PERGAMON

Computers & Chemistry 23 (1999) 153–163

Computers & Chemistry

Determining protein structure using the distance geometry program APA

R. Reams, G. Chatham, W. Glunt, D. McDonald, T. Hayden*

Department of Mathematics, University of Kentucky, Lexington, KY 40506, USA

Received 1 August 1998; accepted 21 December 1998

Abstract

APA is a computer program, written in C, designed to determine the three-dimensional structure of proteins using distance geometry. We present the sampling and convergence properties of APA, as tested on bovine pancreatic trypsin inhibitor (BPTI). The results confirm the program's earlier success with poly-L-alanine, albeit with some complications. The correct overall orientation of the BPTI conformation is achieved at an early stage in the algorithm. The correct orientations of the α-carbons are obtained by local reflections, instead of a penalty term, resulting in a smoother convergence. Finally, a process of choosing dissimilarities from two reduced data sets resulted in almost all structures converging. In order to compare with Havel's DG-II distance geometry program, the sampling and convergence properties were tested on Havel's 10 data sets. These simulated data sets were generated from the BPTI crystal and kindly provided by Tim Havel. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Protein structure; Distance geometry; Alternating projections; BPTI; Sampling conformation space; APA; DG-II

1. Introduction

Distance geometry algorithms are especially useful in sampling the conformation space of proteins, when upper bounds for interatomic distances in the molecule are obtained from NOESY nuclear magnetic resonance (NMR) experiments. A review of distance geometry algorithms to determine molecular conformations using NMR may be found in Havel (1990). The data input into these algorithms comes from the known covalent geometry of the primary sequence of amino acids, and upper and lower bounds on the distances between atoms in the molecule (as obtained from the covalent geometry, hard sphere radii, and NMR experiments). Some algorithms also incorporate torsion angle bounds, as determined from measuring coupling between spins.

A distance geometry algorithm should provide unbiased sampling of the sterically allowed conformation space of a molecule, constrained by a set of distance bounds. Several measures can be used to ascertain if a distance geometry algorithm is adequately sampling the allowed conformation space. RMSD measures the average deviation of atomic positions between two structures superimposed as much as possible. A large average RMSD between all pairs of structures in a sample is considered good, since this indicates a well-distributed sample.

The alternating projections algorithm APA (Glunt et al., 1990, 1991, 1993, 1994a; Wells et al., 1994) is in a

* Corresponding author.
E-mail address: hayden@ms.uky.edu (T. Hayden)

0097-8485/99/$20.00 © 1999 Elsevier Science Ltd. All rights reserved.
PII: S0097-8485(99)00003-0

group of algorithms (as is DG-II) known as embedding algorithms. These algorithms use the 'embed' algorithm, which produces coordinates in three dimensions for the position of each atom in the conformation, from a matrix of interatomic distances, which are then used in minimization. Another group of algorithms use the torsion angles between adjacent amino acids as the variables in the minimization.

Previously, in Edwards et al. (1997), we obtained good sampling and convergence for the APA algorithm when applied to poly-L-alanine. BPTI is a medium-sized protein (800 atoms) for a distance geometry program. Due to extensive tests of other distance geometry algorithms on BPTI and the availability of crystallographic data for BPTI, it is considered a good standard for testing distance geometry programs. We found it was more difficult to obtain convergence of our algorithm with BPTI than with poly-L-alanine. In order to obtain convergence several innovations were necessary, which are described below. With these additions, APA was found to have a high success rate in determining the structure of BPTI, and similar sampling properties to the DG-II distance geometry program of Havel.

2. The alternating projections algorithm

The APA algorithm consists of a sequence of algorithms and ideas that have been developed over the last ten years. In order for the reader to follow the development we will first label the key objects and substeps.

2.1. Data base

The input data consist of upper and lower bounds on the interatomic distances obtained from:

- Hard sphere radii (Å)—H 0.95; N 1.35; O 1.35; C 1.45; K 1.5; L 1.6; M 1.7; S 1.8.
- Bond lengths—calculated from the covalent geometry for each amino acid (coordinates from ECEPP/2 (Scheraga, 1982)).
- NMR data—simulated NMR data, calculated from the BPTI crystal.

Hard sphere radii were originally discussed as extreme limits of van der Waals radii in Ramachandran and Sasisekharan (1968), and provide lower bounds l_{ij} between each pair of atoms i and j. K, L and M denote pseudo-atoms (Wüthrich et al., 1983) for the non-stereospecific assignable CH, CH_2, and CH_3, respectively. The NMR data provide small upper bounds u_{ij} between certain pairs of atoms i and j, other upper bounds are taken as arbitrarily large distances. For a molecule with n atoms the distance d_{ij} between atoms i and j must satisfy

$$l_{ij} \le d_{ij} \le u_{ij}, \quad 1 \le i,j \le n$$

2.2. Smooth and data box

The upper distance bounds are lowered and the lower bounds raised by applying triangle inequality bound smoothing, an $O(n^3)$ algorithm (Dress and Havel, 1988). One can further smooth the bounds by using the tetrangle inequality, an $O(n^4)$ algorithm (Easthope and Havel, 1989). Due to the high cost of smoothing with the tetrangle inequality we have not used this procedure in APA. The smoothed bounds produce a rectangular parallelepiped of upper and lower bounds, which we call the *data box*.

2.3. Restricted data box and dissimilarity δ_{ij}

In order to obtain convergence and also have good sampling, we had to restrict the size of the original data box by picking a subinterval of the length of each side of the box to produce a smaller box inside the original data box. The exact choices will be described later.

Next, one chooses a matrix $\Delta = (\delta_{ij})$, called a dissimilarity matrix, by randomly selecting for each atom i and j a matrix entry δ_{ij} which lies between the upper and lower distance bounds for this pair of atoms inside the restricted data box. Then APA uses column metrization (Edwards et al., 1997), an $O(n^3)$ algorithm, which ensures that the entries of the dissimilarity matrix satisfy the triangle inequality (see also Havel and Wüthrich, 1984).

2.4. Map

The dissimilarity matrix is then projected onto the cone of matrices which are negative semi-definite on the orthogonal complement of $\mathbf{e} = (1, 1, \ldots, 1)^{\mathrm{T}}$, then projected back onto the data box (and simultaneously making the diagonal entries zero) to obtain a new dissimilarity.

This process, known as 'map', of alternately making these projections is repeated five times, in an attempt at finding the euclidean distance matrix which is nearest to the original dissimilarity matrix. The advantage of 'map' stems from the fact that it is possible to prove that if the alternating projections are allowed to continue, the alternating projections would converge to a distance matrix which satisfy the upper and lower distance bounds. From numerical experience it was found that the number of negative eigenvalues needed to be set equal to zero, as well as the number of posi-

tive eigenvalues used to form the metric matrix and find the closest matrix in the cone, tended to decrease on each projection (loosely, the embedding dimension of the nearest Euclidean distance matrix decreased on each projection). Because this algorithm is $O(n^3)$, as a cost consideration we stop after five repetitions.

2.5. Embed

After 'map' we perform a three-dimensional embedding using 'embed'. This is done by forming an $n \times 3$ matrix X, the columns of which are the three eigenvectors corresponding with the three largest eigenvalues of the metric matrix associated with the last dissimilarity matrix of 'map'. The rows of X are the atomic coordinates in three dimensions (for a description of map and embed, see Glunt et al., 1991).

2.6. Stress (X, δ_{ij})

Next, APA performs stress minimization, i.e. the function $\sigma(X)$, known as stress in the multi-dimensional scaling literature, is minimized.

$$\sigma(X) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d_{ij}(X) - \delta_{ij})^2 \quad (1)$$

where $d_{ij}(X)$ denotes the distance between atom i and atom j in the embedded structure with coordinates X, and $W = (w_{ij})$ is a symmetric non-negative matrix of weights, with zero diagonal.

Let X^0 denote the matrix of coordinates obtained from 'embed', and Δ^0 the dissimilarity obtained from 'map'. Using Δ^0 and $d_{ij}(X^0)$ APA minimizes stress by performing a pass (approximately 30 iterations) of the spectral gradient algorithm (Glunt et al., 1993, 1994a), which finds a new matrix of coordinates X^1. The distance matrix $d_{ij}(X^1)$ is then projected back onto the data box to obtain a new dissimilarity matrix Δ^1, and another pass of stress minimization is performed with Δ^1 and $d_{ij}(X^1)$. These alternate passes and projections onto the data box are repeated until convergence is reached. We begin with a weight matrix that has greater weights applied when the difference between the upper and lower bounds is small (giving greater importance to the more precisely known interatomic distances). The weights are changed in successive passes to add additional weight to those distances that violate the data box constraints the most. This $O(n^2)$ algorithm is described in Glunt et al. (1993, 1994b).

In order to have planarity in the peptide bonds and aromatic rings, and the correct chirality about the α-carbons, volume constraints are added as penalty terms to stress to form a new function to be minimized, which we will call the total stress.

This algorithm (bound smoothing → metrization → map → embed → stress) when applied to BPTI did not work well without the innovations about to be described (it did work well with poly-L-alanine). The main difficulty was that the structures immediately after 'embed', and before stress minimization, would be both compact and tangled, and the minimization procedure was unable to correct this. If allowed to continue, the minimization would sometimes even cause one segment of the backbone to pass through another segment during convergence. Moreover, with compact tangles the barriers were too large for APA to obtain a good conformation.

Motivated by the fact that heat-denatured elongated proteins will fold back to their original conformation when cooled, the following modification was incorporated into our algorithm: bias appropriately the size of the data box in which one randomly chooses the initial dissimilarities, in order to produce untangled expanded initial conformations with 'embed'. Then, after embedding these initial dissimilarities and using them as approximate structures, we randomly choose new dissimilarities in a less biased data box. In our earlier study of poly-L-alanine we used a compactness parameter to multiply the dissimilarities at this stage by a factor (less than one) to produce compact final structures. The idea of using the two stage process to choose a dissimilarity, as described above (and to be described in more detail in Section 4), was motivated by experience obtained with the compactness parameter. We also desired, in order to mimic the in vivo process, a gradual change from the initial conformation to the final conformation.

Since the five projections of 'map', as well as the 'embed' into three dimensions, to generate an expanded initial structure takes time $O(n^3)$, just 10 such expanded structures (i.e. 10 matrices of coordinates X^0) were produced, to which was applied only 15 passes of stress minimization each. For each of these 10 approximate structures we generated a further five dissimilarities in the second stage, and it was these 50 structures which were then metrized and minimized with 500 passes of stress. However, better sampling was achieved when just three dissimilarities were produced for each initial structure. This strategy appears to be a good compromise between the higher cost of generating initial structures and adequate sampling. More detail on the choice of the dissimilarities and its effect on the final structures appears in Sections 4 and 5.

Another modification adopted for BPTI was to check the orientation of the α-carbons in the initial embedded structure. If more than 50% of these local chiralities were oriented incorrectly, then the overall orientation of the molecule was reversed (achieved by simply reversing the sign of the x-coordinate for each atom). The local orientations of the α-carbons always

correctly determined the overall orientation of BPTI, apparently because of the five projections of 'map'. This use of initial expanded structures, with the corrected overall orientation, let to final structures which converged well in almost all our trials. The ability to detect the correct overall orientation at this early stage yields substantial time savings over waiting until the lack of convergence, after hundreds of passes of stress minimization, reveals the wrong overall orientation.

Using the molecular viewer Xmol's 'movie' option, the effect of using a penalty function on peptides to correct the local orientations, revealed an 'explosion' of the atoms around the α-carbon. Thereafter several passes of minimization were required to bring the atoms which were affected back to their correct positions. In order to correct the local orientations, it was found to be more expedient to perform a reflection of just two atoms in a plane through the chiral center (for each chiral center) to correct these orientations. The locally reflected structure was only a slight perturbation from the correct overall structure. Hence, for smoother convergence we used Householder transformations to correct the local orientations, and then turned on a penalty term to maintain these corrected local orientations. Once the correct local orientations were achieved, the penalty term (which gave some freedom of movement of the calculated volume in an interval about the theoretical volume without penalty) was seldom invoked.

To obtain one fully converged final structure of BPTI beginning with data smoothing and including 500 passes of stress minimization takes about 1 h 15 min on the Convex Exemplar 1200 (in serial mode), or about twice that length of time on a Hewlett Packard HP J210. The 500 passes of stress minimization to obtain a second structure take about 50 min.

3. The data sets

The input data sets were prepared by using the BPTI crystal to calculate exact interatomic distances, and then producing distance bounds to simulate various possible sets of NOE constraints from NMR experiments. A more detailed description of these data sets and their interpretation may be found in Havel and Wüthrich (1985), and Havel (1991).

$d_{\alpha N}$, d_{NN} and $d_{\beta N}$ are (respectively) the sequential NOEs between the C^{α} proton and residue i and the amide proton of the next adjoining residue $i + 1$, between NH_i and NH_{i+1}, and between $C^{\beta}H_i$ and NH_{i+1}.

Set 0 contains no NOE constraints. In Set I, 508 distances less than 4 Å are partitioned as follows: distances less than 2.5 Å are constrained to lie between 2.0 and 2.5 Å, those between 2.5 and 3.0 Å are made to lie between 2.0 and 3.0 Å, and the remaining ones between 2.0 and 4.0 Å. This partition corresponds to the classification of experimental NOE values as 'strong', 'medium' or 'weak' range distances.

In Set II, only the NOEs likely to give rise to well resolved NOESY cross peaks are retained i.e. those involving at least one amide or aromatic proton; as well as retaining NOEs likely to be 'strong' between aliphatic protons. Set II includes the $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ NOEs.

In Set III, $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ were retained, along with giving upper bounds of 5 Å to any medium and long range distances. Set IV is the same as Set III except that an upper bound of 4 Å was assigned for the medium and long-range distances. Set V excluded the 122 sequential NOEs $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ from Set IV. Set VI contained the 122 sequential NOEs, but a range of ±0.1 Å about the actual distance in the crystal structure was used for the medium and long-range constraints. For Set VII the sequential constrains $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ are also given a range of ±0.1 Å about their actual distances. Set VIII is the same as Set VII, but the interval about the actual crystal distance was ±0.5 Å. Set IX was obtained from Set IV by eliminating all constraints corresponding to weak NOEs. Set X includes torsion angle bounds, which APA does not use at the present time.

4. Algorithm outline and parameters for BPTI

The basic steps in the algorithm APA will be outlined and the details of the parameters will follow the outline:

I. Collect the Data Base.
II. Apply Smooth to produce the Data Box.
III.
  (a) Obtain a reduced Data Box I.
  (b) Choose a metrized dissimilarity $\delta_{ij}(I)$ from Data Box I.
IV. Apply five alternating projections of MAP.
V. Embed in three dimensions with coordinates in X.
VI.
  (a) Orient overall structure.
  (b) Use last δ_{ij} from map and X from embed to minimize Stress and obtain a new X.
  (c) Project the new X on the Data Box to obtain a new δ_{ij}.
  (d) Repeat step (b) and (c) 15 times, using the new δ_{ij} and X each time.
  (e) Orient the α carbons via a Householder transformation.
  (f) Call the oriented expanded starting structure X^{e}.

VII.
(a) Produce a reduced Data Box II.
(b) Choose a metrized dissimilarity δ_{ij}(II) from Data Box II.
VIII.
(a) Begin Stress minimization with X^e and δ_{ij}(II) to obtain X^1.
(b) Project X^1 onto Data Box to obtain dissimilarity δ_{ij}^1.
(c) For $k=1, 2, \ldots, N$.
(i) Alternate Stress minimization and Box projections with (X^k, δ_{ij}^k) to obtain (X^{k+1}, δ_{ij}^{k+1}).
(ii) Adjust weights.
(iii) For $k=60$ add volume penalty term to stress to force correct α carbon orientation.
(iv) For $k=99$ add volume penalty term to stress to enforce planarity.
(v) Stop at $N=500$ to obtain final structure X^{500}.
IX. Repeat steps VII(b)–VIII to obtain two additional structures.
X. Return to III(b) and repeat until desired number of final structures are obtained.

We found that an optimal set of parameters in our program for one data set might yield poor results on another set, due to the differences between the 10 data sets. However, our goal was to produce a robust program that would perform well over a wide range of data sets, producing reasonable results with just one set of procedures and parameters. The parameters and details of the steps in the above algorithm are given below.

4.1. Reduced Data Box I

Producing a dissimilarity in step III(a) involves randomly picking for each i, j a δ_{ij} between the upper and lower bounds u_{ij} and l_{ij}. Let I_{ij} be the full interval, i.e. $I_{ij}=u_{ij}-l_{ij}$ then I_{ij} is the length of a side of the original smoothed data box. In order to pick an initial expanded structure, for Data Box I, we did not pick δ_{ij} randomly in the *full* interval. Instead each lower bound (l_{ij}) was moved up to 0.4 the full interval length. Similarly, a new upper bound was found by lowering the original upper bound to 0.9 the length of the side I_{ij}. Thus the new interval width was made half the original width, for each interval I_{ij}.

In step III(b) and step IV the metrized dissimilarity from Data Box I was given five alternating projections of 'map' on the cone of matrices which are positive semi-definite on the subspace orthogonal to the vector of all ones and the *original* smoothed Data Box. Projecting back onto the original Data Box allowed greater variability than projecting onto the restricted Data Box and yielded better sampling. Then the structure is embedded in three dimensions to produce an initial structure X. The initial structures have the same overall shape as the crystal but are expanded along each axis, due to the choice of the dissimilarity in the reduced box which favors larger interatomic distances.

In step VI, these initial structures are then subjected to 15 passes of stress minimization to produce X^e. During this stress minimization we determine if more than 50% of the α-carbons of the expanded structure have the wrong chirality. If so, we obtain the mirror image of the entire conformation by a change of sign of one coordinate. In every case this produced the correct overall orientation of the molecule.

4.2. Data Box II

A new metrized dissimilarity δ_{ij}(II) is chosen from Data Box II. Data Box II has a new lower bound which is 0.2 the length of the original smoothed side. Again the upper bound is 0.9 of the full interval length. Hence the new interval width is 0.7 the width of the side of the original Data Box. The metrized dissimilarity is passed directly on to 'stress' (without 'map' or 'embed').

In step VIII, Stress then uses δ_{ij}(II) and X^e and produces a minimized new structure X^1. Stress is always minimized by the use of spectral gradient algorithm. For each pass, at least five iterations of the spectral gradient algorithm were performed, but no more than 30 iterations were allowed. Stress minimization was halted provided

$$|\text{previous stress} - \text{stress}| < 0.0001\,(\text{previous stress}).$$

The full 30 iterations of the spectral gradient algorithm in a pass only occurs during the early passes, or after weight changes. Otherwise, it only requires 7–12 iterations to achieve the stopping criteria.

In step VIII(b), after stress had met these minimization criteria, a projection onto the *original* data box is carried out to obtain a new dissimilarity δ_{ij}^1. In step VIII(c), the α-carbons with the incorrect chirality (usually less than 10%) were corrected by the Householder reflections after the fourth pass of stress minimization (i.e. for $k \geqslant 4$). Seldom would any reversal occur, even though no penalty to enforce the correct chirality was applied until after 60 passes.

In step VIII(d) a crucial part of the algorithm is to adjust the weights in order to increase the speed of convergence (by up to a factor of 10). The weight matrix is updated on every third pass to add additional penalty to those distances d_{ij} that lie outside the data box (i.e. that lie outside the bounds). An initial weight matrix is chosen to weight more heavily those distances that are known precisely, e.g. the covalent bond

lengths. The entries of the initial weight matrix are given by

$$w_{ij} = \frac{1}{1 + 24(u_{ij} - l_{ij})} + \text{factor} \times \text{viol}_{ij}$$

where the factor is small initially, and viol_{ij} is the absolute value of the distance that d_{ij} lies outside the data box (zero violation if inside). The weights are normalized so that their sum is one. 'Factor' is set to 0.03 for the first 40 passes and then is increased linearly up to a maximum of 5.0 according to:

$$\text{factor} = (\#\text{of pass} - 40)/60.$$

Two penalty terms are added to stress to enforce the correct chirality of the α-carbons and the planarity of the peptide bond. These terms must also be administered gradually or else they can become very large, causing large movements of the molecule. A cutoff value is enforced to keep these terms within reasonable bounds. Other precautions taken include no penalty for chirality until after 60 passes, and zero planarity penalty for the first 99 passes. This allows the conformation to assume its correct approximate shape before activating these additional terms.

These penalty terms are computed using the orientated volume as computed from a certain determinant (Havel and Crippen, 1988). The approximate volume of a tetrahedron with vertices the four atoms adjoining the α-carbon is 1.37 Å^3. The penalty for each volume is made zero if the calculated volume lies in the interval [1.20, 1.54]. Any value outside this interval is penalized by a quadratic function of the extent to which the volume lies outside the interval, multiplied by the factor 2×10^{-8}. (In fact, APA calculates the volumes around each chiral center of each individual amino acid at the very beginning of the program, and uses these volumes as target volumes, i.e. as the mid-points of the intervals of length 0.34 above.) When this penalty was invoked after 60 passes it would go to zero in a few passes and stay small for the remainder of the 500 passes.

The planarity penalty was added at pass 99. To obtain a plane the determinant volume is targeted to zero (plus or minus 0.17, as above), and the quadratic penalty for planarity was the volume squared multiplied by the factor 3×10^{-8}.

The flow chart for APA lists the main steps in the algorithm for APA (Fig. 1). Begin with the data base and first proceed to Data Box I. The starting dissimilarity and starting coordinates in order to minimize stress are indicated by dashes. Double arrows indicate projection steps in the algorithm with new dissimilarities being generated for the next step. For each set of expanded coordinates from the first application of Stress, we obtain three final structures. After three

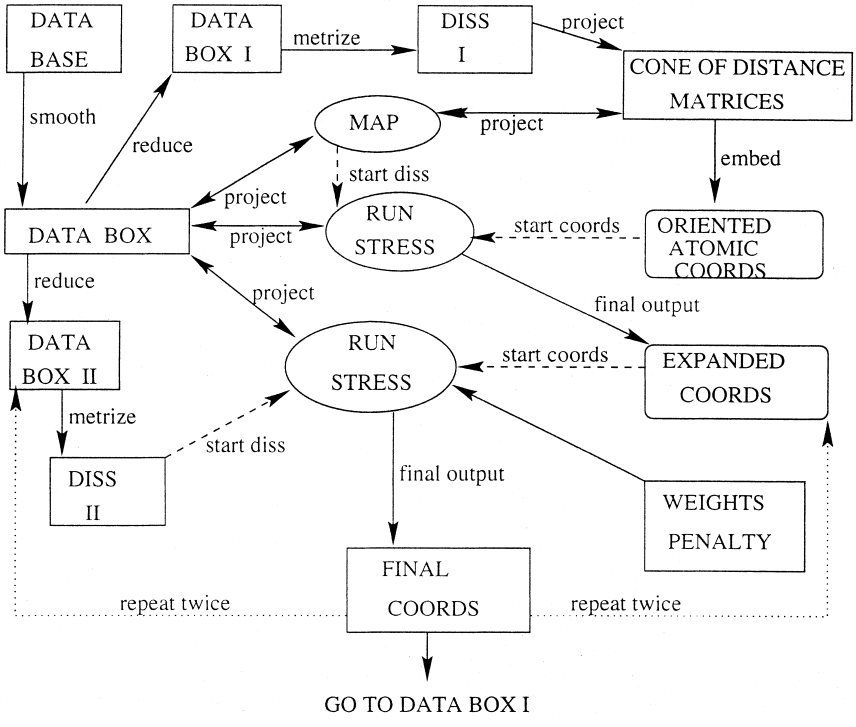

Fig. 1

final structures are obtained, one returns to Data Box I to repeat the process to obtain the desired number of final structures.

5. Results and discussion

We first present the results of a baseline study that will demonstrate the robustness of the APA algorithm. These baseline results were obtained using the same parameters and procedures, as described in Section 4, for each of the 10 data sets.

For each data set we found 10 initial dissimilarities from Data Box I and applied steps III–VI of the algorithm to obtain 10 expanded starting structures X^e. For each of these expanded structures we obtained three final conformations. Thus we computed 30 final conformations for each of the 10 data sets. No conformation is rejected in the baseline study; in which all final structures were found to have the same global orientation as the BPTI crystal structure. After presenting the results of this baseline study, we show that certain subsets of the conformations achieved better convergence, and we show some of the effects of varying the parameters. The principal aspects to be investigated for each set of computed structures is the convergence and sampling of the algorithm.

In the left half of Fig. 2 we report the average of the number of violations for 30 final conformations, in 0.1 Å intervals, for each data set. In the right half of Fig.

	Thirty Structures per Data Set vs						Five Best Structures			
(0)										
Tight	2202	18	0	0	0	0	2218	2	0	0
All	2655	35	0	0	0	0	2639	10	0	0
(I)										
Tight	2183	33	3	0	0	0	2212	8	0	0
All	2741	86	12	1	0	0	2746	31	1	0
(II)										
Tight	2155	58	7	0	0	0	2194	24	2	0
All	2739	145	21	2	0	0	2728	80	6	0
(III)										
Tight	2177	38	4	0	0	0	2201	18	0	0
All	2686	94	12	2	0	0	2667	50	4	0
(IV)										
Tight	2173	42	5	0	0	0	2191	28	1	0
All	2716	103	13	1	0	0	2745	75	7	0
(V)										
Tight	2191	27	2	0	0	0	2216	4	0	0
All	2719	73	6	0	0	0	2689	19	0	0
(VI)										
Tight	2159	54	6	1	0	0	2197	23	0	0
All	2849	159	22	4	1	0	2859	85	5	0
(VII)										
Tight	2152	57	9	1	0	0	2186	32	2	0
All	2879	160	25	6	2	0	2903	95	9	1
(VIII)										
Tight	2174	40	4	0	0	0	2202	18	0	0
All	2773	112	15	4	0	0	2275	54	3	0
(IX)										
Tight	2185	31	4	0	0	0	2208	12	0	0
All	2695	83	9	1	0	0	2697	43	1	0
Distance	0.1		0.3		0.5		0.1		0.3	

Fig. 2. A comparison of the number of interatomic distance bound violations for each of the 10 data sets, and their magnitudes in Ångstroms. The left half reports the number of violations for each data set as an average over 30 computed structures. The right half is the average over the best five structures from each data set, among 30 structures. A 'Tight' bound is taken to mean an exact interatomic distance. The violations are computed by comparison with either an exact interatomic distance (known from the covalent geometry), or an upper or lower interatomic distance bound as computed by the triangle inequality smoothing algorithm. 'All' indicates that the violation may be a tight, lower or upper bound violation.

2, we report similar data except the average is over the five best structures (as measured by the five smallest values of the maximal violations of the tight bounds at the 500th pass) from the 30 previous conformations in each data set. In each case we report the violations of tight bounds and for all bound violations.

A comparison with Fig. 4 in Havel and Wüthrich (1985), which presents a similar histogram for DISGEO, indicates that the general trends of the results are similar. For example, those data sets for which it was more difficult for DISGEO to obtain good convergence were also more difficult for APA. For larger values of the violations, APA has significantly fewer violations. This difference appears to be due to the fact that APA heavily penalized large violations as the number of passes increased. This weighting allowed almost no violations greater than 0.3 Å, and usually less than 20 violations between 0.2 and 0.3 Å. Note that this convergence was obtained over all 300 structures using the same parameter set. DG-II has superior convergence to DISGEO and hence has better convergence than APA, especially for the number of violations less than 0.2 Å, when compared to our average over all 30 structures. In order to observe the effects of better convergence, we chose subsets of the 30 structures which had the best convergence, and looked at their sampling properties. We chose the five best converged structures to be those five that had the smallest maximum tight violation on the 500th pass. In this case, only data set VII had a violation larger than 0.3 Å.

Detailed information was available at each pass to examine the number and extent of the bound violations and the planarity and chirality violations as a percentage of the total stress. Thus it was possible to adjust the parameters and note the effect on various terms. For example, after 496 passes for a structure in set VIII only seven spectral gradient iterations were needed to minimize the total stress (which includes the penalty terms for chirality and planarity). The value of the total stress only varied (with four significant figures reported) between 1.286×10^{-5} and 1.287×10^{-5} during the seven spectral gradient iterations. The chirality penalty term was at 0.0% and the planarity penalty term at 3.9%. The maximum tight violation was 0.18 Å, the maximum lower bound violation was 0.17 Å and the maximum upper bound violation was 0.27 Å. Graphical views of the final conformations indicated that both the aromatic rings and the peptide bonds had been flattened.

We now consider the sampling properties of APA, using three standard measures. The first is the root mean square deviation (RMSD), which reflects the degree of variability between structures. A large RMSD indicates that the algorithm is sampling a large part of the allowed conformation space, RMSD is proportional to the minimum achievable Fröbenius norm of the difference between the coordinate matrices of two structures, where the minimum is taken over all possible rotations of the second coordinate matrix. We used the singular value decomposition, as described in Hanson and Norris (1981), to find an analytic formulation of this minimization problem. A second measure we report is the root mean square dihedral angle difference (DHAD) between the ϕ and ψ angles (see Havel and Wüthrich (1984) for a detailed discussion of this measure and its relation to the variability of the conformations). The final standard measure reported is the radius of gyration (R_g), which is a measure of the overall size and shape of the computed structures.

In Table 1 the values to the left of the '/' are averages of differences between all pairs of the computed structures. The values to the right of the '/' are the averages of differences between the crystal structure and each computed structure. Our Table 1 then reports the same information as Table 2 in Havel (1991), which contains results obtained from DISGEO and DG-II. Comparing the two tables one observes that the general trends are similar for all three programs. That is, when the RMSD for a given set is relatively large for APA, the same is true for DG-II. Secondly, one observes that the RMSD to the right of the '/' is reduced from the value on the left of the '/'. Thus we conclude that the crystal structure, from which the data is derived, is a representative member of the entire set. This result is indicative of good sampling. The average radius of gyration is smaller in Table 1 for APA than that obtained by DG-II. We will show in Table 3, that increasing the upper bound of Data Box II will cause R_g to increase.

Table 2 presents for each of the 10 data sets the sampling properties of the five best converged structures, that Table 1 presented for the 30 structures. One observes that, in general, among the 10 data sets the RMSDa, RMSDh and DHADb are all decreased from Table 1. Thus, structures which turn out to be more tightly confined to the data box (as determined by seeing a smaller number of violations) will have less room to differ, and hence their sampling measures will decrease. Our values are still larger than those reported by Havel, although not significantly larger. Since smoothing with the tetrangle inequality was applied to Havel's BPTI data with the DG-II program, and we did not use tetrangular smoothing, our (original) data box is larger. Thus, since our allowed conformation space is larger, one would expect larger values for the reported sampling measures. Taking into consideration our larger data box, the values we obtained are similar to those of DG-II and indicate good sampling of the conformation space. Again, all the trends in variation among data sets are similar to the previous trends in Table 1.

Table 1
The sampling properties of APA among the 10 data sets are measured using the RMSDa, which is the RMSD of the α-carbons; the RMSDh, which is the RMSD of the heavy (non-hydrogen) atoms; and the DHADb which is the dihedral angle deviation along the backbone, i.e. among the ϕ and ψ dihedral angles. Each of these measures is given as the average over all pairs of the 30 computed structures for the number to the left of the '/', and by comparison with the crystal, for the number to the right of the '/'. R_g denotes the average radius of gyration for the 30 computed structures. The radius of gyration of the crystal was 10.6 Å. The convergence measurement given is the maximum tight bound violation in Ångstroms after 500 passes of stress

Data set	RMSDa	RMSDh	DHADb	R_g	Convergence		
					Minimum	Average	Maximum
(0)	10.22/12.06	11.31/13.49	72.4/71.2	16.09	0.11	0.17	0.22
(I)	2.16/1.85	2.69/2.45	60.6/57.6	10.21	0.12	0.23	0.35
(II)	2.87/2.44	3.47/3.09	64.2/63.1	10.27	0.11	0.26	0.45
(III)	2.84/2.84	3.43/3.47	65.3/64.3	10.92	0.12	0.27	0.41
(IV)	2.65/2.35	3.21/3.03	63.7/62.8	10.60	0.15	0.25	0.38
(V)	2.63/2.34	3.22/3.03	62.8/60.9	10.52	0.10	0.20	0.29
(VI)	1.68/1.51	2.33/2.24	54.3/52.6	10.24	0.13	0.28	0.45
(VII)	1.43/1.26	2.09/2.00	47.5/45.2	10.33	0.19	0.31	0.43
(VIII)	2.01/1.87	2.56/2.53	57.9/55.9	10.12	0.15	0.25	0.54
(IX)	4.07/3.62	4.66/4.15	66.6/65.3	11.08	0.10	0.22	0.37

With the five best converged structures the RMSD with the crystal structure is less than the RMSD among the computed structures. As mentioned earlier, this is a desirable feature. It is somewhat surprising to us that the radius of gyration increased for the best five structures. We found from numerical experience that when we had compact (and tangled) structures it was more difficult to obtain good convergence, no doubt due to the interference of side groups. Thus we see that better convergence tends to occur with more expanded structures.

Fig. 3 presents a graphical comparison of the backbone of the computed structures with the backbone of the crystal structure of BPTI. We made one change in the parameters to illustrate the effects on convergence and sampling. In particular, since the average radius of gyration was smaller than the crystal for several of the data sets with the parameters used to produce the data in Table 1, we made a change to increase this property. After obtaining initial elongated structures from dissimilarities picked in Data Box I of our algorithm as previously described, for Data Box II we increased the width of the intervals by 0.03, where the dissimilarities are chosen. That is, the lower bound is again 0.2 the length of the original smoothed lower bound, and the new upper bound for each pair of atoms is 0.93 times

Table 2
The sampling properties of APA, where the measurements are performed only on the best five converged computed structures from the 30, for each data set given in Table 1. The best converged structures were considered to be those with smallest maximum tight bound violation

Data set	RMSDa	RMSDh	DHADb	R_g	Convergence		
					Minimum	Average	Maximum
(0)	11.14/10.99	12.29/12.31	68.0/69.3	14.54	0.11	0.12	0.12
(I)	2.15/1.77	2.72/2.39	56.0/53.0	10.22	0.12	0.16	0.21
(II)	2.30/1.94	3.02/2.67	60.4/59.1	10.80	0.11	0.19	0.23
(III)	2.55/2.69	3.16/3.29	61.6/60.0	11.09	0.12	0.19	0.22
(IV)	2.81/2.50	3.38/3.15	63.9/62.6	10.18	0.15	0.19	0.22
(V)	1.99/1.90	2.57/2.56	60.1/61.0	10.76	0.10	0.12	0.15
(VI)	1.49/1.31	2.15/2.00	47.7/47.9	10.32	0.13	0.17	0.21
(VII)	1.10/0.95	1.87/1.76	44.3/39.9	10.44	0.19	0.24	0.27
(VIII)	2.08/1.87	2.57/2.49	53.5/56.6	10.19	0.15	0.18	0.20
(IX)	2.88/3.11	3.49/3.58	64.1/66.6	11.37	0.10	0.17	0.18

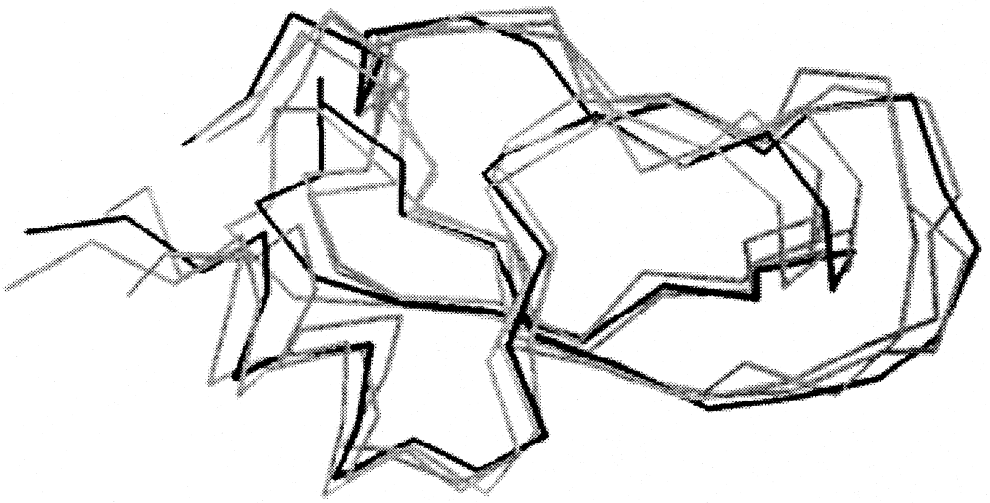

Fig. 3. The black strand is the backbone of the BPTI crystal. The three gray strands are three of the best converged structures from Data Set I computed using APA.

the length of the side of the original data box, whereas it was 0.9 times the length for the earlier 30 computed structures. All other parameters were left unchanged.

With this change, a comparison of Table 1 with Table 3 indicates improved convergence for about half of the data sets, while the convergence is similar for the other data sets. Hence the correlation between improved convergence and expanded structures again appears.

Table 3 reports the same sampling information with the new Data Box II data as in our Tables 1 and 2. The RMSD and DHAD values are smaller, which is probably due to the better convergence that was obtained. The radius of gyration was increased and is similar to the radius of gyration obtained with the best five structures from the first parameter set.

We also ran structures with the upper bound for each interval set at the original upper limit, i.e. picking dissimilarities in Data Box II of our algorithm with upper bound parameter 1.0 the full length of the side of the original data box, and lower bound 0.2 the full length of the interval. This tended to produce structures with excessively large average radius of gyration by comparison with the crystal.

Table 3
The sampling properties of APA, where the measurements are performed on 30 computed structures. Each upper bound for each interatomic distance in Data Box II is increased from 0.9 (in Table 1) to 0.93 of the full interval length for this table. The dissimilarity matrix is picked randomly (and uniformly) in the interval between the given lower and upper distance bounds. Note that the subsequent alternating projections of stress minimization are performed using the *original* Data Box

Data set	RMSDa	RMSDh	DHADb	R_g	Convergence		
					Minimum	Average	Maximum
(0)	10.85/13.54	11.95/15.10	73.5/72.1	17.01	0.11	0.18	0.22
(I)	2.12/1.75	2.61/2.37	57.3/57.8	10.43	0.10	0.22	0.45
(II)	2.16/1.93	2.72/2.56	60.3/61.0	10.65	0.13	0.25	0.42
(III)	2.56/2.70	3.09/3.31	64.4/63.8	11.15	0.13	0.25	0.29
(IV)	2.47/2.29	3.09/3.00	63.6/62.7	10.71	0.14	0.25	0.54
(V)	2.45/2.29	3.06/2.92	60.8/60.0	10.78	0.09	0.18	0.36
(VI)	1.78/1.55	2.37/2.25	54.4/51.1	10.27	0.16	0.28	0.51
(VII)	1.56/1.37	2.21/2.11	49.6/46.0	10.39	0.12	0.33	0.47
(VIII)	1.89/1.65	2.53/2.33	55.8/54.0	10.41	0.08	0.23	0.44
(IX)	2.85/3.21	3.38/3.70	61.0/62.8	11.53	0.11	0.20	0.28

6. Conclusion

We have modified the alternating projections algorithm, which worked well with poly-L-alanine, to produce good sampling and convergence over 10 widely differing sets of simulated data, generated from the crystal structure of BPTI. The main features are the ability to predict the correct overall orientation of the molecule from the orientation of the α-carbons, and one choice of parameters which yields reasonable convergence and sampling properties over the 10 data sets.

The ability to produce a robust program using the same parameters is mainly due to the use of Data Box I and Data Box II to pick dissimilarities with different properties. We first found dissimilarities in a Data Box I that led to an expanded initial (approximate) structure, which had the correct strand crossings. From each of these initial structures, we minimized stress by choosing new dissimilarities in Data Box II which was less restricted and had larger intervals than Data Box I. Each initial structure was used to compute three final structures thereby saving the cost of producing initial structures by approximately one-third. By varying the size of these restricted data boxes in which we picked our initial dissimilarities (note that stress performed minimization with the original triangle smoothed data box), we computed structures with good sampling properties, and also the crystal structures was a representative member of the full set of conformations. Moreover, we found that expanded structures correlated with better convergence properties.

The main disadvantage of APA as compared with DG-II is that DG-II is about seven times faster and has fewer number of small violations. The weighting scheme for APA heavily penalized large bound-violations, and hence the control of these large errors was good in APA. In the future we plan to investigate some other weighting strategies, and the introduction of other error functions in the later stages of convergence in order to further decrease the number of bound violations and increase the speed of convergence.

Acknowledgements

NSF grant CHE-9301120 provided partial support. Robert Reams was supported as a postdoctoral scholar by the Center for Computational Sciences, University of Kentucky.

References

Dress, A., Havel, T., 1988. Discrete and Applied Mathematics 19, 129.
Edwards, J., Chatham, G., Glunt, W., McDonald, D., Wells, C., Hayden, T., 1997. Computers and Chemistry 21, 115.
Easthope, P., Havel, T., 1989. Bulletin of Mathematical Biology 51, 173.
Glunt, W., Hayden, T., Hong, S., Wells, J., 1990. SIAM Journal of Matrix Analysis and Applications 11, 589.
Glunt, W., Hayden, T., Liu, W-M., 1991. Bulletin of Mathematical Biology 53, 769.
Glunt, W., Hayden, T., Rayden, M., 1993. Journal of Computational Chemistry 14, 114.
Glunt, W., Hayden, T., Raydan, M., 1994a. Journal of Computational Chemistry 15, 353.
Glunt, W., Hayden, T., Shelling, J., Ward, D., Wells, C., 1994b. Journal of Mathematical Chemistry 15, 353.
Hanson, R., Norris, M., 1981. SIAM Journal of Science Statistics and Computers 2, 363.
Havel, T., 1990. Biopolymers 29, 1565.
Havel, T., 1991. Progress in Biophysics and Molecular Biology 56, 43.
Havel, T., Wüthrich, K., 1984. Bulletin of Mathematical Biology 46, 673.
Havel, T., Wüthrich, K., 1985. Journal of Molecular Biology 182, 281.
Havel, T., Crippen, G., 1988. Distance geometry and molecular conformation. Research Studies Press, Somerset.
Ramachandran, S., Sasisekharan, V., 1968. Advances in Protein Chemistry 23, 284.
Scheraga, H. 1982 The ECEPP/2 Program 454, QCPE, University of Indiana, Bloomington IN.
Wells, C., Glunt, W., Hayden, T., 1994. Journal of Molecular Structure (Theochemistry) 308, 263.
Wüthrich, K., Billeter, M., Braun, W., 1983. Journal of Molecular Biology 169, 949.