Published in final edited form as:
J Bioinform Comput Biol. 2014 February ; 12(1): 1450002. doi:10.1142/S0219720014500024.

DYNAFOLD: A DYNAMIC PROGRAMMING APPROACH TO PROTEIN BACKBONE STRUCTURE DETERMINATION FROM MINIMAL SETS OF RESIDUAL DIPOLAR COUPLINGS

RISHI MUKHOPADHYAY[*], **STEPHANIE IRAUSQUIN**[†], **CHRISTOPHER SCHMIDT**[‡], and **HOMAYOUN VALAFAR**[§]
Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

Abstract

Residual Dipolar Couplings (RDCs) are a source of NMR data that can provide a powerful set of constraints on the orientation of inter-nuclear vectors, and are quickly becoming a larger part of the experimental toolset for molecular biologists. However, few reliable protocols exist for the determination of protein backbone structures from small sets of RDCs. DynaFold is a new dynamic programming algorithm designed specifically for this task, using minimal sets of RDCs collected in multiple alignment media. DynaFold was first tested utilizing synthetic data generated for the N-H, C_α-H_α, and C-N vectors of 1BRF, 1F53, 110M and 3LAY proteins, with up to ±1 Hz error in 3 alignment media, and was able to produce structures with less than 1.9Å of the original structures. DynaFold was then tested using experimental data, obtained from the Biological Magnetic Resonance Bank, for proteins PDBID:1P7E and PDBID:1D3Z using RDC data from two alignment media. This exercise yielded structures within 1.0Å of their respective published structures in segments with high data density, and less than 1.9Å over the entire protein. The same sets of RDC data were also used in comparisons with traditional methods for analysis of RDCs, which failed to match the accuracy of DynaFold's approach to structure determination.

Keywords

protein; backbone; structure; residual dipolar coupling; RDC; dynamic programming

Introduction

Traditional experimental methods for protein structure determination include X-ray crystallography and Nuclear Magnetic Resonance spectroscopy (NMR). Presently, X-ray crystallography is considered the more mature approach[1] and continues to be the most applied technique[2]. NMR, on the other hand, is still a relatively young approach[3] that allows for solution-state study of the structure of interest and typically focuses on collecting Nuclear Overhauser Effect (NOE)[4] data as structural restraints. However, not every arbitrary

[*]rishi.mukhopadhyay@gmail.com. [†]smjimene@mailbox.sc.edu. [‡]schmidtc@email.sc.edu. [§]homayoun@cec.sc.edu.

target is amenable to either of these experimental methods. This is particularly true for membrane proteins, which seem to resist either crystallization, or the acquisition of sufficient numbers of NOEs to uniquely determine their structure. As a result, this important class of proteins is highly under-represented in the Protein Data Bank (PDB)[5]. Whereas most estimates place membrane proteins at between 20 and 30 percent of all open reading frames in all genomes[6], the number of membrane proteins whose structures have been determined is fewer than 300[7,8] as compared to the 82235 structures deposited in the PDB as of March 2013[5].

Residual Dipolar Couplings (RDCs) have recently emerged as a promising alternative for structural study of challenging macromolecules. Although first observed in the 1960's[9], it is only recently that RDCs have been developed as a new, powerful, and cost-effective NMR-based source of information. In addition, structure determination by RDCs requires less data acquisition than NOE-based methods. For example, NOEs report mostly on atoms located in the side-chains; this adds many more dimensions to the search space and is one reason why significantly more NOE restraints[10–12] are required to uniquely determine a protein structure. RDCs on the other hand can be reported on backbone atoms, allowing for direct investigation of protein backbone structure. Subsequently, other computational means can then be deployed for calculating the atomic coordinates of the side-chains[13], allowing for fewer RDC restraints to be acquired for structure determination. RDCs have been the subject of a number of reviews[14–20], and have been used in the study of carbohydrates[21–24], nucleic acids[14,25–28] and proteins[29–35]. More importantly, they have recently demonstrated promise as a viable approach to the structural characterization of challenging proteins such as membrane proteins[36].

Nearly all NMR analysis programs such as Xplor-NIH[37], CNS[38], Gromacs[39] and C_{yana} [40] have been modified to include analysis of RDCs. These methods are powerful in instances where large heterogeneous data sets, combining many different kinds of structural restraints, are used. When restricted to only RDC data however, without a good starting structure, these methods typically are susceptible to entrapment in local minima and usually never recover. While a Monte-Carlo approach to starting structures may be used, in practice this is rarely feasible without depending on less reliable information such as secondary structure assignment or torsion angle restraints. Other programs such as RDC-analytic[41], Meccano[42], and REDCRAFT[10,43–45] incorporate a more systematic approach when fitting protein structures to experimentally determined RDCs. RDC-Analytic is used as only one part of the larger RDC-Panda protocol, which uses NOE data and other information to compute protein folds, and is therefore not intended as an RDC-only method. It utilizes N-H and C_α-H_α RDCs in one alignment medium and secondary structure assignment, with the option to include sparse NOEs and hydrogen bonding restraints. The use of only one alignment precludes it from being robust to experimental error[46], and it is not always possible to predict secondary structure or collect the NOEs and hydrogen bond restraints needed to perform the final assembly. Meccano requires N-H, C-N, C-H, C_α-C, C_α-H_α and C_α-C_β RDCs collected in two alignment media. It can be expensive and time consuming to perform the labeling and assignment of resonances for that many RDC vector types, and in cases where data is missing, Meccano has to perform a local optimization over ϕ and ψ torsion

angles to produce the best fitting structure. REDCRAFT is another program designed for structure determination primarily from orientational restraints and is more relaxed in terms of its data requirements. It searches the space for all possible combinations of ϕ and ψ angles at a particular resolution (filtering for Ramachandran or secondary structure constraints, if applicable), and prunes this search tree heuristically. As a result, noisy or missing data can cause REDCRAFT to eliminate the branch of the tree containing the optimal solution. Within the last decade, some of the computational modeling tools have been modified to include RDCs[12] or other experimental data[47]. These methods generally utilize a very small portion of RDCs (N-H from one or two alignment media) in order to guide the computational modeling of structures, which have produced very exciting results. A number of other research efforts[10,11,41,48–50], have demonstrated the possibility of De Novo structure determination based on slightly more experimental data than what is required by hybrid approaches. However these methods employ stochastic search approaches that do not provide any upper bound in computation time or quality of the final structure. While structure determination methods based purely on experimental data, continue to appeal to the community of structural biologist. Therefore the topic of protein structure determination from a minimum set of RDC data is of interest and actively pursued. Here we present a new method, DynaFold, for protein structure calculation that is capable of using as few as three RDC restraints per residue from two or more alignment media to guarantee a complete search over the solution space at a particular resolution. DynaFold utilizes a Dynamic Programming[51] algorithm, which guarantees global optimality of the final solution given its parameterization of the problem.

Methods

The presented work relies heavily on the utility of Residual Dipolar Coupling (RDC) data that can be acquired by NMR spectroscopy. To better facilitate our introduction and discussion of DynaFold, we first begin with a brief discussion of the theoretical aspects related to RDCs. This is then followed by a detailed explanation of the DynaFold method.

1.1 Residual dipolar couplings

Just as all chemical elements have the fundamental properties of charge and mass, all chemical elements have the property of magnetic spin, which leads to the emergence of a magnetic dipole moment. Dipolar couplings arise from the interaction between two magnetic dipoles and the external magnetic field of a Nuclear Magnetic Resonance (NMR) spectrometer. Although atomic nuclei may possess a spin number equal to any natural number divided by two, for the remainder of this manuscript any discussion of dipolar couplings will be limited to atomic nuclei of spin -½.

The scalar coupling constants (*J*-coupling) measured for a particular dipole-dipole interaction can be split into anisotropic and isotropic terms. In solution state, when a molecule is tumbling unhindered, the isotropic term given in Eq. (1) averages to zero over time. However, when anisotropic tumbling can be induced, Eq. (1) time-averages to a value known as the Residual Dipolar Coupling (RDC) which is a function of the orientation of the unit vector between the two interacting atomic nuclei:

$$RDC_{ij} = \left\langle -\frac{\mu_0 \gamma_i \gamma_j \hbar}{(2\pi r_{ij})^3} \times \left[\frac{3\, cos^2\theta\,(t) - 1}{2} \right] \right\rangle \quad (1)$$

In Eq. (1), RDC_{ij} denotes the experimentally observed RDC between nuclei *i* and *j*, μ_0 is the magnetic permeability of free space, $\hbar$ is the reduced Planck's constant, γ_i and γ_j are the gyromagnetic ratios for nuclei of type *i* and *j*, r_{ij} is the distance between nuclei *i* and *j*, and θ is the angle between the magnetic field of the NMR instrument and the vector joining nuclei *i* and *j*. Given the relative timescale of the molecular tumbling and the acquisition time of the NMR instrument, only the time average of this quantity is observed[18] which is signified by the angle brackets in Eq.(1). When the effects of this time averaging are subsumed by designated constants, the RDC equation simplifies to Eq. (2):

$$RDC^{ij} = \frac{D_{max}}{\left(r_{eff}^{ij}\right)^3} v \times S \times v' \quad (2)$$

In Eq. (2), *v* is the unit vector in the direction of a particular intramolecular vector (v' signifies the transposed vector), $r^{ij}{}_{eff}$ is the time-averaged length of the intramolecular vector, and *S* is the Saupe Order Tensor Matrix (OTM), which contains the constants that subsume the effects of time averaging. The elements of *S* are named according to Eq. (3):

$$S = \begin{bmatrix} s_{xx} & s_{xy} & s_{xz} \\ s_{xy} & s_{yy} & s_{yz} \\ s_{xz} & s_{yz} & s_{zz} \end{bmatrix} \quad (3)$$

Though many molecules in solution may exhibit some degree of anisotropy in orientation when placed in a large magnetic field[17], that anisotropy is generally too small to be measured and therefore elements of *S* will equal zero. However, certain lipid-based crystalline solutions exhibit a strong degree of alignment in magnetic fields[45,52]. Partial alignment of many proteins can be accomplished by reconstituting a target protein in these liquid-crystalline environments. Some common alignment media include bicelles, filamentous bacteriophages and purple membrane fragments. Additionally, proteins can be aligned mechanically using stressed polyacrylamide gels[17,53]. Frequently, a protein sample is aligned in multiple alignment media to generate multiple channels of RDC data. These datasets will be referred to as n-D RDC data. In particular, for each alignment, there is a different order tensor matrix corresponding to different degrees of anisotropy in different alignment media.

Calculating the protein structure, therefore, reduces to finding a valid protein structure that satisfies a series of constraints in the form of Eq. (2) for each vector (on which RDCs are collected), and for each order tensor matrix (*S*). Due to the presence of error, a practical method should solve the problem in a least-squares sense or according to some other objective function that calculates a penalty for deviations from the experimental data. Recent work has shown that the statistical distribution of an n-D RDC data set can be used to

estimate the values of the order tensor matrices for each alignment medium[54–57]. These estimates have been shown to be of sufficiently high quality as to not distort protein structures significantly[36]. Therefore, the order tensors that describe the alignment of a protein can be assumed to have been determined. The main challenge then consists of folding the target protein such that the corresponding vector orientations are consistent with the RDC constraints. The remainder of this article will introduce DynaFold, a novel algorithm for discovery of the optimal protein structure for a given set of RDC data and order tensor estimates at a specified search resolution. Our presented method is unique in two ways: discovery of globally optimal structure, and its linear complexity in computation time as a function of protein size.

1.2. DynaFold

The problem of calculating a protein structure from RDCs can be presented as a search through the space of all possible protein structures and all possible order tensors where the goal is to find a member of that space which minimizes the discrepancy between the experimental data and the back-computed RDCs as computed by Eq. (2). Utilizing the fact that there are efficient (singular value decomposition-based) methods to find the best order tensor in a least-squares sense for a given structure and set of RDC data[58–60], this search problem is often simplified to a search over all protein structures with n peptides (where n denotes the protein size in number of amino acids) . This search space grows in size as an exponential function of the protein size (denoted by n). Even for extremely low-resolution discretizations of the search space, for a protein of size n=75 residues, there are more structural conformers than there are atoms in the observable universe (i.e. significantly greater than 10^{80} elements). Therefore, for even the smallest proteins, a brute-force discretized search of the space of all proteins is computationally intractable.

There are typically two major categories of algorithms for finding a solution by selectively sampling the solution space. The first approach is to search the solution space sub-optimally by using heuristics to guide the search from some set of starting points through the space of all possible protein structures. Simulated annealing and gradient descent-based optimizers fall into this category. The pitfall with this class of algorithms is that if the solution space is riddled with local minima, the probability of success in a reasonable amount of search time is limited. In particular, unless the starting point of the search is near the globally optimal solution, these methods rarely succeed using only RDC data. The only way to reduce the number of local minima is to provide these methods with more information than is mathematically necessary. Collecting this extra data is expensive and time consuming and, in this sense, there is a significant real-world cost associated with the weaknesses of these methods.

The second approach is to exploit some mathematical relationship in the solution space (as shown by[10,41,48]) so that the problem can be refactored into recursive sub-problems such that only a small number of conformations need to be sampled in the solution space to provide sufficient information to calculate the optimal solution. In particular, for some problems, dynamic programming algorithms[61,62] can be used to produce an optimal result from a search space that is exponentially large in the input size n, using a number of

calculations that grows polynomial in *n*. Examples of dynamic programming include the Smith-Waterman and Needleman-Wunsch sequence alignment algorithms[63] and many optimization and planning algorithms. The key to every dynamic programming algorithm is to express the optimal solution to the problem in terms of recursive sub-problems. Each of these recursive sub-problems must have the property that they can be computed in polynomial time by utilizing the calculations of other sub problems and that the total number of sub problems to solve is polynomial in the size of the input (in the case of protein folding, the number of peptides). Obviously not every problem can be rewritten with this structure and the main difficulty in finding dynamic programming solutions, when they exist, is to find the right parameterization of the search space.

The most obvious way to fractionate this problem into recursive sub-problems is by defining protein structures of length *i*, $1 \leq i \leq n$. This redefinition of the problem is possible since the sum-of-squared-error objective function can be expressed recursively in terms of error accrued on each residue. The first obstacle encountered in this approach is that determination of the optimal order tensor for the protein cannot be expressed recursively. Fortunately, recent introduction of the methods of *a priori* order tensor estimation in[54,55] can address this issue. In practice, the order tensors produced by these methods are known to be of sufficiently high quality for protein structure determination[36]. The following subsections detail DynaFold, a novel dynamic programming solution to find the optimal protein structure for a given order tensor estimate within a discretized approximation to the space of all geometrically valid protein structures.

1.2.1. Parameterization of the search space by DynaFold—The prevalent parameterization of protein structures consists of describing the coordinates of the backbone atoms in Cartesian space. However, since protein structures are required to conform to an accepted model of peptide geometry, its structural parameterization can be described more succinctly. Parameterization of protein structure in terms of its dihedral angles[40,64,65] would lead to a significant reduction of the search variables and therefore, improved computation time. When limiting the scope of structure determination to the backbone atoms represented in the rotamer-space reduces the set of parameters to the ϕ and ψ backbone torsion angles. Although this is a more natural parameterization of the problem, it is still not optimal since there is no way to calculate the fitness of a particular RDC data point without knowledge of all of the preceding ϕ and ψ torsion angles. This violates the requirement that the recursive sub-problems have inputs of fixed size.

Therefore, because of the shortcomings of the established protein structure parameterization schemes, a new parameterization of the solution space is required to find the optimal protein structure for a given order tensor estimate and set of RDC data, subject to the resolution of the search space. DynaFold's parameterization of the search space is expressed in terms of choices of orientation for the N-H and C_α-H_α vector orientations. In order to prove that a parameterization scheme based on these two vectors can unambiguously parameterize the search space, it needs to be shown that such a description can be translated into the (ϕ, ψ) parameterization scheme. Examination of ideal peptide geometry can provide the needed translation. The relationship between ϕ, ψ and the i^{th} internuclear vector orientations v^i_{NH} and v^i_{CaHa} can be established as shown in Eqs. (4) and (5):

$$v^i_{NH} \cdot v^i_{C_\alpha H_\alpha} = -0.1541 + 0.8324 \cdot cos\,(\phi - 60.70^\circ) \quad (4)$$

$$v^i_{C_\alpha H_\alpha} \cdot v^{i+1}_{NH} = 0.1796 + 0.7884 \cdot cos\,(\psi - 119.12^\circ) \quad (5)$$

Allowing substitutions $\bar{\Phi} = \phi - 60.70^\circ$ and $\bar{\Psi} = \psi - 119.12^\circ$ for the sake of convenience, two solutions for Eqs. (4) and (5) can be obtained (in terms of $\bar{\Phi}$ and $\bar{\Psi}$) as shown in Eqs. (6) and (7). It should be noted that the two possible solutions are related simply based on degeneracies of trigonometric functions. Considering the sign degeneracy that is noted in Eqs.(6) and (7), the complete parameterization scheme then consists of v_{NH} followed by the sign of $\bar{\Phi}$ followed by v_{CaHa} followed by the sign of $\bar{\Psi}$ repeated for each peptide in the protein: v^1_{NH}, ±, v^1_{CaHa} , ±, ..., v^i_{NH}, ±, v^i_{CaHa}, ±, ..., v^n_{NH}, ±, v^n_{CaHa}, ±. Critically, Eqs. (6) and (7) establish a one-to-one and onto relationship between DynaFold's parameterization of protein structure and the traditional dihedral parameterization (ϕ, ψ).

$$\bar{\phi} = \pm\, arccos\left(\frac{v^i_{NH} \cdot v^i_{C_\alpha H_\alpha} + 0.1541}{0.8324}\right), \quad (6)$$

$$\bar{\psi} = \pm\, arccos\left(\frac{v^i_{C_\alpha H_\alpha} \cdot v^{i+1}_{NH} - 0.1796}{0.7884}\right). \quad (7)$$

In principle the values of (ϕ, ψ) can be reconstructed by examining the corresponding triple. For example, given a valid triple such (v^i_{NH},±,v^i_{CaHa}), the corresponding ϕ angle can be reconstructed by Eq. (6). Of course, under experimental conditions, not every arbitrary string of vector orientations and torsion angle signs corresponds to a protein with a valid geometry. For example, Eqs. (4) and (5) may not have a solution for every combination of orientations for v_{NH} and v_{CaHa} that adhere to valid peptide geometry. It is therefore prudent to identify the optimal ϕ angle through a search that produces v^i_{NH}, and v^i_{CaHa}, orientations that most resembles that of the triple (v^i_{NH},±,v^i_{CaHa}) as shown in Fig. 1. DynaFold addresses this issue through a search over a discretized set of torsion angles that is described in the following section.

1.2.2. Discretization of the search space—In order to solve the problem of least-squares fitting of a protein structure to RDC data as a combinatorial optimization problem, the set of unit vectors from which to choose the orientations of v^i_{NH} and v^i_{CaHa} needs to be discretized. The natural choice for doing so is to describe the vectors in terms of their spherical coordinates (θ, φ) and appropriately discretizing those variables. Here, a set of isotropic vectors (denoted as U_k) are generated at resolution k, by partitioning θ into k equal intervals in the range [0–π] and φ into [2·$k$sinθ] equal intervals in the range [0–2π] as illustrated in Fig. 2.

The discretization of the vector set poses some challenges in terms of defining legal strings in the vector orientation description of the protein structure that adhere to ideal peptide geometry. While legal ($v^i{}_{NH}$,±,$v^i{}_{C\alpha H\alpha}$) and ($v^i{}_{C\alpha H\alpha}$,±, $v^{i+1}{}_{NH}$) triples may be defined from elements of U_k, in general, the legal successor pairs, (±,$v^{i+1}{}_{NH}$) and (±,$v^{i+1}{}_{C\alpha H\alpha}$) may not be defined when restricted to members of U_k. However, the method defined above for discretizing the vector set has the property that any vector on the unit sphere is no more than $180°/(k\sqrt{2})$ degrees away from its nearest representative in the set U_k. Therefore, by extending the rule for successor pairs to include any vectors within $180°/(k\sqrt{2})$ degrees of a true successor pair, the grammar can be approximated to arbitrarily high accuracy by increasing the value of k.

1.2.3. Graph theoretic view of the search space—Each ($v^i{}_{NH}$,±,$v^i{}_{C\alpha H\alpha}$) and ($v^i{}_{C\alpha H\alpha}$,±,$v^{i+1}{}_{NH}$) triple contains all of the state information required to determine which vector choices can precede or succeed the current set of symbols in the protein structure parameterization string. Therefore, traversing a valid protein structure parameter string can be thought of as state transitions between these triples. In particular, a graph can be constructed of the form shown in Fig. 3, so that every possible triple for every value of i from 1 to n (n is the length of the protein) corresponds to a node in the graph. Each ($v^i{}_{NH}$, ±,$v^i{}_{C\alpha H\alpha}$) triple defines a fragment of the form shown in Fig. 4A, and each ($v^i{}_{C\alpha H\alpha}$, ±,$v^{i+1}{}_{NH}$) triple defines a fragment of the form shown in Fig. 4B. Traversing the graph consists of picking local fragments at each position in the protein such that overlapping atoms from consecutive fragment choices fit (i.e. the overlapping portions can be aligned using only translation and no rotation). Therefore, edges in the graph connect nodes with properly overlapping atoms such that any path through the graph always describes a coherent protein structure as shown for an example node in Fig. 5.

More precisely, directed edges in this graph will exist from nodes of the form ($v^i{}_{NH}$, ±,$v^i{}_{C\alpha H\alpha}$) to nodes of the form ($v^i{}_{C\alpha H\alpha}$,±,$v^{i+1}{}_{NH}$) if and only if the orientation of $v^i{}_{C\alpha H\alpha}$ is the same for both nodes and the pair (±,$v^{i+1}{}_{NH}$) is a valid successor to the first triple. Likewise, directed edges will exist from nodes of the form ($v^i{}_{C\alpha H\alpha}$,±,$v^{i+1}{}_{NH}$) to nodes of the form ($v^{i+1}{}_{NH}$,±,$v^{i+1}{}_{C\alpha H\alpha}$) if and only if they agree in the value of $v^{i+1}{}_{NH}$ and the pair (±, $v^{i+1}{}_{C\alpha H\alpha}$) is a valid successor to the first triple. Additionally, there will be a start node with directed edges to every triple of the form ($v^1{}_{NH}$,±,$v^1{}_{C\alpha H\alpha}$) and an end node with edges directed into it from every triple of the form ($v^n{}_{NH}$,±,$v^n{}_{C\alpha H\alpha}$).

1.2.4. Dynamic programming-based solution—The problem of finding a shortest path through a directed graph has a well-known polynomial-time dynamic programming-based solution. Therefore, the problem of finding the protein that best fits the RDC data can be solved by encoding the fitness function in terms of the edge weights of the state transition graph. Each edge weight (other than those to and from the start and end nodes) can be based on information contained in the two nodes it connects. Any RDC restraint can be incorporated into the edge weights so that the path through the graph corresponding to a particular conformer has length equal to the square of the error function. Additionally, it is possible to include other terms in the objective function such as penalties for ϕ and ψ torsion angles that fall outside of the Ramachandran space for that amino acid type. In general, any

translationally-invariant structural restraint that only depends on “local” information can be included in this objective function.

In this manuscript, the implementation has focused on the inclusion of N-H, $C_\alpha H_\alpha$ and C-N RDCs collected in two or more alignment media and the use of Ramachandran constraints. The objective function, *F* which DynaFold seeks to minimize, is shown in Eq. (8):

$$F(E)=\sum_{i=1}^{n}\left(R(\phi_i,\psi_i,i)+\sum_{t\in\{NH,C_\alpha H_\alpha,CN\}}\sum_{m=1}^{M}\left(RDC\left(v_t^i,S^m\right)-E_{t,i}^m\right)^2\cdot w_{t,i}^m\right). \quad (8)$$

In this equation, *m* iterates over the *M* alignment media, *t* iterates through the three vector types and *i* iterates over the *n* peptides in the protein. v_t^i is the internuclear vector of type *t* on the i^{th} peptide, S^m is the order tensor estimate for alignment medium *m* and $E^m_{t,i}$ is the experimental RDC data in alignment *m* for the i^{th} vector of type *t*. w^m_{ti} is a normalization constant to compensate for the magnitude of RDC differences based on gyromagnetic ratios of different nuclei. The weights w^m_{ti} can be set to 1 for traditional least squares fitting. However, in our work these weights have been selected such that the range of RDCs for each vector type in each alignment spans a range of 1. In this case, the RDC portion of the objective function is minimizing the deviation from the experimental data as a percentage of the total possible numerical range of the values. In the case where experimental data is missing, w^m_{ti} is set to zero. The entity *R* in Eq. (8) is a user-supplied function that calculates penalties for torsion angle restraints for each peptide. In this work, torsion angle restraints were limited to a general Ramachandran map for each residue except for glycine and proline residues, for which amino acid-specific Ramachandran maps were utilized. Ramachandran space data was taken from Lovell et al.,[66] and with a cutoff threshold of 99.5 percentile.

To facilitate a better discussion, Eq. (8) has been expanded and restated as Eqs. 9-13.

$$F(E)=\sum_{i=1}^{n}P_i^R+\sum_{i=1}^{n}P_i^{NH}+\sum_{i=1}^{n-1}P_i^{CN}+\sum_{i=1}^{n}P_i^{C_\alpha H_\alpha}, \quad (9)$$

$$P_i^R=R(\phi_i,\psi_i,i), \quad (10)$$

$$P_i^{NH}=\sum_{m=1}^{M}\left(RDC\left(v_{NH}^i,S^m\right)-E_{NH,i}^m\right)^2\cdot w_{NH,i}^m, \quad (11)$$

$$P_i^{CN}=\sum_{m=1}^{M}\left(RDC\left(v_{CN}^i,S^m\right)-E_{CN,i}^m\right)^2\cdot w_{CN,i}^m, \quad (12)$$

$$P_i^{C_\alpha H_\alpha}=\sum_{m=1}^{M}\left(RDC\left(v_{C_\alpha H_\alpha}^i,S^m\right)-E_{C_\alpha H_\alpha,i}^m\right)^2\cdot w_{C_\alpha H_\alpha,i}^m, \quad (13)$$

The term P_n^{CN} has been omitted from Eq. (9) since $v^i{}_{CN}$ is defined as the vector from C^i to N^{i+1} and the nucleus N^{n+1} does not exist. Edges into nodes of the form ($v^i{}_{NH}$,±,$v^i{}_{C\alpha H\alpha}$) were weighted by$P_i^{C\alpha H\alpha}$. Edges into nodes of the form ($v^i{}_{C\alpha H\alpha}$,±, $v^{i+1}{}_{NH}$) were weighted by $P_i^R+P_{i+1}^{NH}+P_i^{C\alpha H\alpha}$. Edges from the start node to nodes of the form ($v^1{}_{NH}$,±,$v^1{}_{C\alpha H\alpha}$) were additionally weighted by P_1^{NH} since that term is not accounted for by any of the other edge weights. Each of these edge weights can be calculated from the peptide fragments that correspond to the nodes on each side of a given edge.

The dynamic programming-based approach to finding the shortest path through this graph starts by creating a table entry for each node in the graph and initializing the start node's score to 0. Each node's score is calculated by summing the edge weight and the score of the originating node for that edge from each incoming edge, taking the minimum of these values and noting which edge produced that minimum score. In this way, the dynamic programming algorithm can calculate the optimal structure for the first *n* peptides ending in the fragment represented by a particular node in the graph. The overlapping component of dynamic programming is satisfied in a Bottom-Up form by storing the optimal solution to a sub problem in a table that can be used in subsequent steps. Leveraging the results of the sub problems a constant computational time can be achieved for a given discretization of the vector space. Put another way, DynaFold's formulation of the RDC fitness problem reduces finding the optimal fitness of protein structure to the problem of finding the shortest path in the graph illustrated in Figures 3 and 5, which is solvable in polynomial time with a well-known dynamic programming solution[62]. Completing the entries of this table can proceed methodically by alternating between filling in table entries for nodes of type ($v^i{}_{NH}$,±,$v^i{}_{C\alpha H\alpha}$) and nodes of type ($v^i{}_{C\alpha H\alpha}$,±, $v^{i+1}{}_{NH}$) and proceeding from i=1 to i=n and concluding with the end node. Since each node has recorded which incoming edge provided the least-cost path to that node, finding the best protein structure consists of tracing back those edge choices from the end node to the start node.

1.2.5. Production of the final protein structure—Since the discretized vector orientation description of the protein structure is an approximation to the continuous case, the translation from the parameterization to an actual protein structure requires some additional analysis. The approach taken here proceeds in two steps. First, an estimate of the $\bar{\Phi}$ and $\bar{\Psi}$ angles are made from the vector choices using Eq. (4), Eq. (5) and a protein with those torsion angles is constructed in the principal alignment frame of the first alignment medium. Then, since errors accrue from these approximations, a non-linear least squares optimization is used to fine-tune the $\bar{\Phi}$ and $\bar{\Psi}$ angles to get the N-H and C_α-H_α vectors in the final protein as close as possible in the angle-distance sense to the N-H and C_α-H_α vectors chosen in the solution from U_k. After an initial solution is produced in this manner from DynaFold's output, a final round of non-linear least-squares minimization is performed to refine the structure where ϕ, ψ, ω, and the order tensor estimates are all optimized to best fit the RDC data. All of the least square minimizations have been conducted using the *fmin* function within Matlab.

1.2.6. Validation procedure—Validating the effectiveness of DynaFold proceeded in two stages. In the first stage, DynaFold was run on synthetic data generated for protein

1BRF[67]. Using the 20th model (the last model) in the PDB file, RDCs were simulated for N-H, C_α-H_α and C-N vectors in three alignment media using the following three order tensors:

$$S_1 = \begin{bmatrix} -1.67 & 0 & 0 \\ 0 & -7.41 & 0 \\ 0 & 0 & 9.07 \end{bmatrix} \cdot 10^{-4},$$

$$S_2 = \begin{bmatrix} -5.02 & 0.80 & 1.14 \\ 0.80 & 4.96 & -1.66 \\ 1.14 & -1.66 & 0.06 \end{bmatrix} \cdot 10^{-4},$$

$$S_3 = \begin{bmatrix} -1.62 & -1.07 & 4.33 \\ -1.07 & -1.39 & 1.95 \\ 4.33 & 1.95 & 3..01 \end{bmatrix} \cdot 10^{-4}$$

These order tensors were selected to have values that approximately match those that are observed experimentally. To better represent experimental conditions, uniformly distributed noise in the range of ± 1 Hz was added to the simulated data.

In the second stage of validation, DynaFold was tested using experimental data for proteins 1P7E[53] and 1D3Z[68] downloaded from the BMRB[69]. 1P7E is the third IgG-binding domain of Protein G (GB3) and 1D3Z is a human ubiquitin protein. Although the deposited data set for 1P7E contains four RDC vector types collected in 5 alignment media for a total of 20 sets of restraints over 56 residues, we have utilized only N-H, C_α-H_α and C-N vector types in three alignment media for our testing of DynaFold. This reduction is an exercise in establishing the success of DynaFold in the structure determination of more challenging data-sparse cases.

The data set available for 1D3Z in the BMRB contains RDC data for two alignments with seven RDC vector types in the first alignment and six RDC vector types in the second alignment over 76 residues. In our testing of DynaFold, only N-H, C_α-H_α and C-N vector types were used in two alignments.

In order to provide a reasonable comparison of DynaFold to current structural determination methods, Xplor-NIH was used to attempt to determine the structures of 1P7E and 1D3Z using the same data provided to DynaFold. An Xplor-NIH structural determination script was created that uses RDC data, along with the Ramachandran database potential and standard molecular geometry constraints (scripts are available as part of the downloadable DynaFold software package from http://ifestos.cse.sc.edu). Fig. 6 illustrates the control flow of the script for structure determination with an extended structure as the starting point.

Results and Discussion

We present here the evaluations of DynaFold, with our results and their discussion arranged from the least challenging to most challenging case. First, we present results for synthetic data. This is followed by analysis of experimental RDCs from three alignment media, and finally experimental RDCs from two alignment media. This section is then concluded by comparisons between DynaFold and other traditional approaches for analysis of RDCs, in order to establish the success of the DynaFold method.

1.3. Simulated data for proteins 1BRF, 110M, 3LAY, and 1F53

Synthetic data for proteins 1BRF, 110M, 3LAY and 1F53 were generated according to the procedure described previously (refer to section 1.2.6). Next, DynaFold was run at low resolution (k=36) using N-H, C_α-H_α and C-N RDCs in 3 alignment media. The results are displayed in Fig. 7 for each protein. The backbone RMSD with respect to their published structures consist of less than 1.5Å for proteins 110M, 3LAY and 1BRF, 1.9Å for the protein 1F53. Therefore, it is reasonable to conclude that DynaFold was able to successfully fold these proteins from the synthetically generated data.

1.4. 1P7E: three vector types, three alignment media

Because the data set used for 1P7E contained a large number of restraints, this run of DynaFold was conducted at low resolution (k=36). However, there were several regions where data was either missing or marked as excluded from analysis for being extremely noisy. These data points were included for the DynaFold analysis since DynaFold does a complete search of the solution space and even noisy data has some information content. The overall alignment of the output structure and the published structure of 1P7E (which was refined from the crystal structure 1IGD) had a deviation of 1.817Å backbone RMSD. However, if 1P7E is segmented into the regions of high quality data-density the backbone RMSD to the published structure was , 0.276Å for residues 3-9 (first β strand), 0.704Å for residues 12-23 (second β strand), 0.241Å for residues 28-38 (α helix) and 0.454Å for residues 42-55 (last two β sheets). The alignment for residues 11-55, shown in Fig. 8 was 1.255Å backbone RMSD. Therefore, it is reasonable to conclude that DynaFold was able to find the right structure for the regions that had high quality data and was able to make a decent guess for the regions with low quality data so that the different segments of the structure were oriented and roughly translated correctly with respect to each other.

1.5. 1D3Z: three vector types, two alignment media

Because this data set contained data from only two alignment media, this run of DynaFold was conducted at high resolution (k=72). There were regions of low data density between residues 8-10, 19-24, 36-38 and 52-53. Only the missing data between residues 36-38 seemed to cause a major translational shift in the structure. The overall alignment between the structure output from DynaFold and the 10 models for 1D3Z submitted to the PDB were 1.89-1.95Å backbone RMSD; alignment of model 10 of the published 1D3Z structure to DynaFold's output is displayed in Fig. 9A. The backbone RMSD for residues 1-35 was between 0.83-0.91Å backbone RMSD, with alignment of these residues for model 10 of the published 1D3Z structure to DynaFold's output displayed in Fig. 9B. While the backbone RMSD for residues 39-70 was 0.80-0.85Å, with alignment of these residues for model 10 of the published 1D3Z structure to DynaFold's output displayed in Fig. 9C. Once again, it is reasonable to conclude that DynaFold was able to find the right structure for the regions that had high quality data and was able to make a decent guess for the regions with low quality data so that the different segments of the structure were oriented and roughly translated correctly with respect to each other.

1.6. Comparison of DynaFold to Xplor-NIH

The results of using the Xplor-NIH protocol outlined previously in Fig. 6, on the data for 1P7E, were that total backbone RMSD of the output structure compared with the published structure was 27.595Å. This indicates that general minimization routines (such as those used in Xplor-NIH) are insufficient to navigate a complex energy landscape defined purely on the basis of RDC data in order to achieve the global or near-global optimal point. The piecewise backbone RMSD for the structure was 1.375Å for residues 3-9 (β strand), 3.455Å for residues 12-23 (β strand), 5.658Å for residues 28-38 (α helix), 0.821 for residues 42-46 (β strand), and 0.67 for residues 51-55 (β strand). The extended structure had a total backbone RMSD of 54.95Å, giving the total change in RMSD as 27.355Å. It is clear from Table 1 that the RDC term for the output structure is nearly double that of the published structures, the standard atomic geometric model is slightly violated and the Ramachandran energy term minimizes well. Given the energy profile along with the RDC RMSD from Table 2, it is likely that the simulated annealing protocol found a deep local minimum, from which it was unable to recover using only the constraints provided by standard geometry, the Ramachandran database potential, and the provided RDC data.

Residual dipolar couplings from two alignment media for three vector types were used with the Xplor-NIH protocol previously described to attempt to fold protein 1D3Z. Total backbone RMSD of the output structure to the published structure was 45.291Å. Again, this leads to the conclusion that the amount of data was insufficient for Xplor-NIH to navigate the large search space. The piecewise backbone RMSD for the structure was 0.67Å for residues 1-6 (β strand), 1.299Å residues 12-17 (β strand), 6.295Å for residues 23-34 (α helix), 2.371Å residues 37-40 (α helix), 2.398 for residues 41-47 (β strand), 0.172Å for residues 48-49 (β strand), 1.931Å for residues 57-59 (α helix), and 0.63 for residues 66-71 (β strand). The extended structure had a total backbone RMSD of 49.899Å, giving the total change in RMSD as 4.608Å. The energy terms for the output, extended, and published structures can be seen in Table 3. This shows that while the molecular geometry and the RDC term remains on par with results obtained with 1P7E, RDC RMSD for media one and two for the output structure (seen in Table 4) seem to be around three times that of the published structures. Bearing in mind that there is less data available for 1D3Z, the small change in overall backbone RMSD seems to strengthen the argument that the data provided to DynaFold, along with standard geometry, and Ramachandran database potential is insufficient for Xplor-NIH to robustly navigate such a large search space.

Conclusion

This manuscript has validated DynaFold's approach to searching the space of protein conformations for the best match to the experimental data, and demonstrated its performance with RDCs acquired in two or three alignment media. Overall, even with a relatively small set of experimental RDC restraints, DynaFold is able to fold proteins with high accuracy. Our target proteins consisted of α, β, and α/β proteins to illustrate the versatility of our presented work. In addition, we have demonstrated the success of accurate order tensor estimation in absence of a structure, and structure determination of purely α proteins, which are deemed problematic for study by RDCs. In contrast to typical approaches to structure

determination by NMR spectroscopy where larger proteins pose more challenges, DynaFold's performance improves as a function of increasing protein size. This is due to the fact that the accuracy of order tensor estimation will increase for larger proteins due to better sampling of the RDC space. More accurate estimation of order tensors that is used during the course of DynaFold's calculation will lead to more accurate structure determination.

While it is unsurprising that missing data do influence the final structure, this effect is localized to the regions of low data density without any distortion of the rest of the structure. The areas of high data density were folded with high accuracy in each case, and oriented correctly with respect to each other. Since RDCs are translationally invariant, that is the best outcome possible when RDCs are the only structural restraints. In this manuscript we have refrained from the use of additional experimental data in order to study and demonstrate the raw information content of RDCs alone. Although it is clear that inclusion of additional experimental data such as NOEs or dihedral restraints can improve the quality of the final structure, any incidental NOEs (or other types of experimental data) can be incorporated in a subsequent refinement step using any of the existing and well established software packages such as Xplor-NIH. Using the high-resolution structure that is produced by DynaFold as the starting point of a refinement, only a small number of distance-based restraints such as NOEs, PREs or hydrogen bond restraints can make the final improvements in the structure calculation process. In addition to utilizing other types of NMR data, this refinement process can be applied in order to create an ensemble of viable structures via low-temperature annealing of a starting structure. Finally during such a refinement process, the approximated order tensors that were utilized by DynaFold can be permitted to be further optimized. Using a more relaxed final refinement process will improve the overall fitness of the structure and order tensors to the experimental data.

Although a refinement process following DynaFold analysis can in practice utilize all experimental data, we are still interested in providing one complete package in the future. In principle all experimental data can be incorporated into DynaFold's objective function. Therefore our future work will consist of extending this framework to utilize all RDC types, J-couplings, chemical shift anisotropy, and NOE data. Furthermore, the Ramachandran based dihedral restraints that DynaFold currently uses will be set to optionally incorporate predictions from TALOS[71] or other methods of predicting dihedral angles or secondary structures. Once DynaFold is extended to utilize all of these data types, researchers will have tremendous flexibility in choosing the set of restraints that are inexpensive, experimentally feasible and reliable for the given protein target. In particular, DynaFold's mathematical guarantee about the optimality of the result enables researchers to only collect as much data as is absolutely necessary instead of having to compile the large sets of restraints that traditional optimizers tend to require.

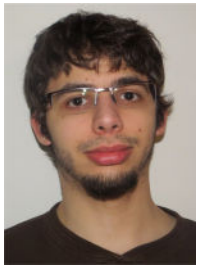
Acknowledgments

This research was made possible by NIH Grant Number P20 RR-016461 from the National Center for Research Resources and NSF Career grant MCB-0644195.

Biography

Rishi Mukhopadhyay received his Bachelors of Science in Computer Science from Cornell University's College of Engineering in 2005. He received his Ph.D. in Computer Science from the University of South Carolina's College of Engineering and Computing in December 2010.

Stephanie Irausquin received her Bachelor's Degree with honors in Exercise Science from the University of South Carolina in 2002. She returned to the University of South Carolina in 2005 as a student in the Professional Science Master's Program and subsequently obtained a P.S.M. Degree with an emphasis in Bioinformatics in 2007, as well as a Doctoral Degree in Biological Sciences in 2010.



Chris Schmidt received his B.S. in Computer Information Systems from the University of South Carolina in 2009. He is currently pursuing a M.S. in Computer Science and Engineering. Chris's work focuses on maintaining and updating REDCAT as well as integrating REDCAT with VMD. Chris has also worked on a java-based interface for Xplor-NIH scripting.

Dr. Homayoun Valafar received his B.S., M.S. and Ph.D. in Electrical and computer engineering from Michigan technological university and Purdue (M.S. & Ph.D.) in 1988, 1990 and 1995 respectively. Following his graduation he received two post doctoral fellowship appointments with Drs. Peter Albersheim and James Prestegard in the areas of structure determination of carbohydrates and computational methods of protein structure determination. Before joining USC as an assistant professor he served as the bioinformatics project coordinator at the Southeast Collaboratory for Structural Genomics (SECSG).

References

1. Adams, P.; Grosse-Kunstleve, R.; Brunger, A. Struct. Bioinforma. Bourne, PE.; Weissig, H., editors. Wiley-Liss, Inc; 2003. p. 75-87.
2. Sali A, Glaeser R, Earnest T, Baumeister W. From words to literature in structural proteomics. Nature. 2003; 422:216–225. [PubMed: 12634795]
3. Markley, J.; Ulrich, E.; Westler, W.; Volkman, B. Struct. Bioinforma. Bourne, PE.; Weissig, H., editors. Wiley-Liss, Inc; 2003. p. 89-113.
4. Kline AD, Braun W, Wüthrich K. Determination of the complete three-dimensional structure of the alpha-amylase inhibitor tendamistat in aqueous solution by nuclear magnetic resonance and distance geometry. J. Mol. Biol. 1988; 204:675–724. [PubMed: 3265733]
5. Berman HM, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
6. Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci. 1998; 7:1029–1038. [PubMed: 9568909]
7. Raman P, Cherezov V, Caffrey M. The Membrane Protein Data Bank. Cell. Mol. Life Sci. 2006; 63:36–51. [PubMed: 16314922]
8. White SH. Biophysical dissection of membrane proteins. Nature. 2009; 459:344–6. [PubMed: 19458709]
9. Saupe A, Englert G. High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules. Phys. Rev. Lett. 1963; 11:462–464.
10. Bryson M, Tian F, Prestegard JH, Valafar H. REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. J. Magn. Reson. 2008; 191:322–34. [PubMed: 18258464]
11. Valafar H, Simin M, Irausquin S. A Review of REDCRAFT: Simultaneous Investigation of Structure and Dynamics of Proteins from RDC Restraints. Annu. Reports NMR Spectrosc. 2012; 76:23–66.
12. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J. Am. Chem. Soc. 2002; 124:2723–9. [PubMed: 11890823]
13. Krivov GG, Shapovalov MV, Dunbrack RLJ. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009; 77:778–795. [PubMed: 19603484]
14. Al-Hashimi HM, Gorin A, Majumdar A, Gosser Y, Patel DJ. Towards structural Genomics of RNA: Rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. J Mol Biol. 2002; 318:637–649. [PubMed: 12054812]

15. De Alba E, Tjandra N. NMR dipolar couplings for the structure determination of biopolymers in solution. Prog. Nucl. Magn. Reson. Spectrosc. 2002; 40:175–197.
16. Bax A, Kontaxis G, Tjandra N. Dipolar couplings in macromolecular structure determination. Methods Enzymol. 2001; 339:127–74. [PubMed: 11462810]
17. Blackledge M. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. Prog. Nucl. Magn. Reson. Spectrosc. 2005; 46:23–61.
18. Prestegard JH, Al-Hashimi HM, Tolman JR. NMR structures of biomolecules using field oriented media and residual dipolar couplings. Q. Rev. Biophys. 2000; 33:371–424. [PubMed: 11233409]
19. Tolman JR. Dipolar couplings as a probe of molecular dynamics and structure in solution. Curr. Opin. Struct. Biol. 2001; 11:532–539. [PubMed: 11785752]
20. Zhou HJ, Vermeulen A, Jucker FM, Pardi A, Zhou Vermeulen A, Jucker FM, Pardi A,HJ. Incorporating residual dipolar couplings into the NMR solution structure determination of nucleic acids. Biopolymers. 1999; 52:168–180. [PubMed: 11295749]
21. Adeyeye J, et al. Conformation of the hexasaccharide repeating subunit from the Vibrio cholerae O139 capsular polysaccharide. Biochemistry. 2003; 42:3979–3988. [PubMed: 12667089]
22. Azurmendi HF, Bush CA. Conformational studies of blood group A and blood group B oligosaccharides using NMR residual dipolar couplings. Carbohydr. Res. 2002; 337:905–915. [PubMed: 12007473]
23. Azurmendi HF, Martin-Pastor M, Bush CA. Conformational studies of Lewis X and Lewis A trisaccharides using NMR residual dipolar couplings. Biopolymers. 2002; 63:89–98. [PubMed: 11786997]
24. Tian F, Al-Hashimi HM, Craighead JL, Prestegard JH. Conformational analysis of a flexible oligosaccharide using residual dipolar couplings. J. Am. Chem. Soc. 2001; 123:485–492. [PubMed: 11456551]
25. Al-Hashimi HM, Bolon PJ, Prestegard JH. Molecular symmetry as an aid to geometry determination in ligand protein complexes. J Magn Reson. 2000; 142:153–158. [PubMed: 10617446]
26. Al-Hashimi HM, et al. Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. J. Mol. Biol. 2002; 315:95–102. [PubMed: 11779230]
27. Tjandra N, Tate S, Ono A, Kainosho M, Bax A. The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase. J. Am. Chem. Soc. 2000; 122:6190–6200.
28. Vermeulen A, Zhou H, Pardi A. Determining DNA Global Structure and DNA Bending by Application of NMR Residual Dipolar Couplings. J. Am. Chem. Soc. 2000; 122:9638–9647.
29. Andrec M, Du PC, Levy RM. Protein backbone structure determination using only residual dipolar couplings from one ordering medium. J. Biomol. NMR. 2001; 21:335–347. [PubMed: 11824753]
30. Assfalg M, et al. 15N-1H Residual dipolar coupling analysis of native and alkaline-K79A Saccharomyces cerevisiae cytochrome c. Biophys. J. 2003; 84:3917–23. [PubMed: 12770897]
31. Bertini I, Luchinat C, Turano P, Battaini G, Casella L. The magnetic properties of myoglobin as studied by NMR spectroscopy. Chem. Eur. J. 2003; 9:2316–2322. [PubMed: 12772306]
32. Clore GM, Bewley CA. Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked protein domains from dipolar couplings. J. Magn. Reson. 2002; 154:329–335. [PubMed: 11846592]
33. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR. 1999; 13:289–302. [PubMed: 10212987]
34. Fowler CA, Tian F, Al-Hashimi HM, Prestegard JH. Rapid determination of protein folds using residual dipolar couplings. J Mol Biol. 2000; 304:447–460. [PubMed: 11090286]
35. Tian F, Valafar H, Prestegard JH. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J. Am. Chem. Soc. 2001; 123:11791–6. [PubMed: 11716736]
36. Park SH, Son WS, Mukhopadhyay R, Valafar H, Opella SJ. Phage-induced alignment of membrane proteins enables the measurement and structural analysis of residual dipolar couplings with dipolar waves and lambda-maps. J. Am. Chem. Soc. 2009; 131:14140–1. [PubMed: 19761238]

37. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. J. Magn. Reson. 2003; 160:65–73. [PubMed: 12565051]
38. Brunger AT, et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D. Biol. Crystallogr. 1998; 54:905–21. [PubMed: 9757107]
39. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. Chem. Theory Comput. 2008; 4:435–447.
40. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. 1997; 273:283–98. [PubMed: 9367762]
41. Zeng J, et al. High-resolution protein structure determination starting with a global fold calculated from exact solutions to the RDC equations. J. Biomol. NMR. 2009; 45:265–81. [PubMed: 19711185]
42. Hus J-C, Marion D, Blackledge M. Determination of protein backbone structure using only residual dipolar couplings. J. Am. Chem. Soc. 2001; 123:1541–2. [PubMed: 11456746]
43. Prestegard JH, Mayer KL, Valafar H, Benison GC. Determination of protein backbone structures from residual dipolar couplings. Methods Enzymol. 2005; 394:175–209. [PubMed: 15808221]
44. Shealy P, Simin M, Park SH, Opella SJ, Valafar H. Simultaneous structure and dynamics of a membrane protein using REDCRAFT: membrane-bound form of Pf1 coat protein. J. Magn. Reson. 2010; 207:8–16. [PubMed: 20829084]
45. Valafar H, et al. Backbone solution structures of proteins using residual dipolar couplings: application to a novel structural genomics target. J Struct Funct Genomics. 2005; 5:241–254. [PubMed: 15704012]
46. Al-Hashimi HM, et al. Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. J. Magn. Reson. 2000; 143:402–6. [PubMed: 10729267]
47. Shen Y, Vernon R, Baker D, Bax A. De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR. 2009; 43:63–78. [PubMed: 19034676]
48. Tripathy C, Zeng J, Zhou P, Donald BR. Protein loop closure using orientational restraints from NMR data. Proteins. 2011
49. Valafar H, Prestegard JH. Rapid classification of a protein fold family using a statistical analysis of dipolar couplings. Bioinformatics. 2003; 19:1549–55. [PubMed: 12912836]
50. Bansal S, Miao X, Adams MWW, Prestegard JH, Valafar H. Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA). J Magn Reson. 2008; 192:60–68. [PubMed: 18321742]
51. Cormen, TH.; Leiserson, CE.; Rivest, RL.; Stein, C. Introduction To Algorithms. MIT Press; 2001. p. 1180
52. Prestegard JH. New techniques in structural NMR--anisotropic interactions. Nat. Struct. Biol. 1998; 5(Suppl):517–22. [PubMed: 9665182]
53. Ulmer TS, Ramirez BE, Delaglio F, Bax A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc. 2003; 125:9179–9191. [PubMed: 15369375]
54. Miao X, Mukhopadhyay R, Valafar H, MR&VH MX. Estimation of Relative Order Tensors, and Reconstruction of Vectors in Space using Unassigned RDC Data and its Application. J. Magn. Reson. 2008; 194:202–211. [PubMed: 18692422]
55. Mukhopadhyay R, Miao X, Shealy P, Valafar H. Efficient and accurate estimation of relative order tensors from lambda-maps. J. Magn. Reson. 2009; 198:236–47. [PubMed: 19345125]
56. Ruan K, Briggman KB, Tolman JR. De novo determination of internuclear vector orientations from residual dipolar couplings measured in three independent alignment media. J Biomol NMR. 2008; 41:61–76. [PubMed: 18478335]
57. Yao L, Vögeli B, Torchia DA, Bax A. Simultaneous NMR study of protein structure and dynamics using conservative mutagenesis. J Phys Chem B. 2008; 112:6045–6056. [PubMed: 18358021]
58. Losonczi, J. a; Andrec, M.; Fischer, MW.; Prestegard, JH. Order matrix analysis of residual dipolar couplings using singular value decomposition. J. Magn. Reson. 1999; 138:334–42. [PubMed: 10341140]

59. Valafar H, Prestegard JH. REDCAT: a residual dipolar coupling analysis tool. J. Magn. Reson. 2004; 167:228–41. [PubMed: 15040978]
60. Zweckstetter M, Bax A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. J Am Chem Soc. 2000; 122:3791–3792.
61. Bellman R. Bottleneck problems and dynamic programming. Proc. Natl. Acad. Sci. U. S. A. 1953; 39:947. [PubMed: 16589356]
62. Bellman RE, Dreyfus SE. Applied dynamic programming. 1962
63. Pevsner, J. Bioinformatics and Functional Genomics. John Wiley & Sons, Inc.; 2009.
64. Fawcett TM, Irausquin SJ, Simin M, Valafar H. An artificial neural network approach to improving the correlation between protein energetics and the backbone structure. Proteomics. 2013; 13:230–8. [PubMed: 23184572]
65. Fawcett, TM.; Irausquin, S.; Simin, M.; Valafar, H. An Artificial Neural Network Based Approach for Identification of Native Protein Structures Using an Extended ForceField.. Proc. 2011 IEEE Int. Conf. Bioinforma. Biomed.; Atlanta, Georg. USA. Novemb. 12-15, 2011; 2011.
66. Lovell SC, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003; 50:437–50. [PubMed: 12557186]
67. Bau R, et al. Crystal structure of rubredoxin from Pyrococcus furiosus at 0.95 Å resolution, and the structures of N-terminal methionine and formylmethionine variants of Pf Rd. Contributions of N-terminal interactions to thermostability. J. Biol. Inorg. Chem. 1998; 3:484–493.
68. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. J. Am. Chem. Soc. 1998; 120:6836–6837.
69. Ulrich EL, et al. BioMagResBank. Nucleic Acids Res. 2008; 36:D402–8. [PubMed: 17984079]
70. Pettersen EF, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. 2004; 25:1605–12. [PubMed: 15264254]
71. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J. Biomol. NMR. 2009; 44:213–223. [PubMed: 19548092]

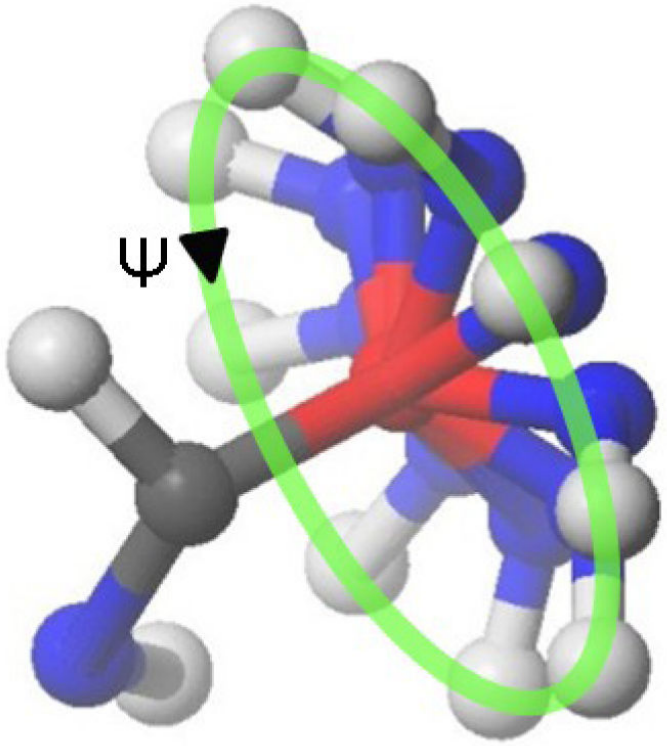


Fig. 1.
Each N-H or C_α-H_α vector is constrained to lie on a cone determined by the preceding two vectors.

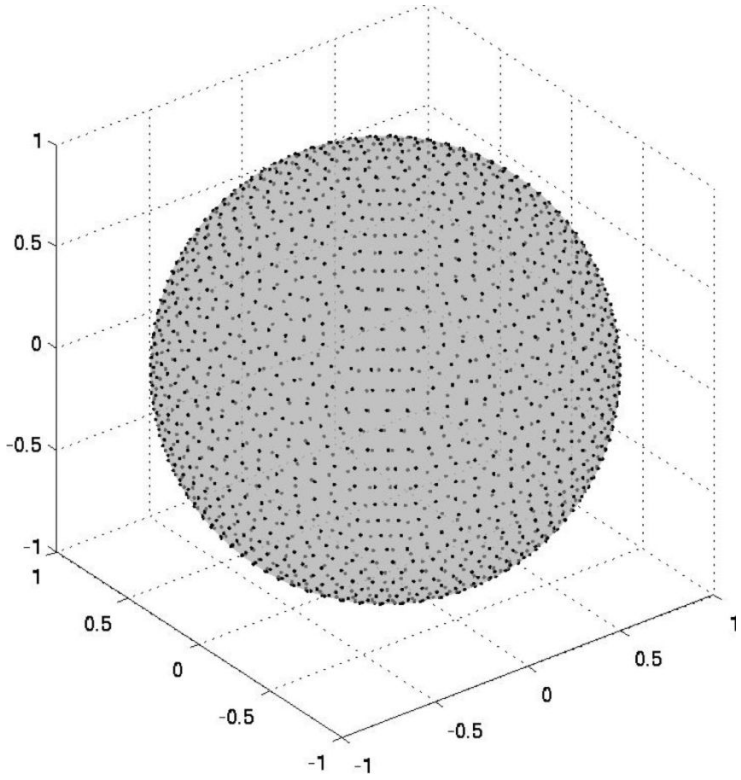


Fig. 2.
Isotropically generated set of vectors at the resolution of 36 (U_{36}).

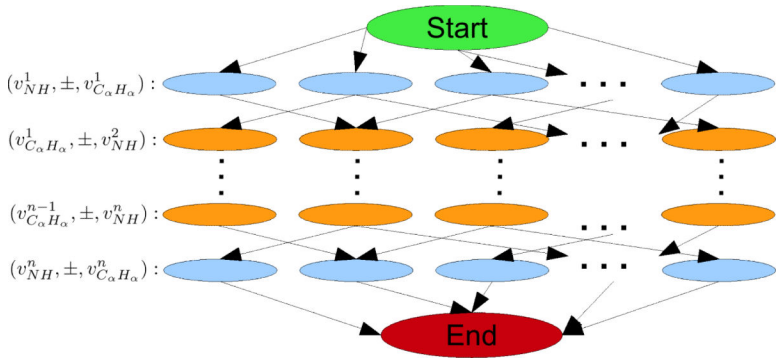


Fig. 3.

Graph structure of state transitions in the vector-orientation-based parameterization of the search space.

Fig. 4.

Protein Fragment defined by A) a ($v^{i}{}_{NH}$,±,$v^{i}{}_{CaHa}$) triple and B) a ($v^{i}{}_{CaHa}$, ±, $v^{i+1}{}_{NH}$) triple.

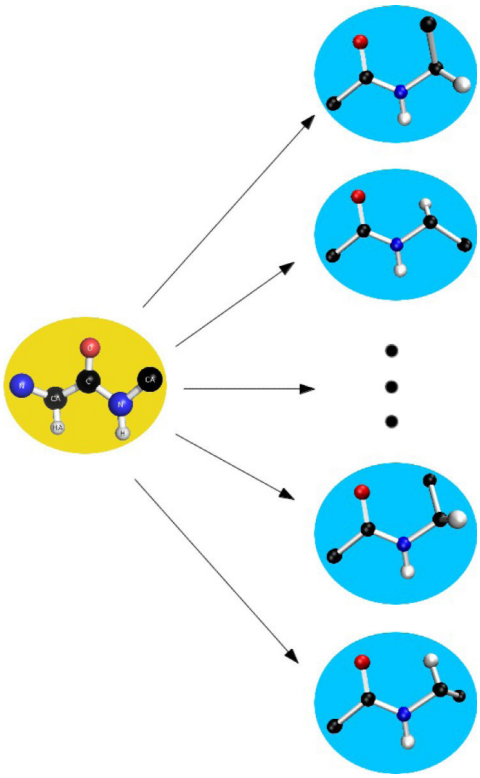


Fig. 5.
Detail of graph connections. Only fragments that overlap without rotation can be connected to form a coherent protein structure.

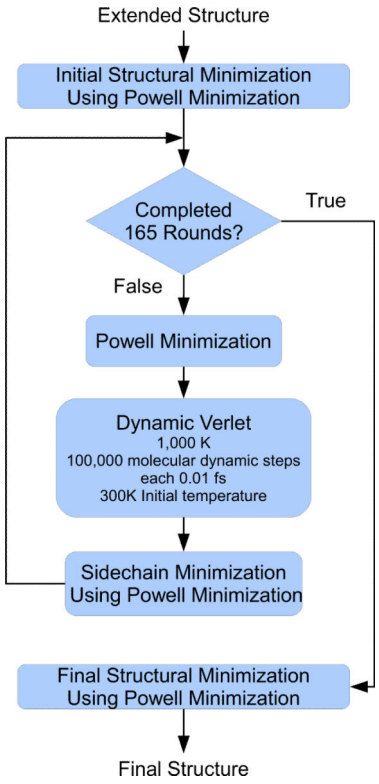


Fig. 6.
Flowchart describing the structure determination protocol used with Xplor-NIH.

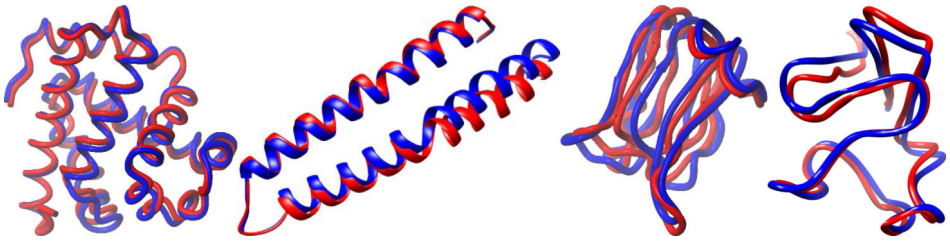


Fig. 7.

Comparison of the backbone structures of proteins 110M, 3LAY, 1F53 and 1BRF (shown in blue from left to right) and the structures produced by DynaFold (shown in red) in each image respectively. The structural alignments to their corresponding published structure are 1.2, 1.3, 1.9 and 1.4Å backbone RMSD respectively. Rendered using Chimera[70].

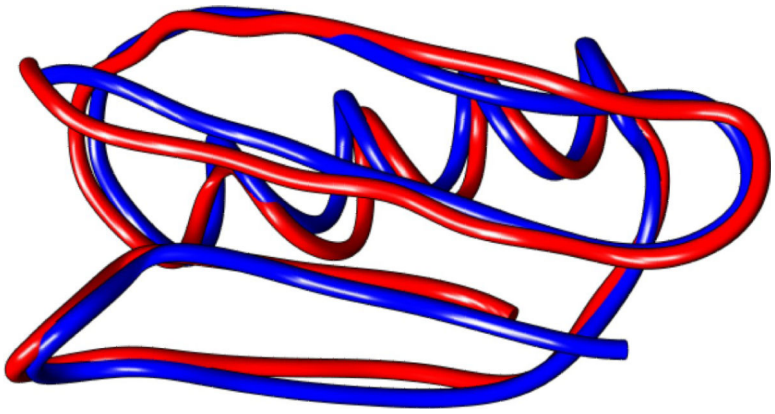


Fig. 8.

Alignment of the published structure of 1P7E (shown in blue) with DynaFold's output (shown in red). Models are aligned on residues 11-55 (1.255Å backbone RMSD). Rendered using Chimera[70].

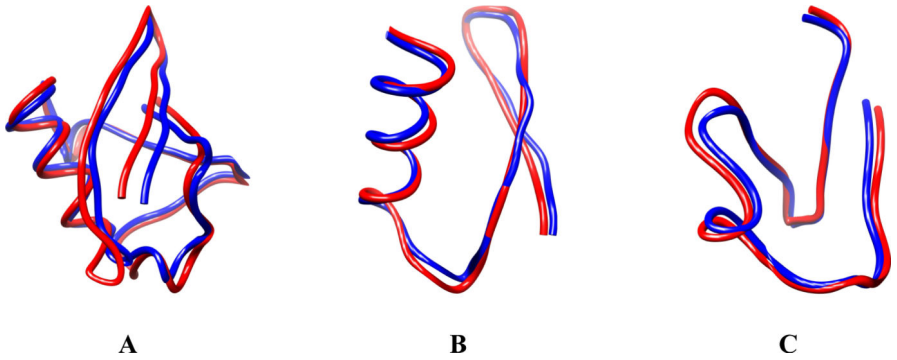


Fig. 9.

Alignment of the published structure of 1D3Z (model 10 shown in blue) with DynaFold's output (shown in red) for: A) residues 1-70 (1.89Å backbone RMSD); B) residues 1-35 (0.835Å backbone RMSD); C) residues 39-70 (0.844Å backbone RMSD); all structures rendered using Chimera[70].

Table 1

1P7E Xplor-NIH energy terms in kilojoules corresponding to three 1P7E structures.

	Angle	Bond	Improper	Vdw	Rama	RDC	Total
Extended Structure	216.674	25.070	169.641	−29.916	−47084.664	19247.096	−27456.099
Output Structure	1734.478	214.069	507.157	−17.434	−0.19E06	2159.607	−0.16E06
Published Structure	213.171	16.742	157.771	−118.036	−0.14E06	1269.918	−0.14E06

Table 2

1P7E RDC RMSD in Hz for three 1P7E structures.

	M1	M2	M3
Extended Structure	20.228	10.916	13.127
Output Structure	8.057	4.669	4.499
Published Structure	2.43	1.512	2.84

Table 3

1D3Z Xplor-NIH energy terms in kilojoules corresponding to three 1D3Z structures.

	Angle	Bond	Improper	Vdw	Rama	RDC	Total
Extended Structure	176.100	33.637	63.905	−48.389	−93390.055	18600.450	−74564.357
Output Structure	1568.123	167.1	576.204	−2.549	−0.25E06	1230.148	−.025E06
Published Structure	171.896	24.382	49.877	−158.065	−0.24E06	253.716	−0.24E06

Table 4

1D3Z RDC RMSD in Hz for three 1D3Z structures.

	M1	M2
Extended Structure	10.033	15.570
Output Structure	3.356	6.010
Published Structure	1.194	1.818