*J. Mol. Biol.* (1991) **217**, 517–530

# Efficient Computation of Three-dimensional Protein Structures in Solution from Nuclear Magnetic Resonance Data Using the Program DIANA and the Supporting Programs CALIBA, HABAS and GLOMSA

**Peter Güntert, Werner Braun and Kurt Wüthrich**

*Institut für Molekularbiologie und Biophysik*
*Eidgenössische Technische Hochschule-Hönggerberg*
*CH-8093 Zürich, Switzerland*

*(Received 22 June 1990; accepted 2 October 1990)*

A novel procedure for efficient computation of three-dimensional protein structures from nuclear magnetic resonance (n.m.r.) data in solution is described, which is based on using the program DIANA in combination with the supporting programs CALIBA, HABAS and GLOMSA. The first part of this paper describes the new programs DIANA, CALIBA and GLOMSA. DIANA is a new, fully vectorized implementation of the variable target function algorithm for the computation of protein structures from n.m.r. data. Its main advantages, when compared to previously available programs using the variable target function algorithm, are a significant reduction of the computation time, and a novel treatment of experimental distance constraints involving diastereotopic groups of hydrogen atoms that were not individually assigned. CALIBA converts the measured nuclear Overhauser effects into upper distance limits and thus prepares the input for the previously described program HABAS and for DIANA. GLOMSA is used for obtaining individual assignments for pairs of diastereotopic substituents by comparison of the experimental constraints with preliminary results of the structure calculations. With its general outlay, the presently used combination of the four programs is particularly user-friendly. In the second part of the paper, initial results are presented on the influence of the novel DIANA treatment of diastereotopic protons on the quality of the structures obtained, and a systematic study of the central processing unit times needed for the same protein structure calculation on a range of different, commonly available computers is described.

## 1. Introduction

The early stages in the development of the presently widely used n.m.r.† method for the determination of three-dimensional biomacromolecular structures in solution (for a review, see Wüthrich, 1989) made it clear that the key data measured by n.m.r. would consist of a network of distance constraints between spatially proximate hydrogen atoms (Gordon & Wüthrich, 1978; Dubs *et al.*, 1979; Keller & Wüthrich, 1980; Wüthrich *et al.*, 1982). It immediately followed that the techniques for structure determination from other experimental data, in particular X-ray diffractions, could not be adapted for the structural analysis of the n.m.r. data, and hence new ways had to be developed. Initially, algorithms were used that combined metric matrix distance geometry (Blumenthal, 1970), which had been applied by the groups of Crippen and Kuntz for systematic studies on protein structures (Crippen, 1977; Kuntz *et al.*, 1976; Havel *et al.*, 1983), with a detailed description of the interplay of constraints imposed by the covalent polypeptide structure and those from the n.m.r. measurements (Braun *et al.*, 1981, 1983; Havel & Wüthrich, 1984, 1985). Subsequent work included a variable target function algorithm (Braun & Gō, 1985), interactive molecular modeling using computer graphics (Billeter *et al.*, 1985), restrained molecular dynamics calculations either applied directly with the n.m.r. data (Brünger *et al.*, 1986) or in conjunction with

† Abbreviations used: n.m.r., nuclear magnetic resonance; BPTI, basic pancreatic trypsin inhibitor; NOE, nuclear Overhauser enhancement; NOESY, 2-dimensional nuclear Overhauser enhancement spectroscopy; c.p.u., central processing unit; MFLOPS, million floating point operations per second; r.m.s.d., root-mean-square deviation.



0022-2836/91/030517-14 $03.00/0

model building (Kaptein *et al.*, 1985) or distance geometry calculations (Clore *et al.*, 1985), and an ellipsoid algorithm (Billeter *et al.*, 1987). Inspection of the recent literature shows that the following procedures are currently mostly employed. (1) Embedding using a metric matrix distance geometry program, e.g. DISGEO (Havel & Wüthrich, 1984) or DSPACE (Hare Research, Woodinville, WA 98072, U.S.A.), followed by simulated annealing using molecular dynamics (Driscoll *et al.*, 1989; Lee *et al.*, 1989). (2) Structure determination using a variable target function algorithm, e.g. DISMAN (Braun & Gō, 1985), which directly generates acceptable structures (Kline *et al.*, 1988; Schultze *et al.*, 1988; Zuiderweg *et al.*, 1989) or can be supplemented by a molecular mechanics energy minimization (e.g. Billeter *et al.*, 1990; Qian *et al.*, 1989; Widmer *et al.*, 1989).

Once it had been established that n.m.r. measurements could provide sufficient data for the determination of globular protein structures at atomic resolution (Havel & Wüthrich, 1984, 1985; Williamson *et al.*, 1985), the main interest shifted to the development of procedures ensuring both high efficiency of structure calculations and minimal bias of the results by the algorithms used. This paper presents a further step in this development by describing a new implementation of the variable target function algorithm of Braun & Gō (1985) in the program DIANA. This program was primarily designed to be efficient with respect to the c.p.u. time used and, in combination with the supporting programs CALIBA, HABAS and GLOMSA, to be user-friendly in routine structure determinations (Güntert *et al.*, 1990, accompanying paper). Furthermore, thanks to the high efficiency of the fully vectorized program, systematic large-scale investigations on the course of variable target function calculations could be started with DIANA (P. Güntert, W. Braun & K. Wüthrich, unpublished results), which should provide a basis for further optimization of structure calculations.

With regard to improving the quality of structure determinations by n.m.r., the treatment of distance constraints with diastereotopic groups of protons (Wüthrich *et al.*, 1983) can be of crucial importance (Güntert *et al.*, 1989). Recent work in this regard focused mainly on establishing individual assignments for diastereotopic pairs of protons (Weber *et al.*, 1988; Neri *et al.*, 1989; Nilges *et al.*, 1990). The program DIANA includes a novel treatment of constraints with prochiral centers for which the diastereotopic ligands were not individually assigned. In calculations with BPTI, the results of this new treatment are compared with corresponding results obtained with the original pseudo-atom concept (Wüthrich *et al.*, 1983).

## 2. Methods

This section describes the 3 programs CALIBA ("*calibra*tion of NOE intensity *versus* distance constraints"), DIANA ("*di*stance geometry *a*lgorithm for *n*.m.r. *a*pplications") and GLOMSA ("*glo*bal *m*ethod for obtaining *s*tereospecific *a*ssignments"), which were all written in Fortran-77. Fig. 1 affords a survey of the functions of these programs, which are used in the order CALIBA, DIANA and then GLOMSA. The program CALIBA accepts the experimental NOE intensities as input and performs the calibration of NOESY cross-peaks, i.e. the conversion from peak volumes to upper distance limits. It thus prepares the principal input for the program HABAS (Güntert *et al.*, 1989), which in turn adds stereospecific assignments to the input for DIANA. DIANA is used for efficient calculation of protein conformations based on distance and dihedral angle constraints that can be obtained by n.m.r. measurements (Wüthrich, 1986). It is a new, improved and vectorized implementation of the variable target function algorithm that has first been used in the program DISMAN (Braun & Gō, 1985) and subsequently in other programs (Vásquez & Scheraga, 1988; Kohda *et al.*, 1988). In addition to the structure calculations, DIANA screens the experimental distance constraints and eliminates irrelevant constraints from the input, and it applies a novel adjustment routine to distance constraints with pairs of diastereotopic protons for which no individual assignments are available. The program GLOMSA accepts as input the structures calculated with DIANA and the conformational constraints list produced by CALIBA. It is used to obtain additional stereospecific assignments based on the comparison of upper distance limit pairs or relational constraints (Güntert *et al.*, 1989) involving the diastereotopic substituents of prochiral centers with a set of preliminary conformers (Kline *et al.*, 1988). The additional stereospecific assignments thus obtained are included in the input for a new DIANA calculation, which in turn

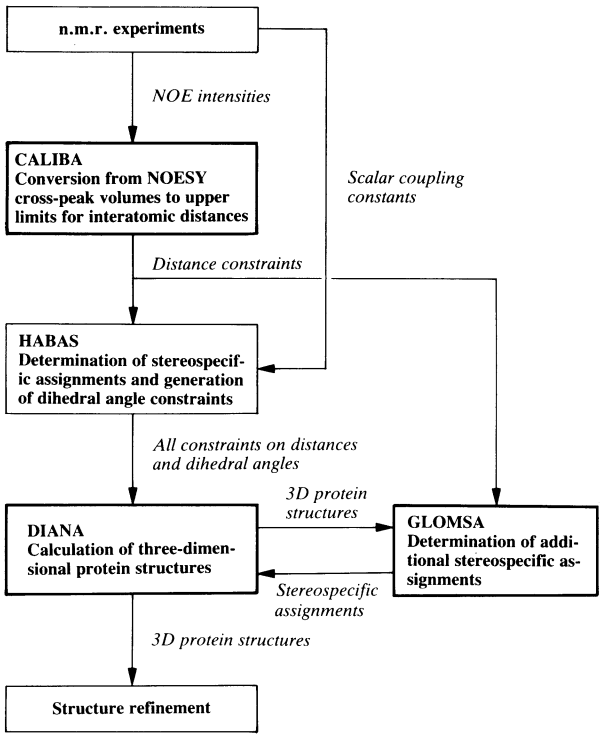



**Figure 1.** Schematic representation of the functions of the programs CALIBA, DIANA and GLOMSA and the input and output of these programs. See the text for details.

produces the structures used for further refinements, e.g. by energy minimization (Brooks *et al.*, 1983; Brünger *et al.*, 1986; Schaumann *et al.*, 1990; Weiner & Kollman, 1981).

(a) *The program CALIBA*

In the present version of the program CALIBA, the calibration of NOE intensity *versus* the corresponding upper distance bound is based on either of 2 model assumptions. The 1st assumes that the NOESY cross-peak volume, $V$, is inversely proportional to the power $n$ of the corresponding upper distance bound $b$:

$$V = \frac{C}{b^n}, \tag{1}$$

where $C$ is a constant, and the values of $n$ are typically in the range from 4 to 6. Clearly the value $n = 6$ is an upper limit for $n$ obtained theoretically by assuming a rigid structure. Exponents $n < 6$ were found empirically to afford improved representations of the relation between cross-peak volumes and distances for peaks that involve peripheral side-chain protons. The 2nd possibility uses the uniform averaging model (Braun *et al.*, 1981):

$$V = \frac{C'}{b - d_0}\left(\frac{1}{d_0^5} - \frac{1}{b^5}\right). \tag{2}$$

Here, $C'$ is a constant, and $d_0 = 1{\cdot}9$ Å is the shortest sterically allowed distance between 2 protons (1 Å = 0·1 nm).

The input for CALIBA consists of cross-peak volumes measured in 1 or several NOESY spectra, for example, with the program EASY (Eccles *et al.*, 1989). The peak volumes from different spectra, which may have been recorded with different experimental conditions, can be multiplied with user-specified weighting factors. The volumes of peaks that correspond to 2 or more protons with degenerate chemical shifts are divided by the number of protons they contain. If more than 1 peak intensity corresponding to the same distance is retained from the analysis of the spectra, only the strongest (after the aforementioned weighting) is considered for the conversion. For an optimal empirical calibration, the curves (1) or (2) and the constants $C$ or $C'$ can be chosen independently for different classes of interatomic distances, e.g. intraresidual, sequential, medium-range and long-range backbone, and long-range constraints (Wüthrich, 1986), and for cross-peaks that do or do not involve methyl groups. To obtain reasonable upper bounds also for very strong or very weak cross-peaks, the values for the upper bounds $b$ are restricted to a limited range $b^{\min} \leq b \leq b^{\max}$, with typical values for $b^{\min}$ of 2·4 Å, and $b^{\max}$ of 5·0 Å. Constraints corresponding to cross-peaks relating resonance lines from multiple protons with degenerate chemical shifts are referred to pseudoatoms, and the appropriate pseudoatom corrections (Wüthrich *et al.*, 1983) are automatically applied. The program produces upper distance limit files that can be read directly by the programs HABAS (Güntert *et al.*, 1989), DIANA and GLOMSA.

To optimize the choice of the calibration curves at different points during a structure calculation where one already has a set of preliminary structures for the protein under investigation, CALIBA has the option to produce a doubly logarithmic plot of peak volumes *versus* the average of the corresponding distances in this set of structures. Using this plot, the calibration curves can then be adjusted such that most of the resulting upper distance limits are fulfilled in the given preliminary structures without being unnecessarily loosened.

(b) *The program DIANA*

The algorithm used by the program DIANA is based on the minimization of a variable target function $T(\phi_1, \ldots, \phi_n)$, where the $n$ degrees of freedom are the dihedral angles $\phi_1, \ldots, \phi_n$ about single (rotatable) bonds of the polypeptide chain. During the calculation the bond lengths, bond angles and chiralities of the covalent structure are kept fixed at the ECEPP standard values (Momany *et al.*, 1975). The target function $T$, with $T \geq 0$, (for an explicit definition see eqn (6) below) is defined such that $T = 0$ if all experimental distance and dihedral angle constraints are fulfilled and all non-bonded atom pairs satisfy a check for the absence of steric overlap. $T(\phi_1, \ldots, \phi_n) \leq T(\theta_1, \ldots, \theta_n)$ if the conformation $(\phi_1, \ldots, \phi_n)$ satisfies the constraints better than the conformation $(\theta_1, \ldots, \theta_n)$. The problem to be solved is to find the values $(\phi_1, \ldots, \phi_n)$ that yield low values of the target function. To reduce the danger of becoming trapped in a local minimum with a function value much higher than the global minimum, the target function is varied during a structure calculation. At the outset, only local constraints with respect to the polypeptide sequence are considered, and in subsequent rounds of calculations, constraints between atoms further apart with respect to the primary structure are included in a stepwise fashion. Consequently, in the 1st stages of a structure calculation, the local features of the conformation will be established, and the global fold of the protein will be obtained only toward the end of the calculation. Similar strategies of avoiding local minima by variation of the pseudoenergy function during a structure calculation have been used with restrained molecular dynamics techniques (Holak *et al.*, 1987; Nilges *et al.*, 1988).

(i) *The variable target function*

Two different kinds of constraints are considered by the target function, i.e. upper and lower bounds on interatomic distances, and restraints on individual dihedral angles in the form of an allowed interval (Wüthrich, 1986; Braun, 1987). An upper or lower limit, $b$, on the distance between the 2 atoms $\alpha$ and $\beta$ is denoted by the triple $(\alpha,\beta,b)$ or, if there is no danger of ambiguity, simply by $b$. A direct constraint on the dihedral angle $a$ that restricts its value $\phi_a$ to an allowed interval $[\phi^{\min}, \phi^{\max}]$, with $\phi^{\min} < \phi^{\max} < \phi^{\min} + 2\pi$, is denoted by $(a, \phi^{\min}, \phi^{\max})$. In the definition of the variable target function we use further the half-width, $\Gamma$, of the forbidden interval of dihedral angle values:

$$\Gamma = \pi - \frac{\phi^{\max} - \phi^{\min}}{2}, \tag{3}$$

and the signed dihedral angle constraint violation

$$\Delta = \begin{cases} 0, & \text{if } \phi_a \in [\phi^{\min}, \phi^{\max}]; \\ -\Delta^{\min}, & \text{if } \phi_a \notin [\phi^{\min}, \phi^{\max}] \text{ and } \Delta^{\min} \leq \Delta^{\max}; \\ \Delta^{\max}, & \text{if } \phi_a \notin [\phi^{\min}, \phi^{\max}] \text{ and } \Delta^{\min} > \Delta^{\max}; \end{cases} \tag{4}$$

with

$$\Delta^{\min} = \min\{|\hat{\phi}^{\min} - \hat{\phi}_a|, 2\pi - |\hat{\phi}^{\min} - \hat{\phi}_a|\},$$

and

$$\Delta^{\max} = \min\{|\hat{\phi}^{\max} - \hat{\phi}_a|, 2\pi - |\hat{\phi}^{\max} - \hat{\phi}_a|\}.$$

$\hat{\phi}$ denotes the equivalent value of $\phi$ in the interval $[0, 2\pi[$, which can be obtained in all instances by the addition of an integer multiple of $2\pi$ to $\phi$. The sign of $\Delta$ will be important only in the calculation of the gradient of the

target function; it is positive if a small increase of $\phi_a$ also increases the violation of the angle constraints, and negative otherwise.

To formulate the target function, we assume that there are $n_u$ experimental upper limits, $n_l$ experimental lower limits, and $n_v$ van der Waals' repulsion lower limits on interatomic distances, and $n_a$ direct dihedral angle constraints:

$$\begin{array}{ll} (\alpha_i^u, \beta_i^u, b_i^u), & i = 1, \ldots, n_u; \\ (\alpha_i^l, \beta_i^l, b_i^l), & i = 1, \ldots, n_l; \\ (\alpha_i^v, \beta_i^v, b_i^v), & i = 1, \ldots, n_v; \\ (a_i, \phi_i^{\min}, \phi_i^{\max}), & i = 1, \ldots, n_a. \end{array} \quad (5)$$

The target function, $T$, then is:

$$T = \sum_{c=u,l,v} w_c \sum_{i \in I_c} \left( \Theta_c \left( \frac{d_i^{c2} - b_i^{c2}}{2 b_i^c} \right) \right)^2 + w_a \sum_{i=1}^{n_a} \left( 1 - \frac{1}{2} \left( \frac{\Delta_i}{\Gamma_i} \right)^2 \right) \Delta_i^2, \quad (6)$$

with:

$$\Theta_c(t) = \begin{cases} \max(0,t), & \text{if } c = u; \\ \min(0,t), & \text{if } c = l,v; \end{cases}$$

Here, $d_i^c$ denotes the distance between the 2 atoms $\alpha_i^c$ and $\beta_i^c$, $w_c \geq 0$ are weighting factors for the 4 types of constraints ($c = u,l,v,a$), and $I_c \subseteq \{1, \ldots, n_c\}$ with $c = u,l,v$ are the subsets of distance constraints included in the target function. In the present version of the program DIANA the subsets $I_c$ cannot be chosen arbitrarily but consist of all distance constraints between the atoms $\alpha$ and $\beta$ in those residues between which the sequence numbers, $R_\alpha$ and $R_\beta$, respectively, differ by not more than a given minimization level $L_c$:

$$I_c = \left\{ i \in \{1, \ldots, n_c\} \middle| |R_{\alpha_i^c} - R_{\beta_i^c}| \leq L_c \right\}, \quad c = u,l,v. \quad (7)$$

It is usual to choose the same minimization level for all 3 kinds of distance constraints; in this case, we denote the common minimization level simply by $L$.

In general, a complete structure calculation with the program DIANA includes several minimization steps (not to be confused with individual iterations of the conjugate gradient minimizer), i.e. the minimization of several forms of the variable target function that differ in the minimization levels $L_c$, and in the weighting factors $w_c$ (see the Appendix). An optimal strategy for selecting the minimization steps is not known, but we found it essential (1) to increase the minimization level gradually in a stepwise fashion, starting with $L_c = 0$ or 1, and (2) to use a weighting factor $w_v$ for steric constraints that is small with respect to the weighting factors $w_u$ and $w_l$ for experimental distance constraints, e.g. $w_v = 0{\cdot}2 w_u$, except toward the end of the calculation, where one usually increases $w_v$ to 2 to 3 times $w_u$ in order to minimize steric overlaps.

The target function of eqn (6) is continuously differentiable over the entire conformation space, and is chosen such that the contribution of a single small violation $\delta_c$ is given by $w_c \delta_c^2$ for all types of constraints ($c = u,l,v,a$). Because only squared interatomic distances and no square-roots have to be computed, the target function can be calculated rapidly.

(ii) *Comparison of the variable target functions used in DIANA and DISMAN*

In the notation used here, the target function in the program DISMAN, $T'$, which corresponds to the target function of eqn (6) used in DIANA, is defined by (Braun & Gō, 1985; Braun, 1987);

$$T' = \sum_{c=u,l} w_c \sum_{i \in I_c} \left( \Theta_c \left( \frac{d_i^{c2} - b_i^{c2}}{2 b_i^c} \right) \right)^2 + \frac{w_v}{4} \sum_{i \in I_v} (\Theta_v (d_i^{v2} - b_i^{v2}))^2 + 4 w_a \sum_{i=1}^{n_a} \left( 1 - \frac{1}{2} \frac{|\Delta_i|}{\Gamma_i} \right)^2 \left( \frac{\Delta_i}{\Gamma_i} \right)^2. \quad (8)$$

The treatment of experimental upper and lower distance constraints is the same in the 2 programs. Steric constraints and dihedral angle constraints, however, are treated somewhat differently; because other normalization factors are used in DISMAN, the contribution of a small violation, $\delta$, of a steric or dihedral angle constraint is not simply equal to the weighting factor multiplied with the squared violation. Rather, this contribution is $w_v b_i^{v2} \delta^2$ for a steric constraint and $4 w_a (\delta/\Gamma_i)^2$ for a dihedral angle constraint. Furthermore, for a dihedral angle constraint violation, the maximal contribution to the DISMAN target function equals $w_a$ and is independent of the width of the forbidden dihedral angle range, whereas in DIANA it equals $(w_a/2)\Gamma_i^2$ and is proportional to the squared width of the forbidden region.

(iii) *The input and output formats*

There are several different input files and some interactively entered parameters. The nomenclature in the standard residue library follows the IUPAC rules (IUPAC-IUB Commission on Biochemical Nomenclature, 1970), and the covalent structure is that of the ECEPP force field (Momany *et al.*, 1975; Némethy *et al.*, 1983) for the 20 proteinogenic amino acid residues. The primary structure is entered in the amino acid sequence file, which also identifies *cis*-peptide bonds. Pairs of diastereotopic substituents for which individual assignments are available are identified in the stereospecific assignments input file. If this file is missing, one has to provide the information that stereospecific assignments are available either for all or for none of the prochiral centers. Upper distance limits, and lower distance limits and dihedral angle constraints are read from input files. The minimization parameters input file contains details about the minimization procedure. The start conformations for the structure calculations can be generated by the program, or read from input files.

The results output file records the interactive input, includes information on the course of the minimization, and lists the constraint violations exceeding given threshold values. The overview output file includes a complete list of the numbers, the sums, and the maximal values of the residual constraint violations for each calculated structure. Furthermore, a table of the important violations in all structures with final target function values less than a user-defined cutoff is written whenever a structure calculation is finished. For the DIANA user, it is important that the overview file can be inspected during the operation of the program, whereas the results file can usually be examined only after completion of the current job. The dihedral angles and Cartesian co-ordinates files of the calculated structures can be written either at intermediate stages or at the end of the minimization. The r.m.s.d. values file includes pairwise global or local r.m.s.d. values between calculated structures, and the modified upper distance limits and modified lower distance limits files list the experimental constraints after processing by DIANA.

(iv) *Identification of irrelevant constraints and too restrictive constraints*

A distance limit is irrelevant if (1) the corresponding interatomic distance is independent of the conformation, (2) there exists no conformation $(\phi_1, \ldots, \phi_n)$ that violates the given limit or (3) a lower distance limit is smaller than the steric limit automatically imposed by DIANA. Conditions (1) and (3) are easy to check, whereas a complete check of condition (2) is difficult. Therefore, this condition is checked only for all constraints on distances that depend on a single dihedral angle, and for some constraints relating to 2 dihedral angles. If the distance between 2 atoms $\alpha$ and $\beta$, $|\mathbf{r}_\alpha - \mathbf{r}_\beta|$, depends on 1 dihedral angle $a$, the range of its values is given by:

$$A - B \leq |\mathbf{r}_\alpha - \mathbf{r}_\beta|^2 \leq A + B, \tag{9}$$

where:

$$A = |\mathbf{d}_\alpha|^2 + |\mathbf{d}_\beta|^2 - 2(\mathbf{e}_a \cdot \mathbf{d}_\alpha)(\mathbf{e}_a \cdot \mathbf{d}_\beta)$$

$$B = 2\sqrt{[\mathbf{d}_\alpha^2 - (\mathbf{e}_a \cdot \mathbf{d}_\alpha)^2][\mathbf{d}_\beta^2 - (\mathbf{e}_a \cdot \mathbf{d}_\beta)^2]}$$

with

$$\mathbf{d}_\alpha = \mathbf{r}_\alpha - \mathbf{r}_a \quad \text{and} \quad \mathbf{d}_\beta = \mathbf{r}_\beta - \mathbf{r}_a.$$

$\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ denote the position vectors of the atoms $\alpha$ and $\beta$ for an arbitrary conformation, $\mathbf{r}_a$ is the position vector of the start point of the rotatable bond $a$, and $\mathbf{e}_a$ is a unit vector along the rotatable bond $a$. In the notation of eqn (9), an upper distance limit $(\alpha,\beta,b)$ is irrelevant if $b \geq A + B$, and too restrictive if $b < A - B$. Irrelevant constraints are removed from the input used for the calculation. Too restrictive distance constraints that cannot be fulfilled by any conformation will thus be identified in the results file, but they will not be removed from the input used for the calculation. For distances depending on 2 dihedral angles, the relation corresponding to eqn (9) is somewhat more complicated.

(v) *Processing of distance constraints involving pairs of diastereotopic substituents without stereospecific resonance assignments*

Because the standard sequential assignment procedure for proteins (Wüthrich, 1986) does not assign individually the diastereotopic substituents of prochiral groups, and additional techniques used for this purpose (e.g. Güntert *et al.*, 1989; Neri *et al.*, 1989) can provide stereospecific assignments for only part of the prochiral centers, programs used for structure calculations from n.m.r. data must contain routines to process distance constraints with pairs of diastereotopic substituents $\beta_1$ and $\beta_2$, $(\alpha,\beta_1,b_1)$ and $(\alpha,\beta_2,b_2)$, to a pseudoatom $\beta_Q$ located centrally with respect to the 2 diastereotopic substituents $\beta_1$ and $\beta_2$, and to add a correction to the distance limit that equals the distance between the diastereotopic substituents and the pseudoatom (Wüthrich *et al.*, 1983):

$$b_Q = \min(b_1, b_2) + |\mathbf{r}_{\beta_1} - \mathbf{r}_{\beta_Q}|. \tag{10}$$

In the program DIANA, we replaced these pseudoatom corrections with a combination of 2 approaches, which has the advantage that a lesser part of the information contained in the experimental data is lost by the data processing.

The 1st approach by DIANA uses the conventional pseudoatom concept, but with variable corrections, depending on the available experimental data. The upper distance limit for the pseudoatom constraint, $(\alpha,\beta_Q,b_Q)$, is then calculated as:

$$b_Q = \sqrt{\frac{b_1^2 + b_2^2}{2} - |\mathbf{r}_{\beta_1} - \mathbf{r}_{\beta_Q}|^2}. \tag{11}$$

If only 1 of the 2 constraints can be measured, say $(\alpha,\beta_1,b_1)$, no improvement of the original pseudoatom correction can be attained, and $b_Q = b_1 + |\mathbf{r}_{\beta_1} - \mathbf{r}_{\beta_Q}|$. If there are 4 constraints between the diastereotopic substituents of 2 prochiral centers, $(\alpha_1,\beta_1,b_{11})$, $(\alpha_1,\beta_2,b_{12})$, $(\alpha_2,\beta_1,b_{21})$, $(\alpha_2,\beta_2,b_{22})$, the corresponding pseudoatom upper distance limit $(\alpha_Q,\beta_Q,b_Q)$ is given by:

$$b_Q = \sqrt{\frac{b_{11}^2 + b_{12}^2 + b_{21}^2 + b_{22}^2}{4} - |\mathbf{r}_{\alpha_1} - \mathbf{r}_{\alpha_Q}|^2 - |\mathbf{r}_{\beta_1} - \mathbf{r}_{\beta_Q}|^2}. \tag{12}$$

In the frequently encountered situation where 1 or several of the 4 constraints are missing, redundant upper distance limits are generated by application of the triangle inequality, so that eqn (12) can still be applied.

In the 2nd approach used by DIANA, no pseudoatom is introduced. The same distance limit:

$$b = \min[\max(b_1, b_2), \min(b_1, b_2) + |\mathbf{r}_{\beta_1} - \mathbf{r}_{\beta_2}|]$$

is applied for both diastereotopic substituents. Obviously, the application of 2 identical limits is, in general, not equivalent to the use of a pseudoatom, and the 2nd approach is applied only if it yields additional information. For example, if only 1 of the 2 diastereotopic substituents has an experimental constraint to an outside proton, only the pseudoatom constraint will be used. On the other hand, if the n.m.r. experiments show that $b_1 \approx b_2$, there is no advantage to the introduction of a pseudoatom distance limit.

Table 1 lists some results obtained by processing distance constraints with prochiral centers with DIANA, or with the conventional pseudoatom correction method (Wüthrich *et al.*, 1983). The 1st example resulted from an input of 2 nearly equal upper distance limits; DIANA imposed the higher limit on both distances, but left the pseudoatom distance unconstrained because the upper bound resulting from eqn (11) would be meaningless besides $b_1$ and $b_2$. The 2nd example involves 2 significantly different upper bounds; DIANA imposes constraints on both individual distances and the pseudoatom distance constraint calculated by eqn (11). In both cases, the resulting constraints are significantly tighter than with the conventional pseudoatom corrections. In contrast, in the 3rd example in Table 1, where there is only 1 experimental upper distance limit, the 2 methods yield nearly equivalent results. In the 4th example, DIANA detected that the pseudoatom constraint would be irrelevant and hence dropped it from the input. The final example shows the result of a treatment of constraints between 2 pairs of diastereotopic protons. Using the triangle inequality, DIANA first generated the smallest possible redundant constraints of the 2 distances that were not constrained by the experimental upper bounds. Next, it imposed the biggest of the 4 upper bounds thus obtained on all four individual distances, and an upper bound on the distance between the 2 pseudoatoms was obtained from eqn (12). Overall, Table 1 confirms that the loss of constraining information is often smaller when the experimental input is processed by DIANA than by the conventional pseudoatom approach. Analogous modifications to those shown here for upper distance constraints result for lower limit distance constraints.

(vi) *Checks for steric overlap*

In molecular mechanics programs, non-bonded interactions between atoms are usually treated by a Lennard-Jones potential (Momany *et al.*, 1975; Weiner & Kollman, 1981; Brooks *et al.*, 1983; van Gunsteren *et al.*,

**Table 1**

*Examples of results obtained by processing experimental distance constraints involving pairs of diastereotopic substituents without individual assignments either by DIANA or with the original pseudoatom concept*

| Constrained distances† | Input upper bounds (Å)‡ | Modified upper bounds (Å) | |
|---|---|---|---|
| | | DIANA§ | Conventional pseudoatom concept¶ |
| $H^{\beta 2,3}_{Arg1}$–$H^{\varepsilon}_{Arg1}$ | $b_1 = 4{\cdot}9$, $b_2 = 5{\cdot}0$ | $b_1 = b_2 = 5{\cdot}0$ | $b_Q = 5{\cdot}9$ |
| $H^{\gamma 2,3}_{Glu7}$–$H^{N}_{Glu7}$ | $b_1 = 3{\cdot}7$, $b_2 = 3{\cdot}1$ | $b_1 = b_2 = 3{\cdot}7$, and $b_Q = 3{\cdot}3$ | $b_Q = 4{\cdot}1$ |
| $H^{\beta 2,3}_{Tyr10}$–$H^{\beta 2}_{Lys41}$ | $b_1 = 3{\cdot}5$ | $b_Q = 4{\cdot}4$ | $b_Q = 4{\cdot}5$ |
| $H^{\beta 2,3}_{Tyr10}$–$H^{N}_{Thr11}$ | $b_1 = 4{\cdot}5$ | | $b_Q = 5{\cdot}5$ |
| $H^{\beta 2,3}_{Tyr10}$–$H^{\delta 2,3}_{Lys42}$ | $b_{11} = 4{\cdot}7$, $b_{12} = 4{\cdot}6$ | $b_{11} = b_{12} = b_{21} = b_{22} = 6{\cdot}5$, and $b_Q = 5{\cdot}5$ | $b_Q = 6{\cdot}6$ |

† Taken from an experimental n.m.r. data set collected with BPTI.

‡ $b_1$ and $b_2$ denote upper distance bounds from the 2 substituents of a diastereotopic pair to the same proton outside the prochiral center. $b_{11}$ denotes an upper distance bound from the 1st atom of one diastereotopic pair to the first atom of another diastereotopic pair, etc.

§ $b_Q$ denotes the upper limit imposed on the pseudoatom distance.

¶ The numbers given result from adding the distance from the pseudoatom to the protons that it replaces to the smaller of the 2 experimental constraints with the prochiral center. For methylene groups, this correction is $m = 1{\cdot}0$ Å (Wüthrich *et al.*, 1983; Wüthrich, 1986).

1983; Wako & Gō, 1987; Schaumann *et al.*, 1990). In a distance geometry approach for structure determination of proteins from n.m.r. data, only the most dominant part of the energy function is kept, i.e. the steric repulsion (Havel & Wüthrich, 1984; Braun & Gō, 1985). In the program DIANA, the steric repulsion between 2 atoms is treated as a lower distance limit for the corresponding interatomic distance, the distance bound being set equal to the sum of the repulsive core radii of the 2 atoms. In the present version of DIANA, the same values for the repulsive core radii are used as by Braun & Gō (1985), i.e. 0·95 Å for amine or amide hydrogen atoms, 1·0 Å for all other hydrogen atoms, 1·35 Å for aromatic carbon atoms, 1·40 Å for all other carbon atoms, 1·30 Å for nitrogen atoms, 1·20 Å for oxygen atoms and 1·60 Å for sulfur atoms. If the distance between 2 atoms exceeds the sum of their repulsive core radii, no contribution to the target function results from this atom pair.

A straightforward implementation of a check for the aforementioned steric overlaps would require the calculation of almost all interatomic distances, and would therefore be very inefficient. Therefore, DIANA stores all atom pairs with reasonably small interatomic distances in a list of potential non-bonded interactions (Verlet, 1967; Allen & Tildesley, 1987). This list is updated only after a notable conformation change or after several iterations of the conjugate gradient minimization, and a fast algorithm for this update ensures that most interatomic distances need not be computed (Hockney & Eastwood, 1981; Braun & Gō, 1985; Grest *et al.*, 1989). In a protein molecule with $m \approx 1000$ atoms, the list of potential non-bonded interactions will usually contain less than 30,000 atom pairs, whereas the total number of atom pairs is of the order of $m(m-1)/2 \approx 500{,}000$. The number of atom pairs that actually give non-vanishing contributions to the target function at the end of the minimization is again much smaller, typically of the order of 100 for a "good" conformation.

In the program DIANA, the list of potential non-bonded contacts is divided into 2 parts. The 1st part is invariant during a structure calculation, i.e. it is set up only once at the start of a calculation and will not be affected by subsequent updates; it comprises all intraresidual and sequential distances (Wüthrich, 1986). Most steric lower limits that are irrelevant will be excluded from the list, which obviously includes all distances that are independent of the conformation. The 2nd part of the list is subject to an updating procedure and includes interatomic distances that are not already included in the invariant part. To create this list, the present conformation of the protein molecule is placed into a cubic lattice with a lattice constraint $g$ equal to twice the biggest repulsive core radius, i.e. $g = 3{\cdot}2$ Å in the present version of the program DIANA, and only distances between atoms located within the same or in neighboring cells of the lattice are added to the list (Hockney & Eastwood, 1981; Braun & Gō, 1985; Grest *et al.*, 1989). Thus, it is ascertained that the list contains all non-bonded contacts that yield a non-vanishing contribution to the target function for the conformation present at the time the list is computed. Slight changes to this conformation will presumably change the list only slightly. Therefore, the list is updated only if, since the last update, at least 1 dihedral angle was changed by more than a preset limit, e.g. 10°, or if this limit is not reached, after a preset number of iterations, e.g. 50.

Special treatments are required for hydrogen bonds, disulfide bridges, and possibly other non-standard covalent links, because these bonds are not represented by the tree structure of rotatable bonds (Abe *et al.*, 1984), since the latter does not allow for flexible, closed rings. As a consequence, the steric lower limits for acceptor–donor distances in potential hydrogen bonds, and the distances between $C^{\beta}_i$ and $S^{\gamma}_j$ in disulfide $Cys_i$–$Cys_j$ are reduced by 1·0 Å, and the steric lower limits between the cysteine sulfur atoms are decreased by 2·0 Å relative to the sum of the corresponding repulsive core radii. The bond lengths and angles of hydrogen bonds and disulfide bridges are fixed by explicit upper and lower distance limits (Williamson *et al.*, 1985). The proline rings are rigid structures in the ECEPP force field (Momany *et al.*, 1975; Némethy *et al.*, 1983) in the sense that there are no internal degrees of freedom within them. In contrast, in the program DIANA, one can allow for flexibility in such rings by "cutting" one of the covalent bonds. In the case of proline, this creates 4 new rotatable bonds. The ring is then closed only by explicit distance constraints, and the necesary elimination or decrease of some steric lower distance limits is done automatically. In the input for DIANA, a flexible proline residue is entered *via* an

additional entry in the residue library, where the new rotatable bonds have to be defined and the closure of the ring is inherent only in the connectivity list, but not in the tree structure of the rotatable bonds.

(vii) *The minimization procedure*

The gradient of the target function defined by eqn (6) can be calculated with a fast algorithm because the target function can be written as a sum of functions of individual interatomic distances and individual dihedral angles (Noguti & Gō, 1983; Abe *et al.*, 1984). The partial derivative of the function $T$ of eqn (4) with respect to a dihedral angle $a'$ is given by:

$$\frac{\partial T}{\partial \phi_{a'}} = -(\mathbf{e}_{a'}, \mathbf{e}_{a'} \wedge \mathbf{r}_{a'}) \cdot \sum_{c=u,l,v} w_c \sum_{\substack{i \in I_c \\ \alpha_i^c \in M_{a'}}} \Theta_c\left(\frac{d_i^{c2} - b_i^{c2}}{b_i^{c2}}\right)\left(\mathbf{r}_{\alpha_i^c} \wedge \mathbf{r}_{\beta_i^c}\right)\left(\mathbf{r}_{\alpha_i^c} - \mathbf{r}_{\beta_i^c}\right) + 2w_a \sum_{i=1}^{n_a}\left(1 - \left(\frac{\Delta_i}{\Gamma_i}\right)^2\right)\Delta_i \delta_{a_i a'}. \quad (13)$$

$r_{\alpha_i^c}$ and $\mathbf{r}_{\beta_i^c}$ are the position vectors of the atoms $\alpha_i^c$ and $\beta_i^c$, respectively, $\mathbf{e}_{a'}$ denotes the unit vector along the rotatable bond $a'$, $\mathbf{r}_{a'}$ the start point of it, and $M_{a'}$ the set of all atoms for which the positions are affected by a change of the dihedral angle $a'$ if the N-terminal part of the protein molecule is kept fixed.

The minimization algorithm used in the program DIANA is the well-known method of conjugate gradients (Powell, 1977). At each minimization step, conjugate gradient iterations are done until either the norm of the gradient vector is smaller than some preset value, or the maximal number of iterations at this step (as given in the minimization parameters input file) is exceeded. Because the minimization routine assumes a continuously differential target function, problems may arise if an update of the list of potential non-bonded contacts (see above) results in a discontinuous change of the target function. Therefore, the conjugate gradient minimization is automatically restarted after premature termination due to a jump in the target function.

(viii) *Optimization of structure calculations with DIANA*

In order to achieve a high level of efficiency of the target function and gradient evaluation, the calculation is divided up into several parts, and each part is executed only if it is necessary. At the outset of a structure calculation, the static list of potential non-bonded contacts is set up, irrelevant constraints are eliminated, and the distance limits with prochiral centers are processed. Then, at the start of each minimization step, a list of all currently used distance constraints according to the subsets $I_u$, $I_l$, $I_v$ and $I_a$ is prepared, where the list of non-static potential non-bonded contacts is subject to the aforementioned updating procedure. The parts of the calculation that have to be excluded once for each combined computation of the target function and its gradient are the generation of the Cartesian atomic coordinates from given dihedral angles, the identification of violated constraints, and the evaluation of some terms that are present in both eqns (6) and (13), and will therefore be needed for the evaluation of the target function and the gradient. Here, the time-limiting step is the computation of the interatomic distances corresponding to all distance constraints in the list, of which the great majority are steric constraints.

The program DIANA has been optimized for the vectorization capabilities of the CRAY X-MP, and it has been implemented on other UNIX machines and on VAX computers (see Table 2). Since standard Fortran-77 has been used as far as possible, the additional implementation of DIANA on other computers will be straightforward.

(c) *The program GLOMSA*

The input for the program GLOMSA includes a group of $m$ 3-dimensional protein structures, and the list of conformational constraints from which these structures were calculated by DIANA. In a typical situation analyzed by GLOMSA, an atom $\alpha$ outside the prochiral center considered has 2 upper distance limits $(\alpha,\beta_1,b_1)$ and $(\alpha,\beta_2,b_2)$ to the diastereotopic substituents $\beta_1$ and $\beta_2$. We denote with $d_{k,l}$ $(k=1,2;\ l=1,\ldots,m)$ the distance between the atoms $\alpha$ and $\beta_k$ in the $l$th conformation. Then the program GLOMSA computes the sum of the residual violation of the 2 upper distance limits in the $m$ conformations, $V$, for either of the 2 possible stereospecific assignments $I$ and $R$ ($I$, is the assignment used arbitrarily in the input, $R$ is the reversed one).

$$V^I = \sum_{l=1}^{m} [\Theta(d_{1,l}-b_1)(d_{1,l}-b_1)+\Theta(d_{2,l}-b_2)(d_{2,l}-b_2)]$$
$$V^R = \sum_{l=1}^{m} [\Theta(d_{1,l}-b_2)(d_{1,l}-b_2)+\Theta(d_{2,l}-b_1)(d_{2,l}-b_1)], \quad (14)$$

and the minimum value, $v$, that the larger of the violations of the 2 constraints $b_1$ and $b_2$ has in any of the conformations:

$$v^I = \min_{l=1,\ldots,m} \max [\Theta(d_{1,l}-b_1)(d_{1,l}-b_1),\ \Theta(d_{2,l}-b_2)(d_{2,l}-b_2)]$$
$$v^R = \min_{l=1,\ldots,m} \max [\Theta(d_{1,l}-b_2)(d_{1,l}-b_2),\ \Theta(d_{2,l}-b_1)(d_{2,l}-b_1)]. \quad (15)$$

In eqns (14) and (15) as well as in eqn (17) below, $\Theta$ denotes the Heaviside function that equals 1 if the argument is positive, and vanishes otherwise. The program GLOMSA further calculates the average of the differences $\Delta d_l = d_{1,l} - d_{2,l}$ in the $m$ structures:

$$\overline{\Delta d} = \frac{1}{m}\sum_{l=1}^{m} \Delta d_l \quad (16)$$

and the signed maximal number of conformations where $\Delta d_l$ has the same sign:

$$n_{\Delta d} = \begin{cases} s & s > m/2; \\ -(m-s) & s \le m/2; \end{cases} \qquad s = \sum_{l=1}^{m} \Theta(\Delta d_l). \quad (17)$$

$n_{\Delta d}$ is constructed such that $|n_{\Delta d}| \ge m/2$, and that its sign is positive if $d_{1,l}$ is bigger than $d_{2,l}$ in the majority of the conformations, and negative otherwise. The sign of $n_{\Delta d}$ is not necessarily the same as the sign of $\overline{\Delta d}$.

To identify stereospecific assignments, GLOMSA correlates the signs $\overline{\Delta d}$ and $n_{\Delta d}$ with the sign of $\Delta b = b_1 - b_2$. Matching signs confirm that the stereospecific assignment $I$ assumed in the input is correct, and opposite signs indicate the stereospecific assignments $R$. In order to exclude cases where the available data do not clearly distinguish between the 2 possibilities, $|\Delta b|$, $|\overline{\Delta d}|$ and $|n_{\Delta d}|$ are further required to exceed user-defined threshold values for an unambiguous stereospecific assignment by the program GLOMSA. If a relation of the type $d_1 > d_2$ or $d_1 < d_2$ was unambiguously established by the

experiments, this relation is accepted without imposing the threshold condition on $|\Delta b|$.

If there are several pairs of upper distance constraints from different hydrogen atoms to the same prochiral center, the above procedure is repeated independently for each distance constraint pair. The user then manually combines the individual stereospecific assignments obtained, since potential inconsistencies could be hidden in the output resulting from an automated combination of the individual stereospecific assignments. On the basis of the quantities defined in eqns (14) and (15), another method for obtaining stereospecific assignments is conceivable, where an unambiguous assignment $I$ would be assumed if $v^I = 0$ and $v^R > 0$. However, it turns out that such a criterion would be very restrictive and would usually not yield a significant number of stereospecific assignments. Therefore, GLOMSA calculates $V$ and $v$ only as informative output.

Examples of the results obtained with the program GLOMSA in an application with experimental n.m.r. data are given elsewhere (Güntert *et al.*, 1990).

## 3. Results and Discussion

### (a) *Computing time for a protein structure calculation with DIANA*

The computation speed of the program DIANA on ten different computers (Table 2) was measured for the calculation of one structure of BPTI using the data set WIST, which was derived from the regularized crystal structure of BPTI (Marquardt *et al.*, 1983) so as to mimic an experimental n.m.r. input (Güntert *et al.*, 1989). The same random start conformation was used on all computers. A total number of 5900 target function evaluations was allowed, and updates of the list of potential non-bonded contacts were made when a dihedral angle was changed by more than 10° since the previous update, or after 50 iterations without update.

**Table 2**

*Central processing unit times required by the program DIANA for the calculation of one BPTI structure*

| Computer type | c.p.u. time (min)† | Factor‡ |
|---|---|---|
| Cray X-MP/28§ | 0·82 (3·2) | 1·0 (3·9) |
| VAX 8650 | 20 | 25 |
| VAX 6000-420 | 18 | 22 |
| SUN 386i | 111 | 136 |
| SUN 3/260¶ | 44 (188) | 54 (229) |
| SUN 4/390 | 14 | 17 |
| SUN 4/60 | 18 | 22 |
| Silicon Graphics Personal IRIS 4D/25 | 13 | 16 |
| Convex C1 | 18 | 22 |
| CDC Cyber 855 | 12 | 15 |

One completely folded conformation of the protein BPTI was calculated starting from random dihedral angles. The same starting structure and the same constraints were used on all computers. The same code, which was optimized for the Cray X-MP, and single precision floating point arithmetics were used on all machines (see the text for further details).

† c.p.u. times were measured using the library routines SECOND on the Cray X-MP and the CDC Cyber, ETIME on the other UNIX computers SUN, Silicon Graphics Personal IRIS, and Convex, and LIB$STAT_TIMER on the VAX machines.

‡ Ratio between the c.p.u. times required on the given machine and on the Cray X-MP using vectorization.

§ The numbers in parentheses were obtained when vectorization was deliberately inhibited. In either case only 1 c.p.u. was used.

¶ The numbers in parentheses were obtained when the floating point coprocessor MC 68881 was used instead of the floating point accelerator.

A clear-cut result from the measurements of the total c.p.u. time used is that the program DIANA runs by more than one order of magnitude faster on the Cray X-MP than on any of the other computers included in the test (Table 2). Even when the vectorization was completely inhibited, which increased the c.p.u. time by a factor of 3·9, the Cray X-MP remained the fastest machine. Following the Cray X-MP, there is a group of computers with similar performances, i.e. the VAX 8650, VAX 6000-420, SUN 4, Silicon Graphics Personal IRIS, Convex C1 and CDC Cyber 855, which used 12 to 20 minutes of c.p.u. time. Finally, the smaller machines SUN 3 and SUN 386i required again significantly more computer time to solve the test problem. It should be pointed out that the same code has been used on all machines, which is optimized with regard to the vectorization capabilities of the Cray X-MP.

### (b) *Comparison of the structures calculated with different computers*

Even though exactly the same input of conformational constraints and identical starting conformations have been supplied to the program on each computer, the final conformation obtained at the end of the minimization was in general different on the different computers (only the SUN 4/60 and the SUN 4/390 produced exactly identical structures). Repeating the calculation on the same computer yielded identical results. On all computers, the target function value of the starting conformation was the same, and the target function values gradually started to diverge with increasing minimization level, with final target function values ranging from 1·28 to 3·81 Å$^2$. These are small target function values when compared to "bad" structures obtained from different starting conformations, which often end up in high local minima with target function values above 100 Å$^2$. The average of the pairwise r.m.s.d. values (McLachlan, 1979) between 11 structures obtained from the same starting conformation on different computers, or on the same computer with different compiler settings (the result of the CDC Cyber was for technical reasons not used in this comparison) are 0·59 and 0·84 Å for all backbone atoms and all heavy atoms, respectively, the maximal values being 0·98 and 1·41 Å. If only the residues 3 to 55 were taken into account (i.e. the loose chain ends are discarded), the average of the pairwise r.m.s.d. values were 0·47 and 0·78 Å for the backbone and heavy atoms, respectively, with maximal values of 0·72 and 1·30 Å. These r.m.s.d. values are comparable to those between the

four conformations obtained with the program DISMAN (Braun & Gō, 1985) starting from widely different initial structures, as described previously (Güntert *et al.*, 1989). Overall, the implication is that the structures obtained with the same starting conformation when using the different computers are distributed within the bounds of the conformation space defined by the experimental constraints.

This result can be rationalized from the following. Because the experimental data restrict distances and dihedral angles to allowed ranges rather than to unique values (Wüthrich, 1986), one cannot expect the target function to have a well-defined global minimum. Instead there is an allowed region in conformation space where the target function has low values throughout, so that conformation changes within this region alter the target function only slightly. Therefore, although the aforementioned allowed region of the conformation space is defined by the experimental constraints (note, however, that the boundaries of this allowed region are not sharp, since they are in practice also influenced by the somewhat arbitrary selection of a set of "good" structures from among those calculated with distance geometry from the same input with different starting conformations), the exact conformation attained within this allowed region is heavily influenced by, for example, round-off errors. Overall, it is thus not really a surprise that the r.m.s.d. values between the structures obtained from different computers and an identical starting conformation are similar to those between structures obtained from widely different starting conformations when using the same computer.

### (c) *Influence of different treatments of distance constraints with pairs of diastereotopic substituents*

In order to assess the influence of different treatments of distance constraints with pairs of diastereotopic substituents on the convergence of the variable target function algorithm and on the quality of the structures obtained, DIANA calculations were carried out using experimental n.m.r. data for BPTI. Four input data sets that differed only in the treatment of the distance constraints with pairs of diastereotopic substituents were compared (Table 3). Datasets I and II include all stereospecific assignments present in the experimental n.m.r. data set (L. Orbons, P. Güntert & K. Wüthrich, unpublished results), whereas datasets III and IV include no stereospecific assignments. Upper distance limits involving pairs of diastereotopic substituents without individual n.m.r. assignments were treated by the method implemented in the program DIANA for datasets I and III, or by the conventional pseudoatom method (Wüthrich *et al.*, 1983) for datasets II and IV. Note that the different numbers of upper distance limits in datasets I to IV (Table 3) are a direct consequence of the different treatments of the NOE distance constraints with prochiral centers. The lower limit distance constraints and dihedral angle constraints were identical in datasets I to IV.

For each of the four datasets, 150 structure calculations were started with random start conformations. The final minimization level was $L = 58$, and the maximal number of target function evaluations per conformation was 9900. The calculations were done on a Cray X-MP computer, and the total c.p.u. time required was 18·6 hours. (Note that the time used per structure calculation is longer than in Table 2, because more target function evaluations were performed.)

A statistical analysis of the residual constraint violations in the 20 conformations with smallest final target function values in each of the groups I to IV is afforded by Table 4. Overall, the structure groups I to IV have similar quality in terms of the target function values and the sums of distance

**Table 3**

*Characterization of the four BPTI data sets used to investigate the influence of different treatments of distance constraints with pairs of diastereotopic substituents*

| Quantity | Datasets | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| Stereospecifically assigned prochiral centers† | 42 | 42 | 0 | 0 |
| Individually assigned $NH_2$ groups of Asn† | 3 | 3 | 0 | 0 |
| Treatment of diastereotopic pairs without individual assignments‡ | DIANA | Conventional pseudoatom concept | DIANA | Conventional pseudoatom concept |
| Upper distance limits | 866 | 781 | 924 | 637 |
| Lower distance limits§ | 31 | 31 | 31 | 31 |
| Dihedral angle restraints¶ | 140 | 140 | 140 | 140 |

† The total number of prochiral centers is 87. 14 $\beta$-methylene groups were stereospecifically assigned using the program HABAS, and all other stereospecific assignments (25 for methylene groups, and 3 for isopropyl groups) were obtained with GLOMSA. In addition, individual assignments for 3 $NH_2$ groups of asparagine were established from the intraresidual NOEs to $C^{\gamma}H_2$.

‡ DIANA refers to the novel treatment of diastereotopic pairs without individual assignments as explained in the text. Conventional pseudoatom concept refers to the method of Wüthrich *et al.* (1983).

§ Exactly the same experimental lower distance limits for disulfide bridges and experimentally established hydrogen bonds (Williamson *et al.*, 1985) were included in the datasets I to IV.

¶ Exactly the same dihedral angle restraints were included in the datasets I to IV. They were derived from a combined analysis of $^3J$ scalar coupling constants and short-range distance constraints using the program HABAS (Güntert *et al.*, 1989).

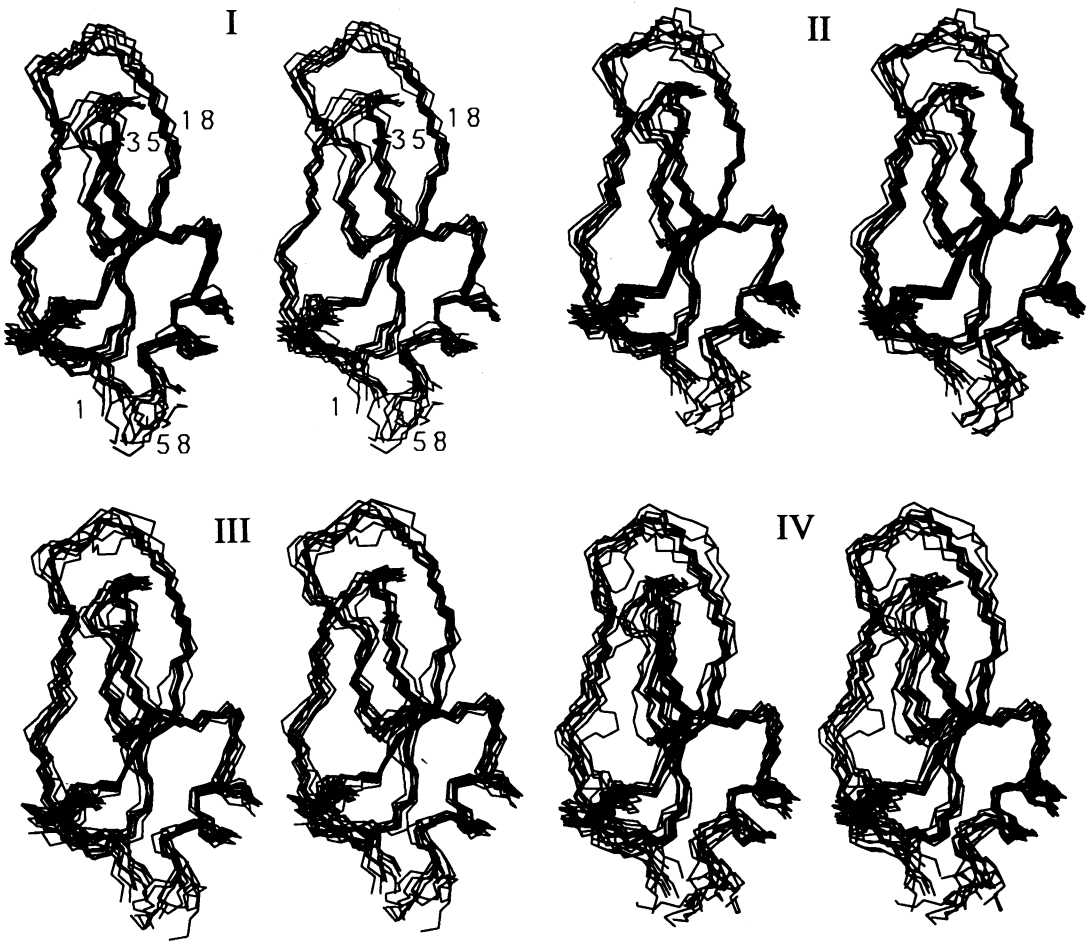
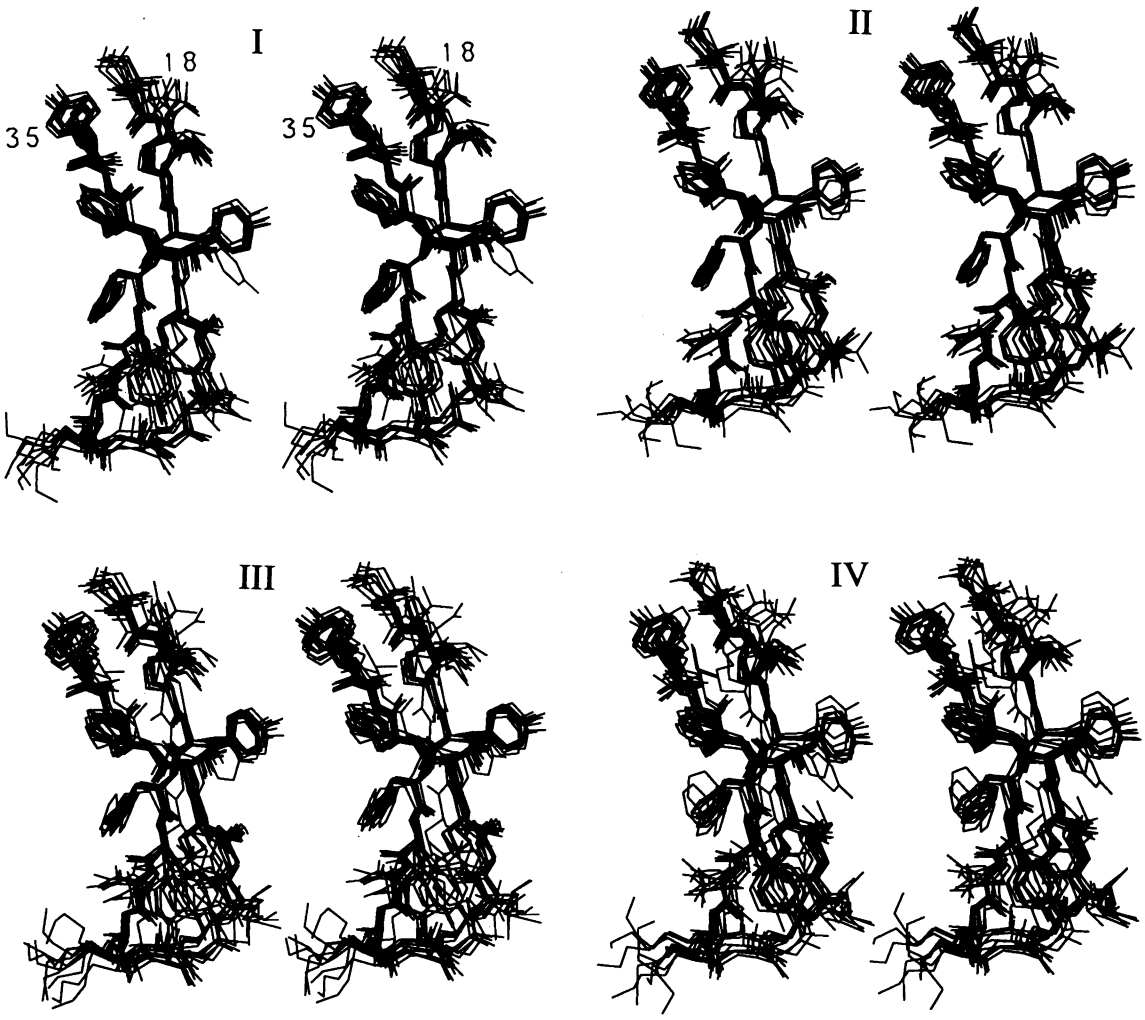



**Figure 2.** Stereo views of the 10 BPTI conformations with smallest target function values in each of the groups I to IV (Table 4). (a) Backbone of the whole protein (residues 1 to 58). (b) Heavy atoms of the β-sheet (residues 18 to 35).

constraint violations, but nonetheless clearly improved convergence of the calculations was obtained for the datasets with stereospecific assignments (I and II). Thus, using a less constraining dataset does not generally yield lower final target function values, even though the global minimum of the target function cannot be higher than for a "better" data set. This observation probably results because the folding pathway is also less determined in the absence of stereospecific assignments.

The different precision of the structure determinations with datasets I to IV is visualized with the aid of stereoviews of superpositions of the ten conformations with smallest final target function values in each group (Fig. 2). These images were produced with the program CONFOR (Billeter *et al.*, 1985). For the backbone of the whole protein (Fig. 2(a)) as well as for the all heavy-atom presentation of the β-sheet (Fig. 2(b)) more precisely defined structures were obtained in groups I and II, which include stereospecific assignments, than in III and IV. The fact that the structures from group IV are clearly less well determined than the structures from group III demonstrates the advantage of the novel DIANA processing of distance constraints with pairs of diastereotopic substituents that were not individually assigned.

For a more quantitative assessment of the observations in Figure 2, we calculated r.m.s.d. values (McLachlan, 1979) among the conformations of each group (Table 5). In a first comparison we included the 20 conformations with smallest final target function values in each of the four structure groups I to IV (Table 3). Because conformations with higher target function values tend to exhibit higher r.m.s.d. values, we made a second comparison using only the conformations with final target function values less than 5 Å$^2$. In this second comparison, possible effects from the poorer convergence found for groups III and IV, which use no stereospecific assignments, are eliminated. There are 17, 19, 10 and 6 such conformations in structure groups I, II, III and IV, respectively. In both comparisons, the average pairwise r.m.s.d. values given in Table 5 show a clear tendency to increase from the structure groups with stereospecific assign-




(b)

**Fig. 2.**

ments (I and II) to those without stereospecific assignments. A smaller increase of the r.m.s.d. is obtained also when going from a structure group calculated using the novel DIANA processing of distance constraints with pairs of diastereotopic substituents to the corresponding group calculated with conventional pseudoatom concept (Wüthrich *et al.*, 1983). Very similar results were obtained with the best 20 conformations from each group, or only those with final target function values less than 5 $Å^2$ (Table 5).

As was shown previously (Güntert *et al.*, 1989; Nilges *et al.*, 1990) the precision of the distance geometry structures can be significantly improved when stereospecific assignments are available. This observation is confirmed by the data in Table 5, which were obtained using experimental input data rather than test situations derived from known structures. The results obtained encourage continuation of studies on experimental methods, e.g. biosynthetically directed fractional $^{13}C$ labeling (Senn *et al.*, 1989; Neri *et al.*, 1989), for obtaining stereospecific assignments prior to the structure calculation. When the novel processing of upper distance constraints with pairs of diastereotopic substituents implemented in DIANA is compared to the more conservative pseudoatom concept (Wüthrich *et al.*, 1983), only a slight improvement of the calculated structures can be registered (Table 5 and Fig. 2, compare I and II, or III and IV). From a practice-oriented viewpoint, it is perhaps a more important advantage of the DIANA processing that it is fully automated, and integrated into the structure calculations. It is thus less laborious than obtaining a maximum number of stereospecific assignments. Therefore, the combined use of the automated HABAS routine (Güntert *et al.*, 1989), which can provide a limited number of stereospecific assignments (see the footnote to Table 3), and the DIANA processing of the distance constraints might become a viable alternative. This idea receives support from the observation that the advantages of the DIANA processing relative to the conventional pseudoatom treatment (Wüthrich *et*

**Table 4**

*Statistics of the final target function values and residual constraint violations of the 20 BPTI structures with smallest final target function values in each of the four structure groups calculated from datasets I to IV*

| Quantity | Average value ± standard deviation† | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| Target function value (Å$^2$)‡ | 2·9±1·3 | 2·6±1·3 | 4·8±2·5 | 6·0±3·1 |
| Upper distance limits: | | | | |
| Violations > 0·2 Å | 8·0±4·2 | 7·0±3·4 | 11·6±6·9 | 11·2±6·0 |
| Sum of violations (Å) | 8·1±2·4 | 7·1±2·0 | 9·4±3·2 | 8·3±3·5 |
| Maximal violation (Å) | 0·51±0·16 | 0·48±0·17 | 0·71±0·26 | 0·78±0·32 |
| Lower distance limits: | | | | |
| Violations > 0·2 Å | 0·9±1·0 | 0·4±0·6 | 0·6±0·7 | 0·8±0·9 |
| Sum of violations (Å) | 0·7±0·3 | 0·5±0·2 | 0·7±0·2 | 0·6±0·3 |
| Maximal violation (Å) | 0·24±0·12 | 0·17±0·06 | 0·22±0·08 | 0·19±0·10 |
| Steric constraints: | | | | |
| Violations > 0·2 Å | 1·0±1·3 | 1·5±1·3 | 3·7±3·1 | 5·2±3·7 |
| Sum of violations (Å) | 3·6±1·0 | 3·8±1·2 | 5·2±1·8 | 6·3±2·5 |
| Maximal violation (Å) | 0·28±0·19 | 0·27±0·11 | 0·39±0·20 | 0·43±0·16 |
| Dihedral angle restraints: | | | | |
| Violations > 5° | 1·2±1·2 | 1·1±1·5 | 1·6±1·1 | 2·7±1·8 |
| Sum of violations (°) | 27±13 | 29±15 | 40±13 | 50±24 |
| Maximal violation (°) | 6·9±2·6 | 7·3±4·2 | 9·2±4·4 | 14·3±10·7 |

The data sets I to IV are defined in Table 3. A total of 150 structures were calculated with each of the datasets I to IV, and the 20 structures with smallest final target function value were included in this analysis.

† Of the individual values for the 20 conformations of each structure group.

‡ The weighting factors for experimental upper and lower limit distance constraints were $w_u = w_l = 1$, the weighting factor for steric lower distance limits was $w_v = 2$, and the weighting factor for dihedral angle restraints was $w_a = 5$ Å$^2$ for the final minimization step.

**Table 5**

*Average values and empirical standard deviations of the pairwise r.m.s.d. within the BPTI structure groups calculated from datasets I to IV to monitor the influence of stereospecific assignments*

| Atom set | r.m.s.d. within structure group (Å) | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| A. *The 20 conformations with smallest final target function* | | | | |
| Backbone 1–58 | 1·12±0·22 | 1·15±0·19 | 1·29±0·22 | 1·52±0·24 |
| Backbone 3–55† | 0·89±0·16 | 0·93±0·15 | 1·09±0·18 | 1·32±0·20 |
| Backbone 18–35 (β-sheet) | 0·54±0·16 | 0·57±0·16 | 0·79±0·28 | 0·75±0·24 |
| Backbone 48–55 (α-helix) | 0·27±0·10 | 0·26±0·09 | 0·32±0·10 | 0·36±0·11 |
| Heavy atoms 1–58 | 1·80±0·18 | 1·87±0·19 | 2·10±0·23 | 2·24±0·22 |
| Heavy atoms 3–55 | 1·72±0·17 | 1·80±0·19 | 2·04±0·24 | 2·16±0·22 |
| Heavy atoms 18–35 (β-sheet) | 1·10±0·16 | 1·12±0·14 | 1·46±0·29 | 1·41±0·28 |
| Heavy atoms 48–55 (α-helix) | 1·28±0·32 | 1·19±0·29 | 1·44±0·33 | 1·39±0·32 |
| B. *Conformations with final target function value <5·0 Å$^2$*‡ | | | | |
| Backbone 1–58 | 1·10±0·23 | 1·16±0·19 | 1·14±0·15 | 1·43±0·27 |
| Backbone 3–55 | 0·86±0·16 | 0·94±0·16 | 0·97±0·15 | 1·27±0·26 |
| Backbone 18–35 (β-sheet) | 0·55±0·16 | 0·56±0·16 | 0·64±0·23 | 0·76±0·28 |
| Backbone 48–55 (α-helix) | 0·27±0·10 | 0·26±0·09 | 0·29±0·08 | 0·38±0·09 |
| Heavy atoms 1–58 | 1·79±0·18 | 1·88±0·19 | 1·96±0·18 | 2·18±0·24 |
| Heavy atoms 3–55 | 1·72±0·17 | 1·81±0·20 | 1·93±0·20 | 2·13±0·23 |
| Heavy atoms 18–35 (β-sheet) | 1·10±0·16 | 1·11±0·14 | 1·33±0·25 | 1·39±0·28 |
| Heavy atoms 48–55 (α-helix) | 1·30±0·31 | 1·17±0·26 | 1·36±0·27 | 1·59±0·40 |

The data sets I to IV that were used to calculate the structure groups I to IV are described in Table 3 (see the text for further details).

† Residues 3 to 55 are chosen in order to exclude the less well determined terminal parts of the polypeptide chain.

‡ We obtained 17, 19, 10 and 6 conformations with final target function values less than 5 Å$^2$ from datasets I, II, III and IV, respectively.

*al.*, 1983) are most clear-cut when no stereospecific assignments are available (III and IV in Table 5 and Fig. 2).

## Appendix

### Choice of Parameters for Protein Structure Determinations with DIANA

Table A1 affords a complete list of the minimization levels, the weighting factors for steric constraints and the iteration limits used for the protein structure calculations described in Results and Discussion, sections (a) and (b). The choice of these parameters is typical for structure calculations with proteins of the size of BPTI, and agrees with the general recommendations given in the main text following equation (7).

**Table A1**

*Minimization steps used for a structure calculation of BPTI*

| Step | Minimization level, $L$† | Weight of steric constraints, $w_v$‡ | Maximal number of iterations§ |
|---|---|---|---|
| 1 | 0 | 0·2 | 300 |
| 2 | 1 | 0·2 | 500 |
| 3 | 2 | 0·2 | 300 |
| 4 | 3 | 0·2 | 200 |
| 5 | 4 | 0·2 | 100 |
| 6 | 5 | 0·2 | 100 |
| 7 | 6 | 0·2 | 100 |
| 8 | 7 | 0·2 | 100 |
| 9 | 9 | 0·2 | 200 |
| 10 | 10 | 0·2 | 100 |
| 11 | 11 | 0·2 | 100 |
| 12 | 12 | 0·2 | 100 |
| 13 | 13 | 0·2 | 200 |
| 14 | 15 | 0·2 | 100 |
| 15 | 17 | 0·2 | 100 |
| 16 | 18 | 0·2 | 100 |
| 17 | 19 | 0·2 | 100 |
| 18 | 21 | 0·2 | 300 |
| 19 | 22 | 0·2 | 100 |
| 20 | 23 | 0·2 | 100 |
| 21 | 24 | 0·2 | 300 |
| 22 | 25 | 0·2 | 100 |
| 23 | 26 | 0·2 | 200 |
| 24 | 27 | 0·2 | 100 |
| 25 | 30 | 0·2 | 100 |
| 26 | 31 | 0·2 | 100 |
| 27 | 32 | 0·2 | 100 |
| 28 | 36 | 0·2 | 100 |
| 29 | 38 | 0·2 | 100 |
| 30 | 39 | 0·2 | 100 |
| 31 | 45 | 0·2 | 100 |
| 32 | 50 | 0·2 | 100 |
| 33 | 54 | 0·2 | 100 |
| 34 | 56 | 0·2 | 100 |
| 35 | 58 | 0·2 | 100 |
| 36 | 58 | 0·6 | 300 |
| 37 | 58 | 2·0 | 500 |

BPTI consists of a polypeptide chain with 58 residues.

† For each minimization step, the same minimization levels were used for the 3 kinds of distance constraints, $L = L_u = L_l = L_v$.

‡ Throughout the structure calculation, the weighting factors for experimental distance constraints were $w_u = w_l = 1$, and the weighting factor for dihedral angle restraints was $w_a = 5$ Å$^2$.

§ The numbers of iterations are usually chosen based on the number of upper distance limits on the given level that were not already included at the preceding level.

We thank Dr M. Billeter for helpful discussions, Mr F. Suter for the use of a SUN 386i work station, and Mr R. Marani for the careful processing of the typescript. We acknowledge financial support by the Schweizerischer Nationalfonds (project 31.25174.88), the use of the Cray X-MP/28 of the ETH Zürich, and the use of a Silicon Graphics Personal IRIS of the Institut für Zellbiologie of the ETH Zürich.

## References

Abe, H., Braun, W., Noguti, T. & Gō, N. (1984). *Comput. Chem.* **8**, 239–247.

Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids*, pp. 147–152, Clarendon Press, Oxford.

Billeter, M., Engeli, M. & Wüthrich, K. (1985). *J. Mol. Graphics*, **3**, 79–83 and 97–98.

Billeter, M., Havel, T. F. & Wüthrich, K. (1987). *J. Comp. Chem.* **8**, 132–141.

Billeter, M., Schaumann, T., Braun, W. & Wüthrich, K. (1990). *Biopolymers*, **29**, 695–706.

Blumenthal, L. M. (1970). *Theory and Applications of Distance Geometry*, Chelsea, New York.

Braun, W. (1987). *Quart. Rev. Biophys.* **19**, 115–157.

Braun, W. & Gō, N. (1985). *J. Mol. Biol.* **186**, 611–626.

Braun, W., Bösch, C., Brown, L. R., Gō, N. & Wüthrich, K. (1981). *Biochim. Biophys. Acta*, **667**, 377–396.

Braun, W., Wider, G., Lee, K. H. & Wüthrich, K. (1983). *J. Mol. Biol.* **169**, 921–948.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comp. Chem.* **4**, 187–217.

Brünger, A. T., Clore, G. M., Gronenborn, A. M. & Karplus, M. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 3801–3805.

Clore, G. M., Gronenborn, A. M., Brünger, A. T. & Karplus, M. (1985). *J. Mol. Biol.* **186**, 435–455.

Crippen, G. M. (1977). *J. Comp. Phys.* **24**, 96–107.

Driscoll, P. C., Gronenborn, A. M., Beress, L. & Clore, G. M. (1989). *Biochemistry*, **28**, 2188–2198.

Dubs, A., Wagner, G. & Wüthrich, K. (1979). *Biochim. Biophys. Acta*, **577**, 177–194.

Eccles, C., Billeter, M., Güntert, P. & Wüthrich, K. (1989). *Abstracts Xth Meeting of the International Society of Magnetic Resonance*, Morzine, France, July 16–21, 1989, p. S50.

Gordon, S. L. & Wüthrich, K. (1978). *J. Amer. Chem. Soc.* **100**, 7094–7096.

Grest, G. S., Dünweg, B. & Kremer, K. (1989). *Comput. Phys. Comm.* **55**, 269–285.

Güntert, P., Braun, W., Billeter, M. & Wüthrich, K. (1989). *J. Amer. Chem. Soc.* **111**, 3997–4004.

Güntert, P., Qian, Y. Q., Otting, G., Müller, M., Gehring, W. & Wüthrich, K. (1991). *J. Mol. Biol.* **217**, 531–540.

Havel, T. F. & Wüthrich, K. (1984). *Bull. Math. Biol.* **46**, 673–698.

Havel, T. F. & Wüthrich, K. (1985). *J. Mol. Biol.* **182**, 281–294.

Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1983). *Bull. Math. Biol.* **45**, 665–720.

Hockney, R. W. & Eastwood, J. W. (1981). *Computer Simulations Using Particles*, McGraw-Hill, New York.

Holak, T. A., Prestegard, J. H. & Forman, J. D. (1987). *Biochemistry*, **26**, 4652–4660.
IUPAC-IUB Commission on Biochemical Nomenclature (1970). *J. Mol. Biol.* **52**, 1–17.
Kaptein, R., Zuiderweg, E. R. P., Scheek, R. M., Boelens, R. & van Gunsteren, W. F. (1985). *J. Mol. Biol.* **182**, 179–182.
Keller, R. M. & Wüthrich, K. (1980). *Biochim. Biophys. Acta*, **621**, 204–217.
Kline, A. D., Braun, W. & Wüthrich, K. (1988). *J. Mol. Biol.* **204**, 675–724.
Kohda, D., Gō, N., Hayashi, K. & Inagaki, F. (1988). *J. Biochem.* **103**, 741–743.
Kuntz, I. D., Crippen, G. M., Kollman, P. A. & Kimmelman, D. (1976). *J. Mol. Biol.* **106**, 983–994.
Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A. & Wright, P. E. (1989). *Science*, **245**, 635–637.
Marquardt, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). *Acta Crystallogr. sect. B*, **39**, 480–490.
McLachlan, A. D. (1979). *J. Mol. Biol.* **128**, 49–79.
Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). *J. Phys. Chem.* **79**, 2361–2381.
Némethy, G., Pottle, M. S. & Scheraga, H. A. (1983). *J. Phys. Chem.* **87**, 1883–1887.
Neri, D., Szyperski, T., Otting, G., Senn, H. & Wüthrich, K. (1989). *Biochemistry*, **28**, 7510–7516.
Nilges, M., Clore, G. M. & Gronenborn, A. M. (1988). *FEBS Letters*, **239**, 129–136.
Nilges, M., Clore, G. M. & Gronenborn, A. M. (1990). *Biopolymers*, **29**, 813–822.
Noguti, T. & Gō, N. (1983). *J. Phys. Soc. Jpn*, **52**, 3685–3690.
Powell, M. J. D. (1977). *Math. Program.* **12**, 241–254.
Qian, Y. Q., Billeter, M., Otting, O., Müller, M., Gehring, W. J. & Wüthrich, K. (1989). *Cell*, **59**, 573–580.
Schaumann, T., Braun, W. & Wüthrich, K. (1990). *Biopolymers*, **29**, 679–694.
Schultze, P., Wörgötter, E., Braun, W., Wagner, G., Vašák, M., Kägi, J. H. R. & Wüthrich, K. (1988). *J. Mol. Biol.* **203**, 251–268.
Senn, H., Werner, B., Messerle, B. A., Weber, C., Traber, R. & Wüthrich, K. (1989). *FEBS Letters*, **249**, 113–118.
van Gunsteren, W. F., Berendsen, H. J. C., Hermans, J., Hol, W. G. J. & Postma, J. P. M. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 4315–4319.
Vásquez, M. & Scheraga, H. A. (1988). *J. Biomol. Struct. Dynam.* **5**, 757–784.
Verlet, L. (1967). *Phys. Rev.* **159**, 98–103.
Wagner, G., Braun, W., Havel, T. F., Shaumann, T., Gō, N. & Wüthrich, K. (1987). *J. Mol. Biol.* **196**, 611–639.
Wako, H. & Gō, N. (1987). *J. Comp. Chem.* **8**, 625–635.
Weber, P. L. Morrison, R. & Hare, D. (1988). *J. Mol. Biol.* **204**, 483–487.
Weiner, P. K. & Kollman, P. A. (1981). *J. Comp. Chem.* **2**, 287–303.
Widmer, H., Billeter, M. & Wüthrich, K. (1989). *Proteins*, **6**, 357–371.
Williamson, M. P., Havel, T. F. & Wüthrich, K. (1985). *J. Mol. Biol.* **182**, 295–315.
Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, Wiley, New York.
Wüthrich, K. (1989). *Acc. Chem. Res.* **22**, 36–44.
Wüthrich, K., Wider, G., Wagner, G. & Braun, W. (1982). *J. Mol. Biol.* **155**, 311–319.
Wüthrich, K., Billeter, M. & Braun, W. (1983). *J. Mol. Biol.* **169**, 949–961.
Zuiderweg, E. R. P., Nettesheim, D. G., Mollison, K. W. & Carter, G. W. (1989). *Biochemistry*, **28**, 175–185.

*Edited by P. E. Wright*