate on one of these steps, others tackle several steps at once. The underlying tools and procedures to accomplish these tasks include tree search algorithms, best-first deterministic approaches, exhaustive searches, genetic algorithms, threshold accepting, Monte Carlo simulations coupled with energy minimization algorithms, neural networks, and others. Most of the programs, in particular those for assignment of larger proteins, rely on a specific set of (numerous) two-dimensional (2D) and 3D NMR spectra or a considerable minimum amount of NMR data to produce reliable results. Therefore, assignment programs are often quite demanding in terms of spectrometer time necessary to acquire sufficient input data. In addition, existing assignment programs are sensitive to missing or incorrect input data (resulting from signal overlap, relaxation processes, noise, and artifacts) and fail to find the correct assignment under nonideal conditions.

In this chapter, we present a new tool for structure-based signal assignment and protein fold validation. It requires minimal NMR data input and the existence of a structure homologue. Although it is reminiscent of existing NMR software packages (Hitchens *et al.*, 2003), it is novel as it also allows for screening for structural fragments (e.g., template scanning).

Materials and Methods

Fold Stability Analysis

The following protein samples were used in this study: α-lactalbumin (Acharya *et al.*, 1991), lysozyme (Diamond, 1974), MutS (Tollinger *et al.*, 1998), creatine kinase (Rao *et al.*, 1998), ubiquitin (Vijay-Kumar *et al.*, 1987), bovine pancreatic trypsin inhibitor (BPTI) (Parkin *et al.*, 1996), myoglobin

(Maurus *et al.*, 1998), v-Myc (Fieber *et al.*, 2001), and bovine serum albumin (BSA) (Janatova *et al.*, 1968). Lysozyme, α-lactalbumin, creatine kinase, BPTI, myoglobin, and BSA were purchased from Sigma (St. Louis, MO) and used without further purification, while v-Myc (Fieber *et al.*, 2001) and MutS (Tollinger *et al.*, 1998) were prepared as described previously. Ca^{2+}-depleted α-lactalbumin was prepared by overnight dialysis using a buffer at pH 1.5 and subsequently refolded by adjusting the pH to 6.5. The molten globule state of α-lactalbumin was prepared by adjusting the pH of the protein solution to 2.5. The pH of the protein solution was carefully controlled with a pH meter. All NMR experiments were performed on a Varian UNITY*Plus* 500-MHz spectrometer equipped with a pulse field gradient unit and a triple resonance probe with actively shielded z gradients. All spectra were recorded at 26°. Water suppression was achieved with a presaturation and WATERGATE detection scheme. For the ^{13}C,^{15}N-labeled proteins ubiquitin and MutS the first trace [omitting the nuclear Overhauser enhancement spectroscopy (NOESY) mixing period and with ^{13}C, ^{15}N-decoupling during acquisition] of a ^{13}C,^{15}N-NOESY-hetero nuclear single-quantum correlation (HSQC) (Pascal *et al.*, 1994) spectrum was used.

In contrast to the previously published application of random matrix theory to the statistical analysis of protein ^1H chemical shifts (Lacelle, 1984), the applied statistical analysis used the autocorrelation function of protein one-dimensional (1D) ^1H spectra. NMR spectra were processed and analyzed using NMRPipe (Delaglio *et al.*, 1995) software. Acquisition parameters were as follows: spectral width, 12,000 Hz; number of spectral points, 11,392 for 1D ^1H spectra and 1536 for spectra acquired using a ^{13}C,^{15}N-NOESY-HSQC, respectively; zero filling, 24 K. However, we have demonstrated that the exact number of spectral points does not influence the outcome of the statistical analysis (data not shown). Residual water was eliminated by deleting the spectral region 4.90–4.55 ppm. To eliminate possible errors introduced by the elimination of the spectral region around the water resonance, the autocorrelation function was calculated for several spectra in which different spectral regions (around the water resonance frequency) were eliminated. No changes in the autocorrelation function $C(\omega)$ were observed. Intensities were extracted from the 1D ^1H spectra with a perl script using the function nLinLS provided with NMRPipe (Delaglio *et al.*, 1995) (and calculated as integrals over 10 data points). The ^1H spectrum for the theoretical random coil peptide was calculated using the sequence of α-lactalbumin and the published random coil shifts for short peptides (Wishart *et al.*, 1995). Shifts for each proton were additionally randomized within ± 0.02 ppm. The resulting data files were used to calculate the autocorrelation functions. The obtained autocorrelation functions,

$C(\omega)$, were normalized to the value at the smallest available frequency difference (0.01 ppm). The raw data were incorporated into the program package xmgr and numerically averaged (averaging window, 50 data points). The values of the autocorrelation function at frequency 0.5 ppm, $C(0.5)$ were used as measures of protein structural stability.

Contact orders were determined from structural coordinates in the Protein Data Bank (Berman *et al.*, 2000). Relative contact orders were calculated according to the published procedure by Baker and co-workers (Plaxco *et al.*, 1998) (see also http://depts.washington.edu/bakerpg). The contact order for the partially folded v-Myc protein was calculated based on the solution structure, which revealed an α-helical conformation for the leucine zipper region comprising residues 384–411 (Fieber *et al.*, 2001). The unfolded segments of v-Myc were taken as random coils and thus neglected for the calculation of the relative contact order.

Fold Validation

Cross-peaks are automatically picked in the HNCA and ^{15}N-NOESY-HSQC (Cavanagh *et al.*, 1996) spectra employing Nmrview software (Johnson and Blevins, 1994). The peak picking in the ^{15}N-NOESY-HSQC is restricted to the H^N–H^N NOEs in the ^{1}H spectral window from 6 to 12 ppm. Artifacts and noise peaks are deleted manually. In addition, an in-house written software tool is used to filter out cross-peaks arising from J-coupled asparagine and glutamine side chain amide resonances.

HNCA cross-peaks are then grouped into individual spin systems (i), with (i) being an arbitrary reference number. Cross-peaks that are separated by less than the digital resolution ($\sim$0.2 ppm in the ^{15}N dimension and less than $\sim$0.02 ppm in the direct ^{1}H dimension) are assumed to belong to the same spin system. The more intense cross-peak is assigned to the C$\alpha(i)$ nucleus, whereas the less intense signal is attributed to the C$\alpha(i-1)$ nucleus. The observation of more than two aligned cross-peaks is indicative for overlapping residues with degenerate ^{1}H and ^{15}N backbone amide frequencies. In these cases C$\alpha(i)$ and C$\alpha(i-1)$ resonances cannot be distinguished and hence these chemical shifts are not included in the input shift table. If only one (^{1}H, ^{15}N, ^{13}C) correlation is observed within boundaries of digital resolution, the C$\alpha(i-1)$ chemical shift is assumed to coincide with the C$\alpha(i)$ chemical shift. The collection of a supplementary HN(CO)CA (Cavanagh *et al.*, 1996) dataset is recommended to obtain a complete and correct input shift list with clear discrimination of C$\alpha(i)$ and C$\alpha(i-1)$ resonances.

The arbitrary reference numbers attributed to spin systems detected in the HNCA experiment are transferred to residues observed in the

^{15}N-NOESY-HSQC spectrum and a list including all potential H^N–H^N NOEs is generated. The identification of the dipolar coupling partner of a specific H^N–H^N NOE (preliminary assignment in F2) is achieved in the following fully automated manner: (1) It is checked for each individual H^N–H^N NOE arising from dipolar interaction between residues *i* and *j* and with chemical shift coordinates ($^1H_i/^1H_j/^{15}N_i$) whether a symmetric NOE exists at the position ($^1H_j \pm 0.03/^1H_i \pm 0.03/^{15}N_j$). If only one symmetric NOE partner is found in the NOE list, the residue *j* is in all likelihood the dipolar coupling partner of residue *i*. (2) If multiple symmetric NOEs are found, no clear assignment is feasible in the indirect dimension F2; all residues giving rise to the symmetric NOE represent potential dipolar coupling partners. The intensity of the symmetric NOE (*j*/*i*) with respect to the NOE (*i*/*j*) is neglected in this analysis for the sake of simplicity. The NOE (*i*/*j*) is then duplicated according to the number of potential coupling partners *j*, whose preliminary reference numbers are assigned to the F2 dimension of each of these duplicated NOEs. Although, with this procedure, wrong NOEs are included in the NOE input list, at least one of the ''cloned'' NOEs will have the correct assignment. (3) If for a specific NOE at the position ($^1H_i/^1H_j/^{15}N_i$) no symmetric NOE ($^1H_j/^1H_i/^{15}N_j$) is found, potential coupling partners can be unraveled by inspecting the $^1H^N$ chemical shifts of all experimentally observed amino acids. The NOE (*i*/*j*) is again multiplied in the NOE input list according to the number of residues *j* having an H^N chemical shift of $^1H_j \pm 0.03$ ppm and the reference numbers of these residues are assigned to the F2 dimension of the NOE (*i*/*j*). (4) If no potential dipolar coupling partner is found in steps (1) to (3) (i.e., NOE between a backbone amide proton and an aromatic side chain proton), the NOE (*i*/*j*) is omitted from the H^N–H^N NOE input list.

As a result of this procedure, two input files are obtained. The first one contains all experimentally observed residues with their arbitrary reference numbers as well as their backbone C$\alpha(i)$ and C$\alpha(i-1)$ chemical shifts. The second input file lists all potential H^N–H^N NOEs that are observed among the query protein residues.

Prediction of Query Protein Chemical Shifts and NOEs

Chemical shifts of the homology model were either obtained by taking chemical shift values deposited in the BMRB database (Doreleijers *et al.*, 2003) or by shift prediction employing ShiftX software (Neal *et al.*, 2003) (see Table II). Homology model secondary chemical shifts are calculated by subtracting random coil shifts from H, N, and Cα chemical shifts. These secondary shifts are subsequently added sequence specifically to the query protein random coil shifts according to the sequence alignment between

homology model and query protein. This yields chemical shift predictions for the query protein. Query protein H^N–H^N NOEs are predicted by computing all pairs of backbone amide protons with distances shorter than 5 Å from the homology model atom coordinates.

Monte Carlo Simulation

The Monte Carlo simulation (Metropolis *et al.*, 1953) attempts to find the best global mapping of experimentally observed spin systems onto the query protein primary sequence. A start configuration is generated by randomly assigning experimentally observed residues to residue positions in the primary sequence. The program is able to handle unoccupied sequence positions that occur when the number of experimentally observed residues is smaller than the total number of query protein residues. Multiple random changes are generated by choosing two query protein sequence positions A and B and by exchanging the experimentally observed residues characterized by their chemical shift values and H^N–H^N NOEs between both positions. After each Monte Carlo step the objective function E (analog of energy) is evaluated with respect to its value before the rearrangement. The random change proposed to the system is accepted or rejected according to the Metropolis criterion, i.e., if $E_2 \leq E_1$, the step is necessarily accepted; if $E_2 > E_1$, the step is accepted with the probability of $p = \exp(E_1 - E_2)/kT$, with $k = 1$. The start temperature T is set to a value that is considerably larger than the largest ΔE normally encountered. The temperature is held constant for several thousand Monte Carlo steps and is then lowered in multiplicative steps, each amounting to a 1–5% decrease in T with respect to the previous temperature value. When T has reached a value where further efforts to reduce the objective function E become inconclusive, the first cycle of the Monte Carlo simulation is stopped. The uniqueness of assignment is assessed by running 10–20 independent Monte Carlo assignment cycles.

In our approach the objective function E is defined as $E = -\log P$, with P being an overall probability scoring value:

$$P = \mathrm{TAN} \cdot \exp(-f_1 \cdot \phi\mathrm{RMSD}_1) \cdot \exp(-f_2 \cdot \phi\mathrm{RMSD}_2) \cdot \mathrm{CA} \qquad (1)$$

The Tanimoto coefficient TAN is a measure of the number of experimentally observed NOEs that coincide with predicted NOEs for a given tentative assignment. It is defined as $c \cdot w/(a + b - c)$, where a and b are the number of experimentally observed and predicted NOEs, respectively, and c is the number of matching NOEs in both input lists A and B. The weighing factor $w = b/a$ ensures that TAN can reach its maximum value of 1 even if a does not equal b. Although the Tanimoto coefficient

forces the system into configurations with a maximum number of coinciding experimental and predicted H^N–H^N NOEs, the second term $\exp(-f_1 \cdot \phi\text{RMSD}_1)$ with

$$\phi\text{RMSD}_1 = \left\{\sum(\Delta C\alpha^2_{l,k} + \Delta C\alpha^2_{i-1,k-1} + \Delta C\alpha^2_{j,i} + \Delta C\alpha^2_{j-1,l-1})\right\}^{1/2}/c \quad (2)$$

ensures that the average root mean square deviation (RMSD) of the four query protein Cα chemical shifts $C\alpha_i$, $C\alpha_{i-1}$, $C\alpha_j$, and $C\alpha_{j-1}$ and the corresponding predicted shifts $C\alpha_k$, $C\alpha_{k-1}$, $C\alpha_l$, and $C\alpha_{l-1}$ is minimized for all c coinciding pairs of experimental and predicted NOEs, with i/j and k/l being dipolar-coupled partners in the query protein and homology model, respectively. Query protein residues whose H^N–H^N NOEs do not coincide with predicted NOEs at a specific Monte Carlo step or with no detectable H^N–H^N NOEs do not contribute to the term ϕRMSD_1. To account for these residues, the third factor $\exp(-f_2 \cdot \phi\text{RMSD}_2)$ with

$$\phi\text{RMSD}_2 = \left\{\sum(\Delta C\alpha^2_{m,n} + \Delta C\alpha^2_{m-1,n-1})\right\}^{1/2}/z \quad (3)$$

is introduced into the probability scoring function [Eq. (1)]. This term is a measure for the overall matching of experimental shifts $C\alpha_m$ and $C\alpha_{m-1}$ with the corresponding predicted shifts $C\alpha_n$ and $C\alpha_{n-1}$ in a specific configuration. The expression ϕRMSD_2 forces the z experimentally observed spin systems to move toward configurations with an overall good match of experimental and predicted Cα chemical shifts. The factors f_1 and f_2 in Eq. (1) represent empirically determined weighing factors and were set to $f_1 = f_2$ in our test runs. The factor CA in Eq. (1) represents the percentage of residues whose $C\alpha_{i-1}$ chemical shifts match the $C\alpha_l$ shifts of the predecessor within a user-defined tolerance value (0.15 ppm). If the total number of experimentally observed residues is smaller than the number of residues in the query protein sequence, a constant number of residue positions remain ''unoccupied'' throughout the Monte Carlo simulation. Spin systems with no predecessor or successor are treated as adjacent residues with matching sequential Cα chemical shifts.

Our assignment and structure validation software is written in the programming language ''C.'' The program is streamlined with respect to CPU time requirements. This is achieved by avoiding noninteger arithmetic and by outsourcing the most time-consuming computational steps from the actual Monte Carlo/simulated annealing procedure. To this end, look-up values contributing to the ϕRMSD_1, ϕRMSD_2, and CA terms of the probability scoring function [Eq. (1)] are computed in advance for all combinations of scalar- and/or dipolar-coupled residue pairs that may be encountered at any of the query protein sequence positions in the course of

the subsequent Monte Carlo run. In addition to these measures, after each Monte Carlo step, the scoring function is not evaluated from scratch, i.e., by summing up the contributions of all residues in that particular tentative assignment. Instead, only changes induced by those residues subjected to the random change are calculated. Our test calculations were executed on a Pentium-grade Linux PC performing ~30,000 Monte Carlo steps per second. The required CPU time for 20 independent assignment runs ranges from ~30 min for medium sized proteins (~150 residues) to 2 h for MBP.

Results

Fold Stability

An outline of the method for rapid assessment of protein stability is illustrated in Fig. 1. The starting point is a conventional protein ^{1}H 1D spectrum in which the residual water is eliminated by simply deleting the spectral region 4.90–4.55 ppm. This data file is used to calculate the autocorrelation function $C(\omega)$. The autocorrelation function is the Fourier transform (FT) of the product between the free induction decay (FID) and its complex conjugate and is thus related to the distribution function of the frequency and relaxation rate differences, respectively. It is important to realize that the lack of specific long-range contacts in unfolded states compared to well-structured proteins leads to a significant narrowing of the distribution function. The obtained autocorrelation functions $C(\omega)$ are normalized to the value at the smallest available frequency difference (0.01 ppm) and numerically smoothed. In Fig. 2 typical (smoothed) autocorrelation functions $C(\omega)$ are shown. From inspection of Fig. 2, it is obvious that there is a clear distinction between a well-folded protein with pronounced structural stability and partially folded or unfolded states. The α-lactalbumin molten globule state at pH 2.5 was chosen as an example of a partially folded state (Fig. 2, blue line), and the dashed black line indicates a theoretical autocorrelation function $C(\omega)$ assuming ^{1}H random coil shifts for the protons of α-lactalbumin. We have found that the primary sequence of the protein does not significantly influence $C(\omega)$; thus the random coil data presented in Fig. 2 can be regarded as representative for a completely unfolded protein in solution. It is evident from Fig. 2 that partially folded proteins as evidenced by the α-lactalbumin molten globule at pH 2.5 or v-Myc (Fieber *et al.*, 2001) are remarkably different from native proteins and display a significant reduction of the autocorrelation function $C(\omega)$.

A closer inspection of Fig. 2 reveals that although the overall appearances of the various $C(\omega)$ for folded proteins (Fig. 2, black, red, and green

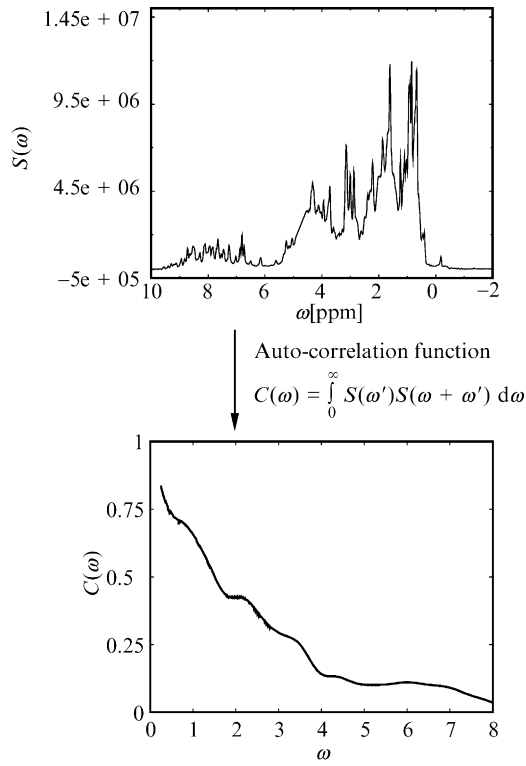


FIG. 1. Outline of the statistical analysis. The starting point of the method is the experimental protein 1D ^{1}H spectra (consisting typically of 2400 data points or 0.01 spectral resolution), in which a residual water signal is eliminated by discarding the spectral region between 4.55 and 4.90 ppm. The obtained autocorrelation function, $C(\omega)$, is normalized to the value at the smallest available energy difference (0.01 ppm), numerically smoothed (typically by averaging over 50 data points). The value of the autocorrelation function at a frequency difference of 0.5 ppm, $C(0.5)$, is taken as a measure for cooperative structural properties of the proteins and can be used as a quantitative measure of protein structural stability.

lines) are remarkably similar, there are noticeable differences. Specifically, the slight additional maxima for $C(\omega)$ at larger frequencies (around 6 ppm) suggest the possibility to extract structural features of proteins from the autocorrelation function, which is reminiscent of CD spectroscopy. It is interesting to note that the α-lactalbumin molten globule (Fig. 2, blue line) displays this slight additional maximum in the autocorrelation function $C(\omega)$, suggesting that the partly folded molten globule state comprises polypeptide fragments with extended chain conformations. We have made

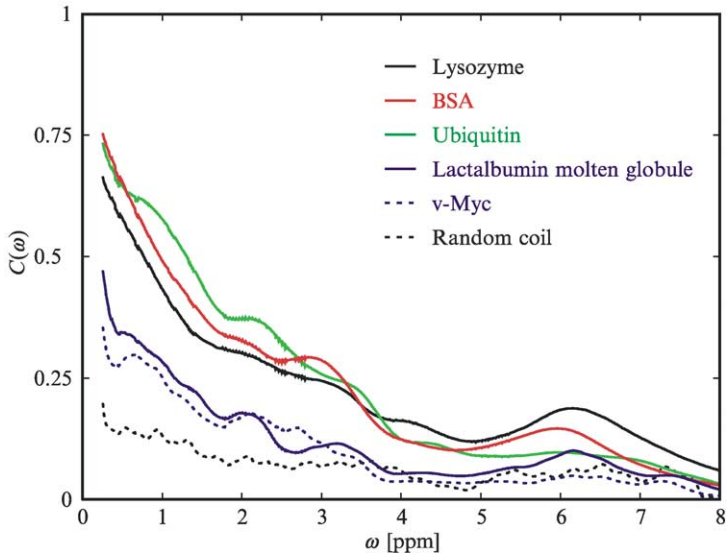


FIG. 2. Autocorrelation functions of protein 1D ^{1}H spectra. The following proteins are shown: lysozyme (black), BSA (red), ubiquitin (green), the molten globule state of α-lactalbumin (blue), the partially folded protein v-Myc (blue, dashed line), and a theoretical random coil polypeptide assuming random coil ^{1}H chemical shifts (black, dashed line). Only energy difference data with $\Delta\omega > 0.25$ ppm are shown (see text).

similar observations (e.g., additional maxima at larger frequencies) for the β-catenin binding fragment of the T-cell factor-4 (TCF4) for which an extended conformation (in additional to a C-terminal α-helix) was observed in the crystal structure. CD spectroscopy of apo-TCF4, however, indicated a random coil in solution. Preliminary NMR data obtained for ^{13}C,^{15}N-labeled apo-TCF4 also provided evidence for the prevalence of extended local structure elements in solution (data not shown). It thus may be feasible to study partially folded protein states by means of the proposed autocorrelation function analysis. For a completely unfolded state, however (Fig. 2, dashed black line), no additional maxima are observed, which again results from the significantly reduced dispersion observed in 1D ^{1}H spectra of unfolded proteins (Wishart *et al.*, 1995).

To derive an unbiased measure for protein structural stability in solution and given the fact that the autocorrelation function is unknown, we propose the autocorrelation function value $C(\omega)$ at 0.5 ppm, $C(0.5)$, as a benchmark of fold stability. Although we have also tested alternative measures such as information theory, methods of moments, and nonlinear curve-fitting, we prefer to use the $C(0.5)$ value, partly because the $C(0.5)$

value can be related to the heterogeneity of the individual protein ^{1}H resonances. Table I lists $C(0.5)$ values obtained for the various proteins. It can be seen that there is a significant difference between proteins that exhibit a well-defined solution structure (e.g., lysozyme, myoglobin, ubiquitin, creatine kinase, MutS) and proteins that exist in partly folded states (e.g., the α-lactalbumin molten globule at pH 2.5, the oncogenic transcription factor v-Myc). Whereas natively folded proteins display $C(0.5)$ values >0.5, partially folded or unfolded proteins have values of <0.4. A $C(0.5)$ threshold value of 0.4–0.5 thus significantly discriminates between these two regimes.

We then explored whether there is a quantifiable relationship between the native state topology of a protein and the statistics of the ^{1}H chemical shift distribution obtained from the autocorrelation analysis of protein 1D ^{1}H spectra. The topological complexity was specified numerically according to a procedure proposed by Plaxco *et al.* (1998). We have used the relative contact order, which reflects the relative importance of local and nonlocal residue contacts to the global fold of a protein. The relative contact order, CO, can be interpreted as the average primary sequence

TABLE I
STATISTICAL ANALYSIS OF PROTEIN ^{1}H CHEMICAL SHIFT DISTRIBUTIONS[a]

Protein (PDB code)	$C(0.5)$	CO (%)
Lysozyme (6LYZ)	0.58	11.1
Creatine kinase (2CRK)	0.54	7.5
α-Lactalbumin (1A4V)	0.62	9.7
BPTI (1BPI)	0.60	15.9
Myoglobin (1AZI)	0.72	7.9
MutS (1BE1)	0.57	9.1
Ubiquitin (1UBQ)	0.65	14.9
BSA	0.66	—[b]
v-Myc	0.28	2.0[c]
α-Lactalbumin molten globule	0.34	—[b]
Ubiquitin in 10 *M* urea	0.32	—[b]
Random coil	0.14	—[b]

[a] The decay of the autocorrelation function $C(\omega)$ is described by its value at a frequency difference of 0.5 ppm (see Materials and Methods). The relative contact order (Plaxco *et al.*, 1998) (CO) is taken as a measure of the topological complexity of proteins.
[b] No structure/contact order available.
[c] Calculated based on the solution structure of v-Myc (Fieber *et al.*, 2001). Only the C-terminal α-helix comprising residues 384–411 were considered.

distance between all pairs of contacting residues along the polypeptide chain (normalized by the total number of residues in the protein). Figure 3 shows the relationship between the topological complexity of the proteins (described by the relative contact order) and the ^{1}H chemical shift distribution described by the $C(0.5)$ value. For example, proteins with compact 3D structures (large relative contact order) display $C(0.5)$ values between about 0.54 and 0.72, respectively. From Fig. 3 it can be seen that the $C(0.5)$ values for the various natively folded proteins are below 0.75, the only exception being myoglobin, which has an attached heme moiety and is thus different compared to the other unligated proteins. Additionally, in the 1D ^{1}H spectra of myoglobin, signals from the bound heme moiety

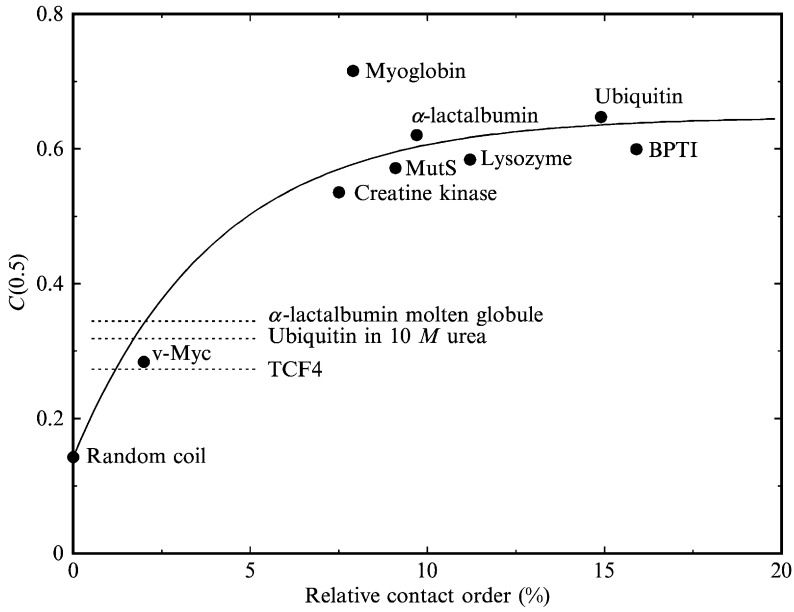


FIG. 3. The relationship between the statistics of protein ^{1}H chemical shifts and protein topology. The autocorrelation function $C(\omega)$ (Fig. 2) was approximated by its value at a frequency of 0.5 ppm, $C(0.5)$. The correlation between the relative contact order of proteins and $C(0.5)$ defines a criterion for the identification of a folded protein. Black symbols depict experimental values; the gray circle denotes the calculated value for a random coil peptide; and the dotted horizontal lines indicate experimental $C(0.5)$ values for the α-lactalbumin molten globule (pH 2.5) and ubiquitin denatured in 10 M urea. For these conformationally flexible proteins no contact order could be calculated. Proteins with $C(0.5) > 0.5$ exhibit a well-defined global fold and exist in a well-structured form in solution. The solid line represents a fit to the experimental data using the analytical function $C(0.5) = A_0 + A_\infty[1 - \exp(-A_1 \cdot \mathrm{CO})]$.

are not suppressed and thus contribute to the observed ^{1}H autocorrelation function. The relationship $C(0.5)$ vs. relative contact order, CO (Fig. 3), presumably reflects the common folding principles of proteins that are based on the chemical similarities of the amino acid building blocks. It also indicates the existence of a threshold value for $C(0.5)$. Most likely, this reflects the upper limit of structural or topological complexity in proteins, which is due to the avoidance of steric clashes of amino acid side chains upon contraction of the polypeptide chain. A detailed understanding of the true relationship between the correlation of energy levels in proteins, as reflected in $C(\omega)$, and the topological complexity of proteins (relative contact order) will provide some (qualitative) insight into the cooperativity of protein structures but is beyond the scope (and not of particular relevance to the proposed applications) of this chapter.

Partially folded proteins or proteins with molten globule-like behavior, however, display significantly smaller $C(0.5)$ values (<0.4). The larger $C(0.5)$ value (0.34) observed for the α-lactalbumin molten globule compared to the partially folded oncogenic transcription factor v-Myc (0.28) suggests more cooperative long-range interactions in this dynamic protein state. Indeed, there is evidence that the molten globule of α-lactalbumin has a native-like overall fold with weak but reasonably well-defined tertiary interactions (Alexandrescu *et al*., 1993; Baum *et al*., 1989; Chakraborty *et al*., 2001; Chyan *et al*., 1993; Dobson, 1994; Peng and Kim, 1994; Peng *et al*., 1995; Redfield *et al*., 1999; Wu *et al*., 1995). In contrast, v-Myc exists as a partially folded protein displaying a well-defined C-terminal α-helix and a ''nascent'' helix in the N-terminal basic domain with no evidence for significant long-range order (Fieber *et al*., 2001).

It is also illuminating to compare our findings on the molten globule state of α-lactalbumin with data obtained using NMR spin diffusion as a probe for protein compactness and residual structure in molten globule states (Griko and Kutyshenko, 1994; Kutyshenko and Cortijo, 2000). The rigidity parameter (G) was introduced as a measure for residual structure in proteins subjected to denaturing conditions, such as temperature, denaturing agents, and changes in pH. G is defined as the intensity ratio between conventional 1D ^{1}H spectra and spin diffusion spectra for certain spectral regions of proteins (e.g., amide, aromatic, and or aliphatic). G values of $\sim$0.1 were obtained for denatured (unfolded) proteins, whereas values of $\sim$0.5 have been found for native and, surprisingly, molten globule states (Griko and Kutyshenko, 1994; Kutyshenko and Cortijo, 2000). This was suggestive of the existence of native-like tertiary structures in molten globules. NOEs and other data also supported the notion that molten globules exist in significantly compact structural ensembles (Balbach *et al*., 1997; Choy *et al*., 2001). However, the fact that a native-like spectral

appearance is observed does not imply that a molten globule exists as a compact, impermeable sphere (Griko and Kutyshenko, 1994; Kutyshenko and Cortijo, 2000). Our finding that the α-lactalbumin molten globule is significantly less compact and less ordered than well-structured native proteins emphasizes the notion that a molten globule is best described as a native-like but noncooperative assembly of the constituent core regions of the polypeptide chain (Schulman and Kim, 1996; Schulman *et al.*, 1997). The lack of cooperativity in molten globules is observed as a significantly faster decay of the autocorrelation function, described with $C(0.5)$, compared to native proteins (see Fig. 2), which exist as densely packed polypeptide chains of a highly cooperative nature. Interestingly, our findings are also consistent with recent NMR experiments that also demonstrated a noncooperative unfolding of the α-lactalbumin molten globule by probing unfolding events at individual residues (Schulman and Kim, 1996; Schulman *et al.*, 1997). Finally, the dynamic nature of the transiently formed structural ensemble of a molten globule is indicated by effective transverse spin relaxation (e.g., extreme line-broadening due to motional dynamics in the millisecond to microsecond time scale), which typically precludes direct NMR studies of molten globules (Last *et al.*, 2001).

Interestingly, ubiquitin denatured in 10 *M* urea displays a $C(0.5)$ value of 0.32 similar to the values of the molten globule of α-lactalbumin and v-Myc. The observation of small $C(0.5)$ values is consistent with the notion that the auto correlation function predominantly probes cooperative long-range interactions in well-defined protein folds. It also suggests, however, that urea-denatured ubiquitin exhibits some residual structure. Similar observations (e.g., the prevalence of residual structure in denatured proteins) have been made for the denatured forms of 434-repressor (Neri *et al.*, 1992) and the fragment $\Delta 131\Delta$ of staphylococcal nuclease (Shortle and Ackerman, 2001).

Encouraged by the quality of the data, we investigated the possibility of using this analysis to probe protein stability in general and to determine whether the accuracy of the method is sufficiently high to monitor subtle changes of protein structural stability (foldedness) upon, for example, ligand binding. As a first example, we present data obtained on monitoring stability changes of α-lactalbumin upon Ca^{2+} binding. α-Lactalbumin is the regulatory component of the lactose synthase complex that catalyzes the biosynthesis of lactose. It has a bipartite structure and consists of two lobes. The α-domain is composed of four α-helices (and two short 3_{10} helices), whereas the smaller β-domain consists of a triple-stranded antiparallel β-sheet and a 3_{10} helix, linked by a series of loops (Acharya *et al.*, 1991; Calderone *et al.*, 1996; Pike *et al.*, 1996). All known α-lactalbumin crystal structures revealed a conserved Ca^{2+}-binding site, formed by the side chain

β-carboxylate groups of three aspartic acid residues, two backbone carbonyl oxygens, and two bound water molecules (contributing two oxygens to the metal coordination site), which are arranged in a distorted pentagonal bipyramidal coordination sphere (Acharya *et al.*, 1991; Anderson *et al.*, 1997; Calderone *et al.*, 1996; Pike *et al.*, 1996). The apparent K_{Ca} of α-lactalbumin (Wijesinha-Bettoni *et al.*, 2001) is of the order of 10^6–10^7 M^{-1} at physiological pH levels. The impact of Ca^{2+} on α-lactalbumin protein folding has been investigated (Anderson *et al.*, 1997; Troullier *et al.*, 2000; Wijesinha-Bettoni *et al.*, 2001). Upon formation of a loosely defined protein state, Ca^{2+} binding drives the formation of the α-lactalbumin native state, presumably in a cooperative manner (Forge *et al.*, 1999; Kuwajima *et al.*, 1989; Troullier *et al.*, 2000). The structural role of Ca^{2+} and its influence on the stability of α-lactalbumin were also demonstrated by means of hydrogen exchange protection (Wijesinha-Bettoni *et al.*, 2001). It was observed that Ca^{2+} binding stabilizes the structure of native bovine α-lactalbumin; at pH 8 the Ca^{2+}-depleted (apo) form of has a melting point T_m of 34°, compared to 64° for the Ca^{2+}-loaded (holo) form. Although apo α-lactalbumin displays a native-like structure, as inferred from CD, fluorescence, and low-resolution NMR data, and the helical content of apo α-lactalbumin is equal to (or even slightly greater than) holo α-lactalbumin, the hydrogen-exchange results indicated that the Ca^{2+}-binding loop and the C-helix are stabilized in the holo form. Recently, the crystal structure of apo α-lactalbumin was solved, and the X-ray data additionally corroborated the previous finding that Ca^{2+} causes an increase in stability but little structural change (Chrysina *et al.*, 2000; Wijesinha-Bettoni *et al.*, 2001).

The Ca^{2+}-depleted α-lactalbumin was titrated with a concentrated stock solution of $CaCl_2$ until a 10-fold molar excess of Ca^{2+} over α-lactalbumin was reached. Each solution of varying Ca^{2+}/α-lactalbumin concentration ratio was subjected to the analysis of the ^1H chemical shift distribution. The titration curve that is obtained is shown in Fig. 4. It can be seen that the elimination of Ca^{2+} resulted in a significant reduction in the $C(0.5)$ value. The addition of Ca^{2+} leads to an increase in the $C(0.5)$ value. Given the structural similarities of apo and holo α-lactalbumin, the significant increase of $C(0.5)$ reflects the increased protein stability of the Ca^{2+}-loaded form compared to the Ca^{2+}-depleted form of α-lactalbumin. It should be noted that this change in protein stability is not obvious from a simple inspection of the 1D ^1H spectra. The dashed line in Fig. 4 was calculated using the well-known Ca^{2+} association constant of α-lactalbumin, $K_{Ca} = 10^6$ M^{-1} (literature value of $K_{Ca} = 10^6$–10^7 M^{-1} at physiological pH) (Wijesinha-Bettoni *et al.*, 2001). The agreement between the theoretical curve [by using the experimentally obtained $C(0.5)$ values for the Ca^{2+}-depleted and for the Ca^{2+}-saturated form, respectively] and the

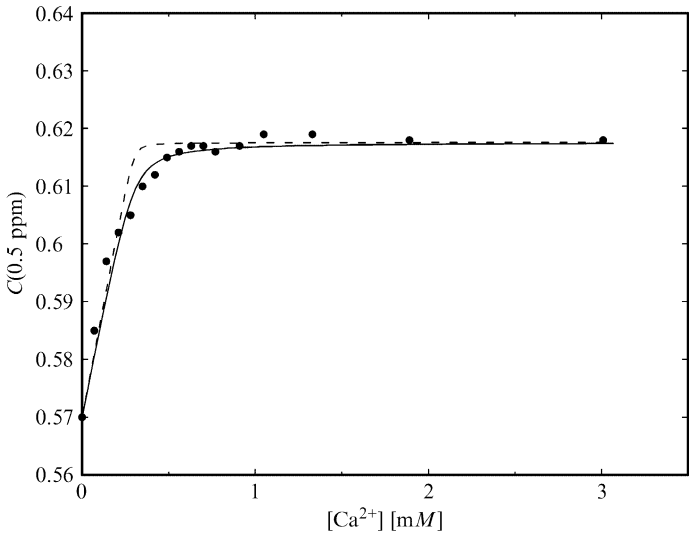


FIG. 4. Ca^{2+}-induced fold stabilization of α-lactalbumin monitored by statistical analysis of ^{1}H chemical shift distribution. The value of the autocorrelation function at 0.5 ppm, $C(0.5)$, was taken as a measure for protein structural stability (e.g., foldedness, see text). The black line represents the titration curve fitted to the experimental values ($K_d = 1.2 \times 10^{-5}$ M), the dashed black line the theoretical titration curve using the published Ca^{2+} dissociation constant K_d of α-lactalbumin (Wijesinha-Bettoni *et al.*, 2001), $K_a = 10^{-6}$ M.

experimental $C(0.5)$ values *vs.* Ca^{2+} concentration is remarkably good. The experimentally determined dissociation constant (1.2×10^{-5} M, Fig. 4 solid line) at pH 6.5 convincingly demonstrates that the proposed method can be used to study subtle changes of protein structural stability caused by binding of metals and/or small-molecular-weight ligands. The small reduction of the observed K_d is presumably due to the lower pH used in the present study.

Fold Validation

We present here a Monte Carlo/Simulated Annealing (MC/SA) program for automated backbone H^N, N, and Cα chemical shift assignment and structure validation. The program requires minimal NMR data input and the existence of a 3D structure of a homology model. Arbitrary reference numbers are attributed to experimentally observed residues in 3D HNCA and 3D ^{15}N-NOESY-HSQC spectra. Thus, each residue of the protein is represented by four resonance frequencies [^{15}N, $^{1}H^N$, $^{13}C\alpha(i)$, and $^{13}C\alpha(i-1)$, respectively]. Input lists containing query protein

$C\alpha(i)$ and $C\alpha(i-1)$ chemical shifts of these residues as well as backbone amide H^N–H^N NOEs are generated. Based on a precise sequence alignment between query protein and homology model, the homology model is used to predict $C\alpha$ chemical shifts and H^N–H^N NOEs for stretches of structurally equivalent query protein residues. Starting from an arbitrary start configuration, the Monte Carlo algorithm picks randomly pairs of sequence positions and swaps the residues tentatively assigned to these positions. An overall scoring value is computed after each of the multiple Monte Carlo steps to determine whether the proposed random change is a step toward the correct assignment or not. The random changes proposed to the system are accepted or rejected according to the Metropolis criterion. The Monte Carlo algorithm is coupled with a Simulated Annealing protocol that forces the system into a low-energy configuration in which experimentally observed and predicted NMR shifts and NOEs match best. An in-depth description of the objective function is provided in Materials and Methods. The objective function includes mathematical terms accounting for matching of experimentally observed and predicted NMR parameters such as $C\alpha$ shifts and H^N–H^N NOEs as well as sequential $C\alpha(i)/C\alpha(i-1)$ shift matching along the query protein sequence.

We have checked the performance of our Monte Carlo–based assignment and structure validation program with five query proteins (calmodulin, 150 residues; MBP, 370 residues; Q83, 150 residues; ICIn, 168 residues; and CypD, 165 residues) differing in size and tertiary structure in a series of 15 test runs (Table II, Fig. 5). Whereas calmodulin is a purely α-helical protein, both Q83 and ICln feature β-barrel structures surrounded by a varying number of helices. MBP is a large two-domain protein with each domain being made up of numerous strands and helices. The CypD structure is so far unknown. However, CypD shares a high degree of sequence similarity with its homologue CypA, which is made up of an eight-stranded barrel surrounded by three helices and various extended loop segments.

In the case of MBP and calmodulin, we have chosen their X-ray structures as a homology model. Under these idealized conditions (1) the homology model sequence covers the entire query protein sequence, (2) shift and NOE predictions are available for all query protein residue positions, and (3) sets of ''experimental'' and predicted H^N–H^N NOEs, which were both calculated from the MBP and calmodulin atom coordinates assuming a 5 Å distance cutoff, are identical for NMR-observable residues with available chemical shifts in the BMRB database. Atom coordinate–derived ''experimental'' NOEs for NMR-unobservable residues (i.e., with no chemical shifts reported in the BMRB data bank) were omitted from the input file. Therefore, the input list of ''experimental''

TABLE II
DATA INPUT AND ASSIGNMENT ACCURACY OF 15 MONTE CARLO–BASED ASSIGNMENT TEST RUNS[a]

Test run	Query protein (QP)/homology model (HM)	Non-Pro residues in QP sequence (*n*)	Non-Pro residues in alignment QP/HM (*n*)	Experimentally observed QP residues within alignment (*n*)	HM-based NOE predictions (+sequential NOEs) (*n*)	Experimentally observed QP NOEs (*n*)	Correct/erroneous assignment (*n*) (% correct assignments)
1	Calmodulin/ calmodulin	146	146	139 BMRB 4284	323 PDB 1CFF	300 PDB 1CFF	145/1(99%)
2	MBP/MBP	349	349	330 BMRB 4354	697 PDB 1EZO	656 PDB 1EZO	334/15 (96%)
3	Q83/NGL	151	102	99 BMRB 4664	203 (+43) PDB 1NGL	243 PDB 1JZU	96/6 (94%)
4	Q83/NGL	151	85	82 BMRB 4664	171 (+53) PDB 1NGL	243 PDB 1JZU	81/4 (95%)
5	Q83/NGL	151	33	33 BMRB 4664	59 (+109) PDB 1NGL	243 PDB 1JZU	33/0 (100%)
6	Q83/NGL	151	35	32 BMRB 4664	68 (+103) PDB 1NGL	243 PDB 1JZU	33/2 (94%)
7	Q83/NGL	151	27	24 BMRB 4664	48 (+109) PDB 1NGL	243 PDB 1JZU	24/3 (89%)
8	Q83/NGL	151	26	26 BMRB 4664	51 (+113) PDB 1NGL	243 PDB 1JZU	23/3 (88%)
9	Q83/NGL	151	26	26 BMRB 4664	41 (+114) PDB 1NGL	243 PDB 1JZU	26/0 (100%)
10	ICln/UNC-89	158	74	73	170 (+87) PDB 1FHO	163[b]	69/5 (93%)
11	ICln/UNC-89	158	38	34	87 (+117) PDB 1FHO	163[b]	34/4 (89%)
12	ICln/UNC-89	158	36	36	67 (+119) PDB 1FHO	163[b]	35/1 (97%)
13	CypD/CypA	158	158	153	321 PDB 1CWB	791[c]	152/1 (99%)
14	CypD/CypA	158	54	54	101 (+104) PDB 1CWB	791[c]	49/5 (91%)
15	CypD/CypA	158	55	55	108 (+101) PDB 1CWB	791[c]	55/0 (100%)

[a] Column 1, number of test run. Column 2, query protein and structure homologue. Column 3, number of nonproline residues in the query protein sequence. Column 4, number of query protein residues that are aligned with and structurally similar to the homology model. Column 5, number of NMR-detectable residues of column 4; if chemical shift information was obtained from the BMRB database, the corresponding query protein entry code is provided. Column 6, number of homology model-based H^N–H^N NOE predictions as derived from homology model atom coordinates; homology model PDB entry codes are provided; the numbers in parentheses refer to additional sequential H^N–H^N NOE predictions that were introduced for query protein residues with no structure similarity to the homology model. Column 7, number of experimentally observed query protein H^N–H^N NOEs; the query protein PDB entry codes is provided, if ''experimental'' NOEs were calculated from query protein atom coordinates. Column 8, number of residues in column 4 that were correctly and erroneously assigned and percentage of correctly assigned residues.

[b] Synthetic H^N–H^N NOEs derived from query protein atom coordinates.

[c] H^N–H^N NOE input list results from a manual signal assignment performed with a ^{15}N-NOESY-HSQC and was generated as outlined in Materials and Methods.

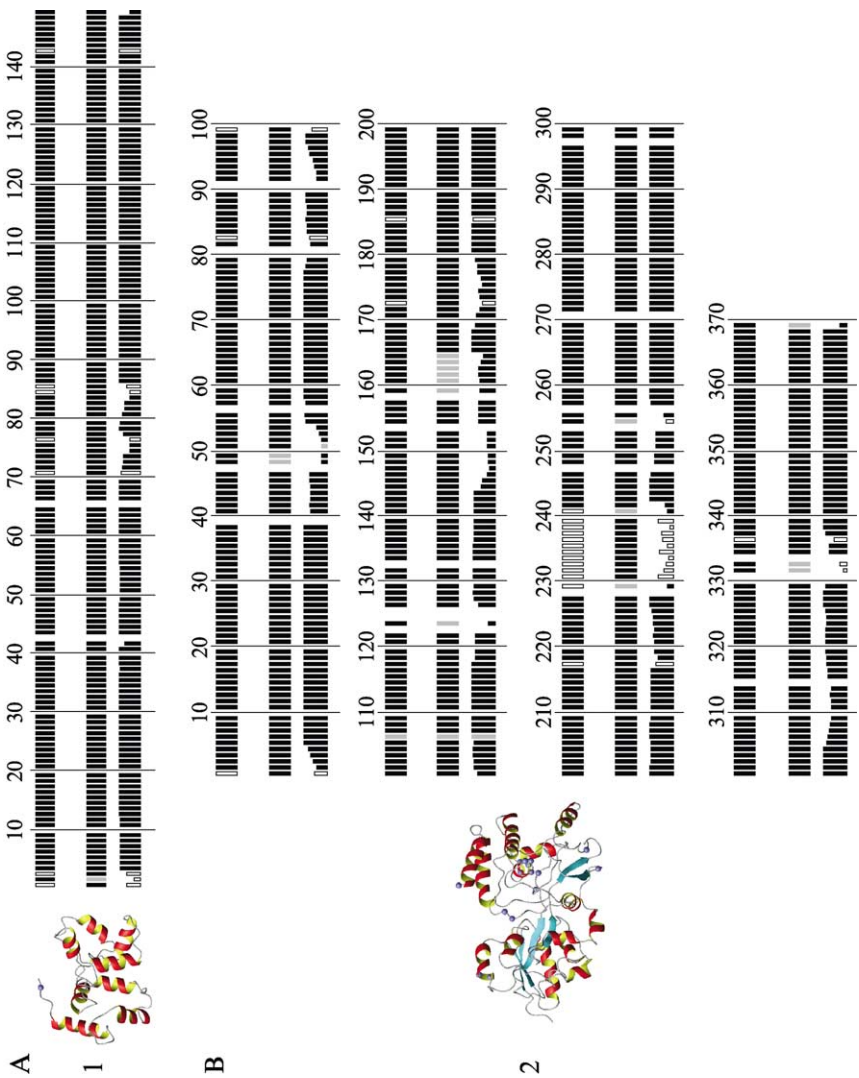
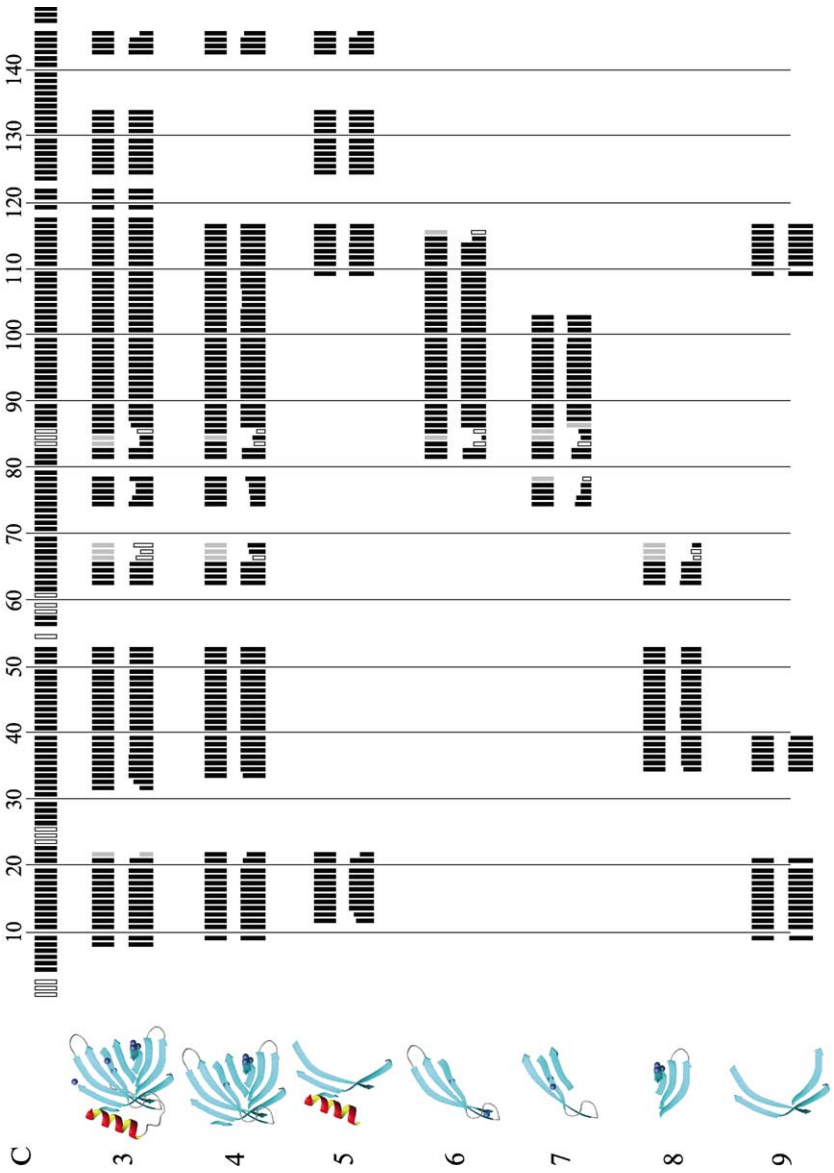


FIG. 5. *(continued)*


10
20
30
40
50
60
70
80
90
100
110
120
130
140
C
3
4
5
6
7
8
9

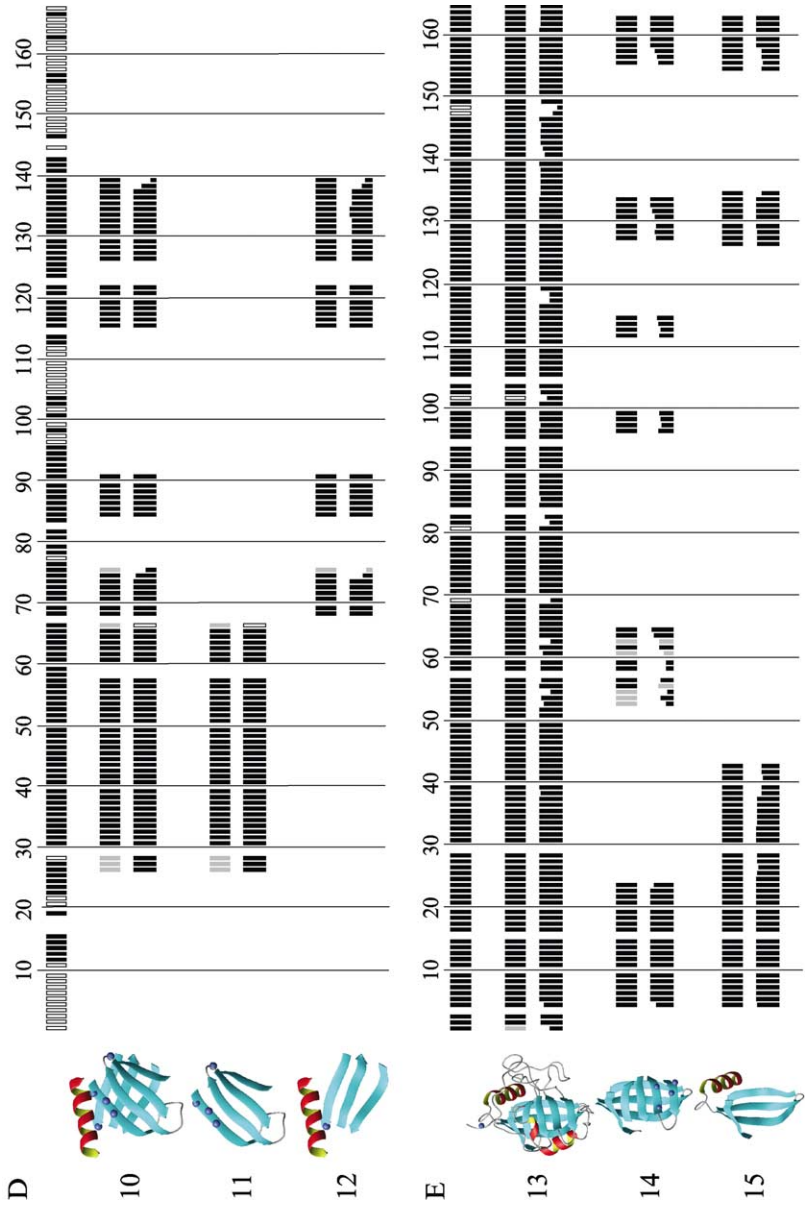


FIG. 5. *(continued)*

NOEs comprises 93% (calmodulin) and 94% (MBP) of the predicted NOEs. Homology model shift estimates were generated by using ShiftX (Neal *et al*., 2003) software and thus differ from experimental chemical shifts obtained from the BMRB data bank.

In the case of the query proteins ICln and Q83, test conditions were much closer to real situations. ICln and Q83 structure homologues were identified through knowledge-based potential methodology (Domingues *et al*., 1999; Sippl, 1993, 1995). Segments of the PH domain from the *Caenorhabditis elegans* muscle protein UNC-89 (PDB-ID, 1FHO: BMRB accession number, 4373) and of human neutrophil gelatinase-associated lipocalin NGL (PDB-ID, 1NGL; BMRB accession number, 4267) were revealed to share strong structural similarities with 47% of the ICln and 71% of the Q83 primary sequence, respectively. Therefore, these homology segments were used for shift and NOE predictions for structurally equivalent query protein segments. For the remaining Q83 and ICln residues with no structure similarity with NGL and UNC-89, respectively, random coil shifts and sequential H^N–H^N NOEs were added to the input lists. A synthetic set of Q83 H^N–H^N NOEs computed from Q83 atom coordinates represents again a complete and unambiguous input list. The ratio of Q83 ''experimental'' NOEs and predicted NOEs based on the structural model (1NGL) is close to 1. In contrast to Q83, the list of experimental ICln NOEs was obtained from a manually picked and edited ^{15}N-NOESY-HSQC spectrum of the protein. This results in a decreased ratio in experimental to predicted NOEs of ~0.7, since complete observation of dipolar-coupled

FIG. 5. Graphic representation of assignment results obtained with our Monte Carlo–based approach for calmodulin (A), maltose-binding protein (B), Q83 (C), Icln (D), and CypD (E). The upper row in (A–E) represents the entire query protein primary sequence with each bar symbolizing one residue. Missing bars indicate residue positions occupied by prolines. NMR-detectable residues are shown as black bars and NMR-unobservable residues as unfilled black bars. Each individual test run is numbered as in Table I. The result of each test run is summarized by two rows of bars. Those query protein residues that are not part of homology segments (which have no structural equivalent in the homology model) as well as prolines were omitted from both rows. Upper row: correct assignment, black; erroneous assignment, gray. Lower row: residue with $C\alpha(i-1)$ chemical shift matching/mismatching with $C\alpha(i)$ chemical shifts of its predecessor, black/gray, respectively. Filled bars represent NMR-observable residues; unfilled bars indicate NMR-unobservable residues. Note that no statement about interresidue $C\alpha$ chemical shift matching/mismatching can be made for residues adjacent (C-terminal) to NMR-unobservable residues. Maximal bar height symbolizes 100% assignment reproducibility after 20 independent assignment cycles. Reduced reproducibility is accordingly indicated by lower bar heights. The ribbon drawings display those parts of the query proteins that are made of the residues shown in the upper and lower rows. The positions of erroneously assigned residues are shown as small spheres. All ribbon drawings were generated with MOLMOL (Koradi *et al*., 1996).

backbone amide protons is hampered by shift degeneracies and relaxation processes.

Cyclophilin P chemical shift and NOE input lists originated exclusively from NMR spectra and, therefore, test conditions were most demanding in calculations performed with these data sets. CypD shares ~90% sequence identity with human cyclophilin A (PDB ID, 1cwb, X-ray structure, complexed with cyclosporin; BMRB entry code, 2208). However, no structural information is yet available for CypD. Input lists containing experimental as well as predicted Cαchemical shifts and H^N–H^N NOEs were generated as described in Material and Methods. A total of 412 H^N–H^N NOEs were peak picked in the ^{15}N-NOESY-HSQC spectrum of CypD in the spectral window from 6 to 12 ppm. For 183 NOEs, a single dipolar coupling partner was identified. For the remaining 229 NOEs either more than one or no symmetric cross-peak was detected. By taking into consideration all potential dipolar coupling partners, these 229 NOEs were duplicated to a total of 776 NOEs (see Materials and Methods). Under the assumption that the correct dipolar coupling partner is assigned to at least one of the duplicated NOEs, this resulted in adding 547 wrong NOEs to the input file list. Redundant NOEs (*i*/*j j*/*i*) were filtered out, leaving 791 experimentally observed dipolar coupling interactions in the NOE input file list. Using a 5 Å distance cutoff, 321 H^N–H^N NOEs were predicted for CypD based on the analysis of the atom coordinates of the CypA structure homologue. Thus, the ratio of experimental to predicted NOEs amounts to 2.5 and is considerably higher than for the previously mentioned test query proteins.

In our test examples, 72% (ICln) to 97% (CypD) of the nonproline residues were NMR detectable. The large majority of query residues within homology segments were NMR detectable; most of the undetected residues fall into sequence regions having no sequence alignment with the homology model and for which no homology model–based shift and NOE predictions are available (Fig. 5).

A first series of test runs (Table II and Fig. 5) was performed for all query proteins with the entirety of available input data (test runs 1, 2, 3, 10, and 13). In subsequent test calculations (4–9, 11–12, and 14–15) test conditions were artificially rendered more demanding. The goal of subsequent test runs was to check whether our Monte Carlo–based assignment procedure produces satisfying results if shift and NOE predictions are available only for smaller structural elements, i.e., if the homology model covers only smaller building blocks of the query protein. To this end, the homology model–based shift and NOE predictions for the query proteins ICln, Q83, and CypD were deleted for varying homology segments. Predicted NMR data were retained for smaller structural subunits comprising only 17–59% of the query protein primary sequence. Shift and NOE

predictions that were deleted for certain query protein segments were replaced by random coil chemical shifts and sequential H^N–H^N NOEs. This results in a considerable decrease in the total number of homology model–based NOE predictions and in a change in the ratio of experimental to predicted NOEs.

In each test run, the program was allowed to assign all experimentally observed query protein spin systems to all nonproline residues of the query protein primary sequence, whether homology model–derived shift and NOE predictions were available for a specific sequence position or not. To assess the reproducibility of the assignment, each test run was performed 20 times. If the assignment for a specific query protein residue position was ambiguous, i.e., if after 20 assignment cycles more than one spin system was attributed to that position, the spin system that occurred most often was chosen for the final assignment. (This method does not necessarily result in a unique assignment, since a specific spin system might be retrieved at more than one residue position of the query protein.)

In spite of the diversity of test conditions, the assignment accuracy is satisfying in all cases. In the first series of test runs performed with the entirety of available homology model–based shift and NOE predictions, the percentage of correctly assigned residues within homology segments ranges from 93% (ICln) to 99% (calmodulin, CypD). The slightly lowered value of successfully assigned ICln residues in test run 10 may be due to the fact that an increased number of ICln residues (28%) was not detected by NMR and that homology model–derived shift and NOE predictions were available for only 47% of the ICln primary sequence. Surprisingly, our Monte Carlo assignment algorithm performed equally well in the second series of test calculations in which test conditions were rendered more demanding by retaining homology model–based shift and NOE predictions for smaller structural motifs. In these test runs, the percentage of correctly assigned residues ranges from 88% to 100% within these smaller building blocks.

In our test runs 75–100% of residues are assigned with a reproducibility of 75% or higher. The vast majority of these residues are correctly assigned. Based on our results, we can define the rule of thumb that the assignment of a specific residue is correct if it is part of a stretch of four or more consecutive residues that do not have any $C\alpha$ chemical shift mismatches and display an assignment reproducibility of $>75\%$ for each residue. Surprisingly, a clear majority of assignments with considerably higher uncertainties are still correct. Erroneous assignments may become manifested in $C\alpha$ chemical shift mismatches between adjacent residues. In addition, wrong assignments are evident if a certain experimentally observed residue appears (in rare cases) more than once in the final

assignment list as a result of the final selection procedure performed after multiple independent Monte Carlo cycles as described above. (These are residues with an assignment reproducibility <50%.) Certain query protein residue positions appear to be more prone to erroneous assignments than others, and special care should be taken in the evaluation of these positions. Within homology segments, most of the erroneous assignments occur (1) at residue positions adjacent to prolines (e.g., MBP residues Pro-48-49-50), (2) at the N- or C-terminal ends of homology segments (e.g., ICln residues 27–29), (3) at sequence positions whose corresponding residues are not detected (e.g., Q83 residue position 85), or (4) in combinations of these situations.

Discussion

The statistical interpretation of chemical shifts was pioneered in the late 1960s by Schaefer and Yaris (1969) when they demonstrated that the complicated ^{13}C and ^{1}H NMR spectra of the cyclic tetramer of polypropylene oxide can be interpreted by an analysis of the spin hamiltonian in terms of the statistical theory of energy levels. Later, their suggestions were taken up by Lacelle (1984), who applied the approach to a vitamin (vitamin B_{12}), an antibiotic (alamethicin), and a protein (trypsin inhibitor homologue K) and showed that the method indeed provides, at least, a qualitative estimation of the degree of correlation between energy levels via a characterization of the spacing distribution of energy levels.

Here we systematically studied a diverse selection of proteins, comprising pure α-helical as well as α/β proteins, the relationship between the ^{1}H chemical shift distribution and protein structural stability in solution (e.g., foldedness and/or topological complexity of protein). The strategy was initiated by the idea of developing a robust, straightforward method to analyze protein spectra and to investigate the possibility of probing protein structural stability by a general method without the need of time-consuming assignment strategies, as this would be of significance to ongoing large-scale structural genomics efforts devoted to structural characterization of a vast number of proteins. In the analysis of the ^{1}H chemical shift distribution, we calculate the autocorrelation function $C(\omega)$ of the 1D ^{1}H spectra and take the value of $C(\omega)$ at 0.5 ppm as a quantitative benchmark to discriminate between folded, partially folded, and random-coil proteins. We do not attempt to physically interpret this parameter but rather use it as a quantitative means to probe fold stability. The analysis of the protein set convincingly demonstrated that it is indeed possible to probe protein structural stability through this simple analysis of 1D ^{1}H protein NMR spectra. There is a significant difference between proteins

exhibiting a well-defined 3D structure and proteins that appear to be unfolded, partially folded, or that exist in a molten globule state.

The particular merits of the method are the ease of implementation, small amount of material (given the advent of more sensitive NMR detection schemes, e.g., cryoprobes), the high-throughput capability, and the fact that no isotope labeling is necessary. We thus foresee several obvious applications. First, recent genomic sequencing efforts have provided the coding DNA sequences of a large number of unknown genes and structural genomics or structural proteomics (Prestegard *et al*., 2001) attempts to provide 3D structural information of proteins encoded by the sequenced genes. Irrespective of the method of structure determination (X-ray or NMR spectroscopy), NMR is expected to play a significant role in structural genomics activities (Prestegard *et al*., 2001), as, for example, ^{15}N-filtered H/D exchange-based NMR experiments (Prestegard *et al*., 2001) (e.g., the identification of rapidly exchanging amide protons) and simple 1D experiments (Rehm *et al*., 2002) have already been demonstrated to be very effective to screen expressed and purified proteins for stability, structural disorder, and/or sample conditions that are favorable for crystallization. The data presented in this chapter suggest that this spectral autocorrelation method will be very valuable for this purpose, as the method does not require isotope labeling and also provides a means to identify metals and/or small ligands as well as macromolecular interactions that may be relevant for fold stabilization and function.

In contrast to structural genomics efforts that aim at characterizing folded proteins, a recently proposed target selection strategy focuses on unusual and uncharacterized soluble proteins in *Mycoplasma genitalium*, the smallest autonomously replicating organism (Balasubramanian *et al*., 2000). The aim of this approach was to identify proteins that show atypical behavior in terms of structural stability (foldedness), for example, proteins that are ''unstructured'' in the absence of a binding partner or that exhibit unusual thermodynamic properties. In this study, CD spectroscopy was used to probe the integrity of folding and to investigate the thermodynamic stability. As an alternative to optical methods, a mass spectrometry–based approach for protein stability screening was recently designed, which can even be extended to *in vivo* studies (Ghaemmaghami and Oas, 2001; Ghaemmaghami *et al*., 2000). With its ease of implementation, numerical analysis, and high-throughput capability, the proposed method should prove to be an additional important element of modern proteomic technology.

Second, the results obtained on the titration of α-lactalbumin with Ca^{2+} show that the proposed method can detect binding through changes of the ^{1}H chemical shift distribution, which in turn reflect protein stability changes. Given the well-established link between thermodynamic protein

stability and ligand binding (Pace and McGrath, 1980), it may also be possible to use the high-throughput capability of the proposed method to screen large ligand libraries (Diercks *et al*., 2001; Moore, 1999). If ^{13}C, ^{15}N-labeling of the protein is available, the method can be applied equally to protein–protein and protein–nucleic acid complexes. In particular, the approach can be applied to identify proteins that are only loosely defined structurally and undergo conformational restructuring or even adopt a well-defined native structure only upon binding to their authentic binding partners (for a review, see Wright and Dyson, 1999), a phenomenon that remarkably and unexpectedly is even more pronounced in higher organisms (Dunker and Obradovic, 2001).

Finally, data obtained on partially folded proteins (the native-like α-lactalbumin molten globule and the partially folded oncogenic transcription factor v-Myc) suggest fruitful applications of the proposed method to studies of molten globules and protein folding (Dolgikh *et al*., 1981; Kuwajima, 1989; Ptitsyn, 1995). For example, site-directed mutagenesis has been successfully applied to obtain a quantitative measurement of the contributions of individual residues to the stability of molten globules (Hughson *et al*., 1991). Additionally, studying the contribution of individual residues to the protein structural stability of molten globules may be valuable for understanding this important protein state.

The tremendous advance in the large-scale gene sequencing of whole genomes poses an enormous challenge to NMR spectroscopy. New integrated approaches are necessary to enable NMR spectroscopy to solve protein structures in a high-throughput manner and thus to keep pace with the generation of huge amounts of sequence information. In this context many research groups, including ours, have focused their efforts on the development of new programs devised to speed up the process of NMR structure elucidation. The program presented here is a powerful new tool for rapid sequence-specific assignment of backbone resonances of uniformly ^{13}C- and ^{15}N-labeled globular proteins and structure validation. Our approach requires minimal NMR data input from two 3D spectra and therefore a reduced amount of spectrometer time. The need for more extensive data collection is circumvented by using chemical shift and NOE predictions derived from a 3D structure of a query protein homologue. Although additional data input is not mandatory for obtaining correct assignments, further chemical shift and interresidue connectivity information can easily be included for Hα, Cβ, and C$'$ nuclei. It is, however, important to note that the performance of our program in its present form depends on the sequence alignment accuracy of structurally equivalent blocks of the query protein and its homologue. We have observed that the assignment accuracy deteriorates in particular if query protein segments

predicted to form a β-sheet structure are inaccurately aligned with the homology model sequence (data not shown). A modification of the current form of the objective function, in particular the replacement of the Tanimoto coefficient by a more sophisticated expression, might help to eliminate the pitfall of improper sequence alignment. As test runs with CypD input data have clearly demonstrated, our program is robust enough to tolerate numerous ambiguous H^N–H^N NOEs resulting from the inability to clearly identify the majority of dipolar-coupled pairs of backbone amide protons on the sole basis of a 3D ^{15}N-NOESY-HSQC spectrum. In addition, even if homology model–based shift and NOE predictions are missing for certain residue stretches, the algorithm is still able to find the correct assignments for the remaining protein segments. The latter feature represents a distinct advantage of the program described in this chapter over existing assignment and structure validation programs and suggests fruitful applications in the scanning of query proteins for the presence of structure templates. Template scanning and motif recognition are useful when complete homology model covering the entire query protein sequence is not available and/or to study protein modules in the context of multidomain proteins.

Conclusions

We have demonstrated that protein structural stability is reflected in the distribution of protein ^{1}H chemical shifts. A method was proposed that does not require isotope labeling but instead uses easily obtainable 1D ^{1}H spectra, from which the spectral autocorrelation function is calculated. The method allows a significant and reliable distinction between unfolded or partially folded proteins and proteins with well-defined global folds. Additionally, the precision of the method is sufficient to discern subtle differences in protein structural stability between, for example, the molten globule state of α-lactalbumin with a native-like overall fold and the partially folded (displaying a single α-helix and lacking long-range tertiary interactions) oncogenic transcription factor v-Myc with possible applications to protein folding studies. Data obtained on the Ca^{2+}-depleted apo and the Ca^{2+}-loaded holo form of α-lactalbumin additionally suggest that the method is able to detect subtle changes in protein stability caused by ligand binding. The method can easily be adjusted for screening purposes using NMR flow probes and micromanipulator robots and should consequently prove useful for target selection in high-throughput structural genomics and the identification of experimental conditions to optimize protein stability and crystal formation.

As the number of experimental protein structures is expected to significantly increase in the foreseeable future, comparative structure prediction

will become an essential tool in structural genomics. Although the reliability of structural modeling approaches is well documented and the precision (and accuracy) of predicted structures is sufficiently high to draw conclusions about putative biochemical functionality, there is still a demand for experimental verification and/or subsequent structural refinement. Given the robustness and reliability of our proposed strategy, we anticipate fruitful applications of the methodology in ongoing structural genomics efforts.

References

Acharya, K. R., Ren, J. S., Stuart, D. I., Phillips, D. C., and Fenna, R. E. (1991). Crystal structure of human alpha-lactalbumin at 1.7 Å resolution. *J. Mol. Biol.* **221,** 571–581.

Alexandrescu, A. T., Evans, P. A., Pitkeathly, M., Baum, J., and Dobson, C. M. (1993). Structure and dynamics of the acid-denatured molten globule state of alpha-lactalbumin: A two-dimensional NMR study. *Biochemistry* **32,** 1707–1718.

Anderson, P. J., Brooks, C. L., and Berliner, L. J. (1997). Functional identification of calcium binding residues in bovine alpha-lactalbumin. *Biochemistry* **36,** 11648–11654.

Balasubramanian, S., Schneider, T., Gerstein, M., and Regan, L. (2000). Proteomics of *Mycoplasma genitalium*: Identification and characterization of unannotated and atypical proteins in a small model genome. *Nucleic Acids Res.* **28,** 3075–3082.

Balbach, J., Forge, V., Lau, W. S., Jones, J. A., van Nuland, N. A., and Dobson, C. M. (1997). Detection of residue contacts in a protein folding intermediate. *Proc. Natl. Acad. Sci. USA* **94,** 7182–7185.

Baum, J., Dobson, C. M., Evans, P. A., and Hanley, C. (1989). Characterization of a partly folded protein by NMR methods: Studies on the molten globule state of guinea pig alpha-lactalbumin. *Biochemistry* **28,** 7–13.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242.

Calderone, V., Giuffrida, M. G., Viterbo, D., Napolitano, L., Fortunato, D., Conti, A., and Acharya, K. R. (1996). Amino acid sequence and crystal structure of buffalo alpha-lactalbumin. *FEBS Lett.* **394,** 91–95.

Cavanagh, J., Fairbrother, W. J., Palmer, A. G., and Skelton, N. G. (1996). ''Protein NMR Spectroscopy: Principles and Practice.'' Academic Press, San Diego, CA.

Chakraborty, S., Ittah, V., Bai, P., Luo, L., Haas, E., and Peng, Z. (2001). Structure and dynamics of the alpha-lactalbumin molten globule: Fluorescence studies using proteins containing a single tryptophan residue. *Biochemistry* **40,** 7228–7238.

Chrysina, E. D., Brew, K., and Acharya, K. R. (2000). Crystal structures of apo- and holo-bovine alpha-lactalbumin at 2.2-Å resolution reveal an effect of calcium on inter-lobe interactions. *J. Biol. Chem.* **275,** 37021–37029.

Choy, W. Y., and Forman-Kay, J. D. (2001). Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **308,** 1011–1032.

Chyan, C. L., Wormald, C., Dobson, C. M., Evans, P. A., and Baum, J. (1993). Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: A hydrogen exchange study. *Biochemistry* **32,** 5681–5691.

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

Diamond, R. (1974). Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.* **82,** 371–391.

Diercks, T., Coles, M., and Kessler, H. (2001). Applications of NMR in drug discovery. *Curr. Opin. Chem. Biol.* **5,** 285–291.

Dobson, C. M. (1994). Protein folding. Solid evidence for molten globules. *Curr. Biol.* **4,** 636–640.

Dolgikh, D. A., Gilmanshin, R. I., Brazhnikov, E. V., Bychkova, V. E., Semisotnov, G. V., Venyaminov, S., and Ptitsyn, O. B. (1981). Alpha-Lactalbumin: Compact state with fluctuating tertiary structure? *FEBS Lett.* **136,** 311–315.

Domingues, F., Koppensteiner, W. A., Jaritz, M., Prlic, A., Weichenberger, C., Wiederstein, M., Wiederstein, M., Flöckner, H., Lackner, P., and Sippl, M. J. (1999). Sustained performance of knowledge-based potentials in fold recognition. *Proteins* **3,** 112–120.

Doreleijers, J. F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J. L., and Ulrich, E. L. (2003). BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR* **26,** 139–146.

Dunker, A. K., and Obradovic, Z. (2001). The protein trinity—linking function and disorder. *Nat. Biotechnol.* **19,** 805–806.

Fieber, W., Schneider, M. L., Matt, T., Krautler, B., Konrat, R., and Bister, K. (2001). Structure, function, and dynamics of the dimerization and DNA-binding domain of oncogenic transcription factor v-Myc. *J. Mol. Biol.* **307,** 1395–1410.

Forge, V., Wijesinha, R. T., Balbach, J., Brew, K., Robinson, C. V., Redfield, C., and Dobson, C. M. (1999). Rapid collapse and slow structural reorganisation during the refolding of bovine alpha-lactalbumin. *J. Mol. Biol.* **288,** 673–688.

Ghaemmaghami, S., and Oas, T. G. (2001). Quantitative protein stability measurement *in vivo*. *Nat. Struct. Biol.* **8,** 879–882.

Ghaemmaghami, S., Fitzgerald, M. C., and Oas, T. G. (2000). A quantitative, high-throughput screen for protein stability. *Proc. Natl. Acad. Sci. USA* **97,** 8296–8301.

Griko, Y. V., and Kutyshenko, V. P. (1994). Differences in the processes of beta-lactoglobulin cold and heat denaturations. *Biophys. J.* **67,** 356–363.

Gronwald, W., and Kalbitzer, H. R. (2004). Automated structure determination of proteins by NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **44,** 33–96.

Heinemann, U. (2000). Structural genomics in Europe: Slow start, strong finish? *Nat. Struct. Biol.* **7**(Suppl.), 940–942.

Hitchens, T. K., Lukin, J. A., Zhan, Y. P., McCallum, S. A., and Rule, G. S. (2003). MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *J. Biomol. NMR* **25,** 1–9.

Hughson, F. M., Barrick, D., and Baldwin, R. L. (1991). Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. *Biochemistry* **30,** 4113–4118.

Janatova, J., Fuller, J. K., and Hunter, M. J. (1968). The heterogeneity of bovine albumin with respect to sulfhydryl and dimer content. *J. Biol. Chem.* **243,** 3612–3622.

Johnson, B. A., and Blevins, R. A. (1994). NMRView: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4,** 603–614.

Jones, S., and Thornton, J. M. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272,** 133–143.

Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L., and Hughey, R. (1999). Predicting protein structure using only sequence information. *Proteins* **3**(Suppl.), 121–125.

Kasuya, A., and Thornton, J. M. (1999). Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.* **286,** 1673–1691.

Kim, S. H. (1998). Shining a light on structural genomics. *Nat. Struct. Biol.* **5**(Suppl.), 643–645.

Koppensteiner, W. A., Lackner, P., Wiederstein, M., and Sippl, M. J. (2000). Characterization of novel proteins based on known protein structures. *J. Mol. Biol.* **296,** 1139–1152.

Koradi, R., Billeter, M., and Wuthrich, K. (1996). MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14,** 29–32.

Kutyshenko, V. P., and Cortijo, M. (2000). Water-protein interactions in the molten-globule state of carbonic anhydrase b: An NMR spin-diffusion study. *Protein Sci.* **9,** 1540–1547.

Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins* **6,** 87–103.

Kuwajima, K. (1996). The molten globule state of alpha-lactalbumin. *FASEB J.* **10,** 102–109.

Kuwajima, K., Mitani, M., and Sugai, S. (1989). Characterization of the critical state in protein folding. Effects of guanidine hydrochloride and specific $Ca2^{+}$ binding on the folding kinetics of alpha-lactalbumin. *J. Mol. Biol.* **206,** 547–561.

Lacelle, S. (1984). Random matrix theory in biological nuclear magnetic resonance. *Biophys. J.* **46,** 181–186.

Last, A. M., Schulman, B. A., Robinson, C. V., and Redfield, C. (2001). Probing subtle differences in the hydrogen exchange behavior of variants of the human alpha-lactalbumin molten globule using mass spectrometry. *J. Mol. Biol.* **311,** 909–919.

Maurus, R., Bogumil, R., Nguyen, N. T., Mauk, A. G., and Brayer, G. (1998). Structural and spectroscopic studies of azide complexes of horse heart myoglobin and the His-64 → Thr variant. *Biochem. J.* **332**(Pt. 1), 67–74.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21,** 1087–1092.

Moore, J. M. (1999). NMR screening in drug discovery. *Curr. Opin. Biotechnol.* **10,** 54–58.

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

Neri, D., Billeter, M., Wider, G., and Wüthrich, K. (1992). NMR determination of residual structure in a urea-denatured protein, the 434-respressor. *Science* **257,** 1559–1563.

Ota, M., Kawabata, T., Kinjo, A. R., and Nishikawa, K. (1999). Cooperative approach for the protein fold recognition. *Proteins* **3**(Suppl.), 126–132.

Pace, C. N., and McGrath, T. (1980). Substrate stabilization of lysozyme to thermal and guanidine hydrochloride denaturation. *J. Biol. Chem.* **255,** 3862–3865.

Parkin, S., Rupp, B., and Hope, H. (1996). Structure of bovine pancreatic trypsin inhibitor at 125K: Definition of carboxyl-terminal residues Gly57 and Ala58. *Acta Crystallogr. D Biol. Cryst.* **52,** 18–29.

Pascal, S. M., Muhandiram, D. R., Yamazaki, T., Forman-Kay, J. D., and Kay, L. E. (1994). Simultaneous acquisition of ^{15}N- and ^{13}C-edited NOE spectra of proteins dissolved in H_2O. *J. Magn. Reson.* **103B,** 197–201.

Peng, Z. Y., and Kim, P. S. (1994). A protein dissection study of a molten globule. *Biochemistry* **33,** 2136–2141.

Peng, Z. Y., Wu, L. C., and Kim, P. S. (1995). Local structural preferences in the alpha-lactalbumin molten globule. *Biochemistry* **34,** 3248–3252.

Plaxco, K. W., Simons, K. T., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277,** 985–994.

Pike, A. C., Brew, K., and Acharya, K. R. (1996). Crystal structures of guinea-pig, goat and bovine alpha-lactalbumin highlight the enhanced conformational flexibility of regions that are significant for its action in lactose synthase. *Structure* **4,** 691–703.

Prestegard, J. H., Valafar, H., Glushka, J., and Tian, F. (2001). Nuclear magnetic resonance in the era of structural genomics. *Biochemistry* **40,** 8677–8685.

Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv. Protein Chem.* **47,** 83–229.

Rao, J. K., Bujacz, G., and Wlodawer, A. (1998). Crystal structure of rabbit muscle creatine kinase. *FEBS Lett.* **439,** 133–137.

Redfield, C., Schulman, B. A., Milhollen, M. A., Kim, P. S., and Dobson, C. M. (1999). Alpha-lactalbumin forms a compact molten globule in the absence of disulfide bonds. *Nat. Struct. Biol.* **6,** 948–952.

Rehm, T., Huber, R., and Holak, T. A. (2002). Application of NMR in structural proteomics: Screening for proteins amenable to structural analysis. *Structure* **10,** 1613–1618.

Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution. *J. Mol. Biol.* **279,** 1211–1227.

Russell, R. B., Sasieni, P. D., and Sternberg, M. J. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282,** 903–918.

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9,** 56–68.

Schaefer, J., and Yaris, R. (1969). Random matrix theory and nuclear magnetic resonace spectral distributions. *J. Chem. Phys.* **51,** 4469–4474.

Schulman, B. A., and Kim, P. S. (1996). Proline scanning mutagenesis of a molten globule reveals non-cooperative formation of a protein's overall topology. *Nat. Struct. Biol.* **3,** 682–687.

Schulman, B. A., Kim, P. S., Dobson, C. M., and Redfield, C. (1997). A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nat. Struct. Biol.* **4,** 630–634.

Seavey, B. R., Farr, E. A., Westler, W. M., and Markley, J. L. (1991). A relational database for sequence-specific protein NMR data. *J. Biomol. NMR* **1,** 217–236.

Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242.

Shortle, D., and Ackerman, M. S. (2001). Persistence of native-like topology in a denatured protein in 8M urea. *Science* **293,** 487–489.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins.* **17,** 355–362.

Sippl, M. J. (1995). Knowledge based potentials for proteins. *Curr. Opin. Struct. Biol.* **5,** 229–235.

Sippl, M. J., and Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* **13,** 258–271.

Staunton, D., Owen, J., and Campbell, I. D. (2003). NMR and structural genomics. *Acc. Chem. Res.* **36,** 207–214.

Terwilliger, T. C. (2000). Structural genomics in North America. *Nat. Struct. Biol.* **7**(Suppl.), 935–939.

Thornton, J. M., Flores, T. P., Jones, D. T., and Swindells, M. B. (1991). Protein structure. Prediction of progress at last. *Nature* **354,** 105–106.

Tollinger, M., Konrat, R., Hilbert, B. H., Marsh, E. N., and Krautler, B. (1998). How a protein prepares for B12 binding: Structure and dynamics of the B12-binding subunit of glutamate mutase from *Clostridium tetanomorphum*. *Structure* **6,** 1021–1033.

Troullier, A., Reinstadler, D., Dupont, Y., Naumann, D., and Forge, V. (2000). Transient non-native secondary structures during the refolding of alpha-lactalbumin detected by infrared spectroscopy. *Nat. Struct. Biol.* **7,** 78–86.

Vijay-Kumar, S., Bugg, C. E., and Cook, W. J. (1987). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194,** 531–544.

Wijesinha-Bettoni, R., Dobson, C. M., and Redfield, C. (2001). Comparison of the structural and dynamical properties of holo and apo bovine alpha-lactalbumin by NMR spectroscopy. *J. Mol. Biol.* **307,** 885–898.

Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995). 1H, 13C, and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *J. Biomol. NMR* **5,** 67–81.

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293,** 321–331.

Wu, L. C., Peng, Z. Y., and Kim, P. S. (1995). Bipartite structure of the alpha-lactalbumin molten globule. *Nat. Struct. Biol.* **2,** 281–286.

Wüthrich, K. (1986). "NMR of Proteins and Nucleic Acids." Wiley, New York.

Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T., Kainosho, M., Nishimura, Y., Inoue, Y., and Kuramitsu, S. (2000). Structural genomics projects in Japan. *Prog. Biophys. Mol. Biol.* **73,** 363–376.

[7] Determination of Protein Backbone Structures from Residual Dipolar Couplings

By J. H. PRESTEGARD, K. L. MAYER, H. VALAFAR, and G. C. BENISON

Abstract

There are a number of circumstances in which a focus on determination of the backbone structure of a protein, as opposed to a complete all-atom structure, may be appropriate. This is particularly the case for structures determined as a part of a structural genomics initiative in which computational modeling of many sequentially related structures from the backbone of a single family representative is anticipated. It is, however, also the case when the backbone may be a stepping-stone to more targeted studies of ligand interaction or protein–protein interaction. Here an NMR protocol is described that can produce a backbone structure of a protein without the need for extensive experiments directed at side chain resonance assignment or the collection of structural information on side chains. The procedure relies primarily on orientational constraints from residual dipolar couplings as opposed to distance constraints from NOEs. Procedures for sample preparation, data acquisition, and data analysis are described, along with examples from application to small target proteins of a structural genomics project.

Introduction

Residual dipolar couplings (RDCs) are now widely used as a source of constraints in the determination of the structure of biomolecules. Several reviews on the subject have appeared (Al-Hashimi and Patel, 2002; Bax *et al*., 2001; de Alba and Tjandra, 2002; Prestegard *et al*., 2000; Tolman,

 0076-6879/05 $35.00