J|A|C|S
ARTICLES
Published on Web 04/30/2004

Completely Automated, Highly Error-Tolerant Macromolecular Structure Determination from Multidimensional Nuclear Overhauser Enhancement Spectra and Chemical Shift Assignments

John Kuszewski,*,† Charles D. Schwieters,† Daniel S. Garrett,‡ R. Andrew Byrd,§ Nico Tjandra,|| and G. Marius Clore*,‡

Contribution from the Division of Computational Bioscience, Building 12A, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624, Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, Structural Biophysics Laboratory, National Cancer Institute, P.O. Box B, Frederick, Maryland 21702-1201, and Laboratory of Biophysical Chemistry, Building 50, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892-8013

Received January 13, 2004; E-mail: mariusc@intra.niddk.nih.gov; john.kuszewski@nih.gov

Abstract: The major rate-limiting step in high-throughput NMR protein structure determination involves the calculation of a reliable initial fold, the elimination of incorrect nuclear Overhauser enhancement (NOE) assignments, and the resolution of NOE assignment ambiguities. We present a robust approach to automatically calculate structures with a backbone coordinate accuracy of 1.0−1.5 Å from datasets in which as much as 80% of the long-range NOE information (i.e., between residues separated by more than five positions in the sequence) is incorrect. The current algorithm differs from previously published methods in that it has been expressly designed to ensure that the results from successive cycles are not biased by the global fold of structures generated in preceding cycles. Consequently, the method is highly error tolerant and is not easily funnelled down an incorrect path in either three-dimensional structure or NOE assignment space. The algorithm incorporates three main features: a linear energy function representation of the NOE restraints to allow maximization of the number of simultaneously satisfied restraints during the course of simulated annealing; a method for handling the presence of multiple possible assignments for each NOE cross-peak which avoids local minima by treating each possible assignment as if it were an independent restraint; and a probabilistic method to permit both inactivation and reactivation of all NOE restraints on the fly during the course of simulated annealing. NOE restraints are never removed permanently, thereby significantly reducing the likelihood of becoming trapped in a false minimum of NOE assignment space. The effectiveness of the algorithm is demonstrated using completely automatically peak-picked experimental NOE data from two proteins: interleukin-4 (136 residues) and cyanovirin-N (101 residues). The limits of the method are explored using simulated data on the 56-residue B1 domain of Streptococcal protein G.

Introduction

Despite the introduction of orientational restraints in the form of residual dipolar couplings[1] which can potentially somewhat reduce reliance on nuclear Overhauser enhancement (NOE) data,[2] the principal source of geometric information for any de novo three-dimensional protein structure determination by NMR (except under very special and restrictive circumstances[3]) still resides in NOE-derived short (≤6 Å) interproton distance restraints.[4−6] The connectivities observed in through-bond correlation experiments employed for resonance assignment are precisely defined by the pulse sequence, and hence their interpretation is straightforward even in the presence of significant chemical shift overlap.[7] In contrast, the correlations

† Center for Information Technology, National Institutes of Health.
‡ National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health.
§ National Cancer Institute.
|| National Heart, Lung and Blood Institute, National Institutes of Health.

(1) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114. (b) Clore, G. M.; Starich, M. R.; Gronenborn, A. M. *J. Am. Chem. Soc.* **1998**, *120*, 10571−10572. (c) Hansen, M. R.; Mueller, L.; Pardi, A. *Nat. Struct. Biol.* **1998**, *5*, 1065−1074. (d) Prestegard, J. J.; Al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys.* **2000**, *33*, 371−424. (e) Bax, A.; Kontaxis, G.; Tjandra, N. *Methods Enzymol.* **2001**, *339*, 127−174.
(2) (a) Clore, G. M.; Starich, M. R.; Bewley, C. A.; Cai, M.; Kuszewski, J. *J. Am. Chem. Soc.* **1999**, *121*, 6513−6514. (b) Clore, G. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *97*, 9021−9025. (c) Fowler, C. A.; Tian, F.; Al-Hashimi, H. M.; Prestegard, J. H. *J. Mol. Biol.* **2000**, *304*, 447−460.
(3) (a) Delaglio, F.; Kontaxis, G.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 2142−2143. (b) Hus, J. C.; Marion, D.; Blackledge, M. *J. Am. Chem. Soc.* **2001**, *123*, 1541−1542. (c) Rohl, C.; Baker, D. *J. Am. Chem. Soc.* **2002**, *124*, 2723−2729. (d) Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2003**, *125*, 2902−2912.
(4) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.
(5) (a) Clore, G. M.; Gronenborn, A. M. *Protein Eng.* **1987**, *1*, 275−288. (b) Clore, G. M.; Gronenborn, A. M. *Crit. Rev. Biochem. Mol. Biol.* **1989**, *24*, 479−564.
(6) Clore, G. M.; Gronenborn, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5891−5898.

10.1021/ja049786h

observed in through-space NOE experiments cannot be predicted a priori without knowledge of the structure. Thus, while certain characteristic patterns of NOEs are known to occur in regions of regular secondary structure,[4,5] the presence of chemical shift overlap and degeneracy, which becomes increasingly problematic as the size of the system under consideration increases, precludes the unambiguous and straightforward assignment of many of the NOE cross-peaks even in three- and four-dimensional heteronuclear-separated experiments. Not surprisingly, therefore, one of the most difficult, time-consuming, and error-prone aspects of protein structure determination lies in the determination of an initial low-resolution protein fold. Once such a low-resolution fold has been reliably determined, resolving cross-peak ambiguities and weeding out incorrect NOE cross-peak assignments can be accomplished using an iterative refinement strategy.[8] If NMR structure determination is to become a high-throughput method in structural genomics, it is therefore clear that suitable methodology, involving a high degree of automation, must be introduced to render both the interpretation of the NOE spectra and the initial fold determination faster and less error prone.

It is invariably the case that the initial set of NOE-derived interproton distance restraints includes errors, that is, restraints that are not compatible with the true atomic coordinates. Errors involving NOE restraints between residues separated by ≤ 5 residues in the primary sequence (i.e., short-range) may have only limited consequences in structural terms, but errors involving NOE restraints between distant positions in the sequence (i.e., long-range) have severe consequences, because they preclude the determination of a correct fold. There are several sources of such errors that are particularly prominent when NOE distance restraints are generated automatically from multidimensional NOE spectra without any human intervention: (1) spectral noise and artifacts may be incorrectly interpreted as real NOE cross-peaks; (2) conversely, weak cross-peaks may be interpreted as noise; (3) incorrect NOE cross-peak assignments may arise as a consequence of the presence of either (i) a number of incorrect resonance assignments, (ii) incomplete resonance assignments for certain spin systems, (iii) small inconsistencies between the chemical shift table and the true chemical shifts of the actual sample on which the NOE data were recorded, and (iv) inaccuracies in chemical shift positions for cross-peaks with low signal-to-noise ratios; (4) extensive chemical shift overlap may severely complicate interpretation of the NOE spectra; and, finally, (5) the upper bounds of some restraints may be severely underestimated due to spin-diffusion. In addition, ambiguous NOE restraints, that is, restraints which have several possible assignments, only one of which is compatible with the true structure, are conceptually similar to incorrect restraints in that they only behave as correct restraints when the proper assignment has been selected.

Identifying incorrect NOE restraints by hand is nontrivial and can often require extensive expertise. This is because there is no guarantee, until structural convergence to a unique fold is obtained, that the NOE restraints that are violated are actually incorrect. Thus, it is often the case that structures can be generated in which incorrect NOE restraints are satisfied at the expense of correct ones. In the context of manual analysis of the NOE spectra (which is a highly time-consuming process), the number of errors present in the initial restraints list can be dramatically reduced by applying various common sense strategies. These strategies include carrying out the initial fold determination using a subset of the NOEs whose assignment is completely unambiguous, followed by iterative refinement in which the structure is used to guide the interpretation of the NOE spectra. In addition, knowledge of stereochemistry, covalent geometry, and secondary structure (derived from a qualitative interpretation of the backbone NOE data) can be used to guide some assignments and avoid major pitfalls. In contrast to manual interpretation of the NOE data, automatic peak-picking and assignment based on a chemical shift table can be carried out in a matter of minutes. However, generating a relatively clean restraint dataset from completely automatically peak-picked NOE spectra is highly problematic.[9,10]

Several attempts have been made to automate the structure calculation process from primary NOE data and chemical shift assignments using various iterative strategies,[11,12] and progress in this field has been recently reviewed by Güntert.[13] A widely used strategy is the ARIA[11] (ambiguous restraints for iterative refinement) algorithm which exploits two important techniques. The first is a method for handling ambiguous restraints that treats each NOE cross-peak as if it were the superposition of NOE cross-peaks arising from each of several possible assignments, in the form of a $(\sum r^{-6})^{-1/6}$ sum.[14] The second is the use of an asymptotic-shaped potential energy term to describe the NOE restraints,[15] thereby reducing the structural strain arising from badly violated NOE restraints. Unfortunately, the structure-based filters employed by ARIA to identify and eliminate incorrect NOE restraints require an ensemble of initial structures with the correct polypeptide fold. When an appropriate reference structure is absent (such as that derived from the structure of a highly homologous protein), obtaining a suitable initial ensemble with approximately the correct polypeptide fold from a dataset that contains a significant proportion of incorrect NOE restraints is therefore difficult.[11d,13] For this reason, ARIA has principally been employed as an efficient means of speeding up iterative refinement once an initial fold has been established.[11,13] Another recently introduced approach is afforded by the CANDID algorithm[12a] in conjunction with the automated peak-picking

(7) (a) Clore, G. M.; Gronenborn, A. M. *Science* **1991**, *252*, 1390−1399. (b) Bax, A.; Grzesiek, S. *Acc. Chem. Res.* **1993**, *26*, 131−138. (c) Cavanagh, J.; Fairbrother, W. J.; Palmer, A. J., III; Skelton, N. J. *Protein NMR Spectroscopy: Principles and Practice*; Academic Press: New York, 1996.

(8) (a) Kraulis, P. J.; Clore, G. M.; Nilges, M.; Jones, T. A.; Petersson, G.; Knowles, J.; Gronenborn, A. M. *Biochemistry* **1989**, *28*, 7241−7257. (b) Mumenthaler, C.; Braun, W. *J. Mol. Biol.* **1995**, *254*, 465−480. (c) Mumenthaler, C.; Güntert, P.; Braun, W.; Wüthrich, K. *J. Biomol. NMR* **1997**, *10*, 351−362. (d) Hare, B. J.; Wagner, G. *J. Biomol. NMR* **1999**, *15*, 103−113. Duggan, B. M.; Legge, G. B.; Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **2001**, *19*, 321−329.

(9) Garrett, D. S.; Powers, R.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson.* **1991**, *95*, 214−220.

(10) (a) Kleywegt, G. J.; Boelens, R.; Kaptein, R. *J. Magn. Reson.* **1990**, *88*, 601−608. Antz, C.; Corne, S. A.; Johnson, P. *Neural Networks* **1992**, *6*, 1023−1032. (b) Neidig, K. P.; Kalbitzer, H. R. *J. Biomol. NMR* **1995**, *5*, 287−296. (c) Koradi, R.; Billeter, M.; Engeli, M.; Güntert, P.; Wüthrich, K. *J. Magn. Reson.* **1998**, *135*, 288−297.

(11) (a) Nilges, M. *J. Mol. Biol.* **1995**, *245*, 645−660. (b) Nilges, M.; Macias, M. J.; O'Donoghue, S. I.; Oschkinat, H. *J. Mol. Biol.* **1997**, *269*, 408−422. (c) Nilges, M. *Folding Des.* **1997**, *2*, S53−S57. (d) Pascual, M.; Linge, J. P.; O'Donoghue, S. I.; Nilges, M. *Methods Enzymol.* **2001**, *339*, 71−90. (e) Longe, J. P.; Habeck, M.; Rieping, W.; Nilges, M. *Bioinformatics* **2003**, *19*, 315−316.

(12) (a) Hermann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209−227. (b) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **2002**, *24*, 171−189.

(13) Güntert, P. *Prog. Nucl. Magn. Reson. Spectrosc.* **2003**, *43*, 105−125.

(14) Nilges, M. *Proteins* **1993**, *17*, 297−309.

(15) Nilges, M.; Gronenborn, A. M.; Brünger, A. T.; Clore, G. M. *Protein Eng.* **1988**, *2*, 27−38.

program ATNOS.[12b] CANDID[12] uses the same method as ARIA[11] for handling multiple possible assignments for each NOE cross-peak. However, CANDID also adds two additional features to reduce the complexity of the NOE potential hypersurface and, hence, to significantly improve the convergence rate.[12] These two features are as follows: (1) a sophisticated prefiltering of the NOE assignment lists founded on the concept of "network anchoring" which requires that any given NOE should be part of a self-consistent, relatively dense, subset of NOEs; and (2) restraint combination, which aims to minimize the impact of incorrect restraints at the expense of a temporary loss of information.[12,13] Despite these conceptual improvements over ARIA, CANDID only performs well when the number of incorrect NOE restraints represents a relatively small fraction (ca. 20−25%) of the complete NOE dataset and an ensemble of structures with a backbone coordinate precision better than 3 Å can be obtained after the first cycle of calculations.[13,16] Thus, neither ARIA nor CANDID are generally suitable for handling completely automatically peak-picked multidimensional NOE spectra which invariably contain a large fraction of incorrect assignments.

The common feature of all iterative algorithms developed to date, whether manual, semi-automated, or fully automated, is that they are heavily reliant on and biased by the coordinates of the structures calculated in the preceding refinement cycle.[13] Thus, in the manual and semi-automated cases, ambiguities are iteratively resolved and additional NOEs are assigned on the basis of successively calculated ensembles of structures.[8] In the fully automated case, incorrect NOE restraints are removed and ambiguities in NOE assignments are progressively resolved on the basis of the previously calculated structures. Thus, all published methods require that the structures calculated in the first pass are reasonably precise and accurate. If this is not the case, refinement can readily proceed down an incorrect path with consequent structural drift toward a precise but inaccurate final structure.[13]

In this paper, we introduce a new, highly error-tolerant probabilistic assignment algorithm for automated structure determination (PASD) from completely automatically peak-picked multimensional heteronuclear-separated NOE spectra. The PASD algorithm which has been incorporated into the molecular structure determination package Xplor-NIH[17] is conceptually and philosophically different from previously implemented algorithms in that it has been expressly designed with the aim of ensuring that the results from successive iteration cycles are not biased by the global fold of structures calculated in the preceding cycles. The PASD algorithm combines three features. First, during the initial stages of the calculation, a linear NOE potential energy function is employed that completely eliminates the significance of the size of an NOE distance violation on the magnitude of the atomic forces which it creates. Second, the forces generated by multiple possible assignments for a given NOE cross-peak are treated independently; this feature allows ambiguous restraints to contribute their information more effectively even when the system is far from the correct structure, thereby reducing the number of local energy minima associated with incorrect selections of assignments within an ambiguous restraint. Third, a probabilistic method is employed for the inactivation and reactivation of all NOE assignments on the fly during simulated annealing. No NOE assignment is ever removed permanently, and consequently the likelihood of becoming trapped in a false minimum of NOE assignment space is significantly reduced. All of these various features ensure that the PASD algorithm is highly tolerant of errors in the automatically peak-picked NOE restraints list and refinement is not easily funnelled down an incorrect path in either three-dimensional structure or NOE assignment space.

We demonstrate the applicability of the PASD algorithm using NOE restraint datasets generated from completely automatically peak-picked multidimensional NOE spectra for two proteins: interleukin-4 (IL-4, 136 residues), which is predominantly α-helical,[18] and cyanovirin-N (CVN, 101 residues), which comprises a substantial amount of β-sheet with an unusual topology.[19] Finally, we investigate the tolerance of the method to errors in the NOE restraints using model calculations on the small 56-residue B1 domain of streptococcal protein G (GB1).[20] The results indicate that the PASD algorithm is capable of totally automatic structure determination using multidimensional NOE spectra and chemical shift assignments, and the algorithm converges successfully to the correct structure even in cases where up to 80% of the starting long-range NOE information in a restraint dataset is incorrect.

Theory

Definition of NOE Assignment and Restraints. We first define a formal, hierarchical relationship between NOE cross-peaks, NOE distance restraints, and NOE assignments. Each NOE restraint corresponds to precisely one cross-peak in one NOE spectrum. A NOE restraint has one or more possible assignments. Each NOE assignment consists of two atom selections that are used to calculate the distance (and hence the distance violation) associated with that assignment. The atom selections associated with a NOE assignment need not specify only a single atom. If more than one atom is specified for a single selection, $(\sum r^{-6})^{-1/6}$ summation[14] is used to calculate the distance associated with the corresponding assignment. It is important to stress that $(\sum r^{-6})^{-1/6}$ summation is not used to choose among possible assignments for a given restraint. $(\sum r^{-6})^{-1/6}$ summation is only used for the simple case of nonstereospecific assignments (i.e., assignments involving methylene protons, methyl protons, methyl groups of leucine and valine, and the Hδ and Hϵ aromatic ring protons of Phe and Tyr).

General Overview of the PASD Algorithm. The PASD algorithm involves three successive passes of simulated annealing calculations in torsion angle space, each of which starts from random initial coordinates. The only information that is handed down from one pass to the next consists of estimates of the

(16) Jee, J. G.; Güntert, P. *J. Struct. Funct. Genomics* **2003**, *4*, 179−189.

(17) Schwieters, C. D.; Kuszewski, J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *121*, 65−73.

(18) (a) Powers, R.; Garrett, D. S.; March, C. J.; Frieden, E. A.; Gronenborn, A. M.; Clore, G. M. *Biochemistry* **1992**, *31*, 4334−4346. (b) Garrett, D. S.; Powers, R.; March, C. J.; Frieden, E. A.; Clore, G. M.; Gronenborn, A. M. *Biochemistry* **1992**, *31*, 4347−4353. (c) Powers, R.; Garrett, D. S.; March, C. J.; Frieden, E. A.; Gronenborn, A. M.; Clore, G. M. *Science* **1992**, *256*, 1673−1677. (d) Powers, R.; Garrett, D. S.; March, C. J.; Frieden, E. A.; Gronneborn, A. M.; Clore, G. M. *Biochemistry* **1993**, *32*, 6744−6762.

(19) Bewley, C. A.; Gustafson, K. R.; Boyd, M. R.; Covell, D. G.; Bax, A.; Clore, G. M.; Gronenborn, A. M. *Nat. Struct. Biol.* **1998**, *5*, 571−578.

(20) (a) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657−661. (b) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.

likelihood that each particular assignment within each restraint is correct. In practice, these prior likelihoods are a metric of how consistent a given assignment is with the ensemble of current structures at the end of each calculation pass. The prior likelihoods, in conjunction with instantaneous likelihoods calculated at various times during the course of simulated annealing, are used to inactivate or reactivate assignments, as well as entire NOE restraints, in a probabilistic manner during the course of simulated annealing. The first two passes of calculations make use of a linear potential function to express the NOE information, a representation that is particularly tolerant of incorrect assignments and restraints that are incompatible with the true atomic coordinates. The first and second pass calculations differ insofar that no prior information concerning assignment likelihoods is available for the first pass calculations, whereas the second pass calculations make use of prior likelihoods calculated at the end of the first pass. The third and final pass of calculations employs a quadratic potential term to generate the best possible ensemble of final structures.

Each pass involves the calculation of several hundred structures (typically around 500). However, because each individual simulated annealing calculation, within each pass, is independent of all others, the PASD algorithm lends itself to coarse-grained parallelism yielding linear speed-up with the number of CPU processors (i.e., distributed computing). A cluster of several dozen processors was used to produce the results presented here. It should be emphasized that these calculations are computer intensive. Hence, application of the PASD algorithm is impractical on a single conventional workstation and necessitates the use of an appropriately sized computer cluster.

The Linear NOE Potential Energy Function. The first two passes of the PASD algorithm make use of a linear NOE potential term which uses a novel method to express the existence of multiple assignments for each restraint. The total linear NOE energy, E_{PASDl}, summed over all restraints i, is given by:

$$E_{\mathrm{PASDl}} = \sum_i \sum_{j=1}^{\eta_i} \frac{k_{\mathrm{PASDl}}}{\eta_i} |\Delta_{ij}| \qquad (1)$$

where k_{PASDl} is a force constant (in kcal mol^{-1} Å^{-1}), η_i is the number of possible assignments for restraint i, and the distance violation for assignment j of restraint i, Δ_{ij}, is given by:

$$\Delta_{ij} = \begin{cases} r_{ij} - u_{ij}, & \text{if } r_{ij} > u_{ij} \\ 0, & \text{if } l_{ij} \le r_{ij} \le u_{ij} \\ r_{ij} - l_{ij}, & \text{if } r_{ij} < l_{ij} \end{cases} \qquad (2)$$

where u_{ij} and l_{ij} are the upper and lower distance bounds for assignment j of restraint i, and r_{ij} is the calculated distance between the selected atoms for assignment j of restraint i.

The formulation of the linear NOE potential function, E_{PASDl}, incorporates three important features. First, E_{PASDl} varies linearly with the distance violation so that the magnitude of the atomic forces generated, which depend on the derivative of the energy, is identical for any violated restraint. Second, in the case of restraints with multiple possible assignments, the overall energy associated with that restraint is calculated as the sum of the energies associated with the individual assignments making up that restraint. This helps avoid local minima in assignment space because the correct assignment will always make some contribution to the atomic forces, even if the atomic coordinates are far from the correct structure. Third, scaling the force constant k_{PASDl} by $1/\eta_i$ ensures that restraints with large numbers of possible assignments generate the same total force as restraints with only a single assignment. Taken together, these features of the linear potential energy reduce the dependence of the PASD algorithm on the actual values of the interatomic distances at any particular instant in time during simulated annealing.

The linear NOE potential representation employed here is fundamentally different from $\sum(r^{-6})^{-1/6}$ summation for ambiguous distance restraints employed by both ARIA[11] and CANDID.[12] The landscape of both the conventional quadratic and the soft asymptotic potential[15] energy terms incorporating $\sum(r^{-6})^{-1/6}$ summation is characterized by multiple local minima: whenever the atomic coordinates are relatively close to satisfying one of the possible assignments, $\sum(r^{-6})^{-1/6}$ summation effectively eliminates the atomic forces that would be generated from all other possible assignments. The linear potential energy function, on the other hand, always generates equal forces from all possible assignments, unless they have been temporarily inactivated as described below.

The Quadratic NOE Potential Energy Function for Final Refinement. As will be discussed below, the first two passes of the PASD algorithm are sufficient to generate reasonably accurate estimates of the likelihood that each particular assignment of each restraint is correct. Consequently, in the third and final pass of the PASD algorithm, the error-tolerant features of the linear NOE potential function are no longer required. The third and final pass of the PASD algorithm makes use of a quadratic NOE potential function of the form:

$$E_{\mathrm{PASDq}} = k_{\mathrm{PASDq}} \sum_i \Delta_{ij}^{\,2} \qquad (3)$$

where k_{PASDq} is a force constant (in kcal mol^{-1} Å^{-2}), and Δ_{ij} is defined in eq 2. It is important to note that, in contrast to the equation for the linear NOE potential term E_{PASDl}, there is no summation over all of the activated assignments; rather only a single assignment is active for each restraint i at any given time during simulated annealing. This chosen assignment is reselected using a probabilistic algorithm several times during the course of simulated annealing as discussed below.

Probabilistic Inactivation and Reactivation of Assignments during Pass 1 and Pass 2 Simulated Annealing Calculations. The linear NOE potential function (eq 1) is not sufficient in its own right to allow convergence to the correct structure. This is largely because the presence of forces from incorrect assignments complicates the energy hypersurface. We therefore make use of a probabilistic algorithm to temporarily inactivate individual assignments, based on their distance violation. The schedule for inactivation/reactivation is discussed subsequently (cf. Table 1).

The instantaneous likelihood $\lambda_{\mathrm{v}}(i,j)$ of each assignment j within each restraint i is evaluated using a Boltzmann function of its distance violation at random intervals during the course of simulated annealing:

$$\lambda_{\mathrm{v}}(i,j) = e^{-\Delta_{ij}^2/D_{\mathrm{v}}^2} \qquad (4)$$

Table 1. Simulated Annealing Protocol for the PASD Algorithm[a]

	pass 1	pass 2	pass 3
number of structures calculated	500	500	500
High-Temperature Phase 1			
bath temperature (K)	4000	4000	4000
duration (ps)	40	15	50
k_{PASDl} (kcal mol^{-1} Å^{-1})	1.0	1.0	
k_{PASDq} (kcal mol^{-1} Å^{-2})			3.0
D_v (Å)	∞		
ΔE_c (restraint^{-1})	100	0.1	0.66
w_o (λ_o weighting factor)	0	1	1
w_a (λ_a weighting factor)			1
number of NOE reevaluations	1	10	10
k_{vdw} (kcal mol^{-1} Å^{-4})	1.0	1.0	1.0
s_{vdw}	1.2	1.2	1.2
van der Waals interactions	Cα−Cα only	Cα−Cα only	Cα−Cα only
k_{dihed} (kcal mol^{-1} rad^{-2})	200	200	200
k_{DB}	0.1	0.1	0.1
High-Temperature Phase 2			
bath temperature (K)	4000	4000	
duration (ps)	40	40	
k_{PASDl} (kcal mol^{-1} Å^{-1})	1	1	
D_v (Å)	5.5	5.5	
ΔE_c (restraint^{-1})	100	0.1	
w_o (λ_o weighting factor)	0	0.5	
number of NOE reevaluations	10	10	
k_{vdw} (kcal mol^{-1} Å^{-4})	1	1	
s_{vdw}	1.2	1.2	
van der Waals interactions	Cα−Cα only	Cα−Cα only	
k_{dihed} (kcal mol^{-1} rad^{-2})	200	200	
k_{DB}	0.1	0.1	
Cooling Phase			
bath temperature (K)	4000 → 100	4000 → 100	4000 → 100
duration (ps)	250	250	250
k_{PASDl} (kcal mol^{-1} Å^{-1})	1 → 30	1 → 30	
k_{PASDq} (kcal mol^{-1} Å^{-2})			3 → 30
D_v (Å)	5.5 → 2.0	5.5 → 2.0	2.0 → 0.7
ΔE_c (restraint^{-1})	0.1 → 0.001	0.1 → 0.001	0.33 → 0.0033
w_o (λ_o weighting factor)	0	0.5 → 0	0.5 → 0
w_a (λ_a weighting factor)			0.5 → 0
number of NOE reevaluations	64	64	64
k_{vdw} (kcal mol^{-1} Å^{-4})	0.04 → 4.0	0.04 → 4.0	0.04 → 4.0
s_{vdw}	0.9 → 0.8	0.9 → 0.8	0.9 → 0.8
van der Waals interactions	all atoms	all atoms	all atoms
k_{dihed} (kcal mol^{-1} rad^{-2})	200	200	200
k_{DB}	0.1 → 10.0	0.1 → 10.0	0.1 → 10.0

[a] k_{PASDl}, k_{PASDq}, D_v, w_o, ΔE_c, and w_a are defined in eqs 1, 3, 4, 5, 7, and 8, respectively. k_{db}, k_{dihed}, and k_{vdw} are the force constants for the torsion angle database potential of mean force,[26] the square-well quadratic torsion angle potential,[24] and the quartic van der Waals repulsion potential,[15] respectively. s_{vdw} is the van der Waals radius scale factor used in the quartic van der Waals repulsion term.[15]

where D_v is the characteristic distance, defined as the distance violation (in angstroms) at which the instantaneous likelihood is equal to $1/e$ (Figure 1). The value of D_v is varied using a predefined schedule during simulated annealing: the smaller the value of D_v, the more stringent the selection process.

The instantaneous likelihood, $\lambda_v(i,j)$, is used in combination with the prior likelihood estimate, $\lambda_p(i,j)$, defined below, to determine an overall assignment likelihood, $\lambda_o(i,j)$:

$$\lambda_o(i,j) = (1 - w_o)\lambda_v(i,j) + w_o\lambda_p(i,j) \qquad (5)$$

where w_o is a weighting factor between 0 and 1 which determines the balance between $\lambda_v(i,j)$ and $\lambda_p(i,j)$. For the pass 1 calculations, there is no prior information, and hence w_o is set to zero.

The prior likelihood, $\lambda_p(i,j)$, is determined at the end of each calculation pass and constitutes the only information that is transmitted from one pass of calculations to the next. $\lambda_p(i,j)$,

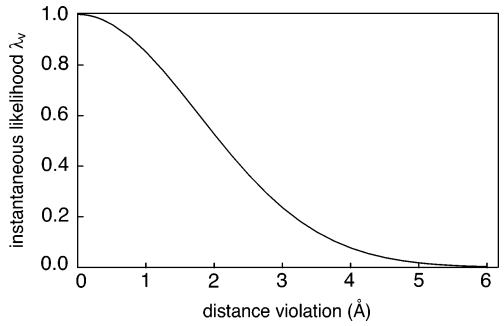

Figure 1. The dependence of the instantaneous likelihood λ_v (eq 4) as a function of distance violation is expressed in terms of a Boltzman probability function. In the example plotted, the characteristic distance, D_v, is equal to 2.5 Å.

which in essence is a metric of how consistent a given assignment is with the ensemble of converged structures at the end of each calculation pass, is given by:

$$\lambda_p(i,j) = \frac{1}{N_c}\sum_{k=1}^{N_c} \mathcal{U}(D_c - \Delta_{ij,k}) \qquad (6)$$

where N_c is the number of converged structures, D_c is the cutoff distance (set to 0.5 Å) above which an assignment is said to be violated, $\Delta_{ij,k}$ is the violation of assignment j within restraint i in converged structure k (see eq 2), and $\mathcal{U}$ is the Heaviside step function.[21] The converged structures at the end of each pass of the PASD algorithm are defined operationally as the 10% of the structures with the fewest long-range NOE distance violations (i.e., involving residues separated by more than five positions in the primary amino acid sequence).

The overall assignment likelihood $\lambda_o(i,j)$ is used in conjunction with a random number generator to determine whether a particular assignment should be inactivated or reactivated. Thus, for each assignment, a random number X is generated between 0 and 1. If $X < \lambda_o(i,j)$, assignment j within restraint i is activated; otherwise it is inactivated. When an assignment of restraint i is inactivated, the number of possible assignments for that restraint, η_i (cf. eq 1), is reduced by one. Consequently, the effective force constant for the remaining active assignments within restraint i is increased (cf. eq 1). If a restraint has no active assignments, the restraint itself is said to be inactivated and there are no forces associated with it. It is important to stress that no assignment (or restraint) is ever inactivated permanently. Thus, if the atomic coordinates become compatible with an inactivated assignment at the next time the instantaneous likelihoods, $\lambda_v(i,j)$, are calculated, then that assignment may be reactivated. The schedule for inactivation/reactivation is discussed below (cf. Table 1).

Because the mechanism for activating or inactivating assignments described above is stochastic, bad choices can occasionally be made that frustrate the optimization of the target function. We therefore combine the assignment activation/inactivation algorithm described above with a complementary Monte Carlo loop[22] that evaluates the entire set of activated assignments as a whole, thereby avoiding gross violations within the entire set

(21) Abramowitz, M.; Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*; Dover: New York, 1972; p 1020.
(22) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, J. H.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

of restraints. The Monte Carlo probability p is given by

$$p = e^{-(E_{new} - E_{orig})/\Delta E_c} \quad (7)$$

where E_{orig} and E_{new} are the total NOE energies (eqs 1 or 3 for the linear or quadratic NOE potentials, respectively) calculated with the currently active assignments and with the new set of active assignments generated using eqs 4−6, respectively, and ΔE_c is the characteristic energy change, defined as the increase in energy at which the probability of accepting the proposed Monte Carlo move is equal to $1/e$. A random number Y between 0 and 1 is generated, and, if $Y < p$, the new set of activated assignments is accepted; otherwise that particular new set of assignments is rejected, and another set of active assignments is generated.

Probabilistic Inactivation and Reactivation of Assignments during the Final Pass 3 Simulated Annealing Calculations. In the case of the pass 3 calculations, only a single assignment j is active at any one time for each restraint i (cf. eq 3).[23] This is equivalent to making the simplifying assumption that a given NOE cross-peak arises from only a single NOE interaction. The assignment choice likelihood, $\lambda_a(i,j)$, of each assignment j within restraint i is therefore given by:

$$\lambda_a(i,j) = (1 - w_a)\lambda_v^{norm}(i,j) + w_a\lambda_p^{norm}(i,j) \quad (8)$$

where w_a is a weighting factor (between 0 and 1), and $\lambda_v^{norm}(i,j)$ and $\lambda_p^{norm}(i,j)$ are the normalized instantaneous and prior assignment likelihoods given by:

$$\lambda_v^{norm}(i,j) = \lambda_v(i,j)/\sum_j \lambda_v(i,j) \quad (9)$$

$$\lambda_p^{norm}(i,j) = \lambda_p(i,j)/\sum_j \lambda_p(i,j) \quad (10)$$

where $\lambda_v(i,j)$ and $\lambda_p(i,j)$ are given by eqs 4 and 6, respectively. The overall assignment likelihoods $\lambda_a(i,j)$ are used in combination with a random number generator, using the same procedure as that described above, to pick a single assignment for each restraint i.

During the pass 3 calculations, restraints can also be temporarily inactivated on the basis of their distance violations. The instantaneous likelihood λ_v is calculated using eq 4 only for the chosen assignment. The overall likelihood λ_o is then calculated from eq 5 using the prior likelihood λ_p of the chosen assignment of the current restraint. As in the case of the pass 1 and pass 2 calculations, the choice of assignments and the activation/inactivation of restraints are made in conjunction with the Monte Carlo algorithm (eq 7) to avoid potential bad choices. The schedule for inactivation/reactivation is discussed below (cf. Table 1).

(23) Given that only a single NOE assignment is active at any point in time during pass 3, the question arises as to how to best treat the target distance in eq 3 in cases where a single NOE cross-peak arises from genuine overlap of two or more NOE interactions. Our choice is to leave the target distance unaltered, because, in most cases, cross-peaks do arise from only a single NOE interaction. If a cross-peak genuinely arises from say two NOE interactions of approximately equal intensity, this will be directly reflected in the final likelihoods computed by the PASD algorithm (see section on Final Assignment and Restraints Likelihoods). The distance correction to the upper distance bound for two equally probable assignments would be 12% at most (corresponding to a reduction by a factor of 2 in NOE intensity). However, because the NOE intensities are converted into loose distance ranges, such a correction, in the context of the PASD calculations, is unnecessary. If, on the other hand, the contribution of one of the two NOE interactions constitutes only a very small proportion of the NOE cross-peak intensity, and the NOE restraint associated with that cross-peak is classified as strong, then the weaker interaction would be excluded by the probabilistic inactivation mechanism employed by the PASD algorithm, with no untoward effect.

Implementation of the PASD Algorithm in Simulated Annealing Calculations

A critical aspect of the PASD algorithm comprises not only the various functions described in the previous section but also the protocol employed for simulated annealing. Three successive passes of simulated annealing calculations are involved. Five hundred independent structures are calculated for each pass, using different random number seeds for the assignment of initial velocities. The target function comprises the following terms: a linear (eq 1) or quadratic (eq 3) NOE potential function, a quartic van der Waals repulsion term to prevent atomic overlap,[15] a square-well torsion angle potential[24] (for loose torsion angle restraints derived from backbone chemical shifts using the program TALOS[25]), and a torsion angle database potential of mean force to bias sampling during simulated annealing to regions of conformational space that are known to be physically realizable from very high-resolution protein crystal structures.[26] The various parameters employed and the manner in which they are varied during simulated annealing are summarized in Table 1. We have found this simulated annealing schedule to be robust, and consequently the application of the PASD algorithm is completely automatic and requires no human intervention.

Starting Coordinates. Each pass of the PASD algorithm begins from a set of randomly generated coordinates (comprising a random selection of torsion angles with intact covalent geometry). These initial coordinates are minimized in torsion angle space against a target function consisting of a radius of gyration restraint to collapse the structure,[20b,27] a repulsive van der Waals interaction term between Cα atoms only, and any backbone torsion angle restraints (derived, for example, from backbone ^{13}C, ^{15}N, and ^{1}H chemical shifts using the database program TALOS[25]) or disulfide bond restraints that are to be used in the structure calculation. The resulting structure has roughly the right overall size for a globular protein and displays good agreement with the applied torsion angle restraints, but is otherwise random.

Pass 1 Protocol. The pass 1 protocol comprises three phases: two high-temperature (4000 K) phases and a slow cooling phase (from 4000 to 100 K). The linear NOE potential (eq 1) is used throughout. During the first high-temperature phase (40 ps), all assignments for all restraints are active, and, to enhance conformational sampling, the repulsive van der Waals interaction term is restricted to Cα atoms only. In the second high-temperature phase (40 ps), assignments are inactivated and reactivated 10 times at random intervals as described above using eqs 4−6, with a value of 5.5 Å for the characteristic distance D_v, and a prior likelihood weight, ω_o, of zero. In the third phase, the system is cooled from 4000 to 100 K over 250

(24) Clore, G. M.; Nilges, M.; Sukuraman, D. K.; Brünger, A. T.; Karplus, M.; Gronenborn, A. M. *EMBO J.* **1986**, *5*, 2729−2735.
(25) Cornilescu, G.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1999**, *13*, 289−302.
(26) (a) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *Protein Sci.* **1996**, *5*, 1067−1080. (b) Clore, G. M.; Kuszewski, J. *J. Am. Chem. Soc.* **2002**, *124*, 2866−2867.
(27) The target radius of gyration R_{gyr}(target) is calculated using $R_{gyr}(\text{target}) = 2.2N^{0.38}$, where N is the number of residues included in the structure determination (ref 20b).

ps, while the force constants for the linear NOE potential term (k_{PASDl}; cf. eq 1) and the quartic van der Waals repulsion term (applied to all atoms) are progressively increased (in a geometric manner). D_v is progressively reduced from 5.5 to 2.0 Å, making the inactivation mechanism more stringent as cooling progresses. Probabilistic activation/inactivation of assignments is carried out 64 times (at random intervals) during the cooling phase. The top 10% of structures, defined operationally as those having the smallest number of NOE distance violations >0.5 Å involving long-range NOE assignments (i.e., between residues separated by more than 5 positions in the linear amino acid sequence), are used to calculate the prior likelihood estimates $\lambda_p(i,j)$ for each assignment j of restraint i (eq 6) for the second pass calculations.

Pass 2 Protocol. As in the case of the pass 1 protocol, the pass 2 protocol comprises two high-temperature (4000 K) phases, and a slow cooling phase (from 4000 to 100 K) with the linear NOE potential (eq 1) used throughout. In the first high-temperature phase (15 ps), probabilistic activation/reactivation of assignments is carried out 10 times at random intervals with the weighting factor w_o (eq 6) set to 1.0. Thus, only the prior likelihoods, $\lambda_p(i,j)$ are employed during this phase. In the second high-temperature phase (40 ps), probabilistic activation/inactivation of assignments is carried out 10 times at random intervals with equal weighting of the prior, $\lambda_p(i,j)$, and instantaneous, $\lambda_v(i,j)$, likelihoods ($w_o = 0.5$; cf. eq 5), using $D_v = 5.5$ Å to calculate the latter (eq 4). In the cooling phase (250 ps), activation/inactivation of assignments is carried out 64 times at random intervals, with the values of D_v and w_o being geometrically reduced from 5.5 to 2.0 Å and from 0.5 to 0, respectively. All other parameters are identical to the cooling phase of pass 1. At the end of pass 2, updated prior assignment likelihood estimates, $\lambda_p(i,j)$, are calculated in the same manner as at the end of the pass 1 calculations.

Pass 3 Protocol. The pass 3 protocol comprises a single high-temperature phase (4000 K) followed by a slow cooling phase with the quadratic NOE potential (eq 7) employed throughout. In the high-temperature phase (50 ps), assignments (one per restraint) are selected (cf. eq 8) and restraints are activated/inactivated (cf. eqs 4−6) 10 times at random intervals. The prior likelihood weights for both the assignment choice (w_a; cf. eq 8) and the restraint evaluation (w_o; cf. eq 5) algorithms are set to 1.0, so that decisions to alter the choice of assignment for a given restraint and to turn restraints on and off are made solely on the basis of the prior likelihoods calculated at the end of the pass 2 calculations. During the cooling phase (from 4000 to 100 K over a duration of 250 ps), assignments are selected/deselected and restraints are activated/inactivated 64 times at random intervals. The value of D_v is reduced from 2.0 to 0.7 Å, the weighting factors w_o and w_a are reduced from 0.5 to 0, and the NOE force constant k_{PASDq} is increased from 3 to 30 kcal mol^{-1} Å^{-2}. All other parameters are the same as those for the cooling phase of the pass 1 calculations.

Final Assignment and Restraints Likelihoods. Final assignment likelihoods are computed at the end of the pass 3 calculations using eq 6. These are readily interpretable by the user as follows. An incorrect restraint will have final likelihoods near zero for each possible assignment. These should reflect artifacts, bad peak-picking, or misassigned cross-peaks. (The latter generally occur as a consequence of either missing assignment or from a combination of small discrepancies between the values of the actual chemical shifts and those present in the chemical shift assignment table, and the limitations of using hard tolerances for automated cross-peak assignment.) A correct restraint will have a final likelihood close to 1 for one assignment and near-zero likelihoods for all of the other assignments within that restraint. Correct restraints with more than one correct assignment, as will be the case if a single NOE cross-peak actually comprises several highly overlapped NOE cross-peaks which cannot be spectrally resolved, will have several assignments with likelihood values close to the reciprocal of the number of correct assignments, with the likelihoods for all other assignments near zero.

Methods

Calculations. The PASD algorithm was written in C++ and incorporated as a module known as MARVIN into the NMR structure determination package Xplor-NIH (available at http://nmr.cit.nih.gov/xplor-nih).[17] All simulated annealing calculations were carried out in torsion angle space using the internal variable (IVM) module[28] of Xplor-NIH which incorporates an automatic variable step integrator. The simulated annealing protocols were written using the Tcl interface of Xplor-NIH. Backbone ϕ/ψ torsion angle restraints were derived from backbone ^{1}H, ^{15}N, and ^{13}C assignments using the program TALOS.[25] All of the ϕ/ψ predictions classified as "good" by TALOS were used to automatically generate an Xplor-NIH torsion angle restraint file with upper and lower bounds given by the mean predicted value ± 1 standard deviation.[29] Structures were displayed using the program VMD-XPLOR.[30]

Automatic NOE Cross-Peak-Picking and Initial NOE Assignment Generation for Interleukin-4. Previously recorded 3D ^{13}C-separated, 3D ^{15}N-separated, and 4D ^{13}C/^{13}C-separated NOE spectra for interleukin-4 (IL-4)[18] and 3D ^{13}C-separated and 3D-^{15}N separated NOE spectra for cyanovirin (CVN)[19] were automatically peak-picked using the program CAPP (with default settings),[9] which generates a list of cross-peak chemical shift coordinates for each spectrum. Possible assignments for each cross-peak were then generated using the program STAPP[9] on the basis of the cross-peak chemical shift coordinates, the identity of each spectral dimension and any chemical bonding constraints between them, a list of published ^{1}H, ^{15}N, and ^{13}C chemical shift assignments obtained from through-bond triple resonance correlation experiments, and two error tolerances along each dimension. The looser error tolerance corresponds to the maximum error (in ppm) tolerance associated with a particular dimension; the tighter error tolerance corresponds to the degeneracy proximity limit (i.e., two chemical shift values within this limit cannot be distinguished). The error tolerances were calculated for each NOE spectrum by selecting 15 cross-peaks that correspond to unambiguous intraresidue NOEs and comparing their chemical shift coordinates with those in the chemical shift table. The maximum error tolerance was set to two standard deviations of these 15 error measurements, while the tighter error tolerance was set to one standard deviation. The values for the maximum error tolerances for the three IL-4 NOE spectra were as follows: 0.06, 0.10, and 0.04 ppm for the ^{1}H (F_1), ^{13}C (F_2), and ^{1}H (F_3) dimensions, respectively, in the 3D ^{13}C-separated NOE spectrum; 0.018, 0.255, and 0.015 ppm for the ^{1}H(F_1), ^{15}N(F_2), and ^1H_N(F_3) dimensions, respectively, in the 3D ^{15}N-separated NOE spectrum; and 0.30, 0.10, 0.22, and 0.04 ppm in the ^{13}C(F_1), ^{1}H(F_2), ^{13}C(F_3), and ^{1}H(F_4) dimensions, respectively, in the 4D ^{13}C/^{13}C-separated NOE spectrum. The values for the

(28) Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* **2001**, *152*, 288−302.
(29) A "good" TALOS prediction is defined as follows: either all 10 best database matches for that residue fall in a consistent region of the Ramachandran map, or 9 out of the 10 best database matches fall in a consistent region with $\phi < 0$, and the one outlier also lies in the $\phi < 0$ half of the Ramachandran map (cf. ref 25).
(30) Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* **2001**, *149*, 239−244.

maximum error tolerances for the two CVN NOE spectra were as follows: 0.033, 0.42, and 0.033 ppm for the ^{1}H (F_1), ^{13}C (F_2), and ^{1}H (F_3) dimensions, respectively, in the 3D ^{13}C-separated NOE spectrum; and 0.036, 0.42, and 0.036 ppm for the ^{1}H(F_1), ^{15}N(F_2), and ^1H_N(F_3) dimensions, respectively, in the 3D ^{15}N-separated NOE spectrum.

For each NOE cross-peak, all possible assignments whose chemical shifts are within the maximum error tolerances are gathered. If, for a given NOE cross-peak, there is only one possible assignment within the tighter error tolerances, only that assignment is reported for that particular cross-peak; otherwise all of the possible assignments within the looser tolerance limits are reported. Two filters were employed by STAPP in the analysis of NOE cross-peaks in the 3D ^{13}C-separated and 4D ^{13}C/^{13}C-separated NOE spectra: a symmetry filter which eliminates a given assignment if the expected symmetry partner is absent from the spectrum, and a sign filter to take into account the sign alternation of the cross-peaks arising from extensive folding employed in the ^{13}C-dimensions.

All stereospecific assignments were eliminated from the output of STAPP and replaced by nonstereospecific assignments. Thus, for example, two NOE cross-peaks corresponding to NOEs from proton *x* to the methylene Hβ1 and Hβ2 protons of a particular residue are represented by two identical restraints involving the same $(\sum r^{-6})^{-1/6}$ sum distance calculated from the corresponding two distances.

Generating Distance Restraints from Spectral Intensities. The output of STAPP consists of a list of NOE cross-peaks with the observed peak-height intensity and the atomic selections to define one or more possible assignments for each cross-peak. Distance bounds were classified into four ranges, 1.8−2.7, 1.8−3.3, 1.8−5.0, and 1.8−6.0 Å,[5] corresponding to strong (the most intense 20%), medium (the next most intense 30%), weak (the following most intense 30%), and very weak (the remaining 20%) cross-peaks. In addition, a correction of 0.5 Å was added to the upper bounds of restraints involving methyl protons to account for their higher apparent intensities in the spectra.[31] It is important to note that any more involved analysis of NOE intensities is not justified due to the presence of incorrect assignments and restraints in the initial restraints file and the absence of a good starting structure.

Combining Restraints from Different Spectra. The automatic cross-peak analysis described above is carried out independently for each NOE spectrum. The restraints derived from all of the spectra are then concatenated together. No effort is made to eliminate duplicate assignments or restraints arising from different NOE spectra for several reasons. First, the presence of incorrect restraints or differing numbers of possible assignments for the restraints arising from the different spectra makes recognizing duplicate restraints nontrivial. Second, if the same restraint arises from cross-peaks in multiple spectra, then it seems reasonable to give this restraint more weight in the structure determination than a restraint that arises from a cross-peak in only a single spectrum, because a multiple-spectra restraint is supported by more experimental data.

Generation of Test Restraint Datasets for GB1. Experimental NOE-derived distance restraints for the B1 domain of streptococcal protein G (GB1) were obtained from the protein data bank (accession code 3gb1.mr).[20b] These comprise 735 interproton distance restraints subdivided into 424 short-range ($1 < |l - m| \leq 5$) and 247 long-range ($|l - m| > 5$) interresidue restraints (where *m* and *l* are the residue positions within the primary sequence). 1.7% and 4% of the published short- and long-range interresidue NOE restraints, respectively, are violated by >0.5 Å in the X-ray structure (PDB accession code 1PGB).[32]

Twenty separate test restraint datasets were generated as follows using the experimental NOE restraints (3gb1.mr)[20b] as a starting point. (Note that all of the experimental restraints each have only a single assignment.) First, all stereospecific assignments were eliminated from the experimental restraints list and replaced by corresponding $(\sum r^{-6})^{-1/6}$ sum restraints. A set of incorrect distance restraints with upper and lower bounds set to 5.0 and 1.8 Å, respectively, was then generated by randomly choosing pairs of nonexchangeable protons and filtering the results to ensure that all selected restraints were violated by >0.5 Å in the X-ray structure. Incorrect NOE restraints were added to each dataset until a given target number of long-range incorrect restraints was attained. Any short-range incorrect distance restraints that happened to be generated using the above algorithm were also included in the dataset. Finally, incorrect NOE assignments were generated using the same algorithm and added to randomly selected NOE restraints (among both the experimental and random incorrect restraints).

Comparing the Incorrect Long-Range Information Content of Different Datasets. In this work, we examine the performance of the PASD algorithm using a variety of datasets. To compare the quality of these datasets, we define a measure of the incorrect long-range information content of a set of initial NOE restraints; that is, the situation present at the very beginning of the pass 1 calculations in which all assignments of all restraints are active and no estimates of the prior likelihoods of the various assignments are available. In the case of the linear NOE potential function (eq 1), the magnitude of the atomic force produced by each assignment of restraint *i* is $1/\eta_i$, where η_i is the number of assignments for restraint *i*. Hence, one can determine the fraction $f_{\text{bad}}^{\text{long}}$ of long-range atomic forces that arise from long-range assignments (i.e., between residues separated by more than 5 amino acids in the primary sequence) that are inconsistent with the known correct structure (as defined by a corresponding distance violation >0.5 Å):

$$f_{\text{bad}}^{\text{long}} = \frac{[\sum_i 1/\eta_i \sum_j \mathscr{U}(S_{ij} - S_c) \cdot \mathscr{U}(\Delta_{ij} - D_c)]}{[\sum_i 1/\eta_i \sum_j \mathscr{U}(S_{ij} - S_c)]} \qquad (11)$$

where S_{ij} is the primary sequence distance between the atoms selected by assignment *j* of restraint *i*, S_c is the sequence distance cutoff (in residues) which in this instance is set to 5, D_c is the distance violation cutoff (in angstroms) which in this case is set to 0.5 Å, $\mathscr{U}$ is the Heaviside step function,[21] and η_i and Δ_{ij} are defined in eq 1. (Note we ignore NOE connectivities between residues separated by ≤5 amino acids in the primary sequence because these do not provide any significant global structural information.)

Results

Quality of the IL-4 and CVN NOE Restraint Datasets Generated by Completely Automatic Peak-Picking of Multidimensional NMR Spectra. A statistical characterization of the automatically peak-picked NOE restraints derived from multimensional heteronuclear-separated NOE experiments on IL-4[18] and CVN[19] is provided in Table 2. It is worth noting that IL-4 presents a challenging system from the perspective of automated NOE cross-peak assignment because IL-4 is a largely helical protein and hence exhibits extensive chemical shift overlap with limited spectral dispersion.[18] Because the structures of IL-4[18c,d,32] and monomeric CVN[19] are known, one can readily evaluate the fraction $f_{\text{bad}}^{\text{long}}$ of the atomic forces arising from "long-range" assignments (i.e., between residues separated by more than 5 positions in the linear amino acid sequence) in the automatically peak-picked restraints list that are violated by >0.5 Å in the crystal (IL-4)[33] or NMR (CVN)[19] structures. The overall

(31) Clore, G. M.; Gronenborn, A. M.; Nilges, M.; Ryan, C. A. *Biochemistry* **1987**, *26*, 8012−8023.

(32) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721−4729.

Table 2. Statistical Characterization of the Automatically Peak-Picked NOE Restraint Dataset for IL-4 and CVN[a]

	number of restraints/average number of assignments per restraint[b]			
	good long-range	good short-range	bad long-range	bad short-range
	(a) IL-4			
3D ^{13}C-separated NOE spectrum ($f_{bad}^{long} = 61.2\%$)	202/1.4 ± 0.7	603/1.2 ± 0.7	219/1.4 ± 0.7	24/2.0 ± 1.6
3D ^{15}N-separated NOE spectrum ($f_{bad}^{long} = 98.1\%$)	5/1.8 ± 0.8	314/1.5 ± 1.0	149/1.9 ± 1.1	43/2.3 ± 1.2
4D ^{13}C/^{13}C-separated NOE spectrum ($f_{bad}^{long} = 79.1\%$)	90/1.9 ± 1.6	276/1.7 ± 1.8	180/2.0 ± 1.6	45/3.2 ± 3.1
overall ($f_{bad}^{long} = 75.1\%$)	297/1.5 ± 1.1	1193/1.4 ± 1.2	548/1.8 ± 1.2	112/2.6 ± 2.3
	(b) CVN			
3D ^{13}C-separated NOE spectrum ($f_{bad}^{long} = 51.8\%$)	208/1.1 ± 0.5	296/1.1 ± 0.3	172/1.2 ± 0.5	27/1.8 ± 1.8
3D ^{15}N-separated NOE spectrum ($f_{bad}^{long} = 88.9\%$)	180/2.5 ± 1.6	571/2.2 ± 1.7	474/2.1 ± 1.3	123/2.6 ± 1.8
overall ($f_{bad}^{long} = 77.5\%$)	388/1.8 ± 1.3	867/1.8 ± 1.5	646/1.8 ± 1.2	150/2.5 ± 1.8

[a] "Long-range" restraints are defined as those restraints which have no assignment with a primary sequence separation ≤5 residues. All other restraints are deemed "short-range". "Good" restraints are defined as restraints which have at least one assignment with a distance violation ≤0.5 Å in the 2.25 Å resolution crystal structure of IL-4 (PDB code 1RCB)[33] or the NMR structure of the monomeric form of CVN (PDB code 2EZM).[19] All other restraints are deemed "bad". f_{bad}^{long} is the fraction of long-range forces that are incorrect (i.e., originating from assignments with distance violations >0.5 Å in the crystal structure of IL-4 or the NMR structure of CVN) at the beginning of the calculations (cf. eq 11). The overall fractions of long-range forces that originate from assignments with distance violations >1, >5, and >10 Å are 73.2%, 63.3%, and 41.9%, respectively, for IL-4, and 75.9%, 66.4%, and 47.1%, respectively, for CVN. [b] 60% of the NOE restraints for IL4 and 72% for CVN have unique assignments; the maximum number of assignments per NOE restraint is 16 for IL-4 and 10 for CVN.

value of f_{bad}^{long} from all spectra combined is 75.1% for IL-4 and 77.5% for CVN. The corresponding values for violations >1, >5, and >10 Å are 73.2%, 63.3%, and 41.9%, respectively, for IL-4, and 75.9%, 66.4%, and 47.1%, respectively, for CVN. Another way of assessing the quality of the automatically peak-picked restraints is to categorize the long- and short-range restraints into "good" and "bad", defined by the presence or absence, respectively, of at least one assignment with a distance violation ≤0.5 Å in the crystal (IL-4) or NMR (CVN) structures. In this context, a restraint is considered to be "long-range" if it contains no assignment between residues separated by 5 or less in the primary amino acid sequence. Overall, ∼65% of the "long-range" restraints (297 out of a total of 845 for IL-4, and 646 out of 1034 for CVN) are "bad". In addition, it is worth noting, contrary to what one might expect, that the quality of the "long-range" restraints generated automatically from the 4D ^{13}C/^{13}C-separated NOE spectrum recorded on IL-4 is significantly worse than that from the 3D ^{13}C-separated NOE spectrum. This is largely due to the lower digital resolution of the 4D spectrum and the concomitant decrease in accuracy of the cross-peak positions. Moreover, the quality of "long-range" restraints generated automatically from the 3D ^{15}N-separated NOE spectra is much worse than that from either the 3D or the 4D ^{13}C-separated NOE spectra. This is due in large part to the fact that the cross-peaks in a 3D ^{15}N-separated NOE spectrum, with the exception of NH−NH cross-peaks, do not have symmetry-related cross-peaks that can be used for filtering (see Methods).

Clearly, the presence of so many "bad" long-range restraints presents a considerable challenge to automatic structure determination. The fraction of "bad" short-range restraints (i.e., which have at least one assignment between residues separated by 5 or less in the primary amino acid sequence) is much smaller, only about 8.5% for IL-4 and 14.7% for CVN.

Two questions arise from the data presented in Table 2. Why does automatic assignment of cross-peaks based on chemical shifts and peak tables result in so many "bad" restraints, and why are the "bad" restraints predominantly "long-range"?

The large number of "bad restraints" that are generated by automated peak-picking and analysis based on a chemical shift table is due to the interplay of several factors. These include tight NOE assignment tolerances, tight 3D symmetry matching tolerances, imprecise values for the chemical shift assignment data, and the simple NOE assignment protocol employed by STAPP.

Tight NOE assignment and 3D symmetry matching tolerances are employed to limit the number of ambiguous NOE assignments per cross-peak. As a consequence, however, the correct assignment can be missed if either the peak lies outside the NOE assignment tolerance or the 3D symmetry peak is outside the 3D symmetry matching tolerance. Obviously, missing the correct NOE assignment for a particular cross-peak yields a "bad" restraint. In the case of IL-4, the chemical shift data exacerbate this problem because they were generated several years ago by manual analysis of double and triple resonance through-bond correlation experiments and the chemical shifts were only reported to within 0.1 ppm for ^{15}N and ^{13}C, and 0.01 ppm for ^{1}H.[18a] Updating and improving the IL4 chemical shift data was not carried out to more accurately reflect the more general case where the chemical shift data may have unassigned or incorrectly assigned resonances. (In the case of both IL-4 and CVN, the resonance assignments were >99.5% complete.) Regardless of the precision of the chemical shift data, there are always small variations and inconsistencies between the shifts in the table and the true shifts for the sample on which the NOE data are actually recorded. This is clearly in evidence for both the IL-4 and the CVN data. These arise from variations in sample conditions that are difficult to control: these include small differences in temperature from one spectrometer to another as well as from one experiment to another (e.g., TOCSY type through-bond correlation experiments used for side-chain assignments invariably cause a small amount of sample heating);

chemical shift differences between samples dissolved in D_2O and H_2O; and small sample differences in pH and concentration.

The NOE cross-peak assignment program STAPP (as described in the Methods section) was primarily designed to be used for the purpose of iterative structure refinement to locate NOE assignments consistent with a postulated structure. STAPP, however, was never designed to find all possible and reasonable NOE assignments for a given peak. The algorithm employed by STAPP is very simple and noniterative. The use of hard tolerances rather than a probability distribution is a major cause of bad assignments, and the noniterative nature of STAPP prevents reassignment of NOE cross-peaks based on later NOE cross-peak assignments.

The large predominance of "bad" long-range restraints over short-range ones is largely structural in origin. First, the maximum distance associated with an incorrect long-range assignment is far greater than that associated with an incorrect short-range assignment. Consequently, the incorrect assignment of a given cross-peak to a short-range interaction (particularly if this happens to be intraresidue or sequential interresidue) is much more likely to be satisfied in the true structure than an incorrect assignment to a long-range interaction, because the distance range for the former is far more limited than for the latter. Indeed, in the case of IL-4, 64% of the short-range restraints are intraresidue, and a further 12% are sequential; the corresponding values for CVN are 47% and 31%, respectively.

Automatic Structure Calculation of IL-4 and CVN Using the PASD Algorithm. The performance of the PASD algorithm was tested on two structurally diverse proteins, IL-4 and CVN, both of which provide distinct challenges for NMR structure determination. IL-4 is predominantly helical, consisting of a left-handed four-helix bundle with two overhand connections.[18] Helical proteins always present a challenge for fold determination by NMR because the long-range NOEs are generally limited to side-chain−side-chain interactions, and, in general, side-chain ^{1}H resonances exhibit a much higher degree of overlap and degeneracy than the ^{1}H backbone resonances. CVN, on the other hand, is an elongated, largely β-sheet protein that displays an unusual topology with structural pseudo-symmetry at two levels:[19] there are two sequential sequence repeats (residues 1−50 and 51−101) which have 32% sequence identity and superimpose with a backbone rms of ∼1.3 Å; and there are two symmetrically related structural domains comprising residues 1−39 and 91−101 and residues 40−90, each of which comprises a triple-stranded β-sheet of one repeat and a β-hairpin from the other repeat, which also superimpose with a backbone rms of ∼1.3 Å. In addition, the loop connecting strand β5 to helix α3 in domain B (which in domain A would be equivalent to a link between the N- and C-termini of the protein) is unusual and strained. Finally, the elongated nature of CVN, which as an aspect ratio of ∼3:1,[19] presents a challenge in itself because the incorrect restraints, which are essentially random in all directions, will favor a spherical structure, and hence may not be canceled out as effectively.

Structures for IL-4 and CVN were computed with the PASD algorithm using the three-pass protocol described in Table 1 and the automatically peak-picked NOE restraint datasets summarized in Table 2. The NOE restraints were supplemented by torsion angle restraints derived from ^{1}H, ^{13}C, ^{15}N backbone chemical shifts using the program TALOS,[25] as described in the Methods section: 101 ϕ/ψ restraints for IL-4 and 57 for CVN. Incorporating torsion angle restraints in this manner is perfectly reasonable because backbone chemical shifts will always be available for any structure determination of a protein of this size determined using heteronuclear multidimensional methods. The much higher percentage of TALOS ϕ/ψ predictions for IL4 (∼80%) relative to CVN (∼55%) reflects the secondary structure makeup of the two proteins: TALOS is readily able to make good ϕ/ψ predictions for regular helical regions, reflecting the small dispersion of ϕ/ψ values in helices, but is less successful for turns, loops, and strands where the ϕ/ψ angles can populate a much wider range of values in the Ramachandran map.

Table 3. Accuracy of the IL-4 Structure Calculated Using the PASD Algorithm from Completely Automatically Peak-Picked 3D and 4D Heteronuclear-Separated NOE Spectra[a]

	atomic rms difference (Å)[b]		
	third pass (average)	NMR (1ITI)	X-ray (1RCB)
third pass (average)	0	1.52	1.62
NMR (1ITI)	2.40	0	1.36
X-ray (1RCB)	2.63	2.33	0

[a] The fraction, f_{bad}^{long}, of long-range forces (eq 11) that arise from incorrect assignments (violations >0.5 Å in the X-ray structure[33]) in the automatically peak-picked restraint dataset is 75.1% (Table 2a). Displayed are the atomic rms differences for residues 7−130 (backbone and all heavy atoms above and below the diagonal, respectively) between the restrained regularized mean structure of IL-4 calculated from the accepted structures after pass 3 with the published NMR (PDB code 1ITI, restrained regularized mean)[18d] and X-ray (PDB code 1RCB)[33] coordinates. [b] Residues 1−6 and 131−133 are disordered in solution[18] and therefore excluded from the comparison.

Table 3 summarizes the atomic rms differences between the mean structure obtained after the third and final pass calculations of the PASD algorithm on IL-4 with the previously published X-ray (1RCB)[33] and refined NMR (restrained regularized mean, 1ITI)[18d] coordinates. A superposition of the three structures is shown in Figure 2. Despite the very large fraction of bad long-range NOE restraints in the automatically peak-picked NOE restraint dataset (overall $f_{bad}^{long} = 75.1\%$), the resulting backbone coordinates are only 1.5 and 1.6 Å away from the NMR and X-ray structures, respectively, as compared to a backbone rms difference of 1.4 Å between the NMR and X-ray coordinates.

Similar quality results (Figure 3 and Table 4) are obtained for CVN where the overall value of f_{bad}^{long} is 77.5%. The backbone atomic rms difference between the mean structure obtained after the third and final pass calculations of the PASD algorithm and the NMR coordinates of the CVN monomer (which was solved using a very extensive set of experimental NMR restraints including a full complement of one- and two-bond backbone dipolar couplings and ^{1}H chemical shift refinement)[19] is only 1.1 Å for residues 3−101. The only significant divergence involves residues 1−2 as a consequence of an incorrect TALOS ϕ/ψ prediction for residue 3 (see Figure 3 and footnote to Table 4). The structure of CVN solved by X-ray crystallography is that of a domain-swapped dimer.[34] The structure of the monomeric form of CVN is essentially identical to that of the AB′ (or A′B) half of the dimer (where A and A′ comprise residues 1−50 of each subunit, and B and B′ residues

(33) Wlodawer, A.; Pavlosky, A.; Gustchina, A. *FEBS Lett.* **1992**, *309*, 59−64.
(34) Yang, F.; Bewley, C. A.; Louis, J. M.; Gustafson, K. R.; Boyd, M. R.; Gronenborn, A. M.; Clore, G. M.; Wlodawer, A. *J. Mol. Biol.* **1999**, *288*, 403−412.

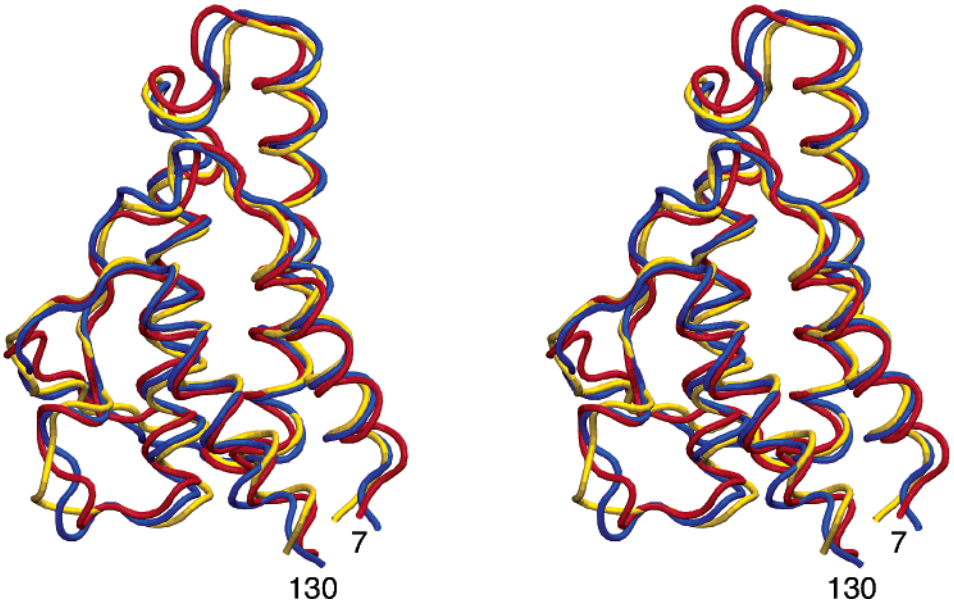

Figure 2. Performance of the PASD algorithm for IL-4 using completely automated peak-picking and analysis of experimental 3D and 4D heteronuclear-separated NOE spectra. Stereoview showing a comparison of the restrained regularized mean structure of IL-4 calculated from the accepted structures upon completion of the pass 3 calculations (red) with the published NMR (blue, PDB code 1ITI[18d]) and X-ray (gold, PDB code 1RCB[33]) coordinates. The starting fraction, f_{bad}^{long}, of long-range forces in the automatically peak-picked NOE data that arise from long-range assignments that are violated by more than 0.5 Å in the X-ray structure is 75.1%. The corresponding values for violations >1, >5, and >10 Å are 73.2%, 63.3%, and 41.9%, respectively. The backbone is represented as a tube, and only residues 7−130 are displayed.

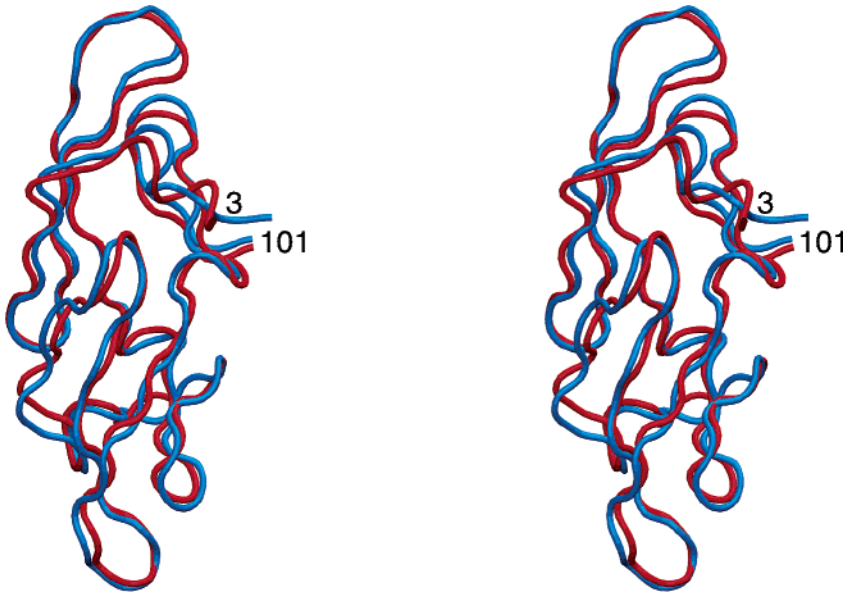

Figure 3. Performance of the PASD algorithm for CVN using completely automated peak-picking and analysis of experimental 3D heteronuclear-separated NOE spectra. Stereoview showing a comparison of the restrained regularized mean structure of CVN calculated from the accepted structures upon completion of the pass 3 calculations (red) with the published NMR (blue, PDB code 2EZM[19]) structure of monomeric CVN. The starting fraction, f_{bad}^{long}, of long-range forces in the automatically peak-picked NOE data that arise from long-range assignments that are violated in the published NMR coordinates[19] by more than 0.5 Å is 77.5%. The corresponding values for violations >1, >5, and >10 Å are 75.9%, 66.4%, and 47.1%, respectively. The backbone is represented as a tube.

52−101 of each subunit).[34] Excluding the linker region (residues 48−55) in the dimer whose conformation is obviously distinct from that in the monomeric form, the backbone rms difference between the PASD structure and the AB′ half of the X-ray structure (residues 1−47 of one subunit and 56−101 of the other) is also only 1.1 Å.

Thus, the results on IL4 and CVN clearly demonstrate that the PASD algorithm is remarkably tolerant of errors in the initial NOE restraints and can generate structures after the final pass 3 calculations that are remarkably close to highly refined, high-resolution structures. Moreover, the precision of the backbone coordinates generated by the PASD algorithm (defined as the average backbone atomic rms difference between the accepted top 10% of structures and the mean coordinates) is similar to the coordinate accuracy (defined as the backbone rms difference between the mean coordinates and the reference X-ray or NMR structure): 1.5 ± 0.5 Å versus 1.5−1.6 Å, respectively, for IL-4; and 1.4 ± 0.3 Å versus 1.1 Å, respectively, for CVN. This indicates that the PASD algorithm efficiently samples the conformational space consistent with the true structure. Thus,

Table 4. Accuracy of the CVN Structure Calculated Using the PASD Algorithm from Completely Automatically Peak-Picked 3D Heteronuclear-Separated NOE Spectra[a]

	atomic rms difference (Å)[b]		
	third pass (average)	NMR (2EZM)	X-ray (3EZM)[c]
third pass (average)	0	1.10	1.09
NMR (2EZM)	1.78	0	0.53
X-ray (3EZM)	1.78	0.99	0

[a] The fraction, f_{bad}^{long}, of long-range forces (eq 11) that arise from incorrect assignments (violations >0.5 Å in the published NMR structure[19]) in the automatically peak-picked restraint dataset is 75.1% (Table 2b). Displayed are the atomic rms differences for residues 3−101 (backbone and all heavy atoms above and below the diagonal, respectively) between the restrained regularized mean structure of CVN calculated from the accepted structures after pass 3 with the published NMR (PDB code 2EZM, restrained regularized mean)[19] coordinates for monomeric CVN and the X-ray (PDB code 3EZM)[34] coordinates of the AB′ half of the domain-swapped dimer. [b] Residues 1−2 are excluded in the comparison because these show a large divergence between the structures calculated using the PASD algorithm and both the NMR and the X-ray coordinates (cf. Figure 4). This originates from an incorrect TALOS ϕ/ψ prediction for Lys3: TALOS predicts that the ϕ/ψ angles of Lys-3 occupy the right-handed helical region of the Ramachadran map (ca. −65°/−40°), whereas in both the NMR and the X-ray coordinates, the ϕ/ψ angles of Lys-3 are actually in the unusual and much rarer left-handed helical region (ca. 70°/20°). In this regard, on average, ∼3% of "good" TALOS ϕ/ψ predictions are found to be incorrect.[25] If residues 1−2 are included, the backbone rms difference between the pass 3 structure and the NMR (and X-ray) coordinates is increased to 1.5 Å. [c] The NMR structure of CVN is that of the monomer,[19] while the X-ray structure is that of the domain swapped dimer.[34] (Note that in solution at neutral pH, the predominant form is monomeric (>90%) and is readily purified to homogeneity[19,34]). The domain-swapped dimer represents an alternative form of CVN which is slowly and irreversibly converted to the monomeric form over time.[34,38] In the domain-swapped dimer, the AB′ and A′B halves of the molecule correspond to the monomeric fold.[34] The X-ray coordinates used for the comparison therefore comprise residues 1−47 of one subunit and 56−101 of the other. (Residues 48−55 are excluded because these link the two halves of the dimer in the X-ray structure, and therefore they adopt a different conformation than in the monomer).

for an unknown structure determination, the coordinate precision obtained with the PASD algorithm is likely to provide a very good estimate of the actual coordinate accuracy.

As discussed in the sections dealing with conceptual design and implementation, a unique feature of the PASD algorithm is that the results achieved in any given iteration cycle are not biased by the global fold of structures calculated in the preceding calculational passes. In other words, the PASD algorithm is not dependent on finding a well-defined ensemble of structures after either the pass 1 or the pass 2 calculations. This is clearly illustrated in the case of both the IL-4 and the CVN calculations. For IL-4, the backbone precision and accuracy of the converged structures have values of 5.0 ± 2.5 and 2.4 Å, respectively, after pass 1, and 2.5 ± 0.8 and 1.6 Å, respectively, after pass 2. The corresponding values for CVN are 8.4 ± 1.5 and 8.5 Å, respectively, after pass 1, and 5.1 ± 1.8 and 3.5 Å, after pass 2. (Note that in a case of a de novo structure determination, the only metric available to judge structural convergence is precision.) The robustness of the PASD algorithm is also reflected by the progressive increase in the fraction of correct long-range NOE assignments with likelihoods greater than 0.9 with each successive pass: for IL-4, this fraction is 37% after pass 1, 78% after pass2, and 88% after pass 3; the corresponding values for CVN are 23%, 40%, and 83%, respectively.

Figure 4 presents an analysis of the distribution of the number of NOE restraints and the percentage of good NOE restraints as a function of the final restraints likelihood after pass 3 (defined as the largest assignment likelihood $\lambda_p(i,j)$ for a restraint i, cf. eq 6; note that in the pass 3 calculations, only one assignment is active per restraint at any time, and that in general only one assignment within that restraint will be of high likelihood). A good NOE restraint i is defined as a NOE restraint for which the assignment j with the largest likelihood, $\lambda_p(i,j)$, has a distance violation <0.5 Å in the X-ray (IL-4) or NMR (CVN) coordinates. For IL-4 and CVN, 53% and 48%, respectively, of all restraints, 14% and 22%, respectively, of the long-range restraints, and 78% and 74%, respectively, of the short-range restraints have final restraint likelihoods of 1.0; and an additional 11.5/8% (IL-4/CVN) of all restraints, 15/9.5% of the long-range restraints and 12/7% of the short-range restraints, have final restraints likelihoods $0.9 < \lambda_p(i,j) < 1.0$ (Figure 4a,b, top panels). 25/33% (IL-4/CVN) of all restraints, 57/57% of the long-range restraints, and 4/9% of the short-range restraints have final restraints likelihoods of 0, and the number of restraints with final assignment likelihoods $0.1 < \lambda_p(i,j) < 0.9$ is extremely small. The fraction of good NOE restraints decreases rapidly as the final restraints likelihood decreases (Figure 4a, b, bottom panels). The percentage of good NOE restraints for IL-4 and CVN with a final likelihood of 1.0 is 99.5% and 99.9%, respectively, overall; 99.1% and 100%, respectively, for the long-range restraints; and 99.5% and 99.9%, respectively, for the short-range restraints. The corresponding values for restraints with $0.9 < \lambda_p(i,j) < 1.0$ are 94.7/92.9% (IL-4/CVN), 94.3/94.9%, and 99.5/90.1%, respectively. Conversely, the percentage of good NOE restraints with a final likelihood of 0 is 0.4/0.2% (IL-4/CVN) overall, 0/0.3% for the long-range restraints, and 3.5/0% for the short-range restraints. Thus, the PASD algorithm is extremely efficient at identifying the correct NOE assignments with a very small false positive rate for final assignment likelihoods greater than 0.9, and a negligible false negative rate for final assignment likelihoods smaller than 0.1.

Probing the Limits of Convergence of the PASD Algorithm Using GB1 as a Model System. The data presented above on IL-4 and CVN represent the results of the PASD algorithm using a single set of experimental NOE restraints generated in a completely automated manner from multidimensional NOE spectra and chemical shift assignments. To probe the performance of the PASD algorithm for a range of different datasets and to assess the limits of its convergence power, we carried out an extensive series of model calculations on the small (56-residue) protein GB1. Twenty separate test datasets were generated, starting from the deposited experimental NOE restraints (3gb1.mr),[20b] as described in the Methods section, with the fraction f_{bad}^{long} of long-range forces arising from incorrect long-range assignments (violated by >0.5 Å in the X-ray coordinates[32]) in the starting restraints varying from 4% (corresponding to the deposited experimental restraints[20b]) up to ∼95% (Table 5). In addition, the datasets permit one to investigate the effects of adding both completely incorrect NOE restraints, which correspond to failure of the automatic peak-picking and assignment procedures, as well as incorrect NOE assignments within restraints that contain a correct NOE assignment, which correspond to the use of loose chemical shift tolerances or ambiguous NOE assignments.

The chemical shifts deposited with the GB1 NMR restraints comprised only ^{1}H and ^{15}N assignments.[20] Because ^{13}C shift

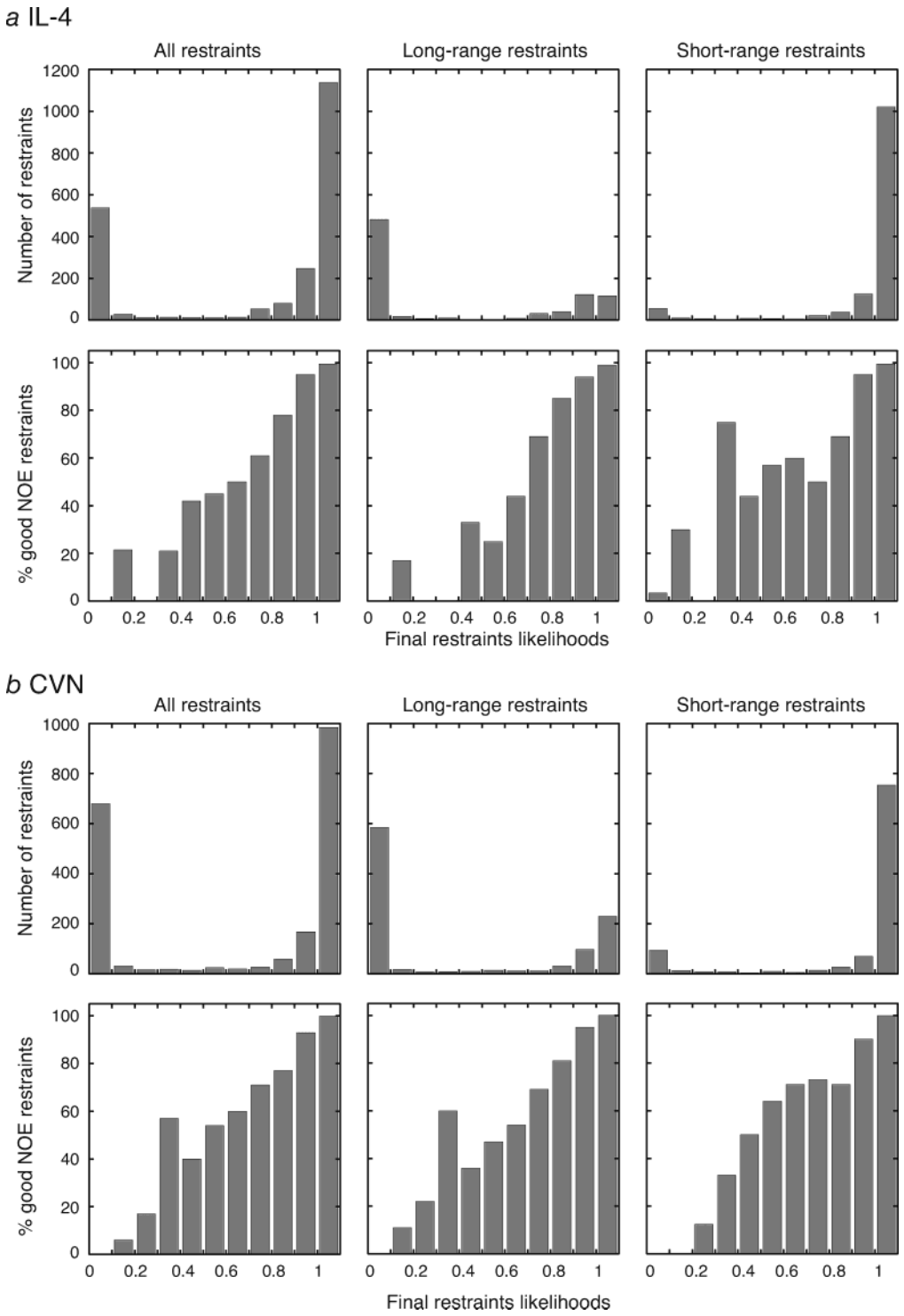

Figure 4. Distribution for (a) IL-4 and (b) CVN of the number of NOE restraints (top) and the percentage of "good" NOE restraints (bottom) as a function of the final restraints likelihood calculated following completion of the pass 3 calculations of the PASD algorithm. A "good" NOE restraint is defined as a NOE restraint for which the assignment with the largest likelihood has a distance violation <0.5 Å in the X-ray (IL-4[33]) or NMR (CVN[19]) coordinates. Long-range restraints are defined as those restraints which have no assignment with a primary sequence separation ≤5 residues; all other NOE restraints are deemed short-range.

assignments were not available, the program TALOS[25] could not be used to generate torsion angle restraints. To simulate such restraints, 23 ϕ and 23 ψ torsion angle restraints were generated from the known structure to restrain roughly half of the residues to either helical or β-sheet regions of Ramachandran space. Helical residues (residues 23−33) were restrained to $-80° < \phi < -40°$ and $-65° < \psi < -25°$, while sheet residues (positions 3−7, 14, 43−45, and 52−54) were restrained to $-170° < \phi < -50°$ and $60° < \psi < 180°$. In this instance, these restraints correspond to residues whose secondary structure class was entirely obvious from the pattern of sequential NOEs and $^3J_{HN\alpha}$ coupling constants; however, if ^{13}C shifts had been

Table 5. Test Restraint Datasets for Model Calculations on GB1[a]

no. of incorrect restraints added	no. of incorrect assignments added	no. of assignments per restraint	f_{bad}^{long} (%)[b]
0	0	1.0 ± 0	4.0
0	671	2.0 ± 1.0	58.1
0	1342	3.0 ± 1.4	75.8
0	2013	4.0 ± 1.7	83.8
225	0	1.0 ± 0	45.5
225	896	2.0 ± 1.0	72.7
225	1792	3.0 ± 1.4	84.0
225	2688	4.0 ± 1.7	88.6
350	0	1.0 ± 0	56.6
350	1021	2.0 ± 1.0	77.3
350	2042	3.0 ± 1.4	85.8
350	3063	4.0 ± 1.8	89.7
600	0	1.0 ± 0.0	68.8
600	1271	2.0 ± 1.0	83.6
600	2542	3.0 ± 1.4	88.7
600	3813	4.0 ± 1.7	92.0
1000	0	1.0 ± 0	78.4
1000	1671	2.0 ± 1.0	86.9
1000	3342	3.0 ± 1.4	91.0
1000	5013	4.0 ± 1.8	93.9

[a] Twenty separate NOE restraint datasets were generated for GB1. In each case, the experimental NOE restraints used in the determination of the refined NMR structure (PDB accession code for the coordinates and restraints, 3GB1 and 3gb1.mr, respectively)[20b] were employed as the starting point. All stereospecific NOE assignments were eliminated. A list of all NMR observable protons was assembled, and various numbers of new, incorrect restraints (i.e., which have distance violations >0.5 Å in the X-ray coordinates 1PGB[32]) were generated by randomly pairing up atom selections. The upper and lower bounds of each new restraint were set at 5.0 and 1.8 Å, respectively. New, incorrect NOE assignments were then generated in the same way and added to randomly selected restraints (among both the experimental and the random incorrect restraints). [b] f_{bad}^{long} (cf. eq 11) is the fraction of long-range forces that originate from incorrect long-range assignments (distance violations >0.5 Å in the 1.92 Å resolution crystal structure; PDB accession code 1PGB)[32] in the restraint dataset at the beginning of the calculations.

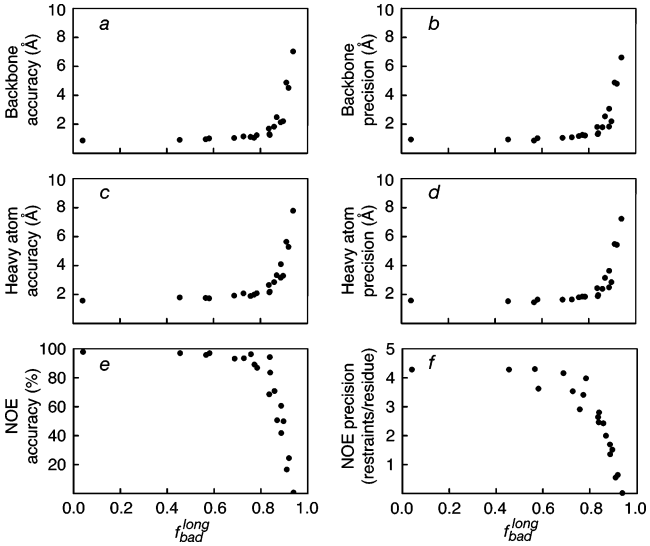

Figure 5. Performance after pass 3 of the PASD algorithm on the GB1 model system as a function of the fraction f_{bad}^{long} of long-range forces arising from incorrect long-range assignments (violations >0.5 Å in the X-ray coordinates) in the starting restraint datasets. (a) Backbone coordinate accuracy, (b) backbone coordinate precision, (c) all-heavy-atom coordinate accuracy, (d) all-heavy-atom coordinate precision, (e) long-range NOE accuracy, and (f) long-range NOE precision. Coordinate accuracy is defined by the atomic rms difference between the mean of the accepted structures calculated for each dataset and the published X-ray coordinates (PDB code 1PGB[32]). Coordinate precision is defined by the average atomic rms difference between the converged structures from each dataset (defined operationally as the top 10% of structures in terms of NOE distance violations) and the corresponding mean coordinates. Long-range NOE accuracy is defined as the fraction of correct long-range NOE restraints (i.e., satisfied in the reference structure) that have an assignment with a final likelihood $\lambda_p > 0.9$. Long-range NOE precision is defined as the number of high-likelihood ($\lambda_p > 0.9$) long-range NOE restraints per residue.

available for GB1, TALOS[25] would have yielded substantially more predictions with tighter tolerances, as assessed from results on the highly homologous GB3 protein where ^{13}C shifts are available.[35]

The performance of the PASD algorithm as a function of f_{bad}^{long} upon completion of the third pass of the calculations is shown in Figure 5. The accuracy and precision of the backbone coordinates as a function of f_{bad}^{long} are displayed in Figure 5a,b, respectively, and equivalent plots for all-heavy-atom accuracy and precision are given in Figure 5c,d, respectively. Coordinate accuracy is defined by the atomic rms difference between the mean of the accepted structures calculated for each dataset and the published X-ray coordinates (1PGB[32]). Coordinate precision is defined by the average atomic rms difference between the converged structures from each dataset and the corresponding mean coordinates. Figure 5e,f displays long-range NOE accuracy and precision, respectively, as a function of f_{bad}^{long}. Long-range NOE accuracy is defined as the fraction of correct long-range NOE restraints (distance violations <0.5 Å in the crystal coordinates[32]) that have an assignment with a final likelihood $\lambda_p > 0.9$ (cf. eq 6); long-range NOE precision is defined as the number of high-likelihood ($\lambda_p > 0.9$) long-range NOE restraints per residue.

The data in Figure 5 reveal the following findings. First, there is essentially no degradation in performance of the PASD algorithm, either in terms of coordinate accuracy (Figure 5a) or NOE assignment accuracy (Figure 5e) up to f_{bad}^{long} values of ~80%. Thus, the backbone coordinate accuracy only decreases from 0.8 Å for $f_{bad}^{long} = 4\%$ to 1.3 Å for $f_{bad}^{long} \approx 80\%$. Likewise, the long-range NOE accuracy only falls from ~98% to ~86% over the same range of f_{bad}^{long}. It is only for values of f_{bad}^{long} beyond 80% that rapid degradation in performance occurs. This is a critical finding because even relatively simplistic automated NOE cross-peak assignment algorithms such as STAPP can generate an initial NOE restraint dataset with $f_{bad}^{long} \leq 80\%$, even in a case such as IL-4 which exhibits extensive chemical shift degeneracy. Second, taking into account the distribution of incorrect restraints and assignments in the various test datasets shown in Table 5, it is evident that bad information arising from added incorrect assignments is equivalent to bad information arising from added incorrect restraints. In other words, the impact on the PASD calculations of ambiguous assignments within a restraint containing a correct assignment is essentially the same as the presence of completely incorrect restraints (i.e., restraints which have no correct assignment). Third, coordinate precision, for both backbone (Figure 5b) and all heavy atoms (Figure 5d), closely mirrors coordinate accuracy (Figure 5a,c, respectively). This is important because it indicates that coordinate precision attained using the PASD algorithm provides a good reflection of coordinate accuracy.

The improvement in long-range NOE accuracy following passes 1, 2, and 3 of the calculations is displayed in Figure 6 as a function of f_{bad}^{long}. As expected, after each pass of the PASD

(35) Chou, J. J.; Case, D. A.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 8959−8966.

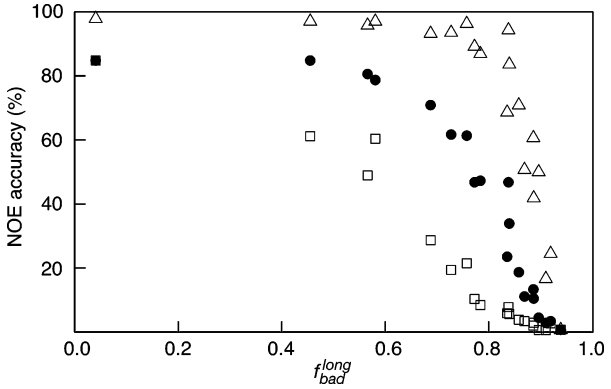

Figure 6. Convergence as a function of iteration cycle of the PASD algorithm for the GB1 model system. The NOE accuracy (defined as the fraction of correct long-range NOE restraints that have an assignment with a final likelihood $\lambda_p > 0.9$), determined from the accepted structures calculated for each dataset, is plotted as a function of the fraction f_{bad}^{long} of long-range forces arising from incorrect long-range assignments in each dataset (i.e., violations >0.5 Å in the X-ray coordinates[32]). The results after the first, second, and third passes are shown as □, ●, and △, respectively.

algorithm, there is a substantial improvement in long-range NOE accuracy. Thus, the value of f_{bad}^{long} at which a long-range NOE accuracy of 50% is achieved is increased from ~0.6 after pass 1, to ~0.8 after pass 2, and finally to ~0.9 after pass 3. Another way of viewing the data is to look at the increase in long-range NOE accuracy achieved for $f_{bad}^{long} = 0.8$ from ~8% after pass 1 to ~50% after pass 2 and finally to ~85% after pass 3.

A direct structural comparison between the restrained regularized mean structure calculated after pass 3 from a restraint dataset with $f_{bad}^{long} \approx 83.6\%$ and the X-ray[32] and refined NMR[20b] coordinates is shown in Figure 7. The fold has been clearly determined using the PASD algorithm, and the overall backbone rms differences to the X-ray and NMR coordinates are only 1.7 and 1.8 Å, respectively. The main areas of discrepancy are restricted to turns and loops, and, in addition, the second strand of the four-stranded β-sheet is a little distorted. In the context of a structure determination from real experimental NOE data, it is evident that a structure of this quality, together with the derived high likelihood NOE assignments, can easily be used as the basis for further iterative refinement.

Concluding Remarks

The results from both the experimental data on IL-4 and CVN and the model calculations on GB1 indicate that the PASD algorithm provides a robust method for automated NMR structure determination that is highly tolerant of errors in the initial NOE restraint dataset and can readily generate reasonably accurate structures even when the NOE restraint datasets contain up to 80% incorrect long-range information. We anticipate that the PASD algorithm will play a major role in high-throughput determination of unrefined protein NMR structures.

The key to any successful algorithm lies in the simple observation that the set of correct restraints are correlated and generate forces that act in concert. In contrast, incorrect restraints are usually uncorrelated, and therefore their associated forces tend to cancel one another out. To this end, the PASD algorithm makes use of two complementary approaches. First, the linear NOE potential that is active in the first two passes of the algorithm eliminates the effects of large distance violations, thereby permitting reasonably efficient cancellation of uncorrelated forces. The hierarchical implementation of the NOE data, in which a single NOE cross-peak corresponds to a unique restraint that can comprise multiple possible assignments, each of which is treated independently, ensures that some forces from the correct assignment(s), within a particular restraint, are always applied and can thereby cooperatively reinforce forces arising from other correct assignments. Second, probabilistic inactivation/reactivation of assignments, applied with increasing stringency during the course of simulated annealing, simplifies the energy hypersurface, thereby facilitating the search for the global minimum region. In this regard, it is important to stress that the PASD algorithm never removes NOE assignments or restraints permanently and, consequently, is unlikely to be trapped in a false minimum in NOE assignment space. By this

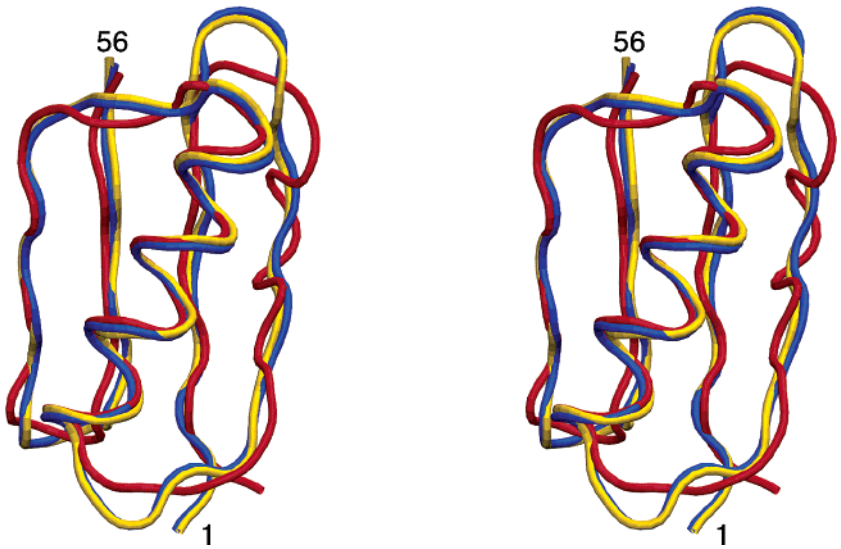

Figure 7. Example of the performance of the PASD algorithm for the GB1 model system. The restrained regularized mean coordinates (red) of the accepted structures after pass 3, calculated from a dataset with $f_{bad}^{long} = 83.6\%$ (>0.5 Å violations in the X-ray coordinates), are superimposed on the published NMR (blue, PDB code 3GB1[20b]) and X-ray (gold, PDB code 1PGB[32]) coordinates. The fractions of long-range forces arising from long-range assignments that are violated by >1, >5, and >10 Å in the X-ray coordinates are 82.3%, 68.6%, and 41.1%. The backbone atomic rms difference from the published NMR and X-ray coordinates is 1.8 and 1.7 Å, respectively; for comparison, the backbone atomic rms difference between the NMR and X-ray coordinates is 0.5 Å. The backbone coordinate precision obtained after passes 1, 2, and 3 is 6.3 ± 1.0, 3.9 ± 0.9, and 1.8 ± 0.5 Å, respectively. The backbones are displayed as tubes.

means, the PASD algorithm can take optimal advantage of the information available in highly ambiguous restraints (that is, restraints which have many possible assignments).

It is of interest to briefly compare the robustness of the PASD algorithm with published data on the CANDID algorithm.[12,13,16] A successful structure calculation using CANDID requires the fulfillment of two criteria:[13,16] (a) fewer than 20−25% of the long-range NOEs should have been discarded at the end of the calculation, and (b) the coordinate precision after the first cycle should not exceed 3 Å. By way of contrast, the results presented here indicate that the PASD algorithm can readily handle up to 80% bad long-range NOE information and the coordinate precision after the first pass is not critical (8.4 ± 1.5 Å for CVN, 6.3 ± 1.0 Å for GB1 using the data set with $f_{\mathrm{bad}}^{\mathrm{long}} =$ 83.6%, and 5.0 ± 2.5 Å for IL-4). The dependence of the CANDID algorithm on the completeness of the chemical shift assignment table has been investigated using model calculations.[16] The performance of CANDID appears to break down rapidly when the missing chemical shift assignments exceed ∼10%.[16] Elimination of 10% of the chemical shift assignments implies that no correct assignment is possible for ∼19% ($1.0-0.9^2$) of the NOE cross-peaks, which therefore yield incorrect restraints.[36] This is in agreement with the first CANDID criterion listed above. The principal mechanism used by CANDID to eliminate the untoward effect of incorrect NOE assignments involves restraint combination.[12] In this procedure NOE restraints are paired up to yield a single restraint which is satisfied whenever either of its constituent restraints is satisfied. Thus, a set of 1000 NOE assignments, 20% of which are incorrect, will yield a set of 500 paired restraints, only 4% of which would be incorrect. If, however, 80% of the 1000 NOE assignments are incorrect, 64% of the 500 paired restraints will also be incorrect. Both CANDID and PASD identify NOE assignments to be inactivated on the basis of their distance violation. The mechanism of restraint combination employed by CANDID, however, is inherently unable to handle large fractions of incorrect NOE assignments. Moreover, restraint combination is intrinsically less flexible than the mechanism employed by the PASD algorithm. This is because the distance violation at which an NOE assignment will be inactivated using restraint combination depends entirely on its combined-restraint partner rather than on a tunable parameter such as the characteristic distance D_v (cf. eq 4) whose stringency can be progressively increased during the course of the calculation (cf. Table 1). As the fraction of incorrect long-range NOE assignments increases, the precision (and accuracy) of the ensemble of structures calculated after the first CANDID cycle will necessarily decrease. Because the outcome of each successive cycle of the CANDID algorithm is dependent on the Cartesian coordinates of the structures calculated in the preceding cycle, it is not surprising that the likelihood of the CANDID algorithm funnelling down an incorrect refinement pathway increases dramatically as the fraction of incorrect restraints exceeds a relatively low threshold (20−25%).

The final set of converged structures obtained using the PASD algorithm with NOE restraints derived by completely automated analysis of NOE spectra do not represent fully refined NMR structures. Further refinement entails improvements in the quality of the NOE restraints list and incorporation of additional experimental NMR restraints (e.g., side-chain torsion angle restraints from coupling constant measurements and analysis of ROE and short mixing time NOE data, chemical shift restraints, dipolar coupling restraints, etc.). Thus, the set of converged structures generated at the end of the pass 3 calculations serve three purposes: first, as starting coordinates for further refinement calculations; second, as a distance filter to repick the NOE spectra automatically with larger chemical shift error ranges, thereby circumventing many of the problems associated with generating the original NOE restraints list; and, third, as a basis for the identification of potentially incorrect high-likelihood NOE assignments within regions of low proton density (e.g., exposed loops and turns) using concepts from information theory[37] because an incorrect NOE assignment in such regions would be expected to result in an undue increase in local coordinate precision.

Acknowledgment. We thank Ad Bax and Dennis Torchia for useful discussions. This work was supported in part by the AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health (G.M.C.). Support from NIDDK, NCI, NHLBI, and NIDCR is acknowledged.

JA049786H

(36) A missing resonance assignment can have two possible consequences with regard to the generation of NOE assignments. If a particular NOE cross-peak cannot be assigned to any pairwise interaction due to a missing resonance assignment, the effect is entirely neutral because no incorrect information is incorporated into the NOE restraints list. If, on the other hand, a given NOE cross-peak has multiple possible assignments, because of chemical shift degeneracy, and the resonance assignment for one of the partners for the correct NOE assignment is absent, the resulting NOE restraint (which may comprise multiple possible NOE assignments) will be incorrect.

(37) Nabuurs, S. B.; Spronk, C. A. E. M.; Krieger, H.; Maasen, G. H.; Vriend, G.; Vuister, G. W. *J. Am. Chem. Soc.* **2003**, *125*, 12026−12034.
(38) (a) Bewley, C. A.; Clore, G. M. *J. Am. Chem. Soc.* **2000**, *122*, 6009−6016. (b) Barrientos, L. G.; Louis, J. M.; Botos, I.; Mori, T.; Han, Z.; O'Keefe, B. R.; Boyd, M. R.; Wlodawer, A.; Gronenborn, A. M. *Structure* **2002**, *10*, 673−686.