ELSEVIER

PIPATH: An optimized algorithm for generating α-helical structures from PISEMA data

T. Asbury [a], J.R. Quine [b,c], S. Achuthan [b], J. Hu [e], M.S. Chapman [a,d],
T.A. Cross [c,d], R. Bertram [a,b,*]

[a] *Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306, USA*
[b] *Department of Mathematics, Florida State University, Tallahassee, FL 32306-4510, USA*
[c] *National High Magnetic Field Laboratory, Tallahassee, FL 32310, USA*
[d] *Department of Chemistry and Biochemistry, Florida State University, Tallahassee, FL 32306, USA*
[e] *NIDDK, National Institutes of Health, Bethesda, MD 20892, USA*

Abstract

An optimized algorithm for finding structures and assignments of solid-state NMR PISEMA data obtained from α-helical membrane proteins is presented. The description of this algorithm, PIPATH, is followed by an analysis of its performance on simulated PISEMA data derived from synthetic and experimental structures. PIPATH transforms the assignment problem into a path-finding problem for a directed graph, and then uses techniques of graph theory to efficiently find candidate assignments from a very large set of possibilities.
© 2006 Elsevier Inc. All rights reserved.

1. Introduction

Membrane proteins exhibit characteristic resonance patterns in two-dimensional solid-state NMR (ssNMR) experiments. In particular, the polar inversion spin exchange at magic angle (PISEMA) experiment [1] on transmembrane proteins gives distinctive polarity index slant angle (PISA) wheels [2], a direct result of a high degree of α-helicity. These patterns are very useful for determining features of the secondary structure, such as helix tilt and rotation angle [3,4].

PISEMA data are typically obtained from proteins or peptides that have been uniformly labeled, or possibly selectively labeled according to residue type. It is highly desirable to assign the resonance data from such a labeled protein, that is, to match each resonance peak with a residue within the amino acid sequence. This assignment problem is formidable since there there is only one correct assignment that must be chosen from a large number of potential assignments. Yet, even without a correct assignment, there is still vital structural information in the data [2].

Here, we present an algorithm to efficiently extract initial models and plausible assignments from PISEMA data sets of uniformly labeled peptides. There are many similar structures which can be constructed to match the data set, and our algorithm systematically orders these structures based on a user-defined metric of α-helicity. Furthermore, the algorithm is capable of finding the most α-helical (as defined by the metric) assignment and structure that agrees with the data.

Recently, Nevzorov and Opella described an algorithm that generates plausible assignments by "structural fitting" [5]. This algorithm builds atomic structures from random assignments and computes model PISEMA data for each residue as the structure is being built. These data are compared with the experimental PISEMA spectrum to

---

* Corresponding author. Fax: +1 850 644 7244.
  E-mail address: bertram@math.fsu.edu (R. Bertram).

determine if the assignment is plausible. In this way, an assignment of the data and an initial model are determined simultaneously. The search is restricted to α-helical structures, thus eliminating many potential assignments. Furthermore, the search algorithm can be computationally expensive, with the bulk of the procedure's computation spent iteratively building structures.

We present an algorithm, PIPATH, that provides plausible assignments and associated structures without the computational cost of atomic structure construction during the search process. This provides a substantial speedup in search time, allowing for a more complete search and application to larger PISEMA data sets. The input to PIPATH is a PISEMA data set, and the output is a set of potential assignments and their most α-helical structures. The algorithm is thus intended as a first step toward model building.

2. The PISEMA search space

2.1. PISEMA data and its degeneracies

The PISEMA experiment measures two physical qualities of the target nuclei which provide orientation information: the anisotropic chemical shift (σ) and dipolar coupling (v). For the protein backbone ^1H–^{15}N interaction, a single data pair (σ, v) is obtained for each labeled peptide plane in the molecule. Plotting all data points on a (σ, v) coordinate system gives the PISEMA spectrum, which falls within the *powder pattern*, bounded by the PISEMA ellipse and triangle in the frequency plane [6,7] (Fig. 1). The powder pattern is the locus of possible measurements and can be determined by forming a "powder" sample containing a large number of randomly distributed orientations.
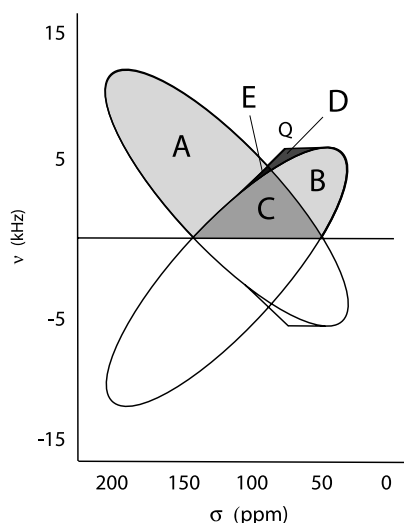

Fig. 1. A typical PISEMA powder pattern bounded by a primary and reflected ellipse and a small extra-elliptical triangle near point Q. The experimental data value (σ, $|v|/2$) falls within one of the shaded regions (A–E). The number of peptide plane degeneracies varies according to the location of the data: points in regions (A) and (B) have 4, (C) and (D) have 8, and (E) has 12.

A PISEMA data set with N data points from a uniformly labeled peptide has $N!$ potential assignments. This is a very large number, even for smaller transmembrane proteins; e.g., data from a 15-residue peptide contain 1,307,674,368,000 (15!) potential assignments. This number can be reduced by using secondary structural characteristics that may be present in the data set, such as helicity.

An assignment is an ordering of the data points in the frequency plane, and two consecutive elements of the assignment correspond to two consecutive residues in the protein, and the two associated peptide planes form a *diplane*. Using any consecutive pair of points, a diplane can be formed and its set of possible ϕ and ψ torsion angles can be determined. This is discussed in detail in [8], and is briefly summarized in Appendix A.

More than one possible torsion angle pair exists for each diplane because of orientational degeneracies contained in the PISEMA data. The number of possible torsion angles varies according to the location of the consecutive data points in the PISEMA powder pattern. For example, a resonance within region A of Fig. 1 followed by another in region C has $4 \times 8 = 32$ possible torsion angle pairs, one for each combination of peptide plane degeneracies between the two peptide planes. Depending on the regions in which the data points lie, there will normally be 16 or 32 torsion angle pairs [8]. Higher degeneracies are possible, but not probable since transmembrane helices often have small helical tilt angles of $20° \pm 10°$ [9,10], resulting in a positive dipolar coupling with resonances predominately in region A of Fig. 1 [11].

An assignment therefore has many possible structures that can match its PISEMA data set. Specifically, for an assignment of an N residue peptide, the number of structures is:

$$\prod_{l=1}^{N-1} T_l, \tag{1}$$

where T_l is the number of possible torsion angles connecting residues r_l and r_{l+1} in the assignment. Since there are $N!$ possible assignments, each with multiple structures, the number of candidate structures that can match a given PISEMA data set of size N is at most:

$$\text{num}(A_N) = \sum_{k=1}^{N!} \prod_{l=1}^{N-1} T_{kl}, \tag{2}$$

where A_N denotes the PISEMA search space, or set of possible structures that agrees with the PISEMA data.

Two separate phenomena contribute to the extremely large size of A_N. The first is the combinatorial nature of possible assignments, and the second is the large number of possible structures per assignment due to degeneracies.

2.2. Reduction of the PISEMA search space

The set of structures that match PISEMA data, A_N, is too large to examine exhaustively. Here, in an effort to

reduce this problem to manageable size, we order the set A_N according to degree of α-helicity.

The key assumption in our ordering of A_N is the regularity of α-helices within the transmembrane environment. There is evidence that transmembrane α-helices are more stable than equivalent α-helices in aqueous environments [14]. This increased stability should result in a higher degree of helical regularity among transmembrane proteins. A PISEMA experiment performed on a transmembrane α-helical structure should thus yield a highly regular α-helix. With this assumption, it is possible to discard a large number of possible yet improbable non-α-helical structures.

A tabulation of transmembrane α-helices currently available in the Protein Data Bank (PDB) shows the mean torsion angle and variance to be $\overline{\phi} = -63.31° \pm 10.94°$ and $\overline{\psi} = -41.99° \pm 11.42°$ (Table 1), which lie between the canonical α-helix model values $-65° \leqslant \phi \leqslant -60°$ and $-45° \leqslant \psi \leqslant -40°$ [14]. We define an α-helical subset of A_N, denoted as A_N^α, as those structures that match the PISEMA data and have (ϕ, ψ) within 10° of the ideal values $(\phi^\alpha = -63°, \psi^\alpha = -42°)$. However, even with this sizable reduction, the search space A_N^α is still quite large and searching the set of candidate structures is very time consuming.

Nevzorov and Opella [5] employed a Monte Carlo search technique to explore a set similar to A_N^α. Here, we describe a new algorithm, PIPATH, that uses graph theoretical techniques to more efficiently search A_N^α, the set of α-helical structures that match a PISEMA data set.

3. The PIPATH algorithm

3.1. The assignment graph

Let $G = (N, E)$ be a graph with N vertices and E edges. Each vertex corresponds to a single PISEMA data point and each pair of vertices is connected by two directed edges, making G a well-connected directed graph with $E = N(N - 1)$.

Let P_k be a path through the graph in which each vertex is visited exactly once (a Hamiltonian Path) [15]. P_k then corresponds to a unique assignment of the data. Since a data set of N points has $N!$ possible assignments, a well-connected graph $G = (N, E)$ has $N!$ Hamiltonian paths. A graph G with N vertices thus contains all possible assignments of the data set. This graph is hereafter referred to as the *assignment graph* (Fig. 2).

An edge e_{ij} connecting two vertices v_i and v_j in the assignment graph is equivalent to a diplane. The set of

Table 1
Transmembrane proteins from the PDB used to calculate α-helical torsion angle statistics

PDB ID	Protein	TM α-helices	$\overline{\phi}$	$\overline{\psi}$	Res(Å)
1C17	F1F0 ATP synthase	7	−64.1 ± 9.2	−42.4 ± 9.0	3.0
1C3W	Bacteriorhodopsin	7	−66.1 ± 8.3	−40.2 ± 7.6	1.9
1E12	Halorhodopsin	7	−64.3 ± 5.9	−41.2 ± 7.5	1.8
1EHK	Ba3 cytochrome-c oxygenase	14	−62.9 ± 8.5	−41.0 ± 10.4	2.4
1EZV	Cytochrome BC1 complex	10	−65.3 ± 7.2	−40.2 ± 7.0	2.3
1FFT	Ubiquinol oxidase	22	−63.2 ± 22.4	−41.7 ± 22.1	3.5
1FX8	*Escherichia coli* glycerol facilitator	6	−65.7 ± 10.6	−40.7 ± 10.4	2.2
1H6I	Aquaporin	6	−61.2 ± 9.4	−44.6 ± 8.2	3.5
1IWG	Multidrug efflux transporter	11	−62.1 ± 18.4	−43.7 ± 17.4	3.5
1JB0	Photosynthetic reaction center	23	−64.6 ± 8.1	−41.0 ± 9.1	2.5
1JGJ	Sensory rhodopsin II	7	−63.6 ± 6.3	−42.1 ± 7.6	2.4
1KQF	Formate dehydrogenase N	5	−63.7 ± 5.8	−43.0 ± 7.1	1.6
1L7V	ABC transporter	8	−54.7 ± 18.8	−43.9 ± 18.0	3.2
1L9H	Bovine rhodopsin	7	−67.0 ± 11.0	−39.5 ± 11.7	2.6
1LGH	Light harvesting complex	2	−64.4 ± 5.1	−40.0 ± 6.5	2.4
1MSL	Large mechanosensitive channel	2	−55.9 ± 10.4	−46.7 ± 13.7	3.5
1MXM	Small mechanosensitive channel	3	−65.3 ± 10.9	−40.7 ± 13.9	3.9
1OCC	aa3 Oxidoreductase cytochrome-c	27	−62.7 ± 8.5	−42.3 ± 9.5	2.8
1OKC	Mitochondrial ADP/ATP carrier	3	−65.0 ± 4.3	−42.4 ± 6.9	2.2
1P7B	Inward rectifier Ka channel	3	−53.8 ± 16.1	−49.8 ± 18.6	3.7
1PRC	Photosynthetic reaction center	10	−63.0 ± 8.5	−42.1 ± 10.2	2.4
1Q16	Nitrate reductase A	5	−65.2 ± 7.3	−43.0 ± 7.5	1.9
1QLA	Fumerate reductase complex	5	−63.5 ± 6.5	−41.6 ± 10.2	2.2
1R3J	KCSA potassium channel	2	−63.8 ± 3.2	−42.3 ± 5.0	1.9
1RHZ	SecYEβ channel	11	−55.5 ± 13.3	−49.2 ± 16.0	3.5
1SU4	Calcium ATPase	7	−67.5 ± 13.0	−38.5 ± 12.9	2.4
1VF5	Cytochrome B6F complex	14	−63.7 ± 16.3	−42.6 ± 15.2	3.0
	Total	234	−63.31° ± 10.94°	−41.99° ± 11.42°	

Helices of length $N \geqslant 20$ were detected using the Kabsch and Sander algorithm [12]. Membrane traversal was verified using TMHMM, a transmembrane hidden Markov model server [13].

possible torsion angles between the peptide planes is determined by the (σ, v) values at each vertex. This set of angles is given in Appendix A. There is a large number of structures $S(P_k)$ corresponding to assignment P_k, due to torsion angle degeneracies.

To specify a structure $S_k \subset S(P_k)$ that is consistent with an assignment represented by path P_k, one (ϕ, ψ) pair must be chosen for each edge. The closeness of a (ϕ, ψ) pair to a canonical α-helix ($\phi^\alpha = -65°$, $\psi^\alpha = -40°$) can be measured using a simple root mean squared deviation (RMSD):

$$\Delta^\alpha(\phi, \psi) = \sqrt{(\phi - \phi^\alpha)^2 + (\psi - \psi^\alpha)^2}. \tag{3}$$

In PIPATH, we choose $(\phi^\star, \psi^\star)$ that minimizes Δ^α. This minimum value of Δ^α is then used as a weight w_{ij} for the edge e_{ij}

$$w(e_{ij}) = \min[\Delta^\alpha\{\phi\psi\}_{ij}] = \Delta^\alpha(\phi^\star, \psi^\star), \tag{4}$$

where $\{\phi\psi\}_{ij}$ is the set of possible torsion angles corresponding to the edge e_{ij}, i.e., connecting vertices i and j.

With $(\phi^\star, \psi^\star)$ chosen in this way, any Hamiltonian path P_k in the assignment graph represents an assignment and its edge weights reflect the closeness of the most α-helical structure in $S(P_k)$, denoted as $S_k^\star$, to a canonical α-helix. It is important to note that $w(e_{ij}) \neq w(e_{ji})$ since the internal torsion angle equations are not commutative (Appendix A), and thus the assignment graph G is a directed graph.

3.2. The continuity conditions

When building atomic models by joining diplanes defined by torsion angles, there are geometric constraints that limit the number of possible internal torsion angles. The process of joining any two diplanes involves the gluing a common internal peptide plane, which must have the same orientation in both diplanes. This orientational restriction propagates through the structure as it is being built. We call these additional constraints *continuity conditions*. They are discussed in [16,17] and are described in detail in [8]. Continuity conditions exist whenever diplanes are glued together, regardless of secondary structure.

PIPATH uses the geometric relations between any two PISEMA data points to compute an α-helical torsion angle for its respective diplane. Since the continuity conditions are dependent on consecutive diplanes, it is not possible to apply the continuity condition to the assignment graph because the ordering of the diplanes (the minimal path) is the unknown being solved. However, once a path has been found, the continuity conditions can be applied as a postprocess.

3.3. Solving for the minimal α-helical structure

Each path P_k in the assignment graph G has a cost $C(P_k)$ that reflects the α-helicity of the structure $S_k^\star$ and is computed by adding the edge weights of P_k (Fig. 3). A useful quantity is the deviation from α-helicity for a given structure S:

$$\Delta^\alpha S = \sum_1^{N-1} \Delta^\alpha(\phi^S, \psi^S)_{i(i+1)}, \tag{5}$$

where $(\phi^S, \psi^S)_{i(i+1)}$ are the internal torsion angles of structure S. For each structure $S \subset S(P_k)$, $\Delta^\alpha S \geqslant C(P_k)$, since the edge weights of P_k are the minimal torsion angle α-helical deviations by definition (4). $S_k^\star$ is the structure in $S(P_k)$ in which the deviation from a canonical α-helix is minimal, i.e., where $\Delta^\alpha S_k^\star = C(P_k)$.

Let $P^\star$ be the path in G that minimizes $C(P)$. The most α-helical structure associated with $P^\star$ is denoted $S^{\star\star}$. If $\Delta^\alpha S^{\star\star} = C(P^\star)$, $S^{\star\star}$ is the most α-helical structure in A_N^α, and $P^\star$ is its associated assignment. If $\Delta^\alpha S^{\star\star} > C(P^\star)$, then the structure whose cost was $C(P^\star)$ did not satisfy the continuity conditions, and $S^{\star\star}$ is the structure with the lowest cost that does satisfy the conditions.

The goal of PIPATH is to find the most α-helical structures which match the PISEMA data set. It does this by searching and ranking structures based on the cost of paths through the assignment graph. This requires finding the Hamiltonian paths P with minimal cost $C(P)$ in the assign-
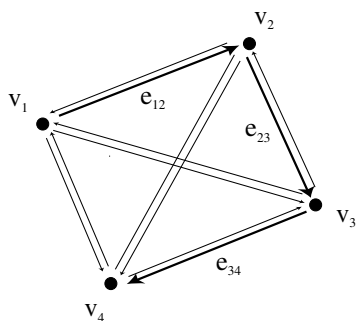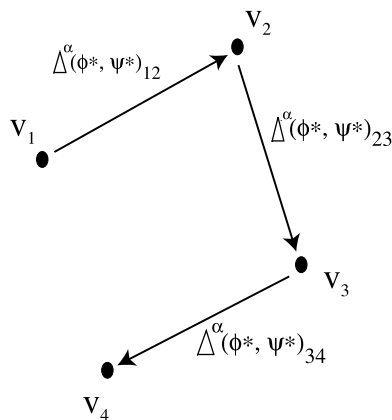

Fig. 2. Assignment graph $G = (4, 12)$ with one path illustrated (bold arrows) corresponding to the assignment (v_1, v_2, v_3, v_4) of data points (d_1, d_2, d_3, d_4) to a 4-residue peptide. The vertex and data point labels are arbitrary but uniquely specify an assignment.

Fig. 3. A path P_k through assignment graph $G = (4, 12)$. The cost of P_k is $C(P_k) = \Delta^\alpha(\phi^\star, \psi^\star)_{12} + \Delta^\alpha(\phi^\star, \psi^\star)_{23} + \Delta^\alpha(\phi^\star, \psi^\star)_{34}$, where $(\phi^\star, \psi^\star)$ is the torsion angle pair for edge e_{ij} that is closest to α-helical.

ment graph G. This is a well-studied problem in graph theory. It is transformable to the Traveling Salesman Problem and a variety of methods are available to solve this problem with reasonable computational cost [15]. The details and implementations of this algorithm are described below.

3.4. Implementation and availability

The algorithm we use is formally described in Appendix B. The input parameters are the primary sequence, the PISEMA data and an α-helicity bound B. The program calculates all possible assignments P_k whose $C(P_k) \leqslant B$.

The path-finding algorithm requires solving the Traveling Salesman Problem. This is done using a branch-and-bound technique [18,19] which limits the search space based on the input α-helicity bound B. Careful choice of B prevents long search times while allowing for sufficient sampling of paths. In our implementation, B was initially set to 0 and slowly increased until a prescribed number of assignments (100) was returned.

PIPATH generates a list of plausible assignments that can yield structures with high α-helicity. For each path, the minimal α-helical structure that satisfies the continuity conditions must be computed. Here, we use an analytic expression of the continuity condition [8] that efficiently determines whether consecutive torsion angles meet the continuity condition.

The algorithm was implemented in the Python programming language [20] and tested on a Linux PC operating at 2.2 GHz. Calculation time for PIPATH has a strong dependence on peptide chain length N. For our tests of generating 100 α-helical structures, with $N \leqslant 15$, run times averaged under 5 min. For longer peptides ($15 < N \leqslant 25$) the average calculation times ranged to several hours. The Python implementation is freely available at http://www.math.fsu.edu/~bertram/software/sb. We request that those who use this software reference this article.

4. Example

As an example of how PIPATH is used, we consider a set of five PISEMA resonances as shown in Fig. 4A. These data were generated by calculating the dipolar coupling and chemical shift from an α-helix with 25° tilt and 5° torsion angle deviation. The associated assignment graph is shown in Fig. 4B. Each edge of this directed graph is weighted according to (4), which is the minimal α-helical deviation for two peptide planes connecting the corresponding vertices. A large edge weight, such as $w(e_{53}) = 30$ indicates that it is unlikely that the peptide plane associated with resonance 3 immediately follows that associated with resonance 5. However, $w(e_{51}) = 4$ is small, so that the peptide plane associated with resonance 1 is more likely to follow that associated with resonance 5. The assignment graph was restricted to include only those torsion angles with $\Delta^\alpha \leqslant 30°$.

PIPATH computes α-helical structures by finding Hamiltonian paths of minimal cost through the assignment graph. The top 10 paths of the assignment graph shown in Fig. 4B are listed in Table 2. Note that for path P_1, $\Delta^\alpha S^\star > C(P_1)$. This indicates that the path with smallest cost did not satisfy the continuity conditions. In contrast, path P_3 generated a minimal structure with $\Delta^\alpha S^\star = C(P_3)$ and thus satisfied these conditions. For each path, the structure with minimal α-helical deviation ($S^\star$) is constructed and its RMSD from the original structure is calculated. In this example, the assignment corresponding to P_3 yields a structure $S^\star$ with deviation $\Delta^\alpha S^\star = 27$. Since this is the minimal α-helical deviation for those structures from paths with $C(P_i) \leqslant 27$ and all remaining paths have $\Delta^\alpha S^\star \geqslant C(P_i) > 27$, $S^\star(P_3)$ is the most α-helical structure within A_N and is denoted $S^{\star\star}$.

5. Algorithm performance

5.1. Performance

We first test the performance of PIPATH using simulated data derived from synthetic model α-helices of varying length, degree of α-helicity, and tilt. For each model, which we call the "generating model", the anisotropic ^{15}N chemical shift and ^{1}H–^{15}N dipolar coupling interaction are computed for all backbone nitrogens to generate a PISEMA resonance set [11]. The success rate of the algorithm is then determined by comparing output structures with the synthetic input models and measuring the root mean-square deviation (RMSD).

Because PIPATH computes all paths below an upper bound B, it is possible to output many α-helical structures for a given data set. In the following tests, we computed multiple α-helical structures to characterize PIPATH performance, and the figures show data for the 100 most α-helical structures that PIPATH computes.

The performance of PIPATH as a function of peptide length N is shown in Fig. 5. Since RMSD magnitude depends on the size of the structures being compared, we use a normalized RMSD [21] whenever comparing PIPATH behavior across peptides of different lengths. The RMSD values were normalized relative to the smallest peptides in our data set ($N = 5$) using the formula:

$$rmsd_5 = \frac{rmsd}{1 + \ln\sqrt{N/5}}, \tag{6}$$

where $rmsd_5$ is the normalized RMSD. For each length N, an ensemble of 100 random α-helices were generated with Gaussian noise in the tilt angle and the α-helicity. If noise were not added to the α-helicity, it would be guaranteed that the minimal PIPATH structure would match the original structure. The mean and standard deviation of the helical tilt were $\mu = 20°$, $\sigma = 10°$, reflecting the naturally occurring distribution for membrane proteins [9,10]. The helices
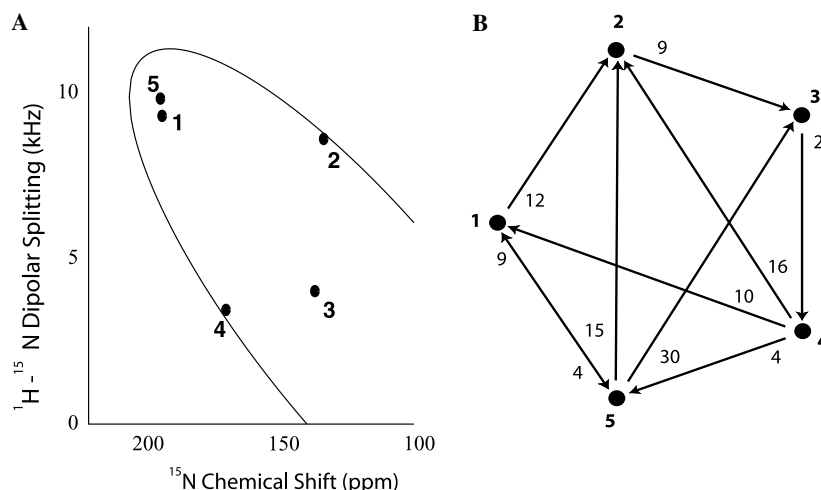
Fig. 4. A sample PISEMA resonance data set (A) and its associated assignment graph (B) for a uniformly labeled 5-residue peptide. The data set shown in the frequency plane has an arbitrary numbering 1 through 5. The assignment graph for the 5 resonance pairs is a directed graph with edge weights equal to the minimal α-deviation of torsion angles between diplanes, as defined by (4). It has been pruned here to only include those edges with $\Delta^\alpha \leqslant 30°$. Edge weight $w(e_{ij})$ is located close to vertex i.

Table 2
The top 10 paths as determined by path cost $C(P_i)$ and their associated minimal α-helical structures $S^\star$ for the assignment graph shown in Fig. 4

	Path	$C(P_i)$	$\Delta^\alpha S^\star$	RMSD (Å)	
P_1	[2,3,4,5,1]	19	47	0.809	
P_2	[3,4,5,1,2]	22	71	0.907	
P_3	[1,2,3,4,5]	27	27	0.015	$S^{\star\star}$
P_4	[5,1,2,3,4]	27	64	0.551	
P_5	[4,5,1,2,3]	29	78	0.999	
P_6	[2,3,4,1,5]	30	54	0.941	
P_7	[1,5,2,3,4]	35	63	0.452	
P_8	[5,2,3,4,1]	36	36	0.182	
P_9	[3,4,1,5,2]	36	76	0.924	
P_{10}	[4,1,5,2,3]	43	83	0.995	

For each structure, the minimal α-helical deviation (Δ^α) and its RMSD from the original structure is calculated. Since the continuity condition can only increase the α-deviation, $\Delta^\alpha S^\star \geqslant C(P_i)$. P_3 is the assignment which generates the optimal α-helical structure ($S^{\star\star}$) which matches the data set.

within the ensemble had torsion angle means of $\phi^\alpha = -63°$, $\psi^\alpha = -42°$, and a standard deviation of $\sigma = 5°$ (Fig. 5).

Fig. 5 shows that top structures determined by PIPATH closely match the structure used to simulate the PISEMA data. For each length N, the top 100 structures output by PIPATH are compared against the generated model using normalized RMSD. Most of the top structures PIPATH generates are typically within 1 Å of each other, and in Fig 5A all have normalized RMSD within 1 Å of the generating model. The most α-helical structures (shown as open diamonds) are generally closer to the generating model. Fig. 5 also shows PIPATH performance is negatively impacted by increasing peptide length, which is expected since the size of the search space A_N^α is proportional to N. In addition, the performance of the algorithm is not as good when the generating model has greater α-helical deviation (data not shown).

The dependence of PIPATH performance upon helical tilt was measured by fixing peptide length and α-helicity and then varying the tilt angle from 0° to 90° (Fig. 6). PIPATH performed best on structures with tilt angles less than ∼30°; fortunately, naturally occurring α-helices in membrane proteins have mean tilt of less than 30° [9,10]. PIPATH performed worse on structures with tilt of 40–70°. This is because ideal α-helices within this tilt range exhibit unresolvable dipolar splitting signs and have data values which typically lie in region C of the powder pattern of Fig. 1 [2]. In this case, the undetermined sign of ϵ_1 generates a second set of torsion angles (Appendix A) which match the data and thus increases the size of A_N^α.

We next tested PIPATH on the high resolution structures (<2.5 Å) of the transmembrane α-helical data set (Table 1) by numerically generating PISEMA data for these structures, applying PIPATH on these data, and then calculating the RMSD of PIPATH's 100 most α-helical structures from the generating structures (Fig. 7). Peptides of varying lengths were generated by truncation of 20-residue α-helices. PIPATH performance for these experimentally determined α-helices is similar to that for the synthetic α-helices (Fig. 5).

5.2. Further reductions to PISEMA search space

The size of the α-helical search space, A_N^α, can be further reduced by applying additional constraints. The most effective way to limit A_N^α is to use non-uniform labeling of residues. In general, the selective labeling of M residues of a single amino acid type reduces the search space size from $\mathrm{num}(A_N^\alpha)$ to $\mathrm{num}(A_{N-M}^\alpha)M!$, compared to $\mathrm{num}(A_{N-M}^\alpha)$ for specific labeling of M individual residues. Increasing M is always advantageous, and consecutive labeling reduces the total number of edges in the assignment graph, resulting in greater algorithmic efficiency. Thus, a selective label
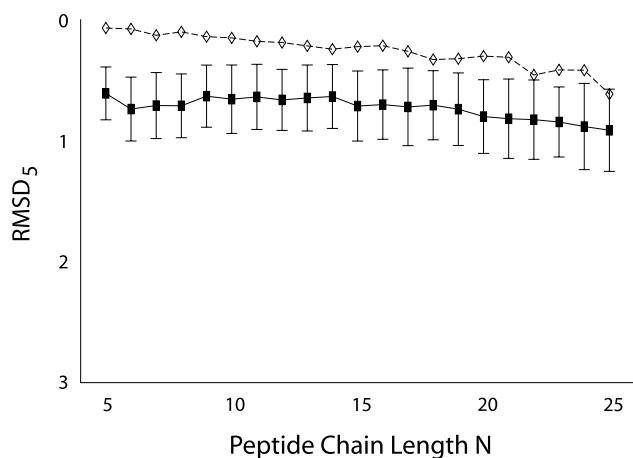
Fig. 5. PIPATH performance using simulated transmembrane α-helices as a function of peptide chain length for different degrees of α-helicity. Torsion angles have mean of $\phi^\alpha = -63°$, $\psi^\alpha = -42°$ and standard deviation of 5°. The black squares with error bars are the normalized RMSD mean and deviation of the top 100 PIPATH structures returned by PIPATH as ranked by assignment path cost. The diamonds represent the average of the most α-helical structures that matches the PISEMA data.

Fig. 7. PIPATH performance as a function of peptide length using 62 transmembrane α-helices from the PDB (see Table 1). PISEMA data were calculated numerically for each structure, and the normalized RMSD was calculated between the top 100 PIPATH output structures and the generating structure. The black squares with error bars are the RMSD mean and deviation of the top 100 PIPATH structures. The diamonds represent the average of the most α-helical structures.
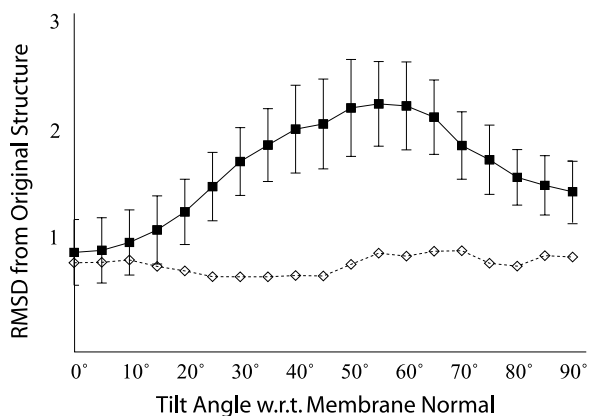

Fig. 6. PIPATH performance as a function of helical tilt angle. At each tilt angle, 100 random peptides ($\bar{\phi} = -63°$, $\bar{\psi} = -42°$, $\sigma = 5°$) of length $N = 10$ were tested. Each random peptide is then compared with the top 100 α-helical structures as computed by PIPATH. The black squares with error bars are the RMSD mean and deviation of the top 100 structures. The diamonds show the averages of the most α-helical structures of each set.

Fig. 8. PIPATH performance on the transmembrane α-helical data set (Table 1) with additional constraints placed on output structures. The white bar shows PIPATH success rate with no filtering on output structures. The black bar removes all structures whose simulated PISEMA data does not form a PISA wheel which rotates in a single direction in the frequency plane. The striped bar removes all structures with no handedness and which have a torsion angle pair with α-helical deviation >30°.

which affects 4 residues is preferred over a specific labeling of 3, and given a choice, labeling 2 residues that are consecutive in primary sequence is preferred over labeling 2 residues that are non-consecutive.

The search space can also be reduced by enforcing a strict upper bound on deviation of torsion angles from canonical values. This generates more regular α-helices, but limits the detection of kinks in the structure. For example, one may wish to include only those torsion angles with $\Delta^\alpha(\phi, \psi) < 30°$ in the assignment graph.

The assignment space can be further reduced by eliminating output structures that do not produce an expected PISA wheel. For example, if part of a structure produces
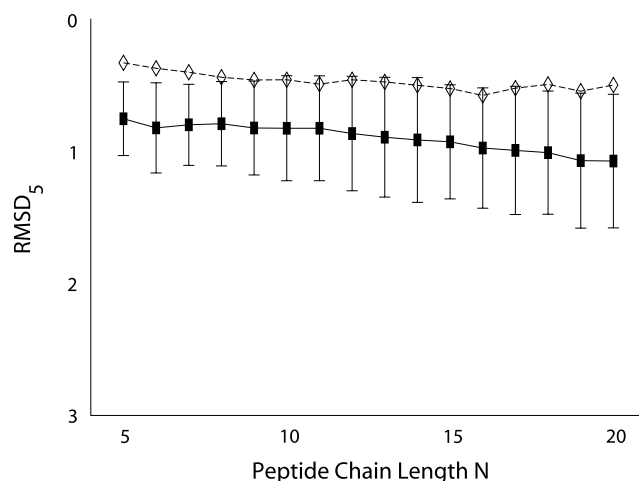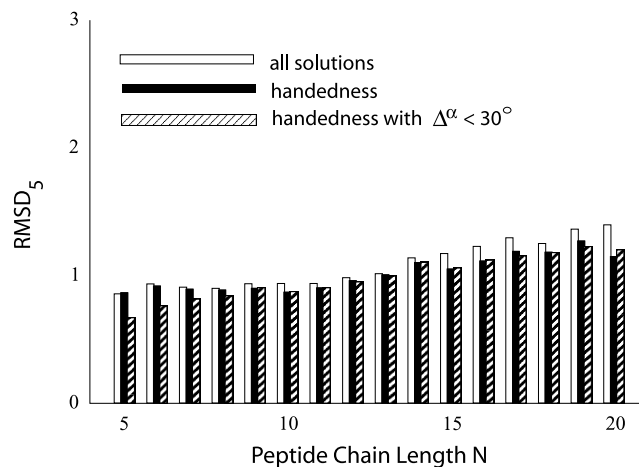
a right-handed PISA wheel while another section produces a left-handed wheel, it may be appropriate to delete the structure [14]. Fig. 8 shows the results of applying these additional constraints to PIPATH performance.

6. Discussion

PIPATH uses principles from graph theory to find plausible initial models and assignments of PISEMA data. It produces an ordered set of structures and assignments ranked by an α-helicity metric. The highest-ranked structures are

those that are closest to a canonical α-helix. PIPATH addresses the same problems (initial modeling and assignment) as the Nevzorov and Opella algorithm [5]. However, PIPATH more efficiently searches for optimal α-helical structures.

Although PIPATH treats data from each residue as a single data point, a typical data set will have data peaks of finite width. The issue of interpreting line shape as a source of experimental error is examined in detail in [7]. There, it is shown that typical error bars within the dipolar coupling and chemical shift dimensions result in small torsion angle variations. Larger variations are possible depending on where the data lies on the frequency plane. Since PIPATH relies on torsion angle calculations to measure α-helicity, the algorithm is in most cases robust to a small amount of experimental error.

It is well known that the PISEMA data set contains much structural information that is available without assignment. Our work with PIPATH confirms this, as the algorithm can generate a large number of structures with different assignments, yet all match the data equally well and are structurally similar. Indeed, for peptide chains of length 20 or greater, there can be thousands of assignments whose optimal α-helical structures deviate <0.5 Å from each other. The set of "top" structures produced by PIPATH are relatively good matches to original structures from which the PISEMA data was generated. Thus, the output from PIPATH is a set of possible initial models for subsequent atomic refinement.

Additional experimental contraints, as long as they can be expressed as orientational constraints, can be incorporated into PIPATH. These new constraints will reduce the number of available torsion angles, affecting the number of degeneracies in Eqs. (7) and (8) of Appendix A. This will result in the reduction of the PISEMA search space and presumably give better results.

The variability of output structures from PIPATH can be controlled in several ways. Output structures can be biased toward greater α-helicity by reducing the upper bound (B), by eliminating edges in the assignment graph when Δ^α is too large, or by post-processing to enforce handedness in the associated PISA wheel for the output structure. Further culling of structures may be possible with additional structural information.

Acknowledgments

Appendix A. Torsion angle equations

A complete derivation of the torsion angle equations is given in [8,22]. Here, we give the equations with only a brief description.

The set of possible torsion angles $\{\phi\psi\}_{ij}$ between residues connecting two consecutive PISEMA data points $(\sigma, v)_i$ and $(\sigma, v)_j$ are given by the following:

$$\phi = \arg\left(-\epsilon_1^i \mu_1 + \epsilon_1^i \mu_2 \kappa_1, \epsilon_2^i g(\mu_1, \mu_2, \kappa_1)^{1/2}\right)$$
$$+ \arg\left(\epsilon_1^j \mu_3 - \epsilon_1^i \mu_2 \kappa_2, -\epsilon_c g(\epsilon_1^i \mu_2, \epsilon_1^j \mu_3, \kappa_2)^{1/2}\right), \quad (7)$$

$$\psi = \arg\left(-\epsilon_1^i \mu_2 + \epsilon_1^j \mu_3 \kappa_2, \epsilon_c g(\epsilon_1^i \mu_2, \epsilon_1^j \mu_3, \kappa_2)^{1/2}\right)$$
$$+ \arg\left(\epsilon_1^j \mu_4 - \epsilon_1^i \mu_3 \kappa_3, -\epsilon_2^j g(\mu_3, \mu_4, \kappa_3)^{1/2}\right), \quad (8)$$

where ϵ_1^i, ϵ_2^i are the degeneracies associated with data point $(\sigma, v)_i$, and ϵ_1^j, ϵ_2^j are the degeneracies for $(\sigma, v)_j$. The constants κ_{1-3} and μ_{1-4} are determined by the peptide geometry and the resonance points, respectively. For a typical peptide geometry [23] with torsion angle $\omega = 180°$,

$$\kappa_1 = \cos 59°, \quad \kappa_2 = \cos 70°, \quad \kappa_3 = \cos 65°. \quad (9)$$

For a ^{15}N chemical shift reference frame, offset by an angle β (typically 17°) from the ^1H–^{15}N bond vector,

$$\mu_1 = \cos(-\beta + 32°) \cdot B_x^i + \cos(\beta + 58°) \cdot B_z^i, \quad (10)$$

$$\mu_2 = \cos(\beta + 27°) \cdot B_x^i + \cos(\beta + 117°) \cdot B_z^i, \quad (11)$$

$$\mu_3 = \cos(\beta + 33°) \cdot B_x^j + \cos(\beta + 123°) \cdot B_z^j, \quad (12)$$

$$\mu_4 = \cos(-\beta + 32°) \cdot B_x^j + \cos(\beta + 58°) \cdot B_z^j, \quad (13)$$

where $\boldsymbol{B}^i = [B_x^i, B_y^i, B_z^i]$, $\boldsymbol{B}^j = [B_x^j, B_y^j, B_z^j]$ are unit direction vectors of the magnetic field in the chemical shift reference frames of $^{15}N^i$ and $^{15}N^j$ residues, respectively. The principal values of the ^{15}N chemical shift tensors are assumed to be constant. Note that the arguments to the cos functions differ from [8], where a dipolar reference frame is used, but the μ values are dot products and thus do not depend upon reference frame.

The gramian function $g(x, y, z)$, used in (7) and (8), is defined as

$$g(x, y, z) = \begin{vmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{vmatrix} = 1 - x^2 - y^2 - z^2 + 2xyz. \quad (14)$$

The gramian should never be negative. However, this may occur if there are errors in the PISEMA data [24]. In such cases, we set $g(x, y, z) = 0$. The arg function is defined as

$$\arg(x, y) = \arctan(y/x). \quad (15)$$

To generate the degeneracies, $\{\epsilon_1^i, \epsilon_2^i, \epsilon_c, \epsilon_1^j, \epsilon_2^j\}$ are permuted over the values $(1, -1)$, which gives $2^5 = 32$ torsion angle pairs, 16 of which are unique.

If the sign of the dipolar coupling is not known for a data point (e.g., the value falls within region C of Fig. 1), two sets of torsion angles are computed, one for (σ, v) and one for $(\sigma, -v)$.

Appendix B. The PIPATH algorithm

DEFINITION: α-helical deviation $\Delta^\alpha(\phi, \psi) = ((\phi - \phi^\alpha)^2 + (\psi - \psi^\alpha)^2)^{1/2}$

INPUT:

a primary peptide sequence $\{r_1, \ldots, r_N\}$
a PISEMA data set $\{d_1, \ldots, d_M\}$ with $M \leqslant N$, $d_i = (\sigma, v)_i$
an upper bound B

OUTPUT: All structures S_k and assignments $a_k = \{r_1 \rightarrow d_i, \ldots, r_N \rightarrow d_j\}$ that match the input PISEMA data and deviate from α-helicity by $\leqslant B$.

(1) Construct the assignment graph: a well-connected directed graph $G = (V, E)$
   (a) Add N vertices to V
   (b) For each vertex v_i in G, arbitrarily associate one unassociated resonance d_i
     (i) If $M < N$, $N - M$ vertices will remain unassociated
   (c) For each vertex pair add edge $e_{ij} = (v_i, v_j)$ to E
   (d) Weight each edge e_{ij}:
     (i) If v_i and v_j have associated resonances,

$$w(e_{ij}) = \min[\Delta^\alpha\{\phi\psi_{ij}\}],$$

      where the minimum is computed over all possible torsion angles connecting d_i to d_j
     (ii) If either v_i or v_j has no associated resonance, $w(e_{ij}) = 0$
   (e) Prune edges
     (i) For each consecutive specifically labeled resonance pair $d_i \rightarrow d_j$, remove associated edges e_{ik} where $k \neq j$ and e_{ki} where $j \neq i$
     (ii) If applicable, remove each edge e_{ij} corresponding to impossible residue connections $d_i \rightarrow d_j$ as determined by non-uniform selective labeling
(2) Find all Hamiltonian Paths of G with a cost $\leqslant B$
(3) For each path P_k:
   (a) Build most α-helical structure that satisfies continuity condition: delete path if $\Delta^\alpha S_k > B$.
(4) Output all remaining paths P_k and structures S_k.

References

[1] C.H. Wu, A. Ramamoorthy, S.J. Opella, High resolution heteronuclear dipolar solid-state NMR spectroscopy, J. Magn. Reson. 109 (1994) 270–282.

[2] J. Wang, J. Denny, C. Tian, S. Kim, Y. Mo, F. Kovacs, Z. Song, K. Nishimura, Z. Gan, R. Fu, J.R. Quine, T.A. Cross, Imaging membrane protein helical wheels, J. Magn. Reson. 144 (2000) 162–167.

[3] F.M. Marassi, S.J. Opella, A solid-state NMR index of helical membrane protein structure and topology, J. Magn. Reson. 144 (2000) 150–155.

[4] M. Schiffer, A.B. Edmundson, Use of helical wheels to respresent the structures of proteins and to identify segments with helical potential, Biophys. J. 7 (1967) 121–135.

[5] A.A. Nevzorov, S.J. Opella, Structural fitting of PISEMA spectra of aligned proteins, J. Magn. Reson. 160 (2003) 33–39.

[6] J. Denny, J. Wang, T.A. Cross, J.R. Quine, PISEMA powder patterns and PISA wheels, J. Magn. Reson. 152 (2001) 217–226.

[7] J.R. Quine, S. Achuthan, T. Asbury, R. Bertram, M.S. Chapman, J. Hu, T.A. Cross, Intensity and mosaic spread analysis from PISEMA tensors in solid-state NMR, J. Magn. Reson. 179 (2006) 190–198.

[8] S. Achuthan, J.R. Quine, T. Asbury, R. Bertram, M.S. Chapman, J. Hu, T.A. Cross, Continuity conditions and torsion angles in protein backbone structure determination with ssNMR data, in preparation.

[9] J.U. Bowie, Helix packing in membrane proteins, J. Mol. Biol. 272 (1997) 780–789.

[10] T.A. Eyre, L. Partridge, J.M. Thornton, Computational analysis of α-helical membrane protein structure: implications for the prediction of 3D structural models, Protein Eng. Des. Sel. 17 (2004) 613–624.

[11] R. Bertram, T. Asbury, F. Fabiola, J.R. Quine, T.A. Cross, M.S. Chapman, Atomic refinement with correlated solid-state NMR restraints, J. Magn. Reson. 163 (2003) 300–309.

[12] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[13] A. Krogh, B. Larsson, G. von Heijne, E.L.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.

[14] S. Kim, T.A. Cross, Uniformity, ideality, and hydrogen bonds in transmembrane α-Helices, Biophys. J. 83 (2002) 2084–2095.

[15] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, D.B. Shmoys, The Traveling Salesman Problem, Wiley, New York, 1985.

[16] F.M. Marassi, S.J. Opella, Simultaneous assignment and structure determination of a membrane protein from NMR orientational restraints, Protein Sci. 12 (2003) 403–411.

[17] R.R. Ketchem, K.C. Lee, S. Huo, T.A. Cross, Macromolecular structural elucidation with solid-state NMR-derived orientational constraints, J. Biomol. NMR 8 (1996) 1–14.

[18] J.D.C. Little, K.G. Murty, D.W. Sweeney, C. Karel, An algorithm for the traveling salesman problem, Oper. Res. 11 (1963) 972–989.

[19] M.M. Syslo, N. Deo, J.S. Kowalik, Discrete Optimization Algorithms, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[20] The Python Programming Language. <http://www.python.org/>.

[21] O. Carugo, S. Pongor, A normalized root-mean-square distance for comparing protein three-dimensional structures, Protein Sci. 10 (2001) 1470–1473.

[22] Z. Sang, F.A. Kovacs, J. Wang, J.K. Denny, S.C. Shekar, J.R. Quine, T.A. Cross, Transmembrane domain of M2 protein from Influenza A virus studied by solid-state N15 polarization inversion spin exchange at magic angle NMR, Biophys. J. 79 (2000) 767–775.

[23] R.A. Engh, R. Huber, Accurate bond and angle parameters for X-ray protein-structure refinement, Acta Crystallogr. A 47 (1991) 392–400.

[24] J.R. Quine, T.A. Cross, M.S. Chapman, R. Bertram, Mathematical aspects of protein structure determination with NMR orientational restraints, Bull. Math. Biol. 66 (2004) 1705–1730.