A DATA-DRIVEN, SYSTEMATIC SEARCH ALGORITHM FOR STRUCTURE DETERMINATION OF DENATURED OR DISORDERED PROTEINS

Lincong Wang
Dartmouth Computer Science Department
Hanover, NH 03755, USA
Email: wlincong@cs.dartmouth.edu

Bruce Randall Donald*†
Dartmouth Computer Science Department, Dartmouth Chemistry Department
Dartmouth Department of Biological Sciences
Hanover, NH 03755, USA
Email: brd@cs.dartmouth.edu

Traditional algorithms for the structure determination of native proteins by solution nuclear magnetic resonance (NMR) spectroscopy require a large number of experimental restraints. These algorithms formulate the structure determination problem as the computation of a structure or a set of similar structures that best fit the restraints. However, for both laboratory-denatured and natively-disordered proteins, the number of restraints measured by the current NMR techniques is well below that required by traditional algorithms. Furthermore, there presumably exists a heterogeneous set of structures in either the denatured or disordered state. We present a data-driven algorithm capable of computing a set of structures (ensemble) directly from sparse experimental restraints. For both denatured and disordered proteins, we formulate the structure determination problem as the computation of an ensemble of structures from the restraints. In this formulation, each experimental restraint is a distribution. Compared with previous algorithms, our algorithm can extract more structural information from the experimental data. In our algorithm, all the backbone conformations consistent with the data are computed by solving a series of low-degree monomials (yielding exact solutions in closed form) and systematic search with pruning. The algorithm has been successfully applied to determine the structural ensembles of two denatured proteins, acyl-coenzyme A binding protein (ACBP) and eglin C, using real experimental NMR data.

1. INTRODUCTION

The protein folding problem is fundamental in structural biology. It can be stated as the problem of elucidating how a protein can fold, in less than a second, from the denatured state to the native state. One challenge to solving the folding problem is the lack of knowledge about the structures of proteins in the *denatured state*. In this paper, "denatured state" means the state in which the backbone NH groups have little protection against ^{1}H / ^{2}H-exchange. It has been estimated that about one-third of eukaryotic proteins are disordered or partially-disordered in their native state in solution. Such natively-disordered proteins play key roles in signal transduction and genetic regulation as well as in human diseases such as Alzheimer's and Parkinson's diseases. Although much progress has been made in understanding how the structure defines the biological function of native proteins, it is not well-known how the structures of disordered proteins determine their function. For denatured proteins, an accurate, quantitative structural distribution is key to solving the protein-folding problem [27, 20], while for disordered proteins, an accurate distribution of the structures is critical for establishing the structure-function relationship. To quantify the structural distribution, it is necessary to compute the ensemble of structures directly from experimental data. At present, NMR[a] is the only available technique that can measure many individual structural restraints for these proteins. However, even using the most advanced NMR techniques, the number of measured restraints is well below that required by traditional

*Corresponding author
†This work is supported by the following grants to B.R.D.: National Institutes of Health (R01 GM 65982) and National Science Foundation (EIA-0305444).
[a]Abbreviations used: NMR, nuclear magnetic resonance; PRE, paramagnetic relaxation enhancement; RDC, residual dipolar coupling; CH, the bond vector between backbone C_α and H_α; NH, the bond vector between backbone amide nitrogen and amide proton; CC′, the bond vector between backbone C_α and C′; NC′, the bond vector between backbone amide nitrogen and C′; RMSD, root-mean squared deviation; NOE, nuclear Overhauser effect; POF, principal order frame; SVD, singular value decomposition; ACBP, acyl-coenzyme A binding protein; vdW, van der Waals; MD, molecular dynamics; SA, simulated annealing.

NMR structure determination methods [5, 11]. Furthermore, these methods formulate the structure determination problem as the computation of a structure (or a set of similar structures) that best fit the restraints. Such a formulation is appropriate for the structure determination of a native protein having a single dominant conformation. However, a new formulation is necessary for computing the structures of either denatured or disordered proteins, which are presumably heterogeneous in solution [19, 24]. In this paper, we first formulate the structure determination problem of both denatured and disordered proteins as the determination of a heterogeneous ensemble of structures, from sparse experimental restraints measured in either the denatured or disordered state. In this formulation, the restraints are *distributions*. We then present a data-driven algorithm capable of accurately-computing denatured backbone structures directly from sparse restraints. In our algorithm, the conformational space consistent with the data is searched systematically, rather than randomly as in previous approaches [14, 4, 16, 6]. The algorithm uses considerably more experimental data than previous approaches for characterizing the denatured state from experimental data [16, 14, 4]. The larger amount of data, together with the systematic search, significantly increase the accuracy of the computed ensembles. In the following, we only present the algorithm and application to denatured proteins. The algorithm can be applied to natively-disordered proteins as well. Our contributions are:

(1) A new formulation of the structure determination problem for denatured proteins.
(2) A data-driven, systematic search algorithm for computing an ensemble of all-atom backbone structures for denatured proteins directly from experimental data.
(3) Successful application of the algorithm to compute the structure ensembles of two denatured proteins from real, biological NMR data.

This paper concentrates on the computer science aspects of the algorithm. We will only describe briefly the applications of the algorithm to two real biological systems. The biological significance of our results and the use of the computed ensembles to understand protein folding will be addressed in detail in another paper.

1.1. Organization of the paper

We begin with a probabilistic interpretation of NMR data in the denatured state in terms of equilibrium statistical physics. Section 3 presents a formulation of the structure determination problem of denatured proteins using experimental NMR data such as the orientational restraints from residual dipolar coupling (RDC) [25, 12] and distance restraints from paramagnetic relaxation enhancement (PRE) [1] experiments. Section 4 reviews existing approaches. Section 5 presents the mathematical basis of the algorithm. Section 6 describes our algorithm for computing an ensemble of structures. Section 7 presents briefly the results of applying our algorithm to compute the structural ensembles of two denatured proteins, acyl coenzyme A binding protein (ACBP) and eglin C, from real, experimental NMR data. Finally, in section 8 we analyze the complexity of the algorithm and describe its performance in practice.

2. A PROBABILISTIC INTERPRETATION OF RESTRAINTS IN THE DENATURED STATE

Our algorithm first computes both the backbone dihedral angles and the orientation of each structural fragment[b] independently using the orientational restraints from RDCs, and then assembles the computed structural fragments into a complete structure using the distance restraints from PREs. Our algorithm is based on a new formulation for structure determination in which each experimental restraint is converted to a distribution. In the following, we present the physical basis for the formulation.

RDCs can be measured on proteins weakly-aligned in a dilute liquid crystal medium [25, 26]. The RDC, r, between two nuclear spins is related to the direction of the corresponding internuclear unit vector $\mathbf{v} = (x, y, z)$ by [22],

$$r = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2 \qquad (1)$$

where S_{xx}, S_{yy} and S_{zz} are the three diagonal elements of a diagonalized Saupe matrix $\mathbf{S}$ (the alignment tensor) specifying the ensemble-averaged anisotropic orientation of a molecule in the laboratory frame; x, y and z are, respectively, the $x, y, z-$components of $\mathbf{v}$ in a *principal*

[b] A structural fragment consists of m-consecutive residues; typically $m \approx 10$.

order frame (POF) which diagonalizes $\mathbf{S}$. Before diagonalization, $\mathbf{S}$ is a 3×3 symmetric, traceless matrix with five independent elements. Note that $x^2 + y^2 + z^2 = 1$ and $S_{xx} + S_{yy} + S_{zz} = 0$. Thus, given both the RDC r and tensor $\mathbf{S}$, Eq. (1) represents the projection onto a 2-sphere of an ellipse of solutions for the orientation of the vector $\mathbf{v}$ with respect to a global frame (POF) *common* to all the RDCs measured on the same aligned protein. Tensor $\mathbf{S}$ must be known first in order to extract orientational restraints from RDC data. RDCs alone or in combination with other NMR-measured geometric restraints have been used extensively to determine and refine the solution structures of native proteins [7, 29].

Recently, it has been shown that RDCs can also be measured accurately on weakly-aligned, denatured proteins [23, 2, 18, 9] and disordered proteins [4]. For a folded, native n-residue protein, a *single* global tensor, $\mathbf{S}$, can be used to interpret all the experimental RDCs by Eq. (1). However, according to equilibrium statistical physics [15], a *set* of tensors, $\mathbf{Q}$, is required to interpret the RDCs measured in the denatured state. Each tensor in the set $\mathbf{Q}$ represents a cluster of denatured structures that have similar structures and align similarly in the medium. The set of RDCs corresponding to each tensor in the set $\mathbf{Q}$ can be sampled from the individual distributions associated with each measured RDC. The distribution for each RDC can be defined by an *RDC random variable* that has as its sampling space the RDCs of all the orientations of the corresponding vector $\mathbf{v}$ in different structures that exist in the denatured state. The experimentally-measured RDC value is the *expectation*. The different tensors in the set $\mathbf{Q}$ represent different conformations in the denatured state that are oriented differently in the aligning medium. The tensor $\mathbf{S}$ is also a random variable.

Paramagnetic relaxation enhancement (PRE) is similar to the nuclear Overhauser effect (NOE) [30] in terms of physics[c]. However, PRE can be observed even in the denatured state between an electron spin and a nuclear spin as far as 20 Å away, while no NOE between two nuclear spins can be observed at such a distance. The reason is that PRE is almost 2,000-fold stronger than the NOE. In fact, long-range NOEs, which are critical for computing structures using traditional methods,[5, 11] are generally too weak to be detected on denatured proteins. The PRE-derived distance, d, in the denatured state, is also a random variable, where the measured value is an average over all the possible structures in the denatured state.

3. THE STRUCTURE DETERMINATION PROBLEM FOR DENATURED PROTEINS

As is well known, given bond length, bond angle and peptide plane ω angle, the backbone conformation of an n-residue protein is completely determined by a $2n$-tuple of backbone dihedral angles, $\mathbf{c}_n = (\phi_1, \psi_1, \ldots, \phi_n, \psi_n)$, where (ϕ_i, ψ_i) are the dihedral angles of residue i. This $2n$-tuple will be called a *conformation vector*, $\mathbf{c}_n$. In fact, the sines and cosines of the $2n$ (ϕ, ψ) angles are sufficient to determine a backbone conformation. The structure determination problem for a denatured protein is to compute an ensemble of presumably heterogeneous structures that are consistent with the experimental data within a relative large range for the data. More precisely, the structure determination problem for denatured proteins can be formulated as the computation of a set of conformation vectors, $\mathbf{c}_n$, given the distributions for all the RDCs r and for all the PREs d.

4. PREVIOUS WORK

Solution NMR spectroscopy is the only experimental technique currently capable of measuring geometric restraints for individual residues of a denatured protein at the atomic level. Traditional NMR structure determination methods [5, 11], developed for computing structures in the native state, require more than 10 restraints per residue, derived mainly from NOE experiments, to compute a well-defined native structure. Recently developed RDC-based approaches for computing native structures rely on either heuristic approaches such as restrained molecular dynamics (MD) and simulated annealing (SA) [10, 13] or a structural database [8, 21]. It is not clear how to extend these native structure determination approaches to compute the desired denatured structures. Traditional NOE-based approaches cannot be used since long-range NOEs, which are critical for applying the traditional approaches to determine NMR structures, are usually too weak to be detected in the denatured

[c]The main difference between PRE and NOE is that PRE results from the dipole-dipole interaction between an electron and a nucleus while the physical basis of NOE is the dipole-dipole interaction between two nuclei. Under the isolated two spin assumption, both PRE and NOE (that is, the observed intensity of cross-peaks in either a PRE or NOE experiment) are proportional to r^{-6} where r is the distance between two spins.

state.[d] Previous RDC-based MD/SA approaches typically require either more than 5 RDCs per residue or at least 3 RDCs and 1 NOE per residue (most of them should be long-range) to compute a well-defined native structure. In the database-based approaches, RDCs are employed to select structural fragments mined from the protein databank (PDB) [3], a database of experimentally-determined *native* structures. A backbone structure for a native protein is, then, constructed by linking together the RDC-selected fragments using a heuristic method. Compared with the MD/SA approaches, the database-based approaches require fewer RDCs. However, these database-based approaches have not been extended to compute structures for denatured proteins. In summary, neither the traditional NOE-based methods nor the above RDC-based approaches can be applied to compute all-atom backbone structures in the denatured state at this time.

Recently, approaches [14, 4] have been developed to build structural models for the denatured state using one RDC per residue. These approaches are generate-and-test. They begin with the construction of a library of backbone (ϕ, ψ) angles using only the angles occurring in the loops of the *native* proteins deposited in the PDB. Then, they *randomly* select (ϕ, ψ) angles from the library to build an ensemble of backbone models. Finally, the models were tested by comparing the experimental RDCs with the average RDCs back-computed from the ensemble of backbone structures. There are three problems with these methods. First, the (ϕ, ψ) angle library is biased since only the (ϕ, ψ) angles from the loops of the *native* proteins are used. Consequently, the models constructed from the library may bias towards the native conformations in the PDB. Second, random selection may miss valid conformations. Third, the agreement of the experimental RDCs with the average RDCs back-computed from the ensemble of structures may result from over-fitting. Over-fitting is likely since one RDC per residue is not enough to restrain the orientation of an internuclear vector (such as the NH bond vector) to a finite set. In fact, given an alignment tensor $\mathbf{S}$, an infinite number of backbone conformations can agree with one RDC per residue, while only a finite number of conformations agree with two RDCs per residue [29, 28, 31].

All-atom models for the denatured state have been computed previously in a generate-and-test manner in [16] by using PREs to select the structures from all-atom MD simulation at high temperature. Due to the data sparsity and large experimental errors, PREs alone are, in general, insufficient to define precisely even the backbone C_α-trace. The generated models have large uncertainty. A generate-and-test approach [6] using mainly NOE distance restraints has been developed to determine the ensemble of all-atom structures of an SH3 domain in the unfolded state in equilibrium with a folded state.[e] However, the relatively large experimental errors as well as the sparsity and locality of NOEs similarly introduce large uncertainty in the resulting ensemble of structures, which was selected mainly by the NOEs.

5. THE MATHEMATICAL BASIS OF OUR ALGORITHM

Our algorithm uses a set of low-degree ($\leq$4) monomials for computing, *exactly* and *in constant time*, the sines and cosines of individual backbone dihedral (ϕ, ψ) angles. These monomials have been derived from the RDC equation (1) and protein backbone kinematics, and have been described in detail elsewhere [29, 28, 31]. In the following, for ease of exposition, we state the monomials for computing, respectively, the sine and cosine of the backbone ϕ angle from a CH RDC and those of ψ angle from an NH RDC [28]. NH and CH RDCs denote, respectively, the RDCs measured on NH and CH bond vectors. Starting with peptide plane i, we can compute the sines and cosines of the ϕ_i, ψ_i angles, respectively, from the CH RDC of residue i and the NH RDC of residue $i+1$ using the following two Propositions:

Proposition 5.1 [28] *Given the orientation of peptide plane i in the POF (see section 2) of RDCs, the x-component of the CH unit vector $\mathbf{u}$ of residue i, in the POF, can be computed from the CH RDC by solving a quartic monomial in x. Given the x-component, the y-component can be computed from Eq. (1), and the z-component from $x^2 + y^2 + z^2 = 1$. Given $\mathbf{u}$, the sine and cosine of the ϕ_i angle can be computed by solving linear equations.*

[d] The denatured state in this paper (see section 1) has been called the "unfolded state" [20].

[e] An unfolded state in equilibrium with a folded state [6] differs from the *denatured state* in this paper. In [6], the observed NOEs result from the equilibrium between the folded and unfolded states, not from the unfolded state alone.

Proposition 5.2 [28] *Given the orientation of peptide plane i in the POF of RDCs, the x-component of the NH unit vector* $\mathbf{v}$ *of residue $i+1$, in the POF, can be computed from the NH RDC by solving a quartic monomial in x. Given the x-component, the y-component can be computed from Eq. (1), and the z-component from $x^2+y^2+z^2=1$. Given* $\mathbf{v}$*, the sine and cosine of the ψ_i angle can be computed by solving linear equations.*

According to Propositions 5.1–5.2, given the orientation of peptide plane i (plane i stands for the peptide plane for residue i in the protein sequence), the sines and cosines of the backbone ϕ_i, ψ_i angles can be computed, *exactly* in closed form, from the CH RDC of residue i and the NH RDC of residue $i+1$. Furthermore, the orientation of the peptide plane for residue $i+1$ can be computed, *exactly* in closed form, from the orientation of the peptide plane for residue i and the sines and cosines of the intervening ϕ_i, ψ_i angles. Thus, given a tensor $\mathbf{S}$, the orientation of the peptide plane for residue 1 (the first peptide plane) of the protein sequence, and CH and NH RDCs, all the sines and cosines of backbone (ϕ, ψ) angles can be computed from RDCs by solving a series of quartic and linear equations. Thus, the set of conformations consistent with two RDCs per residue is finite and algebraic. In conclusion, given bond length, bond angle, peptide plane ω angle, and the orientation of the first peptide plane as well as a tensor $\mathbf{S}$, and a set of two RDCs per residue sampled from the RDC distributions (see section 2), a finite and algebraic set of backbone conformations can be determined exactly. Furthermore, this set of conformations can be computed by a systematic search such as a depth-first search over a k-ary tree where $k \leq 64$, the maximum number of solutions for (ϕ, ψ) angles for a single residue [28, 29]. Taken together, we have stated the mathematical basis of our algorithm, that is, an ensemble of denatured structures can be computed exactly by solving a series of monomials each with degree ≤ 4 using different sets of two RDCs per residue sampled from their distributions and the corresponding tensors $\mathbf{S}$ from the set $\mathbf{Q}$.

6. AN ALGORITHM FOR STRUCTURE DETERMINATION OF DENATURED PROTEINS

Our algorithm for computing the structure ensemble of a *denatured* protein extends but differs substantially from our previous algorithms [29, 28, 31] for computing the backbone structures of *native* proteins. The goal of the present algorithm is to compute a presumably heterogeneous *ensemble* of structures that are consistent with the experimental data within a large range, rather than a *single* structure or a set of similar structures that best fits the data (as in the native state). For the native state, a *single* tensor, $\mathbf{S}$, can be used to interpret all the experimental RDCs by Eq. (1). Moreover, for native proteins, this single tensor can be determined during structure computation (if secondary structure elements are known [29, 28]). However, it is physically infeasible to use a single tensor to interpret all the experimental RDCs on a denatured protein (see section 2). Rather one should use a set, $\mathbf{Q}$, of different tensors to compute all the possible different conformations in the denatured state. This set of tensors $\mathbf{Q}$ is updated continuously during the structure computation. Our algorithm computes the ensemble using a divide-and-conquer strategy for efficiency.

6.1. Divide-and-conquer strategy

The algorithm first divides the entire protein sequence into p fragments, $\mathrm{F}_1, \ldots, \mathrm{F}_p$, and $p-1$ linkers, $\mathrm{L}_1, \ldots, \mathrm{L}_{p-1}$ (Fig. 1). A linker consists of the residues between two neighboring fragments. Next, the algorithm computes, *independently*, an ensemble of structures, $\mathbf{W}_i$, for each fragment i where $i = 1, \ldots, p$. This step is called *Fragment computation* (Fig. 2) and will be detailed in Section 6.2. Next, for each structure in ensemble $\mathbf{W}_i$, $i = 1, \ldots, p$, we compute the corresponding tensor $\mathbf{t}_i$ by singular value decomposition (SVD) [17] and save each $\mathbf{t}_i$ into a set $\mathbf{T}_i$. Given a structure and the experimental RDCs, a tensor $\mathbf{S}$ can be computed by using SVD to minimize the RDC RMSD, $E_r = \sqrt{\frac{\sum_{j=1}^{u}(r'_j - r_j)^2}{u-1}}$, where u is the total number of RDCs for fragment F_i, r_j and r'_j are, respectively, the experimental RDC for residue j of F_i and the RDC back-computed from the structure using the tensor $\mathbf{S}$ by Eq. (1). As shown in Eq. (1), given a structure, r'_j is a function of $\mathbf{S}$ so by minimizing E_r, $\mathbf{S}$ can be computed by SVD [17]. Next, the algorithm merges all the tensors in the sets $\mathbf{T}_i$, $i = 1, \ldots, p$, into p-tuples, $(\mathbf{t}_1, \ldots, \mathbf{t}_p)$, such that $\mathbf{t}_i$ is from the set $\mathbf{T}_i$ and all p tensors in a p-tuple have their S_{yy} and S_{zz} values agree with one another up to the ranges defined by $[S_{yy}-\delta_{yy}, S_{yy}+\delta_{yy}]$ and $[S_{zz}-\delta_{zz}, S_{zz}+\delta_{zz}]$ where δ_{yy} and δ_{zz} are thresholds. For each merged p-tuple, the algo-

rithm then computes their common tensor by SVD using the corresponding structures in $\mathbf{W}_i$, $i = 1, \ldots, p$ and all the experimental RDCs for F_i, $i = 1, \ldots, p$, and saves the common tensors into a set $\mathbf{Q}$. The diagonalization of the tensor returned from SVD gives not only the diagonal elements, S_{xx}, S_{yy} and S_{zz}, but also the orientation for each fragment in the common POF as well. In particular, the orientations of all the peptide planes in the POF are returned from SVD where the first and last peptide planes are used for computing (ϕ, ψ) angles from RDCs by Propositions 5.1–5.2. Finally, the algorithm computes the linkers, $L_1, \ldots, L_{p-1}$, using every common tensor in $\mathbf{Q}$ and assembles the corresponding fragments and linkers into complete backbone structures. This step is called *Linker computation and assembly* (Fig. 3) and will be detailed in Section 6.3.

6.2. Fragment computation

A structure ensemble, $\mathbf{W}_i$, of an m-residue fragment F_i is computed as follows (Fig. 2). First, the algorithm estimates an initial tensor $\mathbf{S}_{0,1}$ by SVD using experimental RDCs and a model built with the backbone (ϕ, ψ) angles for polyprolineII. The algorithm then selects b different sets of RDCs, $R_1, \ldots, R_b$, for the fragment by randomly sampling CH and NH RDC values from their respective normal distributions. Next, for each R_t, $t = 1, \ldots, b$, the algorithm computes an optimal conformation vector, $\mathbf{c}_{1t}$, by systematically searching over all the possible conformation vectors, $\mathbf{c}_m$ of $2m$-tuples $(\phi_1, \psi_1, \ldots, \phi_m, \psi_m)$, computed from R_t where the ϕ_k angle for residue k is computed according to Proposition 5.1 from the sampled CH RDC for residue k, and the ψ_k angle is computed according to Proposition 5.2 from the sampled NH RDC for residue $k+1$. An optimal conformation vector is a vector which has the minimum score under a scoring function $T_{_F}$ defined as

$$T_{_F} = E_r^2 + w_v E_v^2 \tag{2}$$

where $E_r = \sqrt{\frac{\sum_{j=1}^{u}\sum_{k=1}^{m}(r'_{j,k} - r_{j,k})^2}{um-1}}$ is the RDC RMSD, u is the number of RDCs for each residue, $r_{_{j,k}}$ and $r'_{_{j,k}}$ are, respectively, the experimental RDC for RDC j of residue k, and the corresponding RDC back-computed from the structure. The variables w_v and E_v are, respectively, the relative weight and score for van der Waals (vdW) repulsion. For each conformation vector $\mathbf{c}_m$ of a fragment, E_v is computed with respect to a quasi-polyalanine model built with $\mathbf{c}_m$. The quasi-polyalanine model consists of alanine, glycine and proline residues with proton coordinates. If a residue is neither a glycine nor a proline in the protein sequence, it is replaced with an alanine residue. If the vdW distance between two atoms computed from the model is larger than the minimum vdW distance between the two atoms, the contribution of this pair of atoms to E_v is set to zero. Since the (ϕ, ψ) angles are computed from the sampled CH and NH RDCs by exact solution, the back-computed NH and CH RDCs are in fact the same as their sampled values. For additional RDCs (CC′ or NC′ RDCs), E_r is minimized as cross-validation using Eq. (2). For each sampled set of RDCs, $R_t, t = 1, \ldots, b$, the output of this systematic search step is the optimal conformation vector $\mathbf{c}_{1t}$ in Fig. 2. The search step is followed by an SVD step to update tensors, $\mathbf{S}_{1t}$, using the experimental RDCs and the just-computed fragment structure. Next, the algorithm repeats the cycle of systematic search followed by SVD (systematic-search/tensor-update) to compute a new ensemble of structures using each of the newly-computed tensors, $\mathbf{S}_{1t}, t = 1, \ldots, b$. The output of the fragment computation for a fragment i is a set of conformation vectors $\mathbf{c}_{hw}, w = 1, \ldots, b^h$, where h is the number of the cycles of systematic search/tensor-update.

6.3. Linker computation and assembly

Given a common tensor $\mathbf{S}$ in set $\mathbf{Q}$ and the orientations of two fragments F_1 and F_2 in the POF for $\mathbf{S}$, an m-residue linker L_1 between them is computed as shown in Fig. 3. The computation of a linker can start from either its N-terminus as detailed in Fig. 3 or from its C-terminus, depending on the availability of experimental data. For the latter, the interested reader can see the Propositions 10.1 and 10.2 (section 10 of APPENDIX) for the detail. Every two consecutive fragments are assembled (combined), *recursively*, into a single fragment and the process stops when all the fragments have been assembled. The scoring function for the linker computation, $T_{_L}$, is computed similarly to $T_{_F}$.

$$T_{_L} = E_r^2 + w_v E_v^2 + w_p E_p^2 \tag{3}$$

The main difference is that E_v for a linker is computed with respect to an individual structure composing of all the previously-computed and linked fragments, and the current linker built with the backbone (ϕ, ψ) angles computed from RDCs. In addition, the PRE violation, E_p, which is essentially the PRE RMSD for an individual

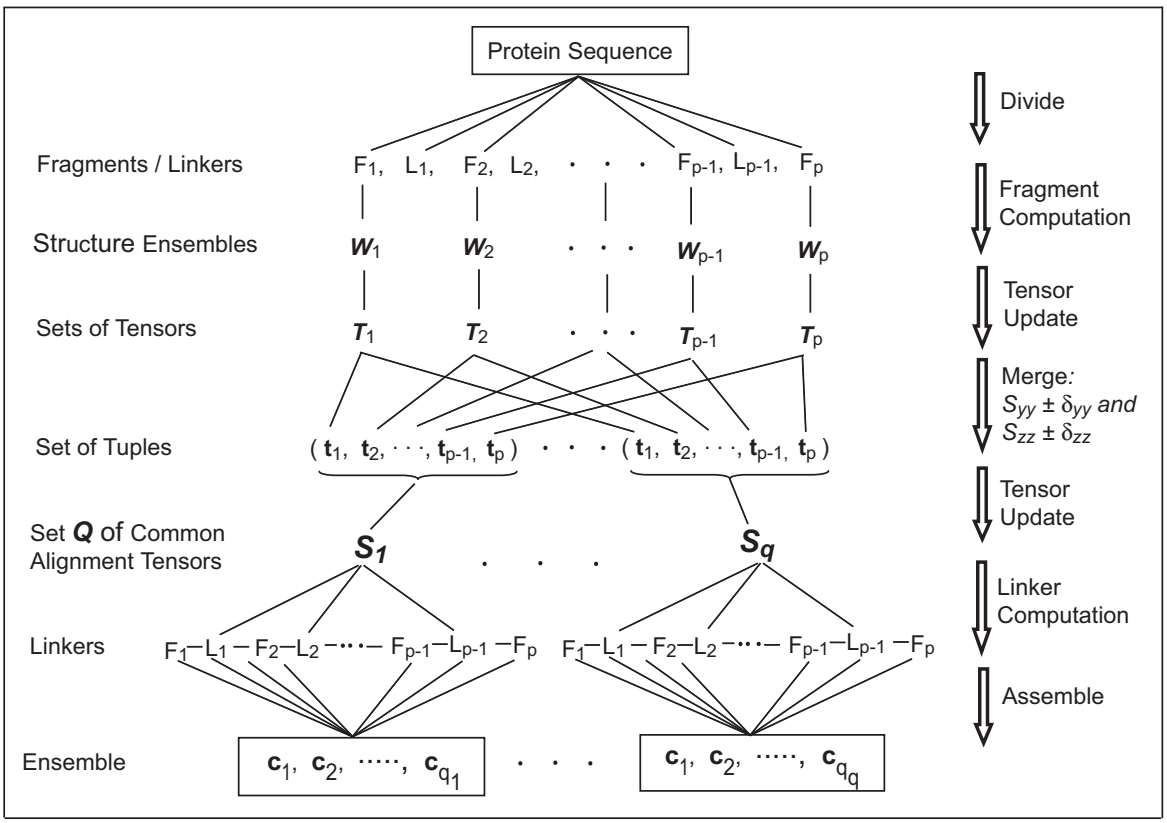

Fig. 1. Divide-and-conquer strategy. The input to the algorithm is: the protein sequence, at least two RDCs per residue in a single medium and PREs (if available). The terms $\mathbf{c}_i$ denote conformation vectors for the complete backbone structure. Please see the text for the definitions of other terms and an explanation of the algorithm.

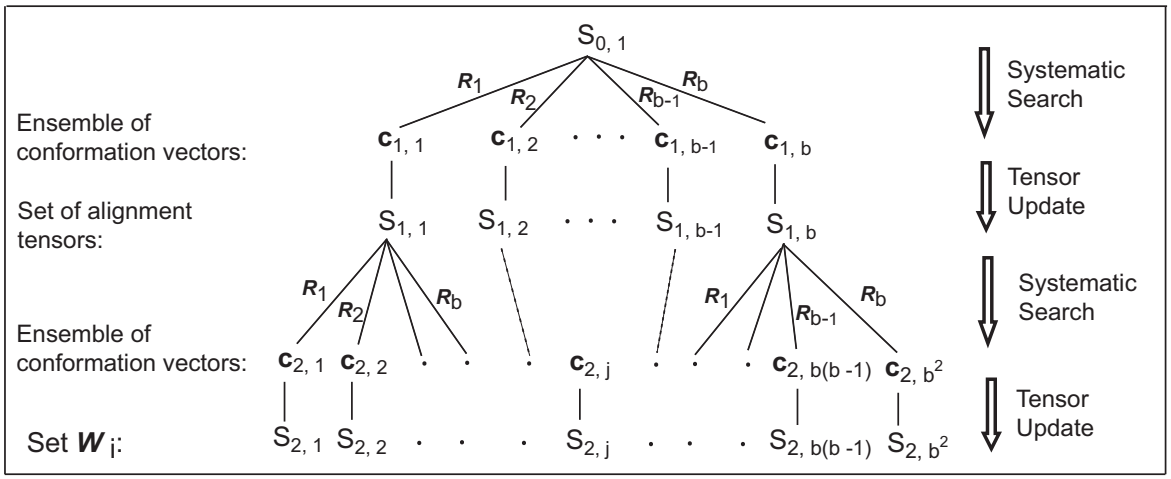

Fig. 2. Fragment computation: the computation of a structure ensemble of a fragment. The figure shows only two cycles of systematic search followed by SVD. Please see the text for the definition of terms and an explanation of the algorithm.

structure composing of all the previously-computed and linked fragments and the current linker, is computed as $E_p = \sqrt{\frac{\sum_{i=1}^{a}(d'_i - d_i)^2}{a-1}}$, where d_i and d'_i are, respectively, the experimental PRE distance and the distance between two C_α atoms back-computed from the model, and a is the number of PRE restraints. An experimental PRE dis-

tance restraint is between two C_α atoms computed from the PRE peak intensity [16]. If $d'_i \leq d_i$, the contribution of PRE violation i to E_p is set to zero. This search step is similar to our previous systematic searches as detailed in [29, 28, 31]. The key difference is that the linker scoring function, Eq. (3), has two new terms: E_v and E_p, and lacks the term in [29, 28, 31] for restraining (ϕ, ψ) angles to the favorable Ramachandran region for a typical α-helix or β-strand.

7. APPLICATION TO REAL BIOLOGICAL SYSTEMS

We have applied our algorithm to compute the structure ensembles of two proteins, an acid-denatured denatured ACBP and a urea-denatured eglin C, from real experimental NMR data.

Application to acid-denatured ACBP. An ensemble of 231 structures has been computed for ACBP denatured at pH 2.3. The experimental NMR data [9] has both PREs and four backbone RDCs per residue: NH, CH, NC′ and CC′. All the 231 structures have no vdW repulsion larger than 0.1Å except for a few vdW violations as large as 0.35 Å between the two nearest neighbors of a proline and the proline itself. These 231 structures satisfy all the experimental RDCs (CH, NH, CC′ and NC′) much better than the native structure, and have PRE violations, E_p, in the range of $4.4 - 7.0$ Å. The native structure also has very different Saupe elements, S_{yy} and S_{zz}. Further analysis of the computed ensemble shows that the acid-denatured ACBP is neither random coil nor native-like.

Application to urea-denatured eglin C. An ensemble of 160 structures were computed for eglin C denatured at 8 M urea. No structures in the ensemble have a vdW violation larger than 0.1Å except for a few vdW violations as large as 0.30 Å. The computed structures satisfy the experimental CH and NH RDCs much better than the native structure. The native structure also has very different Saupe elements, S_{yy} and S_{zz}. Further analysis of the computed ensemble also shows that the acid-denatured ACBP is neither random coil nor native-like.

8. ALGORITHMIC COMPLEXITY AND PRACTICAL PERFORMANCE

The complexity of the algorithm (Fig. 1) can be analyzed as follows. Let the protein sequence be divided into p m-residue fragments and $p-1$ m-residue linkers and let the size of samplings be b. The systematic-search step in Fragment computation takes $O(bpf^m)$ time to compute all the p ensembles for p fragments (Fig. 2) where f is the number of (ϕ, ψ) pairs for each residue computed from two quartic equations (Propositions 5.1–5.2) *and* pruned using a real solution filter as described in [28] and also a vdW filter (repulsion). A single SVD step in Fig. 2 takes $m5^2 + 5^3 = O(m)$ time. Thus, h cycles of systematic-search/SVD take $t_{_F}$ time in the worst-case, where $t_{_F} = \sum_{j=1}^{h} pb^j(f^m + m) = p\frac{b^{h+1}-b}{b-1}(f^m + m) = O(pb^{h+1}(f^m + m) = O(pb^{h+1}f^m)$ since f^m is much larger than m. In implementation, $b = 8 \times 1024$ and $h = 2$ (see section 11 of APPENDIX). In practice, only a small number (about 100) of structures out of all the possible b^h computed structures for fragment i (section 6.2 and Fig. 2), are selected and saved in $\mathbf{W}_i$ (Fig. 1), that is, the selected structures have $T_{_F} \leq T_{max}$ or $T_{_L} \leq T_{max}$ where $T_{_F}$ and $T_{_L}$ are computed, respectively, by Eq. (2) and Eq. (3), and T_{max} is a threshold. The Merge step takes $o(pw^p \log w)$ time, where $w = |W_i|$ is the number of structures in $\mathbf{W}_i$. The Merge step generates q p-tuples of alignment tensors, where $q = \gamma w^p$ and γ is the percentage of p-tuples selected from the Cartesian product of the sets $\mathbf{T}_i, i = 1, \ldots, p$, according to the ranges for S_{yy} and S_{zz} (section 6.1). The SVD step for computing q common tensors from p m-residue fragments takes $q(mp5^2+5^3) = O(mpq)$ time. The linkers are computed and assembled top-down using a binary tree. The Linker computation and assembly step then takes $t_{_L}$ time, where $t_{_L} = bq\sum_{k=1}^{\log p} 2^k f^{(2k+1)m} = bq\frac{(2f^{2m})^{\log p+1}-2f^{2m}}{2f^{2m}-1}f^m$ since at depth k, vdW repulsion and PRE violation are computed for the assembled fragment consisting of $2k$ m-residue fragments and an m-residue linker (Fig. 3). The total time is $O(pb^{h+1}f^m + pw^p \log w + mpq + bqp^{2m+1}f^{2m \log p+m}) = O(pb^{h+1}f^m + pw^p \log w + mpq + bqp^{2(c+1)m+1}f^m)$ where $c = \log f = O(1)$. The largest possible value [28] for f is 16 but on average f is about 2. The largest possible value for γ is 1 but in practice, it is very small, about 10^{-9}, and $q = 10^3$ with $w = 100$. Although the worst-time complexity is exponential in $O(h)$, $O(m)$ and $O(p)$, the parameters for m, h, p are rather small constants in practice with typical values of $m = 10, h = 2, p = 6$ for a 100-residue protein. In practice, on a Linux cluster with 32 2.8GHz Xeon processors, 20 days are required for computing an ensemble of 231 structures for ACBP, and 7 days for computing

For $i \leftarrow 1$ to 4 *// 4-fold degeneracy in relative orientation*

(1) $T_L \leftarrow \infty$
(2) $\mathbf{c}_{m,i} \leftarrow \emptyset$ *// initialize the conformation vector*
(3) For $j \leftarrow 1$ to b *// sampling cycle*

(a) Sample a set of RDCs, R_j, from the normal distributions for RDCs.
(b) Compute an optimal conformation vector $\mathbf{c'}_{m-2,i} \leftarrow (\phi_1, \psi_1, \ldots, \phi_{m-2}, \psi_{m-2})$ by systematic search.
(c) Compute ϕ_{m-1} by Proposition 5.1 using CH RDC for residue $m-1$.
(d) Compute ψ_{m-1}, ϕ_m and ψ_m by Proposition 10.3 (section 10 of APPENDIX).
(e) Build a polyalanine model for linker L_1 using the vector $\mathbf{c'}_{m,i} \leftarrow (\phi_1, \psi_1, \ldots, \phi_m, \psi_m)$
(f) Link L_1 to F_1 and F_2. *// see figure caption for an explanation*
(g) Compute E_p and a new score T'_L by Eq. (3) for the assembled fragment $F_1 \cup L_1 \cup F_2$.
(h) If $T'_L < T_L$ and $E_p < P_{max}$

$T_L \leftarrow T'_L$
$\mathbf{c}_{m,i} \leftarrow \mathbf{c'}_{m,i}$

(4) Return $\mathbf{c}_{m,i}$ *// the optimal conformation vector*

Fig. 3. Linker computation and Assembly. b is the number of sampling cycles. P_{max} is the maximum PRE violation allowed and set to be 7.0Å. The Link step, step (f), is to translate first the N-terminal of L_1 to the C-terminal of F_1, then translate the C-terminal of the fragment $F_1 \cup L_1$ to the N-terminal of F_2. There exists an intrinsic 4-fold degeneracy in the relative orientation between two fragments computed using RDCs measured in a single medium.

an ensemble of 160 structures for eglin C.

9. CONCLUSION AND BIOLOGICAL SIGNIFICANCE

At present, we have only very limited knowledge of the structural distribution of either laboratory-denatured or natively-disordered proteins. The main reason is that the current experimental techniques can only provide a sparse number of restraints, even while the traditional structure determination methods require a large number of them. In this paper, we presented and demonstrated a data-driven, systematic search algorithm capable of computing the ensemble of denatured solution structures directly from sparse experimental restraints. Our algorithm is based on the formulation of structure determination of denatured or disordered proteins as the computation of a set of heterogeneous structures from the distributions for the sparse experimental restraints. We have shown that the ensemble of denatured structures can be computed using the distributions for the orientational restraints from RDCs by solving a series of low-degree monomials. Compared with the previous approaches for characterizing the denatured state from experimental data, the ensemble of structures computed by our algorithm is substantially more accurate. More restraints were used in our algorithm, and most importantly, exact algebraic solutions in combination with systematic search guarantee that all the valid conformations consistent with the experimental restraints are computed. The accurately-computed structure ensemble makes it possible to answer two key questions in protein folding: (a) are the structures in the denatured state random coils? and (b) are the denatured structures native-like? Our quantitative analysis concludes that the denatured states of both ACBP and eglin C are neither random nor native-like.

Acknowledgments

We would like to thank Drs. Kresten Lindorff-Larsen and Flemming Poulsen for NMR data on acid-denatured ACBP, Dr. David Shortle for NMR data on urea-denatured eglin C, and Dr. Jane Dyson for communicating to us the values of backbone (ϕ, ψ) angles for the polyproline II model. We would like to thank Mr. Tony Yan and Drs. Ramgopal Mettu, Kresten Lindorff-Larsen, Andrei Alexandrescu, Mehmet Apaydin and Chris Bailey-Kellogg, and all members of Donald lab for helpful discussions and critical reading of the manuscript.

References

1. A. Abragam. *The Principles of Nuclear Magnetism*. Clarendon Press, Oxford, 1961.
2. M. S. Ackerman and D. Shortle. Molecular alignment of denatured states of staphylococcal nuclease with strained polyacrylamide gels and surfactant liquid crystalline phases. *Biochemistry*, 41:3089–3095, 2002.
3. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N.

Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28:235–242, 2000.
4. P. Bernado, L. Blanchard, P. Timmins, D. Marion, R. W. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. USA*, 102:17002–17007, 2005.
5. A. T. Brünger. *XPLOR: A system for X-ray crystallography and NMR*. Yale University Press: New Haven, 1993.
6. W. Y. Choy and J. D. Forman-Kay. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.*, 308:1011–1023, 2001.
7. G. M. Clore, M. R. Starich, C. A. Bewley, M. L. Cai, and J. Kuszewski. Impact of residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints. *J. Am. Chem. Soc.*, 121(27):6513–6514, 1999.
8. F. Delaglio, G. Kontaxis, and A. Bax. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc.*, 122(9):2142–2143, 2000.
9. W. Fieber, S. Kristjansdottir, and F. M. Poulsen. Short-range, long-range and transition state interactions in the denatured state of ACBP from residual dipolar couplings. *J. Mol. Biol.*, 339:1191–1199, 2004.
10. A. W. Giesen, S. W. Homans, and J. M. Brown. Determination of protein global folds using backbone residual dipolar coupling and long-range NOE restraints. *J. Biomol. NMR*, 25:63–71, 2003.
11. P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.*, 273:283–298, 1997.
12. W. Hu and L. Wang. Residual dipolar couplings: Measurements and applications to biomolecular studies. *Annual Reports on NMR Spectroscopy (In Press)*, 2006.
13. J. C. Hus, D. Marion, and M. Blackledge. Determination of protein backbone using only residual dipolar couplings. *J. Am. Chem. Soc.*, 123:1541–1542, 2001.
14. A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA*, 102:13099–13104, 2005.
15. L. D. Landau and E. M. Lifshitz. *Statistical Physics*. Pergamon Press, Oxford, 1980.
16. K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, M. Poulsen, and M. Vendruscolo. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J. Am. Chem. Soc.*, 126:3291–3299, 2004.
17. J. A. Losonczi, M. Andrec, M. W. Fischer, and J. H. Prestegard. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J. Magn. Reson.*, 138(2):334–342, 1999.
18. R. Mohana-Borges, N. K. Goto, G. J. Kroon, H. J. Dyson, and P. E. Wright. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J. Mol. Biol.*, 340:1131–1142, 2004.
19. P. J. Plory. *Statistical Mechanics of Chain Molecules*. Oxford University Press, New York, 1988.
20. T. L. Religa, J. S. Markson, U. Mayor, S. M. V. Freund, and A. R. Fersht. Solution structure of a protein denatured state and folding intermediate. *Nature*, 437:1053–1056, 2005.
21. C. A. Rohl and D. Baker. De Novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.*, 124(11):2723–2729, 2002.
22. A. Saupe. Recent results in the field of liquid crystals. *Angew. Chem.*, 7:97–112, 1968.
23. D. Shortle and M. S. Ackerman. Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, 293:487–489, 2001.
24. C. Tanford. Protein denaturation. Part C. Theoretical models for the mechanism of denaturation. *Adv. Protein Chem.*, 24:1–95, 1970.
25. N. Tjandra and A. Bax. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, 278:1111–1114, 1997.
26. J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution. *Proc. Natl. Acad. Sci. USA*, 92:9279–9283, 1995.
27. W. F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura. The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew. Chem. Int. Ed.*, 40:351–355, 2001.
28. L. Wang and B. R. Donald. Analysis of a systematic search-based algorithm for determining protein backbone structure from a minimal number of residual dipolar couplings. In *IEEE Computer Society Bioinformatics Conference*, pages 319–330, Stanford University, CA, 2004.
29. L. Wang and B. R. Donald. Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *J. Biomol. NMR*, 29:223–242, 2004.
30. L. Wang and B. R. Donald. An efficient and accurate algorithm for assigning nuclear overhauser effect restraints using a rotamer library ensemble and residual dipolar couplings. In *IEEE Computer Society Bioinformatics Conference*, pages 189–202, Stanford University, CA, 2005.
31. L. Wang, R. Mettu, and B. R. Donald. An algebraic geometry approach to backbone structure determination from NMR data. In *IEEE Computer Society Bioinformatics Conference*, pages 235–246, Stanford University, CA, 2005.

APPENDIX

In this appendix, we first state the polynomials for computing, respectively, the sine and cosine of the backbone ϕ angle from an NH RDC and those of the ψ angle from a CH RDC, starting with the *C-terminus* of a fragment. By comparison, Propositions 5.1–5.2 of the main text compute the (ϕ, ψ) angles from RDCs starting with the *N-terminus*. The proof for these two propositions is very

similar to the proof for lemmas 5.1–5.2 given in [28]. We then present a proof for a new proposition for computing backbone (ϕ, ψ) angles from oriented peptide planes. Finally, we describe the parameters and implementation of the algorithm.

10. LOW-DEGREE POLYNOMIALS FOR COMPUTING BACKBONE DIHEDRAL ANGLES

The following two Propositions, 10.1 and 10.2, are a generalization of Propositions 5.3 and 5.4 of [28] to compute backbone structure from the C-terminus, rather than the N-terminus. Starting with the peptide plane $i+1$, we can compute backbone ϕ_i, ψ_i angles, respectively, from the NH RDC of residue i and CH RDC of residue i as follows:

Proposition 10.1. *Given the orientation of peptide plane $i+1$ in the POF of RDCs, the x-component of the CH unit vector $\mathbf{u}$ of residue i, in the POF, can be computed from the CH RDC for residue i by solving a quartic monomial in x describing the intersection of two ellipses. Given the x-component, the y-component can be computed from Eq. (1) and the z-component from $x^2+y^2+z^2=1$. Given $\mathbf{u}$, the sine and cosine of the ψ_i angle can be computed by solving a linear equation.*

Here, the CH vector ellipse equation is a function of the ψ_i angle. The ellipse equation has been described in detail in [28].

Proposition 10.2. *Given the orientation of peptide plane $i+1$ in the POF of RDCs, the x-component of the NH unit vector $\mathbf{v}$ of residue i, in the POF, can be computed from the NH RDC for residue i by solving a quartic monomial in x describing the intersection of two ellipses. Given the x-component, the y-component can be computed from Eq. (1), and the z-component from $x^2+y^2+z^2=1$. Given $\mathbf{v}$, the sine and cosine of the ϕ_i angle can be computed by solving a linear equation.*

Here, the NH vector ellipse equation is a function of the ϕ_i angle. The ellipse equation has been described in detail previously [28].

The sine and cosine of the backbone (ϕ, ψ) angles of the last two residues linking two oriented fragments can be computed, *exactly* and *in constant time*, by the following Proposition:

Proposition 10.3 *Given the orientation of peptide planes i and $i+2$ and the backbone dihedral angle ϕ_i, the sines and cosine of the backbone dihedral angles ψ_i, ϕ_{i+1} and ψ_{i+1} can be computed exactly and in constant time.*

Proof. In the following, small and capital bold letters denote, respectively, column vectors and matrices. All the vectors are 3D vectors and all the matrices are 3D rotation matrices. Let $\mathbf{v}_1$, $\mathbf{v}_3$ and $\mathbf{w}_1$, $\mathbf{w}_3$ denote, respectively, the NH and C_α vectors of peptide planes i, and $i+2$. From protein backbone kinematics we have

$$\mathbf{L}\mathbf{G}_1\mathbf{w}_3 = \mathbf{R}_z(\psi_i)\mathbf{R}\mathbf{R}_y(\phi_{i+1})\mathbf{R}_x(\theta_3)\mathbf{R}_z(\psi_{i+1})\mathbf{c}_w,$$
$$\mathbf{L}\mathbf{G}_1\mathbf{v}_3 = \mathbf{R}_z(\psi_i)\mathbf{R}\mathbf{R}_y(\phi_{i+1})\mathbf{R}_x(\theta_3)\mathbf{R}_z(\psi_{i+1})\mathbf{c}_v$$

where $\mathbf{R}$ is a constant matrix, and $\mathbf{c}_w$ and $\mathbf{c}_v$ are two constant vectors and θ_3 is a constant angle. Given the backbone angle ϕ_i, the matrix $\mathbf{L}$ is known. The matrix $\mathbf{G}_1$ is the rotation matrix from the POF of RDCs to a coordinate frame defined in the peptide plane i. From Eq. (4), through algebraic manipulation we can derive the following three simple trigonometric equations satisfied by the ψ_i, ϕ_{i+1} and ψ_{i+1} angles

$$a_1 \sin\phi_{i+1} + b_1 \cos\phi_{i+1} = c_1$$
$$a_2 \sin\psi_{i+1} + b_2 \cos\psi_{i+1} = c_2$$
$$a_3 \sin\phi_i + b_3 \cos\phi_i = c_3$$

where a_1, b_1, c_1 are constants derived from the constant matrix $\mathbf{R}$, and the six variables, $a_2, b_2, c_2, a_3, b_3, c_3$, are simple trigonometric function of the ϕ_{i+1} angle. □

11. PARAMETERS AND IMPLEMENTATION OF THE ALGORITHM

Our algorithm (Figs. 1, 2, 3 of the main text) is built upon (a) *exact* solutions for backbone (ϕ, ψ) angles from RDCs, and (b) a *systematic-search* for exploring all the possible solutions consistent with the experimental restraints and biophysical properties (minimum vdW repulsion). However, several parameters must be chosen to ensure the correctness and convergence of the algorithm. We explored via computational experiments the spaces of these parameters to find the proper values that ensure the computed ensembles are stable. The parameters includes:

(1) division of protein sequence into fragments and linkers
(2) initial estimation of alignment tensors

(3) the standard deviations of the probability distributions for convolving the experimental RDCs
(4) the size of sampling, b
(5) the number of systematic-search/SVD cycles, h

In order to see their effects on the computed ensembles, we have run the algorithm with different initial tensors computed by SVD using either an ideal α-helix ($\phi = -64.3°$, $\psi = -39.4°$), or β-strand ($\phi = -120.0°$, $\psi = 138.0°$), or polyProline II model ($\phi = -80.0°$, $\psi = 135.0°$). We have also tested the algorithm using different sizes b of sampling and different numbers h of the systematic-search/SVD cycles. Our computational experiments showed that with an $b = 8 \times 1024$ and $h = 2$, the computed ensemble has already achieved a stable state since further increase in either b or h does not changes the distributions of backbone (ϕ, ψ) angles and pair-wise backbone RMSDs between the structures in the ensembles. The largest effect appears to be how the protein sequence is divided if there are missing RDCs concentrated in a certain region. In the implementation, the division into fragments and linkers is based primarily on the availability of experimental RDCs. In general, the linkers between two fragments have more missing RDCs than the fragments. If no experimental data is available for either CH or NH RDCs, the corresponding ϕ and ψ are selected randomly in the range of $[-\pi, \pi]$. As detailed in section 6.1, the alignment tensor used to compute the linkers is computed from the structures of fragments. Thus, if we exchange a fragment with a linker and if the linker has many missing RDCs, the computed ensemble differs, to some extent, from the original one. Our choice for division emphasizes the experimental data. The standard deviations for RDC random variables are, respectively, 8.0 Hz (Hertz) for CH RDCs and 4.0 Hz for NH RDCs, and both are much larger than the real experimental errors, which are estimated to be less than 1.0 Hz for CH RDCs and 0.50 Hz for NH RDCs. The values of these deviations are, respectively, about one-half of the ranges for all the experimental CH and NH RDC values. The probability distributions used to convolve RDCs are rather broad relative to the experimental values, and thus the algorithm is capable of computing most of structures in the denatured state. The relative weight w_v and w_p in Eq. (2) and Eq. (3) of the main text are set to be 8.0 and 2.0, respectively. The effects on the final ensembles of these weights are minimal, since vdW repulsion is very small in the final structures, and PRE violation is implemented by the requirement that all the final structures have no RMSD in PREs larger than 7.0Å. The function forms for both E_v and E_p in Eq. (3) are flat-bottom-harmonic-walls.