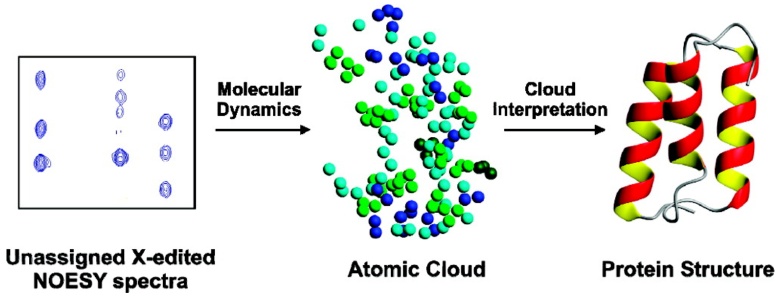
Article

Deuterated Protein Folds Obtained Directly from Unassigned Nuclear Overhauser Effect Data

Guillermo A. Bermejo, and Miguel Llins

J. Am. Chem. Soc., **2008**, 130 (12), 3797-3805 • DOI: 10.1021/ja074836e

Downloaded from http://pubs.acs.org on February 8, 2009



Journal of the American Chemical Society is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

JACS
ARTICLES
Published on Web 03/05/2008

Deuterated Protein Folds Obtained Directly from Unassigned Nuclear Overhauser Effect Data

Guillermo A. Bermejo and Miguel Llinás*

Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Received July 1, 2007; E-mail: llinas@andrew.cmu.edu

Abstract: We demonstrate the feasibility of determining the global fold of a highly deuterated protein from unassigned experimental NMR nuclear Overhauser effect (NOE) data only. The method relies on the calculation of a spatial configuration of covalently unconnected protons—a "cloud"—directly from unassigned distance restraints derived from ^{13}C- and ^{15}N-edited NOESY spectra. Each proton in the cloud, labeled by its chemical shift and that of the directly bound ^{13}C or ^{15}N, is subsequently mapped to specific atoms in the protein. This is achieved via graph-theoretical protocols that search for connectivities in graphs that encode the structural information within the cloud. The peptidyl H^N chain is traced by seeking for all possible routes and selecting the one that yields the minimal sum of sequential distances. Complete proton identification in the cloud is achieved by linking the side-chain protons to proximal main-chain H^Ns via bipartite graph matching. The identified protons automatically yield the NOE assignments, which in turn are used for structure calculation with RosettaNMR, a protocol that incorporates structural bias derived from protein databases. The method, named Sparse-Constraint CLOUDS, was applied to experimental NOESY data on the 58-residue Z domain of staphylococcal protein A. The generated structures are of similar accuracy to those previously reported, which were derived via a conventional approach involving a larger NMR data set. Additional tests were performed on seven reported protein structures of various folds, using restraint lists simulated from the known atomic coordinates.

Introduction

Protein structure elucidation by X-ray diffraction and NMR methods remains relatively slow vis-à-vis the high-throughput generation of genomic protein sequences. In the case of NMR, besides the optimization of protein production, the analysis of spectral data has the highest potential for significantly reducing structure determination time.[1] Within the data analysis stage, the assignment of spectral signals to specific atoms in the molecule prior to structure calculation[2] represents a major bottleneck. This has motivated the development of "top-down" approaches that aim at bypassing the spectral assignment step. Such methods rely on the generation of starting trial structures that are iteratively improved by use of information encoded in NMR spectra, producing the assignments simultaneously.[1]

On the experimental side, an area of active development is that of isotopic labeling techniques, which improve spectral resolution and sensitivity by the introduction of high levels of deuteration. This is achieved at the cost of decreasing the number of detectable nuclear Overhauser effects (NOEs), thus reducing a main source of structural constraints. The NOE sparseness, however, can be partially compensated by protonating specific methyl groups in an otherwise perdeuterated protein.[3] This strategy has been successfully applied to the determination of global folds of proteins in the 7−82 kDa size range via conventional assignment-based approaches.[4−8]

CLOUDS[9,10] is the most recent instance of a set of top-down methods aimed at determining protein structures via a direct implementation of unassigned NOE data only,[11−14] which thus shows the potential to reduce the required experimental data set and, consequently, data acquisition time. Designed to deal with fully protonated proteins, from which abundant NOEs can be obtained, CLOUDS calculates an initial structure by molecular dynamics, enforcing unassigned, precise NOE-derived distance restraints among unconnected protons. The model so

[1] Gronwald, W.; Kalbitzer, H. R. *Prog. Nucl. Magn. Reson. Spectrosc.* **2004**, *44* (1−2), 33−96.
[2] Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986; p 292.
[3] Goto, N. K.; Kay, L. E. *Curr. Opin. Struct. Biol.* **2000**, *10* (5), 585−592.
[4] Gardner, K. H.; Rosen, M. K.; Kay, L. E. *Biochemistry* **1997**, *36* (6), 1389−1401.
[5] Berardi, M. J.; Sun, C. H.; Zehr, M.; Abildgaard, F.; Peng, J.; Speck, N. A.; Bushweller, J. H. *Struct. Fold Des.* **1999**, *7* (10), 1247−1256.
[6] Mueller, G. A.; Choy, W. Y.; Yang, D. W.; Forman-Kay, J. D.; Venters, R. A.; Kay, L. E. *J. Mol. Biol.* **2000**, *300* (1), 197−212.
[7] Zheng, D. Y.; Huang, Y. P. J.; Moseley, H. N. B.; Xiao, R.; Aramini, J.; Swapna, G. V. T.; Montelione, G. T. *Protein Sci.* **2003**, *12* (6), 1232−1246.
[8] Tugarinov, V.; Choy, W. Y.; Orekhov, V. Y.; Kay, L. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (3), 622−627.
[9] Grishaev, A.; Llinás, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (10), 6707−6712.
[10] Grishaev, A.; Llinás, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (10), 6713−6718.
[11] Atkinson, R. A.; Saudek, V., The Direct Determination of Protein Structure from Multidimensional NMR Spectra Without Assignment. In *Dynamics and the Problem of Recognition in Biological Macromolecules*; Jardetzky, O., Lefèvre, J. F., Eds.; Plenum Press: New York, 1996; Vol. 288, pp 49−55.
[12] Atkinson, R. A.; Saudek, V. *J. Chem. Soc., Faraday Trans.* **1997**, *93* (18), 3319−3323.
[13] Kraulis, P. J. *J. Mol. Biol.* **1994**, *243* (4), 696−718.
[14] Malliavin, T. E.; Rouh, A.; Delsuc, M. A.; Lallemand, J. Y. *C. R. Acad. Sci. Paris, Ser.. II* **1992**, *315* (6), 653−659.

10.1021/ja074836e CCC: $40.75 © 2008 American Chemical Society

generated—a “cloud”—is not conventional in the chemical sense as it consists of a spatial distribution of free chemical-shift labeled ^{1}H-atoms, that accounts for the NOE information but lacks heavy atoms, covalent connectivity, and atom—identity assignments to specific counterparts in the protein. The refinement of this starting model is made possible by the subsequent identification of its protons via a Bayesian approach that exploits the structural information in the cloud. This establishes the NOE assignments; along with the addition of chemical prior knowledge (heavy atoms, bond connections, bond distances, etc.), this allows for the calculation of improved structures.

Here we build on the above technical advances by combining the improved spectral features associated with methyl-protonated, deuterated protein samples, with a CLOUDS-like approach. Expectedly, the NOE sparseness adversely affects the accuracy of coordinates within the cloud. The problem was tackled via a two-pronged approach: (i) the accuracy of the clouds was improved by supplementing NOEs with a newly formulated potential (the antidistance constraint or ADC potential), and (ii) an error-tolerant graph-theoretical strategy was developed to identify the cloud atoms. As a test of feasibility, the method—henceforth referred to as “Sparse-Constraint CLOUDS” (SC-CLOUDS)—was applied to experimental ^{13}C- and ^{15}N-edited NOESY data previously recorded from the 58-residue Z domain of staphylococcal protein A,[7] for which NMR models stemming from conventional studies on both fully protonated[15] and highly deuterated[7] samples are available. In addition, the applicability of SC-CLOUDS to other folds, spanning a range of sizes and topological complexities, was assessed by resorting to restraints simulated from seven reported NMR structures.

Although the introduction of high levels of deuteration has emerged mainly as a strategy to study large proteins,[16] the value of this labeling scheme in the high-throughput structure elucidation of other sized proteins has already been recognized.[7] In light of the tests described in this article, the current formulation of SC-CLOUDS suggests an alternative tool for global fold determination of small- to medium-sized proteins, with the potential to further expedite the process by fully exploiting NOESY data. Such folds are likely to be useful for functional annotation, active site detection, or identification of functional specificity determinants.

Methods

Protocol Overview. SC-CLOUDS starts from a random spatial distribution of covalently unconnected protons and proton groups (e.g., methyls), where each proton is labeled by its chemical shift and that of the directly bound ^{13}C or ^{15}N, as observed in isotope-edited NOESY spectra. This initial distribution is subjected to a molecular dynamics/simulated annealing (MD/SA) scheme that incorporates NOE distance restraints, antidistance constraints (ADCs or non-NOEs), and a repulsive van der Waals term. Several thousand atomic configurations, or clouds, are produced in this manner. Individual clouds are selected and their atoms are identified relative to the protein. This analysis proceeds via graph-search algorithms in graphs that encode the structural information in the cloud. Once the cloud atoms are identified, their associated chemical shifts become assigned, which enables interpretation of the NOEs in terms of specific atom pairs in the protein. The cloud-assigned NOEs are used to generate standard structural models of the protein. The protocol is outlined in Figure 1 and described in detail below.

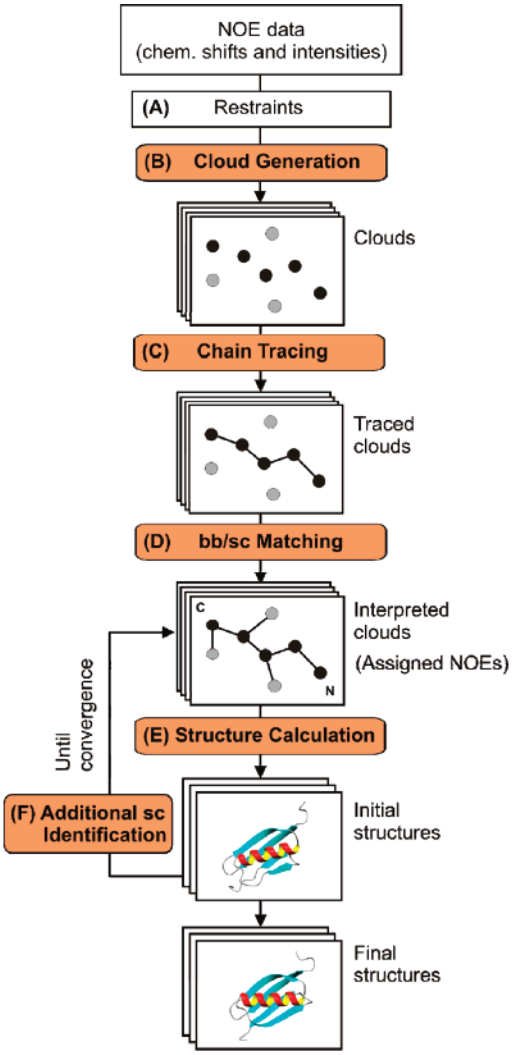


Figure 1. SC-CLOUDS flowchart. (A) Unassigned distance restraints are derived from NOESY spectra. (B) The restraints are used for MD-based cloud generation. Black and gray circles in the cloud denote backbone (bb) H^{N} and side-chain (sc) atoms, respectively. (C) The polypeptide H^{N} chain is traced. (D) Side chains and H^{N} protons belonging to the same amino acid residue are matched. (E) Structures are calculated from the resulting NOE assignments. (F) Initial structures are used for additional side-chain atom identification, leading to more NOE assignments and improved structures.

Input Data. Experimental NOE data from a Val, Leu, Ileδ1 methyl-protonated, ^{15}N-, ^{13}C-, ^{2}H-labeled sample of the Z domain of staphylococcal protein A were obtained from the BioMagResBank (www.bmrb.wisc.edu). The data consist of a list of spectral peaks from 350-ms mixing-time, 3D ^{13}C- and ^{15}N-edited NOESY experiments. Each entry in this list contains a raw NOE intensity, along with its spectral coordinates (δ_i^{H}, δ_j^{H}, δ_k^{X}), where δ_i^{H} and δ_j^{H} are, respectively, the chemical shifts of protons i and j, responsible for the unambiguous NOE, and δ_k^{X} ($X = {}^{13}C, {}^{15}N$) is the chemical shift of heteroatom k, directly bound to proton j. In addition, the published NOE list provides references for the aforementioned atoms to their respective identities within the protein (e.g., $i \rightarrow H^{N}$ of residue 10, $j \rightarrow H^{N}$ of residue 11, $k \rightarrow$ N of residue 11). Such references were replaced with references to “anonymous” atoms, that is, generic atoms whose identities within the protein are unknown, to generate an NOE list that constitutes, along with the protein sequence, the only input to SC-CLOUDS. The NOE information has been previously used as part of a larger data set, which

[15] Tashiro, M.; Tejero, R.; Zimmerman, D. E.; Celda, B.; Nilsson, B.; Montelione, G. T. *J. Mol. Biol.* **1997**, *272* (4), 573−590.
[16] Tugarinov, V.; Hwang, P. M.; Kay, L. E., *Annu. Rev. Biochem.* **2004**, *73*, 107−146.

includes *J*-correlated experiments, to derive the Z domain structure via an assignment-based approach.[7]

Cloud Components and Restraints (Figure 1A). As a result of the isotopic labeling,[17] the only observable Z domain protons are backbone-H^N, Asn/Gln side-chain-NH_2, Ileδ1-CH_3, and Leu/Val isopropyl-CH_3. Protons can be readily classified into these types from their ^1H-X (X = ^{13}C, ^{15}N) chemical shifts. Isopropyl methyl groups attached to common C^γ (Leu) or C^β (Val) atoms are identified by their similar NOE environment as determined by comparison of their NOE partners, a criterion akin to that implemented by Malliavin et al.[18] Specifically, a score, defined as the fraction of common NOE partners, is assigned to all possible isopropyl methyl pairs. Methyls with scores exceeding by 40% the next best pairing are assigned to the same isopropyl group. This “best first” pairing process is continued until all methyls are grouped. Side-chain NH_2 protons are grouped by their $^1H-^{15}N$ HSQC patterns. Isopropyl and NH_2 groups were used as references for ^{13}C- and ^{15}N-NOESY intensity calibration, respectively. Distance upper bounds were set to 1.5 times the isolated-spin-pair-approximation (ISPA) values, and lower bounds to 1.8 Å.

Antidistance constraints (ADCs)[9] are formulated as ad hoc repulsive potentials. In our case, NOE intensities were simulated for a database of protein structures, replicating the conditions of the experimental data on the 58-residue Z domain (350-ms mixing time, isotropic rotational correlation time, INEPT delays, etc.).[7] The database consists of 79 high-resolution X-ray structures (58 ± 15 residue range) from the Protein Data Bank (PDB; www.pdb.org). Protons were attached with the program REDUCE.[19] ^{13}C- and ^{15}N-edited NOESY intensities were simulated with STR2NOE, an in-house relaxation-matrix algorithm that assumes isotropically reorienting molecules which are rigid except for fast methyl rotations accounted for by a three-jump model.[20] The simulation takes into account the fractional occupation of exchangeable sites and differential transfer efficiencies during INEPT and reverse INEPT steps, as well as incomplete magnetization recovery during the duty cycle, as reported elsewhere.[21] Similar to the Z domain, all database proteins were assumed to be deuterated, selectively methyl-protonated,[17] with a correlation time of 3.2 ns (estimated from the molecular mass of the Z domain).

Calculated NOEs were deemed “observable” if their intensities exceeded a threshold determined from the experimental spectra. The fractional probabilities $\mathscr{P}_{noe}$ of observing an NOE as a function of the $^1H-^1H$ distance for the different interactions (H^N-H^N, H^N-CH_3, etc.) were calculated and directly implemented as (repulsive) “ADC potentials” during MD calculations (Figure 2). ADCs were enforced *only* between those protons that failed to yield an experimentally observable NOE, thus biasing them to assume distances corresponding to low $\mathscr{P}_{noe}$ values. Z domain cloud components and restraints are summarized in Table 1.

Cloud Generation (Figure 1B). Clouds are generated via MD/SA, starting from randomly distributed cloud components. H^Ns are treated as free atoms, while proton groups are held together either as rigid bodies in standard conformations (NH_2 and Ileδ1 methyls) or as covalent fragments (isopropyl groups). Force constants for the NOE, ADC, and van der Waals (vdW) repulsion terms are denoted by k_{noe}, k_{adc} [both in kcal•mol^{-1}•Å^{-2}], and k_{vdw} [in kcal•mol^{-1}•Å^{-4}], respectively. The MD/SA protocol consists of (i) 17-ps cooling (2000→0 K) with NOE and vdW terms (k_{noe} = 150 and k_{vdw} 1→4); (ii) 7.5-ps cooling (300→0 K), introducing the ADCs (k_{noe} = 150, k_{vdw} = 4, and k_{adc} 1→100). An ensemble of 4000 clouds is generated. Clouds satisfying NOE violations <0.2 Å, and ≤10 ADCs with energies >90% their maximum (short-distance) value (Figure 2), are selected and ranked according to their total energy.

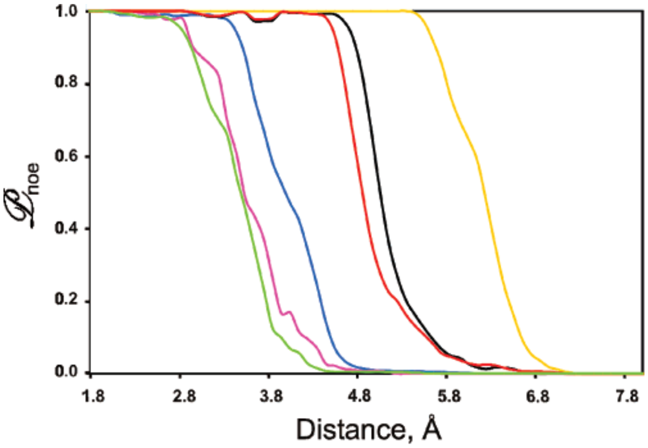


Figure 2. Probability of NOE observation (ADC potential). Curves are for the following interactions: NH_2-NH_2 (green), H^N-NH_2 (magenta), H^N-H^N (blue), CH_3-NH_2 (red), H^N-CH_3 (black), and CH_3-CH_3 (orange).

Table 1. Z Domain Cloud Calculation: Input Data and Cloud Statistics

cloud components	
H^N	55
Asn/Gln NH_2	13
Leu/Val isopropyl	8
Ileδ1 methyl	2
total	78
restraints	
NOEs	234
ADCs	4483
clouds generated with ADCs[a]	
H^N rmsd,[b] Å	6.1 ± 1.3
all-^1H rmsd,[b] Å	6.5 ± 1.2
R_{gyr},[c] Å	11.2 ± 0.5
clouds generated without ADCs[a]	
H^N rmsd,[b] Å	7.0 ± 0.8
all-^1H rmsd,[b] Å	7.4 ± 0.7
R_{gyr},[c] Å	7.6 ± 0.4

[a] Based on the ensemble of 10 lowest-energy clouds. [b] Average pairwise rmsd relative to PDB ID 2spz. [c] Average radius of gyration based on all protons in the cloud.

Chain Tracing (Figure 1C). While cloud atoms can be classified into types (backbone amide, Leu/Val isopropyl methyl, etc.) via their associated chemical shifts, their specific identities within the protein primary structure are unknown. The identification of atoms within a cloud starts by tracing a chain through its H^Ns, thus outlining the sequential array of amide protons in the polypeptide backbone. All possible chains are traced under the assumption that two H^Ns separated farther than a specified cutoff distance in the cloud cannot be adjacent in the chain. The chain yielding the minimum sum of sequential distances is considered the most likely solution.

The problem is formulated as that of finding a minimum-cost chain in an undirected graph. Each vertex in the graph represents an H^N in the cloud, and an edge indicates the possibility that the two linked vertices are adjacent in the chain. Each edge is assigned a cost equal to the corresponding H^N-H^N distance in the cloud. The graph is built including all H^Ns, with edges linking vertices whose associated separation distance lie within the specified cutoff (Figure 3A). Although the cutoff simplifies the subsequent chain search by constraining the set of possible solutions, it may result in graph components unconnected from each other (i.e., an unconnected graph). If so, individual components are treated independently by the chain tracing algorithm (described below), each yielding a chain fragment. Subsequently,

[17] Goto, N. K.; Gardner, K. H.; Mueller, G. A.; Willis, R. C.; Kay, L. E. *J. Biomol. NMR* **1999**, *13* (4), 369−374.

[18] Malliavin, T. E.; Barthe, P.; Delsuc, M. A. *Theor. Chem. Acc.* **2001**, *106* (1−2), 91−97.

[19] Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999**, *285* (4), 1735−1747.

[20] Tropp, J. *J. Chem. Phys.* **1980**, *72* (11), 6035−6043.

[21] Zhu, L. M.; Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **1998**, *11* (1), 17−29.

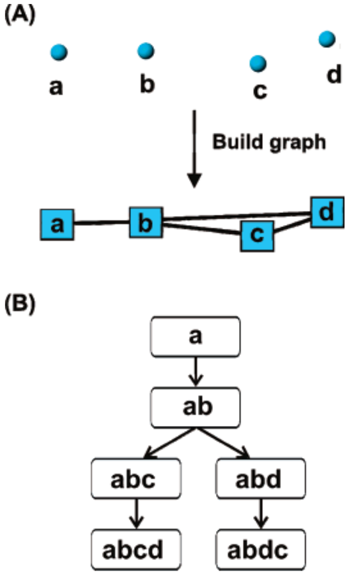


Figure 3. Chain tracing example. (A) Assumed spatial distribution of H^N atoms (labeled a−d) in the cloud, and associated graph built with a specified distance cutoff. Edge lengths are shown proportional to their costs (H^N−H^N distances). (B) Chain search in the graph starting from vertex a. Two possible chains are found, abcd and abdc; the former has the lowest cost (sum of a−b, b−c, and c−d distances) and is saved for further analysis. Similar searches, starting from the other vertices in the graph, eventually determine abcd as the best solution.

fragments are merged into a single chain via a similar tracing strategy on a graph constructed by linking the fragment ends. Without loss of generality, the following description of the chain search strategy assumes that the graph is connected and the complete chain can be directly obtained from it. (For a description of the treatment of unconnected graphs, see Supporting Information.)

The search, illustrated in Figure 3, starts with an initial chain consisting of a single vertex. The chain grows by the stepwise addition of new vertices from the set connected to the vertex at any of the chain ends; growth stops when no more vertices can be added. This search strategy is implemented via the depth-first search algorithm of graph theory.[22] Independent searches are conducted, starting from every vertex in the graph. In all cases, all chains spanning all vertices are sought. Each chain is assigned a cost by summing the costs of edges linking its sequential vertices in the graph. The lowest-cost chains resulting from every starting vertex are listed. The one with the lowest overall cost is interpreted to provide the sought-after H^N sequential information.

Backbone/Side-Chain Matching (Figure 1D). The chain provides only a sequential relationship for the H^Ns, with the location of its N- and C-termini undetermined. A strategy was developed to match H^N and side-chain atoms in the cloud that belong to the same amino acid residue. This identifies the H^Ns given the known protein sequence (i.e., the chain's N → C direction) as well as the side-chain groups.

The problem is cast in terms of bipartite graph matching,[23] which previously has been applied to different aspects of the determination of NMR assignments.[24−28] A bipartite graph, or bigraph, is one whose vertex set can be partitioned into two subsets so that each edge links the two subsets (i.e., no edge links vertices belonging to the same subset). Each edge is assigned a cost. A matching in a bigraph is a set of its edges, with no two edges sharing the same vertex. Figure 4C shows two completely connected bigraphs with edges denoted by full and dashed lines; in each bigraph a matching is indicated by thick edge lines. The sum of edge costs in a matching represents its overall cost. Our aim is to find perfect matchings, which are matchings involving all vertices in a bigraph. (Matchings shown in Figure 4C are perfect.) In addition, the matchings are required to be within the set of the *k* lowest-cost; that is, the *k* best perfect matchings, where *k* is an integer. This combinatorial optimization problem can be solved efficiently via protocols based on the Kuhn−Munkres algorithm.[23] (For a more detailed formulation of the bipartite graph matching problem, see Supporting Information.)

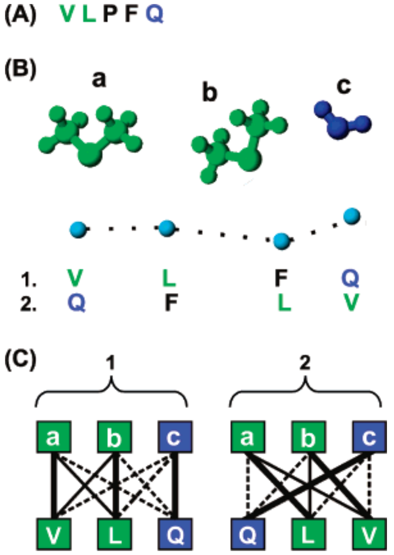


Figure 4. Backbone/side-chain matching example: determination of H^N chain direction. (A) Amino acid sequence. One-letter code labels are colored according to their protonated side-chain groups: isopropyl, green; NH_2, blue; none, black. (B) Assumed spatial distribution in the cloud. Side-chain groups (labeled a−c) are colored as in panel A, and H^Ns are denoted light blue. The chain is indicated by a dotted line, along with its two sets of H^N identities, one for each possible N → C direction (arrays 1 and 2). (C) Bigraphs for the two chain directions. Bigraph 1 is associated with array 1 in panel B, and bigraph 2 with array 2 in panel B. Edge lengths are shown proportional to their costs (distances in the cloud). Edges not considered for matching are shown as dashed lines. Thick edge lines denote the best matchings: the one in bigraph 1 yields a lower cost (sum of edge lengths) than the one in bigraph 2, hence array 1 in panel B represents the correct chain direction with the Val and Leu H^Ns proximal to the isopropyl groups and the Gln H^N to the NH_2. The Pro and Phe residues are absent from panel C as their side chains are completely deuterated; Pro is additionally missing from panel B as it lacks an H^N.

Assuming any of the two possible chain directions, a tentative set of identities for the H^Ns is obtained from the known amino acid sequence. A bigraph is built where H^Ns from residues with protonated side chains (Leu, Val, etc.) are represented by one vertex subset, and the side-chain groups within the cloud are represented by the other subset. An edge in this bigraph denotes the possibility that the paired H^N and side-chain group belong to the same residue and is assigned a cost given by the separation distance in the cloud. While all possible edges are included in the bigraph, only those linking atoms of the appropriate type are considered for matching; for example, a Val H^N should be matched to an isopropyl but not to an NH_2 group. Two bigraphs are built as indicated above, one for each of the two possible chain directions. The best perfect matching is obtained from each of them; the matching with the lowest cost reveals the correct chain direction (hence the correct bigraph), yielding the H^N identities. Figure 4 illustrates the procedure. Side-chain identities are determined by their

[22] Russell, S. J.; Norvig, P. *Artificial Intelligence: a Modern Approach*, 2nd ed.; Prentice Hall/Pearson Education: Upper Saddle River, NJ, 2003; pp xxviii, 1080.

[23] Asratian, A. S.; Denley, T. M. J.; Hèaggkvist, R. *Bipartite Graphs and their Applications*; Cambridge University Press: Cambridge, U. K., and New York, 1998; pp xi, 259.

[24] Hus, J. C.; Prompers, J. J.; Brüschweiler, R. *J. Magn. Reson.* **2002**, *157* (1), 119−123.

[25] Xu, Y.; Xu, D.; Kim, D.; Olman, V.; Razumovskaya, J.; Jiang, T. *Comput. Sci. Eng.* **2002**, *4* (1), 50−62.

[26] Langmead, C. J.; Donald, B. R. *J. Biomol. NMR* **2004**, *29* (2), 111−138.

[27] Langmead, C. J.; Yan, A.; Lilien, R.; Wang, L. C.; Donald, B. R. *J. Comput. Biol.* **2004**, *11* (2−3), 277−298.

[28] Constantine, K. L.; Davis, M. E.; Metzler, W. J.; Mueller, L.; Claus, B. L. *J. Am. Chem. Soc.* **2006**, *128* (22), 7252−7263.

matchings to the identified H^Ns. In order to widen the options, the 10 best matchings in the correct bigraph are considered. H^N/side-chain group pairs that appear with a frequency ≥0.8 within the set of 10 best matchings are assumed to belong to the same residue, thus identifying the side-chain groups. Those groups not satisfying the acceptance criterion are saved for eventual identification once preliminary structures of the protein are available (discussed below).

Structure Calculation (Figure 1E). The identification of cloud protons is formally equivalent to assigning the NOEs,[14] which in turn can be used for structure calculation. However, in practice, deuteration yields sparse NOE lists that are insufficient for generating reliable folds via standard molecular dynamics or distance geometry protocols. This limitation is usually compensated by incorporating extra restraints.[4−8] For instance, secondary structure is predicted mainly from backbone ^{13}C chemical shifts as measured from spectra used for assignment purposes.[29,30] Our approach purposely excludes such information since the goal is to obtain the structure from NOESY data only.

Rosetta[31] is a knowledge-based method that assembles structures with nativelike global properties from fragments of known structures likely to resemble the target protein at each residue position. When Rosetta is combined with experimental restraints, such as assigned NOEs, significant improvements in the accuracy of the computed models result, even for sparse data sets (RosettaNMR).[32] Here, we use Rosetta as an engine for fold generation. In our implementation, a library of protein fragments from the PDB is built from the target sequence (e.g., Z domain) by use of the Fragments module within the Rosetta software package, without relying on homologous proteins, NOEs, or chemical shifts. A total of 2000 structures are calculated with the cloud-assigned NOEs. Structures are selected according to the minimum number of 1-Å NOE violations that yield at least n accepted models, where $n = 10$ for our statistical analysis, and $n = 20$ for additional cloud side-chain identification (see below). After selection, structures are ranked according to their overall Rosetta energy. For comparison, 2000 ab initio models were calculated without NOE restraints and energy-ranked.

Additional Side-Chain Identification (Figure 1F). While the backbone/side-chain matching yields the identities of all peptidyl H^Ns, the pairing frequency cutoff (see above) can result in a number of unidentified side-chain groups. The associated NOEs are, consequently, unassigned and hence ignored in initial rounds of Rosetta structure calculation. However, once structures are obtained from the initially assigned NOEs, they can be used to identify more side-chain groups in the cloud, providing additional assigned NOEs that lead, in turn, to improved structures.

The extra side-chain identification is based on a MD/SA protocol that involves a preliminary structure and the unidentified cloud side-chain groups. The latter interact with the protein and with each other via the unassigned NOEs. All assigned (intraprotein) NOEs are also included. The free-floating groups are made vdW-invisible to protein-attached side-chain atoms; all other vdW repulsions are active. Pending quality of the NOE data, the unidentified side-chain groups—initially randomly distributed in space—are expected to drift during dynamics toward the loci they occupy in the folded protein, thus enabling their identification.[33]

Specifically, the MD/SA protocol consists of a single 100.3-ps cooling stage (1500→25 K) that includes the NOE and vdW repulsion terms, as well as a knowledge-based torsion angle potential[34] (k_{db} force constant) that biases the protein's side chains toward probabilistically favored conformations. Protein backbone segments of regular secondary structure are kept rigid. During a MD/SA run, k_{noe}, k_{vdw}, and k_{db} are increased 1→30 kcal·mol^{-1}·Å^{-2}, 0.1→1 kcal·mol^{-1}·Å^{-4}, and 0.002→1 kcal·mol^{-1}·rad^{-2}, respectively. The accepted 20 lowest-energy Rosetta structures, calculated with the initially assigned NOEs, are used to launch independent runs, each generating 10 final [protein + unidentified cloud side-chain group] configurations. For each of the resulting 200 configurations, a bigraph is built where the two vertex subsets represent the unassigned side-chain protons in the protein and the cloud, respectively. All edges are considered and assigned a cost equal to the corresponding interproton distance in the configuration. The best matching in each bigraph is found as already described, and the frequency of proton pairings is calculated from the set of 200 matchings. Pairings with frequency >0.5 are interpreted to provide the sought-after side-chain group identities. The newly assigned NOEs are added to the initial pool to generate new Rosetta structures. The process is repeated until no additional side-chain groups can be identified.

Tests on Simulated NOE Restraints. In addition to the experimental data on the Z domain, SC-CLOUDS was tested on sparse NOE restraints simulated from seven reported NMR structures of various folds and sizes, determined from fully protonated samples. Cutoff distances, defined from the $\mathscr{P}_{noe}$ curves (Figure 2) at $\mathscr{P}_{noe} = 0.5$, were 4.6 Å (H^N-H^N), 6.0 Å (H^N-CH_3), 4.5 Å (H^N-NH_2), 7.0 Å (CH_3-CH_3), 6.0 Å (CH_3-NH_2), and 4.3 Å (NH_2-NH_2). For each structure, distances shorter than their corresponding cutoffs were input to SC-CLOUDS as upper bounds, after increasing them by 20% to blur the information. All protocols, including ADCs (implemented between protons whose distances did not satisfy the cutoffs), were as described above for the Z domain.

Software and Hardware. Xplor-NIH[35,36] was used for all MD calculations. ADC potentials (Figure 2) were added as a new energy term. Rosetta v2.0 was used for structure generation. The relaxation matrix algorithm STR2NOE is coded in FORTRAN. Chain Tracing, Backbone/Side-Chain Matching, and Additional Side-Chain Identification programs are written in Python. Chain Tracing uses the framework provided in http://aima.cs.berkeley.edu.[22] Molecular graphics were prepared with MOLMOL.[37] Structure calculations with restraint lists simulated from PDB ID 1pc0 and 1k19 (see below) were carried out on IBM Blue Gene supercomputer at the San Diego Supercomputer Center (SDSC). All other computations were performed on PCs with 2.40−3.20-GHz processors.

Results and Discussion

Clouds. All isopropyl methyls in the Z domain were correctly associated with their corresponding isopropyl groups prior to cloud calculation. The lowest-energy Z domain cloud is shown in Figure 5A. Structural statistics for the 10 lowest-energy cloud ensemble are included in Table 1. Rmsds relative to the reported high-resolution NMR structure of the Z domain (PDB ID 2spz; Figure 6A), determined from a fully protonated sample,[15] are taken henceforth as a measure of coordinate accuracy. Relative to 2spz, the average pairwise H^N rmsd of the clouds is 6.1 Å, as high as ∼12 times the rmsds reported for clouds of fully protonated proteins relative to their known structures.[9] This likely stems from: (i) data sparseness, as deuteration results in 25−31% fewer NOEs per proton relative to ref 9, and (ii) the lesser accuracy of our ISPA-derived distances when compared to those obtained via a relaxation matrix formalism in the fully protonated case, afforded by the use of 2D homonuclear data.[9]

[29] Wishart, D. S.; Sykes, B. D. *J. Biomol. NMR* **1994**, *4* (2), 171−180.
[30] Cornilescu, G.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1999**, *13* (3), 289−302.
[31] Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268* (1), 209−225.
[32] Bowers, P. M.; Strauss, C. E. M.; Baker, D. *J. Biomol. NMR* **2000**, *18* (4), 311−318.
[33] AB, E.; Pugh, D. J. R.; Kaptein, R.; Boelens, R.; Bonvin, A. M. J. J. *J. Am. Chem. Soc.* **2006**, *128* (23), 7566−7571.
[34] Clore, G. M.; Kuszewski, J. *J. Am. Chem. Soc.* **2002**, *124* (12), 2866−2867.
[35] Schwieters, C. D.; Kuszewski, J. J.; Clore, G. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48* (1), 47−62.
[36] Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160* (1), 65−73.
[37] Koradi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graphics* **1996**, *14* (1), 51−55.

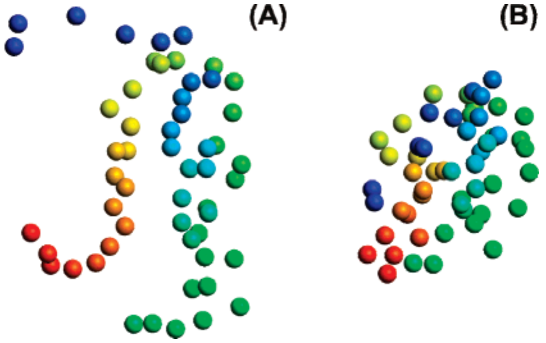


Figure 5. Lowest-energy Z domain clouds (only backbone H^Ns are shown): (A) computed with ADCs; (B) computed without ADCs. Atoms are colored in a blue (N-terminus) to red (C-terminus) gradient.

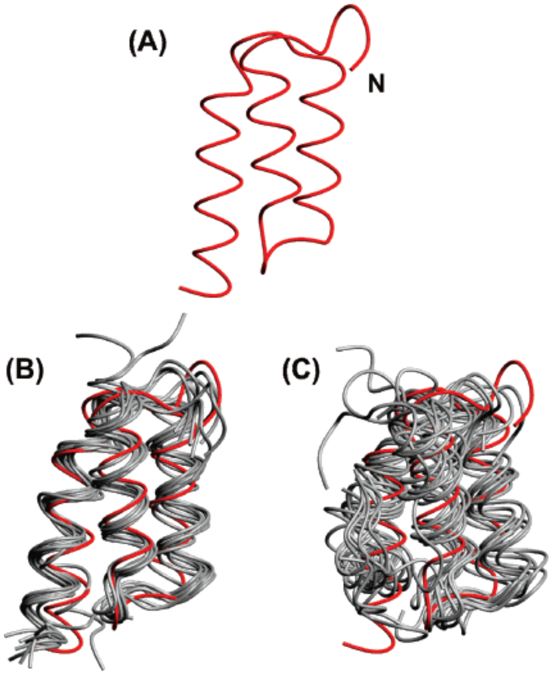


Figure 6. Backbone trace of Z domain structures. (A) Reference high-resolution NMR structure (PDB ID 2spz). (B) Rosetta models generated with NOEs assigned from the lowest-energy cloud. (C) Rosetta ab intio models. Structures are superimposed on the 2spz structure (red).

While providing specific proximity bounds for the various types of $^1H-^1H$ interactions, the ADCs are minimally perturbing as their potential profiles exhibit finite amplitudes and a narrow distance range of nonzero forces (Figure 2). This limits restrictions on atomic motions during dynamics, thus improving conformational sampling.[38] When ADCs were not incorporated in the MD calculations, the H^N rmsd relative to 2spz of the resulting clouds increased to 7.0 Å (Table 1). In addition, the clouds were more compact, with the average radius of gyration decreasing from 11.2 to 7.6 Å, further removed from the 2spz value of 10.0 Å (Figure 5, Table 1). Figure 5B shows the lowest-energy Z domain cloud computed without ADCs. In effect, ADC repulsions counteract the attractive nature of the NOEs, preventing cloud collapse. It is noteworthy that the applied ADCs outnumber the NOEs by >19:1 (Table 1), underscoring their relevance as conveyors of structural information. It is noteworthy that cloud calculations do not involve restraints enforcing chirality, so that topological mirror images are equally likely. This does not pose a problem for the subsequent identification of cloud atoms, as our analysis relies on distances only: both mirror images provide identical distance information.

Cloud Interpretation. In the previously reported CLOUDS analysis of fully protonated proteins, cloud backbone and side-chain protons were identified via a Bayesian protocol based on $^1H-^1H$ distance distributions observed in databases of high-quality protein structures.[10] The analysis implicitly assumed that distances within a cloud abide to such distributions. In contrast, when dealing with sparse NOEs, the computed clouds are of lower accuracy so that reliance on high-quality structure databases for their interpretation is unwarranted. Therefore, we resorted to graph-theoretical protocols involving a sum-of-distances minimization criterion. The analysis proceeded on a selected cloud, yielding NOE assignments that were either directly fed to the reminder of the computational pipeline (Figure 1) or combined with assignments obtained from other clouds to establish a consensus before continuing. In what follows, we focus on the interpretation of the lowest-energy Z domain cloud. Other Z domain clouds are discussed subsequently.

Our interpretation of a cloud hinges on a systematic graph search for a chain of backbone H^N protons. The protocol was inspired by the ARP/wARP method for building a C^α chain from a crystallographic electron density.[39] Exhaustive searches of this kind, however, can be intractable in highly connected graphs, thus our simplification of disregarding unlikely (i.e., “long”) edges by the application of a distance cutoff. (For details on the determination of appropriate distance cutoffs, see Supporting Information.) A 5.0-Å distance cutoff resulted in an H^N graph from the lowest-energy Z domain cloud with five components, leading to chain fragments later determined to involve residues 1−2, 3−6, 7−19, 21−22, and 23−58. The different graph components arise from sets of H^Ns in the cloud farther apart than the specified distance cutoff. This separation is generally caused by weak NOEs, occasionally caused by intercalating proline residues that are 100% deuterated. This is, for example, the case of fragments 7−19 and 21−22, separated by Pro-20. All fragments were merged into a single chain, producing, after backbone/side-chain matching (see below), the identities of all 55 H^Ns. By reference to an assignment-based treatment of a complete NMR data set that includes the NOESY data used here,[7] 43 H^Ns (78%) were correctly identified. The remaining 12 H^Ns (22%) were incorrectly identified due to coordinate errors in the cloud. However, such errors were in all cases local, resulting in the swapping of adjacent H^N identities: Lys-7 ↔ Glu-8, Gln-9 ↔ Gln-10, Asn-11 ↔ Ala-12, Glu-24 ↔ Glu-25, Gln-26 ↔ Arg-27, and Asn-28 ↔ Ala-29.

The subsequent backbone/side-chain matching correctly determined the chain direction and 74% of the side-chain identities. All eight isopropyl groups and two Ileδ1 methyls were correctly identified. Of the 13 NH_2 groups, seven were correctly identified, three were misidentified (Gln-9 → Asn-11, Gln-10 → Gln-9, and Gln-26 → Asn-23), and three remained unidentified. As was the case for the H^Ns discussed above, side-chain identification errors involved atoms proximal within the cloud.

Structure Calculation and Additional Cloud Interpretation. While our strategy for dealing with side-chain groups that

[38] Nilges, M.; Clore, G. M.; Gronenborn, A. M. *FEBS Lett.* **1988**, *239* (1), 129−136.

[39] Morris, R. J.; Perrakis, A.; Lamzin, V. S. *Acta. Crystallogr. D* **2002**, *58*, 968−975.

Table 2. Analysis of Z Domain Structures Derived by Various Methods

		backbone rmsd relative to 2spz, Å	
method	NOEs	all residues	helix bundle[a]
AutoStructure (ref 7)	185[b]	2.9 ± 0.4	1.8 ± 0.4
Rosetta ab initio[c]	0	5.5 ± 1.2	4.1 ± 0.8
SC-CLOUDS (initial)[c,d]	171	4.0 ± 0.7	3.6 ± 0.7
SC-CLOUDS (final)[c,d]	185	2.8 ± 0.5	2.2 ± 0.2

[a] Residues 7−18, 25−36, and 41−55. [b] After removal of redundancies. [c] Rmsds based on the ensemble of 10 lowest-energy structures. [d] Rosetta models were generated with NOEs assigned from the lowest-energy cloud before (initial) and after (final) additional side-chain identification (see text).

escape identification via backbone/side-chain matching is akin to that of AB et al.,[33] aimed at placing unidentified side-chain fragments proximal to their loci in the protein by use of unassigned NOEs, there are certain differences. Within SC-CLOUDS, the implementation of this concept is based on a simpler MD/SA scheme, instead of resorting to ARIA[40] or CANDID[41] calculations. Other differences in our dynamics protocol, devised specifically to cope with sparse constraints, consist in (i) maintaining rigid protein backbone segments within secondary structure elements in the initial Rosetta-derived coordinates and (ii) enforcing a torsion angle knowledge-based potential.[34] While point i avoids the unraveling of secondary structure elements during dynamics, thus counteracting previous Rosetta's effort, point ii biases the conformation of unassigned protein side chains which are not restrained by any NOE. Furthermore, SC-CLOUDS implements a bipartite graph matching protocol to score different side-chain assignment hypotheses. The strategy was applied to the three NH_2 groups in the lowest-energy Z domain cloud, left unidentified by the preceding backbone/side-chain matching stage (see previous paragraph), yielding the identities in the first iteration through the initial structure bundle. Only one misidentification occurred (Asn-23 → Gln-26) as a result of errors carried over from the previous stage. The extra identities produced 14 additional NOE assignments (Table 2).

The ensemble of final Z domain structures is shown in Figure 6B. In order to obtain at least 10 accepted models, a minimum cutoff of 7 (1-Å) NOE violations was required. The 10 lowest-energy structures in this set have 6.9 average NOE violations. Relative to 2spz, the average pairwise backbone (N, C^{α}, C′) rmsd of the final structures is 2.8 Å, which is 1.2 Å lower than that of the initial models generated without the complete side-chain identification (Table 2). It is gratifying that the structures (both initial and final) are more accurate than low-energy Rosetta ab initio models (Figure 6C, Table 2). Furthermore, the final structures are of similar accuracy to those obtained from the same NOE data by the assignment-based method,[7] both when the full backbone and the better-determined helix bundle are considered (Table 2).

Analysis of Multiple Clouds. The above discussion focused on the interpretation of the lowest-energy Z domain cloud. In order to avoid restricting our analysis to a single cloud, the protocols were also applied to the next nine energy-ranked clouds, producing similar results: when only the helix bundle is considered, backbone rmsds relative to 2spz fell within 1.9−2.3 Å, with one exception at 2.7 Å.

We also tested combining atomic identity information obtained from multiple clouds to establish a consensus before structure calculation. Specifically, chain tracing and backbone/side-chain matching were performed on each of the 10 lowest-energy Z domain clouds. The identity of a proton was assumed to be reliably established if obtained from at least six (60%) of the 10 clouds. These consensus identities yielded NOEs used for structure calculation and additional side-chain identification. From this analysis, all H^Ns were correctly identified, with the exception of six misidentified and 10 unidentified (no attempt was made to identify the latter via intermediate structures). No side-chain group remained unidentified after additional side-chain identification, but four were misidentified. As with individual clouds, errors in the consensus identities were local. Final Z domain structures had a full-backbone rmsd relative to 2spz of 3.1 ± 0.7 Å, improving to 2.5 ± 0.5 Å when only the helix bundle was considered.

As observed for the Z domain, the consensus approach may result in an increased number of unidentified protons owing to the application of the 60% confidence threshold. Despite the lesser information, the resulting Z domain structures are comparable to those obtained from single clouds. It is suggested, however, that the single cloud approach ought to be preferred when the number of unidentified protons by consensus becomes too large.

Tests with Other Folds. Results of SC-CLOUDS runs with simulated restraints on seven protein folds with sizes in the 52−112 residue range are shown in Figures 7 and 8 and Table 3. Only the lowest-energy cloud was considered in each case. Distance cutoffs within 5.0−5.5 Å were suitable for H^N graph simplification. The high-resolution structure of the Z domain (PDB ID 2spz) was included as a control. SC-CLOUDS models generated from 2spz are similar to those obtained from the experimental data, with a full-backbone rmsd relative to 2spz of 3.3 ± 0.7 Å, decreasing to 2.5 ± 0.5 Å when only the helix bundle is considered. Overall, SC-CLOUDS structures for the seven folds show correct topologies and fall within 2.4−3.8-Å backbone rmsd from the structures used to derive the input data, with the exception of PDB ID 1k19. In the latter case, SC-CLOUDS models yield a backbone rmsd of 5.8 ± 0.9 Å, which decreases to 3.9 ± 0.4 Å when the relatively long loop and terminal segments are neglected. The above rmsds are consistent with experimental and computational observations on the accuracy of conventionally generated sparse-constraint models relative to their higher resolution counterparts obtained from fully protonated proteins.[4,7] In addition, all SC-CLOUDS structures are more accurate than low-energy Rosetta ab initio models (Table 3 and Figure 8).

Comparison with Other Methods and Possible Improvements. Relative to other NMR structure determination protocols, the main advantage afforded by a direct NOE implementation, such as in SC-CLOUDS, is an alternative to the chemical shift assignment that allows for the possibility of bypassing *J*-correlated experiments, usually performed for this primary purpose. In the case of previous assignment-based work on the highly deuterated Z domain, such experiments accounted for ∼70% of data collection time (ca. 6 days).[7] On the other hand, a common problem that affects “direct” methods, particularly

[40] Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. *Bioinformatics* **2003**, *19* (2), 315−316.

[41] Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319* (1), 209−227.

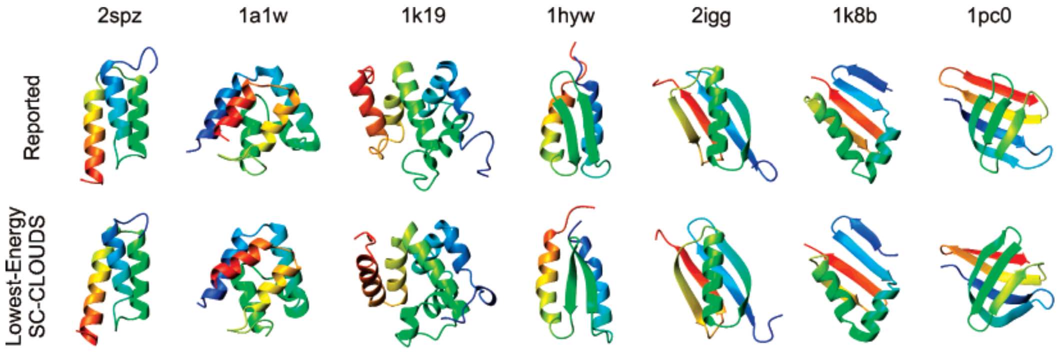


Figure 7. Lowest-energy SC-CLOUDS models generated from simulated restraint lists extracted from reported high-resolution NMR structures (PDB ID indicated).

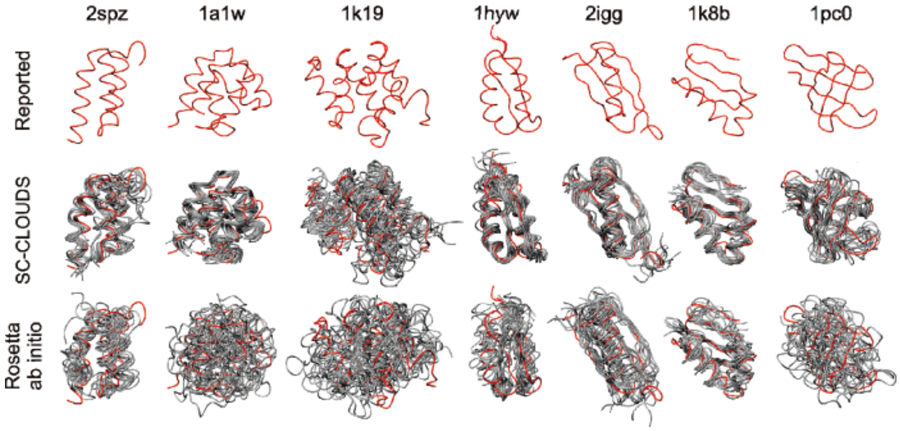


Figure 8. Comparison of Rosetta ab initio and SC-CLOUDS models (generated from simulated restraint lists) relative to the reported high-resolution NMR structures (PDB ID indicated). Each ensemble shows the backbone trace of the 10 lowest-energy models, superimposed on the reference structure (red). The average backbone (N, C$^\alpha$, C′) rmsd for each ensemble is listed in Table 3.

Table 3. SC-CLOUDS on Simulated Restraints from Proteins with Varying Folds and Sizes

protein			input	output
PDB ID	length	fold	NOEs[a]	rmsd,[b] Å
2spz	58	α	211	3.3 ± 0.7 (5.5 ± 1.2)
1a1w	83	α	385	2.4 ± 0.2 (9.7 ± 1.5)
1k19	112	α	406	5.8 ± 0.9 (11.0 ± 1.5)
1hyw	58	αβ	230	2.8 ± 0.3 (5.9 ± 0.6)
2igg	64	αβ	174	2.7 ± 0.4 (4.5 ± 0.9)
1k8b	52	αβ	191	3.1 ± 0.4 (3.9 ± 0.5)
1pc0	61	β	220	3.8 ± 0.7 (10.9 ± 0.6)

[a] Number of NOE restraints simulated from the reported structure (identified by PDB ID in the table). [b] Backbone (N, C$^\alpha$, C′) rmsd of the final 10 lowest-energy SC-CLOUDS models relative to the reported structure. The rmsd obtained via Rosetta ab initio calculations (i.e., without cloud-assigned NOEs) is shown in parentheses.

those based on fully protonated proteins such as ANSRS,[13] CLOUDS,[9,10] and other proposed protocols,[11,14] is that of NOE ambiguity due to chemical shift overlap. Although recently developed algorithms for NOE disambiguation[42,43] have been proposed to produce a suitable input for CLOUDS,[44] they require *J*-correlated information. Thus, the results presented in this paper are encouraging because, while the overlap problem is counteracted by the introduction of high levels of deuteration, SC-CLOUDS is able to exploit the concomitantly sparse NOE data to obtain global folds.

As exemplified by the application of our method to the experimental Z domain data, local cloud imperfections cause a sizable fraction of atoms to be misidentified. Thus, a number of NOEs are misassigned; however, this has a negligible effect on the Rosetta-generated structures. Indeed, the use of correct assignments did not change structure accuracy (not shown). There are two reasons for this resiliency. First, we purposely avoided using NOEs in the selection of protein fragments from the PDB—which determine the local conformational space of Rosetta models—and relied only on sequence information as a selection criterion. Second, the assembly of the fragments into compact structures is a coarse-grained strategy, able to accommodate local misassignments. In other words, although SC-CLOUDS structures display the correct folds, they are not necessarily associated with fully correct chemical shift assignments. The robustness of Rosetta against misassigned data has recently been observed in the context of simultaneous assignment and structure determination of protein backbones via residual dipolar couplings (RDCs).[45] Further evidence for the determination of correct structures despite incorrect assignments is afforded by a Monte Carlo assignment procedure proposed by Meiler and Baker[46,47] for selecting models generated via structure prediction techniques and by a floating chirality approach proposed by Folmer et al.[48] for calculating structures without stereospecific assignments.

(42) Grishaev, A.; Llinás, M. *J. Biomol. NMR* **2004**, *28* (1), 1−10.
(43) Grishaev, A.; Llinás, M. *J. Biomol. NMR* **2002**, *24* (3), 203−213.
(44) Grishaev, A.; Llinás, M. *Methods Enzymol.* **2005**, *394*, 261−295.
(45) Jung, Y. S.; Sharma, M.; Zweckstetter, M. *Angew. Chem., Int. Ed.* **2004**, *43* (26), 3479−3481.
(46) Meiler, J.; Baker, D. *J. Magn. Reson.* **2005**, *173* (2), 310−316.
(47) Meiler, J.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (26), 15404−15409.

The application of SC-CLOUDS to large proteins—including those only amenable to NMR when highly deuterated—requires further development in order to overcome expected chemical shift overlap and Rosetta size limitations.[49] A possible venue to mitigate these restrictions might call for the introduction of a few selected *J*-correlated experiments. HNCA and HN(CO)-CA, for example, ought to help in NOE disambiguation as well as in cloud and structure calculation via the introduction of backbone covalent connectivity and chemical shifts (related to secondary structure). Furthermore, the current SC-CLOUDS implementation can readily incorporate backbone $^1H-^{15}N$ RDC data; after assignment during cloud interpretation, these data can supplement the NOE restraints in subsequent Rosetta structure calculations. This was tested with experimental $^1H-^{15}N$ RDCs available on the Z domain,[7] resulting in structures comparable to those calculated with NOEs only (not shown). Inclusion of RDCs also did not significantly help in the assignment-based structure elucidation of the Z domain.[7] However, the overall usefulness of RDCs in the determination of folds via SC-CLOUDS cannot be ruled out.

Conclusions

We have demonstrated the feasibility of obtaining the global fold of the Z domain of staphylococcal protein A using only the amino-acid sequence and unassigned, unambiguous NOEs obtained from 3D ^{13}C- and ^{15}N-edited NOESY experiments recorded on a selectively methyl-protonated, deuterated sample. The generated structures are comparable to those derived via an assignment-based approach based on a larger NMR data set that includes several *J*-correlated spectra.[7] Tests with simulated restraints from reported small- to medium-sized protein structures—the type expected to yield unambiguous NOEs when highly deuterated—suggest the method is applicable to a variety of folds.

The ADC potentials proved to be a crucial source of structural information, complementary to the NOEs. Although the ADCs used in this work were specific for the particular NOESY data available for the highly deuterated Z domain, they can be readily formulated for other types of NOESY experiments and labeling schemes. Furthermore, the ADC potentials should be useful to strengthen standard structure calculation protocols.

The graph-based interpretation of clouds tolerates inaccurate atomic positions, a problem also encountered in the analysis of low-resolution X-ray diffraction electron-density maps.[39] The combination of the cloud interpretation routines with the Rosetta protocol is highly synergistic. While only low-accuracy structures can be computed with either the cloud-assigned NOEs (subjected to standard calculation protocols) or via Rosetta ab initio, the incorporation of such NOEs within Rosetta significantly improves structure accuracy (Tables 2 and 3 and Figures 6 and 8).

SC-CLOUDS deemphasizes the reliance on *J*-correlated spectra, traditionally used to perform assignment prior to structure calculation,[2] by focusing on experiments rich in structural information. This acknowledges that the goal of protein structural NMR is to obtain not chemical shift assignments but a useful, reasonably accurate 3D model.[1] The shift in paradigm may allow for structure determination to be achieved directly from NOESY spectra.

Software Availability. The ADC potential implementation is available in Xplor-NIH 2.19. In-house developed programs are available upon request from the authors.

Acknowledgment. We thank Alexander Grishaev for valuable discussions and reading the manuscript, Justin T. Douglas and Alfred G. Redfield for reading the manuscript, Charles D. Schwieters and John J. Kuszewski for help with Xplor-NIH, Ross Walker for help with the Rosetta setup at SDSC, and the reviewers of this manuscript for valuable comments. This research was supported by the U.S. Public Health Service, NIH Grant GM-067964, and by the San Diego Supercomputer Center, Grant ASC070011N.

Supporting Information Available: Treatment of unconnected graphs during chain tracing; chain tracing dependence on secondary structure; supplementary figures; determination of the distance cutoff for chain tracing; a detailed formulation of the bipartite matching problem; and treatment of pro-chiral groups within cloud and structure calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

JA074836E

[48] Folmer, R. H. A.; Hilbers, C. W.; Konings, R. N. H.; Nilges, M. *J. Biomol. NMR* **1997**, *9* (3), 245−258.
[49] Rohl, C. A. *Methods Enzymol.* **2005**, *394*, 244−260.