

Protein structure elucidation from NMR proton densities

Alexander Grishaev and Miguel Llinás*

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

Communicated by Gregory A. Petsko, Brandeis University, Waltham, MA, February 26, 2002 (received for review August 3, 2001)

The NMR-generated foc proton density affords a template to which the molecule has to be fitted to derive the structure. Here we present a computational protocol that achieves this goal. H^N atoms are readily recognizable from ¹H/²H exchange or ¹H/¹⁵N heteronuclear single quantum correlation (HSQC) experiments. The primary structure is threaded through the unassigned foc by leapfrogging along peptidyl amide H^Ns and the connected H^αs. Via a Bayesian approach, the probabilities of the sequential connectivity hypotheses are inferred from likelihoods of H^N/H^N, H^N/H^α, and H^α/H^α interatomic distances as well as ¹H NMR chemical shifts, both derived from public databases. Once the polypeptide sequence is identified, directionality becomes established, and the foc N and C termini are recognized. After a similar procedure, side chain H atoms are found, including discriminated cis/trans proline loci. The folded structure then is derived via a direct molecular dynamics embedding into mirror image-related representations of the foc and selected according to a lowest energy criterion. The method was applied to foc densities calculated for two protein domains, col 2 and kringle 2. The obtained structures are within 1.0–1.5 Å (backbone heavy atoms) and 1.5–2.0 Å (all heavy atoms) rms deviations from reported x-ray and/or NMR structures.

NMR direct method | NMR Bayesian analysis | NOE-only molecular structure | proteomics | structural genomics

Notwithstanding continuous advances in NMR instrumentation, experimental design, and data analysis, in most applications the derivation of macromolecular structures via NMR experiments remains slow, mainly because of the requirement to assign signals to individual spins. Consequently, much effort is being devoted to the automation of resonance and nuclear Overhauser effect (NOE) assignments of proteins (1). For the resonance assignment strategies to be successful, an arsenal of triple resonance (¹H/¹³C/¹⁵N) experiments exploring both intra- and interresidue *J* connectivities is required (2). Although such approaches have obvious merits for expediting protein-structure derivation, the question remains as to whether an assignment-independent protocol such as CLOUDS (3) cannot accelerate the procedure further.

As described in the accompanying paper (3), the CLOUDS protocol leads to a family of clouds (foc) or effective “proton density” which conveys a fuzzy image of the underlying macromolecular fold. Here we address the problem of determining the molecular structure from the foc. Unlike electron density maps such as those obtained via x-ray crystallography, the foc is composed of individual atomic densities, which facilitates its analysis. Furthermore, H^N atoms, among the best-localized in the foc, are readily recognized from, e.g., NOE correlated spectroscopy experiments in ¹H₂O/²H₂O or, preferably, from ¹H/¹⁵N HSQC experiments. In the protocol presented in this paper, the atomic H^N focs serve as starting points to identify, via a Bayesian probabilistic approach, sequential polypeptide backbone atomic sites along the amino acid sequence as well as the attached side chains. The molecule then is embedded into the global proton density through a restrained molecular dynamics (MD) optimization procedure. The results show that the

CLOUDS-generated structures are of comparable quality to those obtained via standard NMR protocols.

Methods

CLOUDS starts from unassigned NOEs. Consequently, the foc proton density is also unassigned. As presented below, the foc H atoms can be identified via backbone finder (BAF)/sidechain finder (SIF) protocols based on straightforward Bayesian probabilistic assessments followed by fitting the polypeptide conformation via EMBEDS, a foc-constrained MD procedure. The complete CLOUDS protocol is summarized in Fig. 1. The analysis was applied to col 2 (4) and kringle 2 (5), two globular protein modules of 60 and 83 amino acids, respectively, starting from focs derived in ref. 3. Computational resources are described (3). MOLMOL 2.5.1 (6) and SPOCK 1.0 (7) were used for molecular graphics.

Bayesian Identification of foc Atoms. The interactive NMR assignment procedure amounts to evaluation of the identities of signals obtained from a complete set of mutually exclusive identity “hypotheses” ($\mathcal{H}^1, \dots, \mathcal{H}^m$), by testing the conjecture against a set of available data ($\mathcal{D}_1, \dots, \mathcal{D}_n$), typically extracted from multidimensional NOE- and *J*-correlated spectra. Such analysis is biased by prior knowledge that could include, e.g., crosspeak line shapes, chemical shift patterns, or data (if available) from homologous structures. The hypotheses that best fit both the data and prior knowledge then are chosen as final assignments. Such assessments can be formulated in terms of Bayesian inference. In the standard jargon, $\mathcal{P}(\mathcal{H}^i | \mathcal{D}_1, \dots, \mathcal{D}_n)$, the “posterior” probability of hypothesis \mathcal{H}^i conditional on the data, is estimated from the “prior” probabilities $\mathcal{P}(\mathcal{H}^i)$ as well as the “likelihoods” $\mathcal{P}(\mathcal{D}_k | \mathcal{H}^i)$ of satisfying \mathcal{D}_k conditional on \mathcal{H}^i . For statistically independent \mathcal{D}_k values, Bayes’ theorem (8) states that

$$\mathcal{P}(\mathcal{H}^i | \mathcal{D}_1, \dots, \mathcal{D}_n) = \frac{\mathcal{P}(\mathcal{H}^i) \prod_{k=1}^n \mathcal{P}(\mathcal{D}_k | \mathcal{H}^i)}{\mathcal{P}(\mathcal{D}_1, \dots, \mathcal{D}_n)}, \quad [1]$$

where $\mathcal{P}(\mathcal{D}_1, \dots, \mathcal{D}_n) = \sum_{j=1}^m \mathcal{P}(\mathcal{H}^j) \prod_{k=1}^n \mathcal{P}(\mathcal{D}_k | \mathcal{H}^j)$.

Here our concern is the identification of all atomic components in the foc. Moreover, because the foc atomic densities are frequency-labeled, the obtained identities automatically lead to assignment of the spectrum. The data set pertinent to our assignment hypotheses \mathcal{H} for foc atom H_{*n*} includes its chemical shift δ_n and/or its spatial position relative to other foc atoms. The latter is defined through a number *M* of interproton

Abbreviations: NOE, nuclear Overhauser effect; CLOUDS, computed location of unassigned spins; foc, family of clouds; MD, molecular dynamics; BAF, backbone finder; SIF, sidechain finder; EMBEDS, energy minimization buildup of encysted structure; PD, probability distribution; rmsd, rms deviation.

*To whom reprint requests should be addressed. E-mail: llinas@andrew.cmu.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

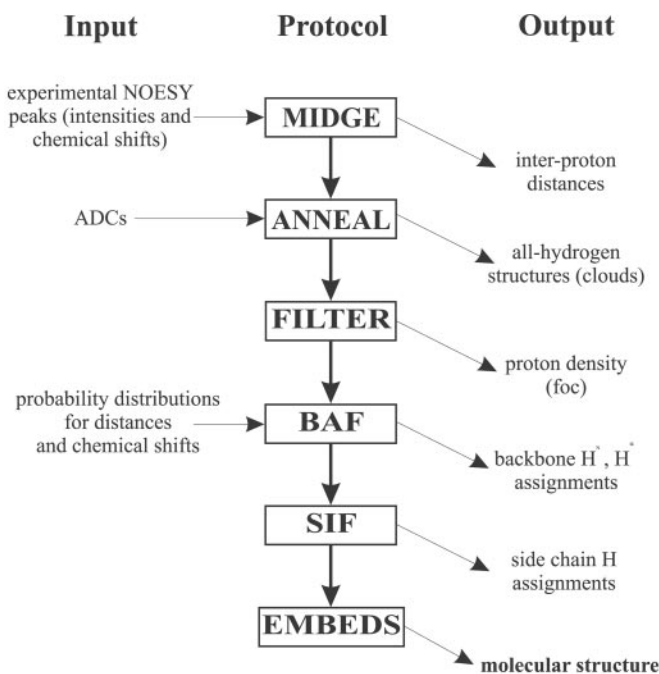


Fig. 1. Flowchart of CLOUDS protocol. NOESY, nuclear Overhauser effect spectroscopy; MIDGE, model-independent distance generation; ADC, anti-distance constraint; ANNEAL, anfractuosity nesting at atomic loci; FILTER, foc identification via lowest error.

distances extracted from each cloud by reference to the set of atoms of currently established, i.e., updated, identity (H_{ξ_j}), where $1 \leq j \leq M$. Within this context, Bayes' ansatz (Eq. 1) becomes:

$$\mathcal{P}(\mathcal{H}^f | \delta_n, x_{\xi_1, n}, \dots, x_{\xi_M, n}) = \frac{\mathcal{P}(\mathcal{H}^f) \cdot \mathcal{P}(\delta_n | \mathcal{H}^f) \prod_{j=1}^M \mathcal{P}(x_{\xi_j, n} | \mathcal{H}^f)}{\mathcal{P}(\delta_n, x_{\xi_1, n}, \dots, x_{\xi_M, n})}, \quad [2]$$

where $x_{\xi_j, n}$ denotes the distance between foc atoms H_{ξ_j} and H_n . In practice, our priors are uniform selection probabilities; e.g., if there are 83 H^N atoms, for $\mathcal{H}^f = H_i^N$, $\mathcal{P}(\mathcal{H}^f) = 1/83$, and the various likelihoods $\mathcal{P}(\delta_n | \mathcal{H}^f)$ and $\mathcal{P}(x_{\xi_j, n} | \mathcal{H}^f)$ are evaluated from empirical databases. The empirical distance probability distributions (PDs) between various H atoms were generated from a set of eight reported high-quality protein structures solved via NMR spectroscopy and x-ray crystallography (9–16), and binned to 0.2 Å. The chemical shift PDs were built from the BioMagRes-Bank database (Madison, WI), approximated by Gaussian functions. All distance likelihoods in Eq. 2 were computed as

$$\mathcal{P}(x_{\xi_j, n} | \mathcal{H}^f) = \int_{x \in x_{\xi_j, n}} dx \mathcal{P}_{\text{loc}}(x) \cdot \mathcal{P}_{\text{db}}(x | \mathcal{H}^f), \quad [3]$$

where the PDs $\mathcal{P}_{\text{loc}}(x_{\xi_j, n})$ and $\mathcal{P}_{\text{db}}(x_{\xi_j, n} | \mathcal{H}^f)$ are over the set of individual clouds and the structure database, respectively. Only hypotheses yielding posteriors that fall within <2 orders of magnitude from their highest calculated values were kept for further analysis.

BAF. The protocol traces the string of backbone atoms within the foc. For each H_i^N atom, BAF finds the most probable intraresidue H_i^α and sequential $H_{i\pm 1}^N$ and $H_{i\pm 1}^\alpha$ atoms. Initially, $H_{i\pm 1}^N$ are inferred from their distances to H_i^N , whereas the H_i^α and

$H_{i\pm 1}^\alpha$ are deduced from both distances to H_i^N and their chemical shifts δ_n .

By reference to a given amide H_i^N , any other H_ℓ^N was considered to be either sequential, $H_{i\pm 1}^N$, or nonsequential, $H_{i\pm 1}^N$. The posteriors for the sequential hypotheses were calculated as in Eq. 2, by using likelihoods of foc H_i^N – H_ℓ^N distances, $x_{i\ell} \equiv x_i$,

$$\mathcal{P}(H_{i\pm 1}^N | x_i) = \frac{\mathcal{P}(H_{i\pm 1}^N) \cdot \mathcal{P}(x_i | H_{i\pm 1}^N)}{\mathcal{P}(H_{i\pm 1}^N) \cdot \mathcal{P}(x_i | H_{i\pm 1}^N) + \mathcal{P}(H_{i\pm 1}^N) \cdot \mathcal{P}(x_i | H_{i\pm 1}^N)}, \quad [4]$$

which corresponds to $M = 1$, $\mathcal{P}(\delta_n | \mathcal{H}^f) = 1.0$, $H_{\xi_1} = H_i^N$ and $H_n = H_\ell^N$. For the nonsequential hypothesis, an expression analogous to Eq. 4 applies. In a similar fashion, to identify the H^α atoms, every nonamide atom m in the foc was considered, relative to H_i^N , to be either non- α , H^α , or one of four types of α : intraresidue H_i^α , sequential downstream or upstream $H_{i\pm 1}^\alpha$, or nonsequential, $H_{i\pm 1}^\alpha$. The posteriors for these five hypotheses were calculated by using likelihoods of H_i^N – H_m^α intracloud distances, x_{im} , and of the δ_m , i.e., according to Eq. 2, setting $M = 1$, $H_{\xi_1} = H_i^N$, and $n = m$.

The sequential amide links were considered unambiguous if two or less high-probability links remained. The individual H_i^α , H_{i+1}^α , and H_{i-1}^α links were assumed unique if only one was left for the particular type. Whenever a given amide showed unique $H_{i\pm 1}^N$ links but nonunique $H_{i\pm 1}^\alpha$ links, the probabilities for the latter were refined by using Eq. 2 with $M = 3$, $H_{\xi_1} = H_i^N$, $H_{\xi_2} = H_{i+1}^N$, and $H_{\xi_3} = H_{i-1}^N$. The choice between H_{ξ_2} and H_{ξ_3} was made by optimally matching the previously selected most likely H_i^α atoms. Whenever either of the uniquely determined $H_{i\pm 1}^N$ links exhibited unambiguous H_i^α links, the posteriors included an additional term $H_{\xi_4} = H_{i\pm 1}^\alpha$ and $M = 4$. Similarly, whenever the H_i^α sites became uniquely identified but the $H_{i\pm 1}^N$ links remained ambiguous, the $H_{i\pm 1}^N$ posteriors in Eq. 4 were recast in line with Eq. 2 with $M = 2$, $\mathcal{P}(\delta_n | \mathcal{H}^f) = 1.0$, $H_{\xi_1} = H_i^N$, and $H_{\xi_2} = H_i^\alpha$. With these extra conditions, the probabilities for the H_{i+1}^N and H_{i-1}^N hypotheses became distinct.

The whole BAF cycle was repeated until convergence. At this point, glycines, which potentially yield two high-scoring H^α matches separated by distances <2.5 Å, were identified. Likewise unique, Pro H^α atoms could be recognized from both their upstream and downstream sequential amide connections. Because BAF determines the directionality of the amide linkages, the polypeptide C and N termini could be identified and the sequential assignment established.

SIF. Once the polypeptide backbone atoms become assigned via BAF, the problem reduces to best-fitting the established side chain covalent structures, as dictated by the sequence, to the remainder proton density. SIF localizes side chain foc H atoms linked to given H_i^N/H_i^α pairs or, in the case of prolines, H_i^α only. The protocol proceeds in a series of cycles, identifying H atoms progressively along the side chain, with probabilities updated from the gained knowledge. The posteriors result from the distance-dependent and chemical shift-dependent likelihoods. Initially, the posteriors for the assignment hypotheses are written according to Eq. 2, with $M = 3$, $H_{\xi_1} = H_{i+1}^N$ (except after Pro residue), $H_{\xi_2} = H_i^N$ (except Pro), and $H_{\xi_3} = H_i^\alpha$. The better defined H^β sites are identified first. In principle, H^β stereospecific assignments are possible, because the empirical PDs for distances between $H_i^{\beta 2}$ and $H_i^{\beta 3}$ to both H_i^N and H_i^α are distinct. In a second round, the probabilities for the remaining assignments are modified by adding $H_{\xi_4} = H_i^{\beta 2}$, $H_{\xi_5} = H_i^{\beta 3}$ terms making $M = 5$, or a $H_{\xi_4} = H_i^\beta$ with $M = 4$, depending on the number of identified H^β atoms. The H^γ assignments then are derived, and the probabilities of the remaining assignments recalculated, along the lines described above. The process is continued until no further assignments result. The strategy for

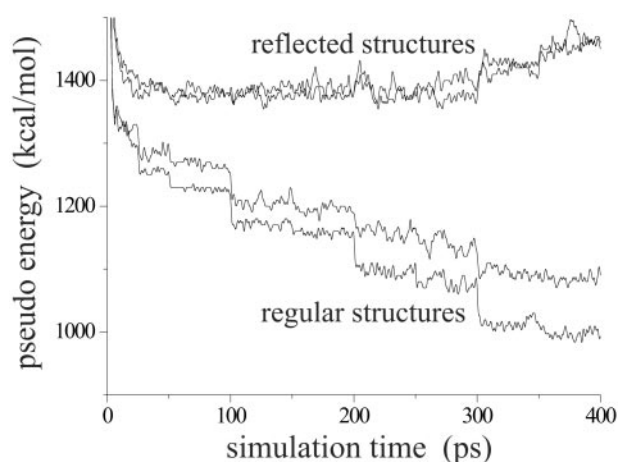


Fig. 2. EMBEDS trajectories for col 2: discrimination between foc and its mirror image. Pairs of MD runs *in vacuo* are shown for fitting to the correct chirality foc and its mirror image. The energy cost function is defined by Eq. 6. Abrupt changes in the cost function stem from stepwise decreases in the values of the Gaussian σ parameter during the protocol (Eq. 5).

the aromatic rings is similar to that applied to aliphatic side chains.

After the atomic foc identification, SIF evaluates cis/trans isomerism of prolyl imino bonds. The expressions for the probabilities are Bayesian as well, with database-estimated priors of 0.15 and 0.85 for cis and trans forms, respectively. The likelihoods are products of likelihoods for the $H_{p-1}^{\alpha}-H_p^{\alpha}$ and $H_{p-1}^{\delta}-H_p^{\delta}$ distances, where the p index identifies the Pro residue.

Structure Computation: EMBEDS. The protein conformation was fitted directly to the foc density; for this, the BAF/SIF-assigned

individual atomic focs $\rho^i(\mathbf{r})$ were digitized on a 0.5-Å cubic grid, normalized to unity, and convolved with a Gaussian function of adjustable width σ , to yield a “continuous” density $\rho_c^i(\mathbf{r})$

$$\rho_c^i(\mathbf{r}) \propto \int d^3\mathbf{r}' \rho^i(\mathbf{r}') \exp\left(-\frac{(\mathbf{r}-\mathbf{r}')^2}{2\sigma^2}\right). \quad [5]$$

A pseudoenergy global cost function, $E(\mathbf{r})$, then was formulated by assuming Boltzmann statistics

$$E(\mathbf{r}) \equiv -k_B T \sum_{i=1}^{N_f} \ln \rho_c^i(\mathbf{r}). \quad [6]$$

with a corresponding force

$$\mathbf{F}(\mathbf{r}) = -\nabla E(\mathbf{r}) = k_B T \sum_{i=1}^{N_f} \frac{\nabla \rho_c^i(\mathbf{r})}{\rho_c^i(\mathbf{r})}. \quad [7]$$

Here, N_f is the number of foc atoms identified by BAF/SIF and incorporated to the structure computation. $\mathbf{F}(\mathbf{r})$ was combined with the PARALLHDG.PRO 4.05 force field, and the geometry was optimized via MD using CNS (17). The empirical force field included bonds, angles, dihedral angles and repulsive-only non-bonded terms.

Smoothing the distribution via application of Eq. 5 facilitates sampling of conformational space as it decreases ruggedness of the pseudoenergy landscape and entrapment in local minima. MD simulations from randomized geometries were carried out for 400 ps at 300 K *in vacuo*, with σ gradually decreased from 2 to 0.5 Å. Thus, by the end of the simulation the target density $\rho_c(\mathbf{r})$ approaches the input foc density $\rho(\mathbf{r})$. During simulation, the radius and force constant of the repulsive nonbonded energy term were increased by an order of magnitude according to a

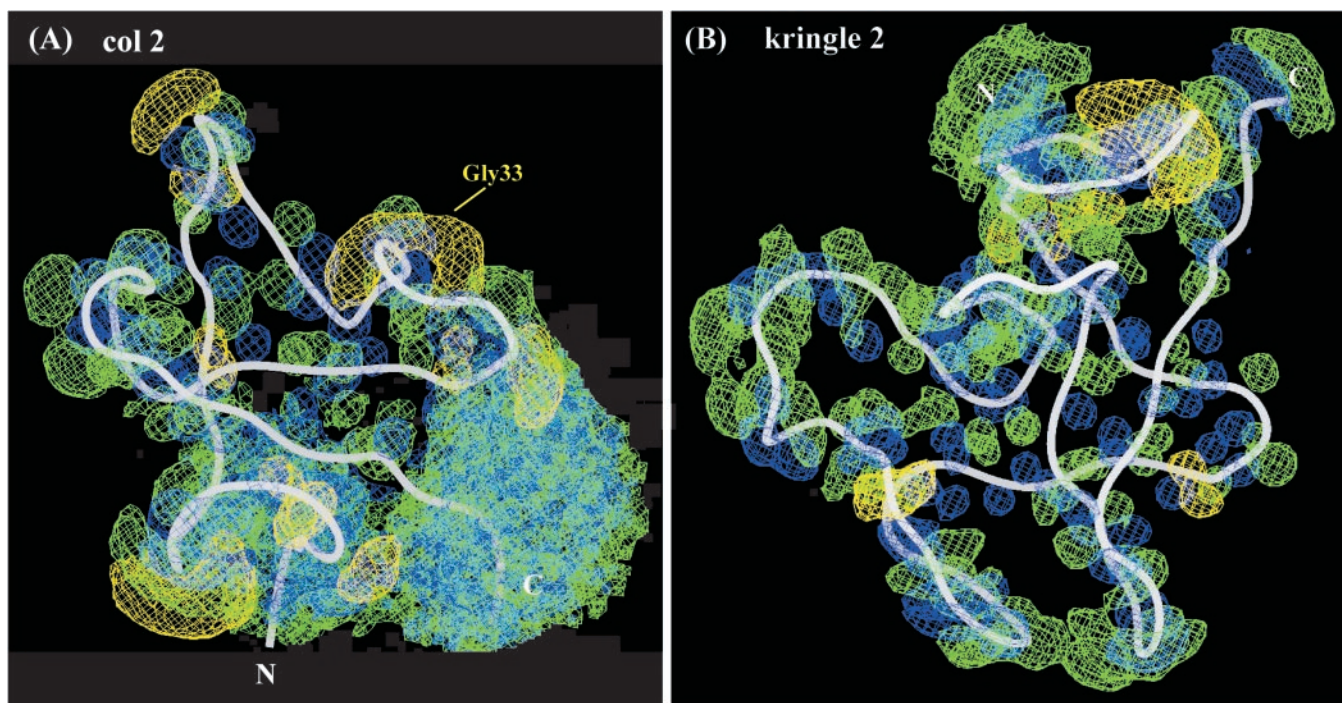


Fig. 3. Gaussian-convolved col 2 (A) and kringle 2 (B) atomic focs: backbone proton density with outline of fitted, energy-minimized structure. The plots were generated by using the program spock (7). Only H^N and H^α focs are shown on a 0.5-Å grid: H^N s are shown in blue, H^α s in green, and Gly H^α s in yellow. The atomic focs were convolved with a $\sigma = 0.5$ Å Gaussian function. The illustrated contour levels enclose 68.5% of each atomic proton density. The col 2 Gly-33 H^α density exhibits two equally populated conformations. The structures shown are those closest to the average.

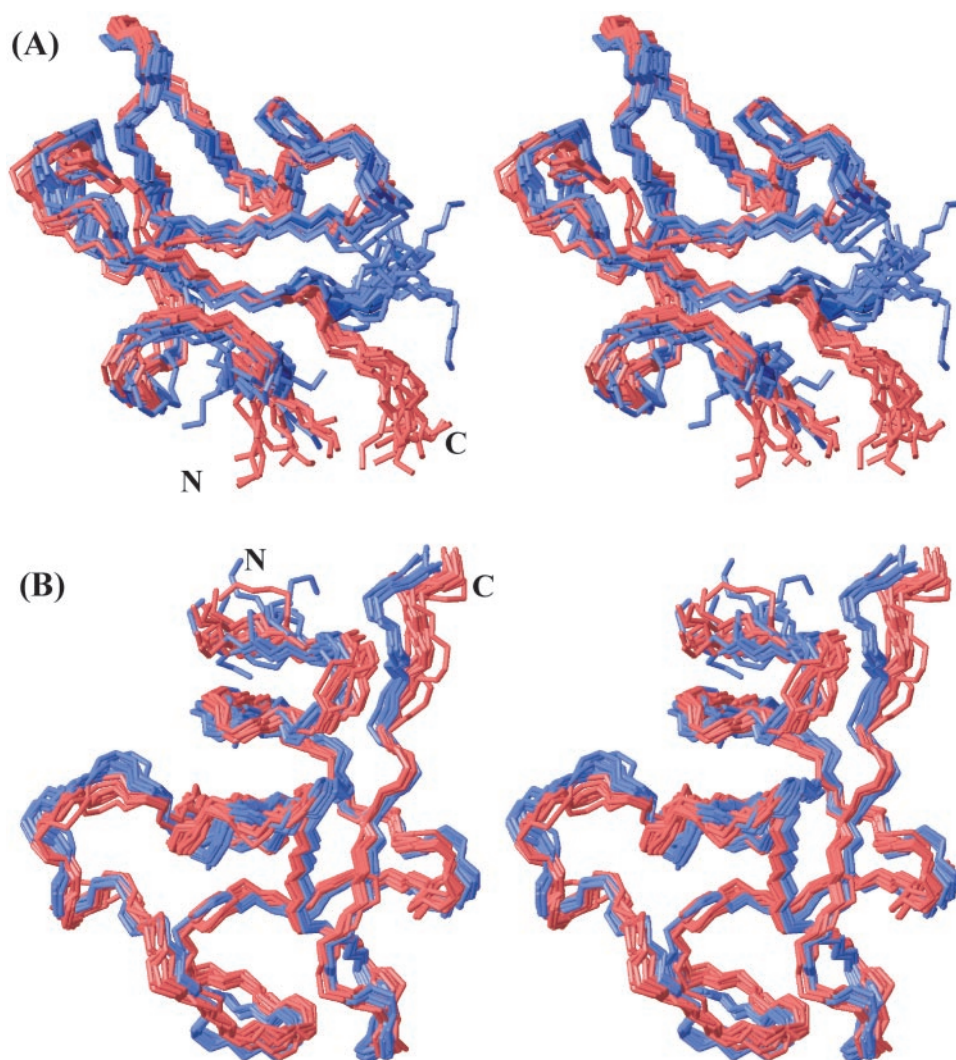


Fig. 4. Comparison of backbone conformation of final EMBEDS-computed CLOUDS structures (red) and previously reported NMR structures (blue; refs. 4 and 5). (A) col 2. (B) kringle 2. Ensembles of 10 structures are shown in stereo.

16-step exponential schedule of the type $y_n = y_0(1 - \alpha^{-n})$, where y_n is the value of the y parameter in n th iteration, y_0 is the target value, and α is chosen for suitable convergence. This schedule was selected on the basis of Ramachandran maps and overall energies of the resulting structures.

Structure Refinement. The *in vacuo* simulation was followed by a 35-ps dynamics in water ($\approx 1,000$ molecules) with pseudoenergy as in Eq. 6, $\sigma = 0.5$ Å, switched electrostatic interactions merged at 8.5 Å, and repulsive/attractive nonbonded potential. In the protocol, the protein was surrounded by an 18.5-Å layer of water molecules as described in ref. 18, heated to 500 K, and annealed to 300 K. To obtain the final structure, the entire system was subjected to 300 cycles of Powell energy minimization with force field (Eq. 7) switched off.

Results and Discussion

Identification of Atomic foci Densities. In the Bayesian identification of the sequential amide links (Eq. 4), the likelihoods strongly favor the sequential hypothesis for $x_{i\ell} < 5$ Å. With regards to H^α atoms, the likelihoods favor both intraresidue and sequential hypotheses whenever $x_n < 7$ Å and $\delta_n \approx 4$ ppm. Characteristically, the distances tend to outweigh the chemical shifts in terms

of their information content; $\mathcal{P}(x_{in} | H_i^\alpha)$ raises the probability of the corresponding hypothesis by a factor of ≈ 50 , whereas $\mathcal{P}(\delta_n | H^\alpha)$ by a factor of only ≈ 3 . However, when only distances-based probabilities were used, some H^α sites were misidentified, a reflection of the high number of non- H^N atoms that tend to derail the selection. Additionally, the inclusion of sequential $H_i^N - H_{i\pm 1}^\alpha$ and $H_i^\alpha - H_{i\pm 1}^\alpha$ distances in deriving the assignment probabilities enabled for a better discrimination between H_i^α and H_{i-1}^α .

Sorting of H^N/H^α pairs via BAF generated 100% correct assignments for both col 2 and kringle 2 foci. The three Pro H^α atoms in col 2 were identified from the immediate upstream and downstream H^N/H^α pairs. For kringle 2, five Pro H^α atoms were detected from both ends, and the remaining two were detected from one end. It is gratifying that for both col 2 and kringle 2, cis and trans isomers of the X-Pro peptide bonds all were recognized correctly: two cis and eight trans.

The analysis of 199 side chain hydrogens in col 2 and 298 protons in kringle 2 via SIF converged in five iterations. In the case of col 2, 10 differed from the reported manual assignments (4). The differences involved $H^{\beta 2} \leftrightarrow H^{\beta 3}$ switches in Phe-4, Pro-14, Phe-17, Pro-18, Cys-29, Arg-34, Arg-39, Asp-48, Lys-51, and Lys-52. Furthermore, in col 2, the Tyr-26 $H^{\beta 3}$ and the Pro-57

H γ were not identified and remained unassigned. In the case of kringle 2, eight side chain assignments differed from the reported assignments (5) because of pairwise switches: H $\beta^2 \leftrightarrow$ H β^3 in Phe-44, Lys-50, Tyr-53, Arg-55, Asp-69, and Asn-71; H $\gamma \leftrightarrow$ H β^2 in Glu-4; H $\epsilon^{21} \leftrightarrow$ H ϵ^{22} in Gln-32; and H $\delta^{21} \leftrightarrow$ H δ^{22} in Asn-42. By reference to previous analyses (4, 5), our success rate for the stereo-specific assignments was 67–86%. Incorporation of additional criteria to discriminate the H β^2 and H β^3 protons such as $^3J_{N-H\beta}$ couplings (19), as used in ref. 5, should be expected to improve the SIF outcome. Similar to the BAF results, the incorporation of the chemical shift-dependent terms in the expressions for the posteriors avoided numerous side chain atom misassignments, which further validates the chemical shift as a powerful criterion for the final selection.

Bayes theorem as formulated by Eq. 1 is valid assuming statistical independence among the various input data ($\mathcal{D}_1, \dots, \mathcal{D}_n$). In our case, although chemical shifts to a large extent can be assumed independent from the interproton distances, the distances can be expected to be correlated weakly among themselves because of the network of constraints imposed by the molecular structure. Despite Eq. 2 ignoring such potential correlations, in practice they did not appear to affect the assignments. A plausible explanation is that the distance PDs derived from the clouds are neither precise nor accurate enough to reflect the interdependencies. Moreover, both clouds-based and database-derived distance PDs were binned to 0.2 Å, which ought to weaken potential correlations further.

Identification of foc Chirality. *A priori*, it is not possible to distinguish a foc proton density from its mirror image, as both are compatible with the NOE-derived distance restraints. The differentiation cannot be obtained during BAF/SIF stages either since these protocols are distance-based as well. Hence, to discriminate between the correct foc and its reflected version, we ran pairs of EMBEDS trajectories *in vacuo* on each of the two proton densities. The EMBEDS cost functions (Eq. 6) for col 2 trajectories are shown in Fig. 2. The results for kringle 2 (not shown) were similar. It is revealing that although the cost function initially drops for the mirror image, it starts to increase as the σ parameter becomes <2 Å (at 100 ps simulation time), indicating a buildup of stress as the L-amino acids fail to accommodate to the reflected proton density. For the nonreflected foc, the fitting score steadily improves as the simulation progresses. Moreover, at any time point in the trajectories, the pseudoenergies are higher for the mirror image. Another parameter that indicates preference for the correct foc chirality is the CNS force field energy, which is 30% lower for the structures fitted to the unreflected foc relative to those fitted to the reflected foc.

When checked against their Ramachandran plots, the foc-fitted structures showed the following statistics (col 2/kringle 2): 53/35% core, 34/42% allowed, 10/15% additional, and 4/7% disallowed. By comparison, the mirror images yielded 14/12% core, 31/38% allowed, 32/31% additional, and 22/19% disallowed. Thus, the quality of both the fits and resulting structures enables for unambiguous discrimination between the two foc chiralities.

Quality of the Fitted Structures. Once the correct foc enantiomer became identified, ensembles of 10 structures were fitted to col 2 and kringle 2 proton densities via a full EMBEDS protocol. The *in vacuo* refinements required 4 h per structure and the water refinement 8 h per structure. Energy minimization in water with the foc potential (Eq. 6) switched off decreased energies by factors of ≈ 2 for bonds and ≈ 1.3 for nonbonded interactions. It should be noted that the procedure generated only minor conformational change (all-atom rmsd ≈ 0.3 Å before/after minimization) or differences in Ramachandran

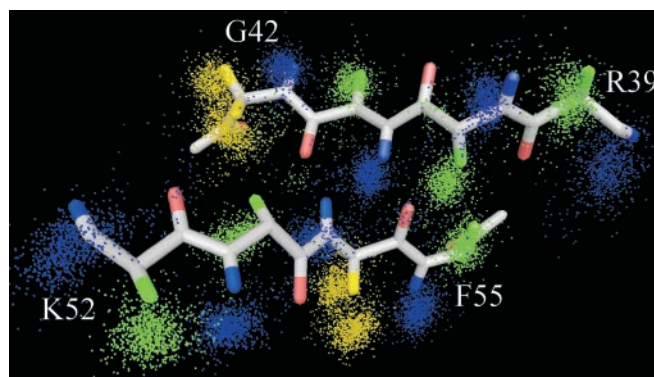


Fig. 5. Fragment of col 2 backbone fitted to the proton density via EMBEDS. H α atoms are shown in blue, non-Gly H α in green, and Gly H α in yellow.

map statistics (1–2%). The polypeptide backbones of energy-minimized structures fitted in water are depicted in Fig. 3, against the backbone H α /H β density.

Fig. 4 shows families of 10 different CLOUDS-derived structures, superimposed onto the previously reported NMR structures (4, 5). In the case of col 2, except for the less determined C-terminal region (residues Pro-57–Ala-60), the differences in backbone conformations are <2.0 Å rmsd (heavy atoms). The differences in the C terminus likely arise from a low number of NOE constraints, which apparently results from enhanced dynamics (backbone order parameters $S^2 = 0.2$ – 0.6 ; ref 4). The col 2 secondary structure elements (two double-stranded antiparallel β sheets, Phe-19–Phe-21/Thr-24–Tyr-26, and Trp-40–Gly-42/Tyr-53–Phe-55 and an α -helical turn, Tyr-47–Asp-50) are present also in the CLOUDS structures. Fig. 5 shows a zoomed view of the Arg-39–Gly-42/Lys-52–Phe-55 β -paired strands. The backbone array is outlined clearly by the atomic foc proton densities, and the secondary structure of the globally fitted conformations closely replicate those shown by the reported NMR (4) or x-ray (20) models. For these segments, the backbone rmsd between the EMBEDS and x-ray geometries is 0.78 Å.

In the case of kringle 2, the differences between CLOUDS and reported structures (5) are mainly in the orientation of the C-terminal region, Cys-77–Thr-83, and in the conformation of the flexible Phe-44–Asn-48 loop. Again, both regions exhibit relatively low numbers of NOEs as well as increased backbone mobilities (5). In contrast, it is gratifying that the left-handed 3_1 -helix delineated by segment 74–79 is also obtained by CLOUDS. Furthermore, notwithstanding that some atomic foci of Trp rings are irregular (3), the geometries of the hydrophobic cluster neighboring the lysine-binding site (side chains of Trp-25, Leu-46, Trp-62, and Trp-72) are remarkably close to those exhibited by the reported structure (ref. 5; Fig. 6).

Although the above results demonstrate that an NOE-only strategy for deriving macromolecular structures is feasible,

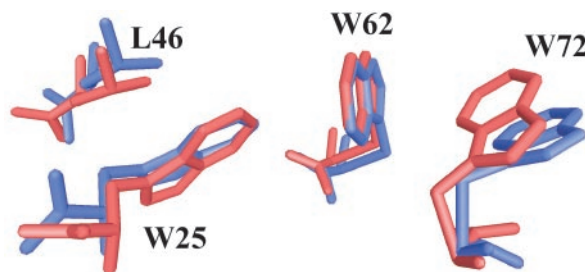


Fig. 6. Hydrophobic cluster neighboring the kringle 2 lysine-binding site: final CLOUDS/EMBEDS (red) versus reported (blue; ref. 5) structures.

Table 1. Structure statistics

	rmsd, Å		Ramachandran map occupancy, %*			
	Backbone heavy atoms	All heavy atoms	Most favored	Additional allowed	Generously allowed	Disallowed
Col 2						
CLOUDS [†]	0.6 ± 0.1	1.0 ± 0.1	68.7	29.8	1.3	0.2
Reported NMR (7) [†]	0.8 ± 0.2	1.3 ± 0.2	73.7	21.6	2.9	1.7
CLOUDS vs. NMR (7) [‡]	1.0 ± 0.3	1.5 ± 0.4				
CLOUDS vs. x-ray (23) [‡]	1.3 ± 0.2	2.1 ± 0.2				
Kringle 2						
CLOUDS*	0.6 ± 0.1	1.0 ± 0.2	55.4	35.1	4.5	5.1
Reported NMR (8) [†]	0.5 ± 0.1	1.0 ± 0.1	61.3	35.9	2.8	0.0
CLOUDS vs. NMR (8) [‡]	1.4 ± 0.4	2.0 ± 0.6				

*Ramachandran map statistics were generated with PROCHECK 3.4.4 (21).

[†]Rmsds are shown relative to the means.

[‡]Rmsds are mean to mean.

recent developments have argued convincingly for the use of residual dipolar couplings (22), chemical shifts indices (23), *J* coupling-derived dihedral angles (24), and pseudocontact paramagnetic shifts (25) to refine NMR structures. Such constraints can be incorporated readily to the CLOUDS protocol to enhance the quality of the EMBEDS-derived structures without compromising the method's main appeal, namely, the avoidance of spectral assignment. Furthermore, when dealing with larger molecules for which spectral resolution can be problematic, the pool of unambiguous distance constraints potentially can be enlarged, e.g., by identifying NOEs through dynamic sampling schemes (26, 27). For this purpose, a Bayesian protocol that hinges on self-consistency of NOE identities may be applicable also.

Summary. Analogous to fitting the protein conformation to the electron density obtained from x-ray diffraction experiments, the global fitting to the complete foci largely compensates for locally irregular or underdefined geometries, particularly crucial in the

case of side chains. The robustness of the protocol is exemplified by the Gly-33 of col 2, where although the foci shows two distinct, equally probable distributions for the H^αs (Fig. 3A), in all 10 computed structures EMBEDS discriminates in favor of the correct geometry vis-à-vis the reported structures.

The quality statistics of the derived col 2 and kringle 2 structures are summarized in Table 1. While by reference to those previously determined by x-ray (20) and NMR (4, 5) the accuracies of CLOUDS/EMBEDS structures are slightly lower than their respective precisions, these structures are comparable to those previously determined by NMR both in terms of their precision and Ramachandran map statistics.

Because it bypasses the assignment bottleneck, an automated CLOUDS protocol should be appealing for high-throughput NMR characterization of novel folds addressed by structural proteomics research.

This research was sponsored by the U.S. Public Health Service, National Institutes of Health Grant HL-29409.

- Moseley, H. & Montelione, G. T. (1999) *Curr. Opin. Struct. Biol.* **9**, 635–642.
- Zimmerman, D. E., Kulikowski, C. A., Huang, Y. P., Feng, W. Q., Tashiro, M., Shimotakahara, S., Chien, C. Y., Powers, R. & Montelione, G. T. (1997) *J. Mol. Biol.* **269**, 592–610.
- Grishaev, A. & Llinás, M. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 6707–6712.
- Briknarová, K., Grishaev, A., Banyai, L., Tordai, H., Patthy, L. & Llinás, M. (1999) *Structure (London)* **7**, 1235–1245.
- Marti, D. N., Schaller, J. & Llinás, M. (1999) *Biochemistry* **38**, 15741–15755.
- Koradi, R., Billeter, M. & Wüthrich, K. (1996) *J. Mol. Graphics* **14**, 51–55.
- Christopher, J. A. & Baldwin, T. O. (1998) *J. Mol. Graphics* **16**, 285–285.
- Kendall, M. G. (1958) *The Advanced Theory of Statistics* (Hafner, New York, NY), Vol. 2.
- James, T. L., Liu, H., Ulyanov, N. B., Farr-Jones, S., Zhang, H., Donne, D. G., Kaneko, K., Groth, D., Mehlhorn, I., Prusiner, S. B. & Cohen, F. E. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 10086–10091.
- Cornilescu, G., Marquardt, J. L., Ottiger, M. & Bax, A. (1998) *J. Am. Chem. Soc.* **120**, 6836–6837.
- Berndt, K. D., Guntert, P., Orbons, L. P. & Wüthrich, K. (1992) *J. Mol. Biol.* **227**, 757–775.
- Zahn, R., Liu, A., Luhrs, T., Calzolari, L., Von Schroetter, C., Garcia, F. L., Riek, R., Wider, G., Billeter, M. & Wüthrich, K. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 145–150.
- Ikura, M., Clore, G. M., Gronenborn, A. M., Zhu, G., Klee, C. B. & Bax, A. (1992) *Science* **256**, 632–638.
- Qi, P. X., Beckman, R. A. & Wand, A. J. (1996) *Biochemistry* **35**, 12275–12286.
- Clore, G. M., Wingfield, P. T. & Gronenborn, A. M. (1991) *Biochemistry* **30**, 2315–2323.
- Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Lesme, V., Blessing, B. & Lecomte, C. (2000) *Proc. Nat. Acad. Sci. USA* **97**, 3171–3176.
- Brünger, A., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Panno, N. S., et al. (1998) *Acta Crystallogr. D* **54**, 905–921.
- Linge, J. P. & Nilges, M. (1999) *J. Biomol. NMR* **13**, 51–59.
- Dux, P., Whitehead, B., Boelens, R., Kaptein, R. & Vuister, G. W. (1997) *J. Biomol. NMR* **10**, 301–306.
- Morgunova, E., Tuuttila, A., Bergmann, U., Isupov, M., Lindqvist, Y., Schneider, G. & Tryggvason, K. (1999) *Science* **284**, 1667–1670.
- Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996) *J. Biomol. NMR* **8**, 477–486.
- Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M. & Bax, A. (1997) *Nat. Struct. Biol.* **4**, 732–738.
- Wishart, D. S., Sykes, B. D. & Richards, F. M. (1992) *Biochemistry* **31**, 1647–1651.
- Schmidt, J. M., Blumel, M., Lohr, F. & Ruterjans, H. (1999) *J. Biomol. NMR* **14**, 1–12.
- Banci, L., Bertini, I., Bren, K. L., Cremonini, M. A., Gray, H. B., Luchinat, C. & Turano, P. (1996) *J. Biol. Inorg. Chem.* **1**, 117–126.
- Nilges, M., Macias, M. J., O'Donoghue, S. I. & Oschkinat, H. (1997) *J. Mol. Biol.* **269**, 408–422.
- Mumenthaler, C., Guntert, P., Braun, W. & Wüthrich, K. (1997) *J. Biomol. NMR* **10**, 351–362.