Journal of Biomolecular NMR, 1 (1991) 447–456
ESCOM

J-Bio NMR 035

Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints

Peter Güntert and Kurt Wüthrich

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

Dedicated to the memory of Professor V.F. Bystrov

Received 29 July 1991
Accepted 5 August 1991

Keywords: NMR structures of proteins; Structure calculation from NMR data; Variable target function method; Program DIANA; Local minimum problem; Dihedral angle constraints

SUMMARY

A new strategy for NMR structure calculations of proteins with the variable target function method (Braun, W. and Gō, N. (1985) *J. Mol. Biol.*, **186**, 611) is described, which makes use of redundant dihedral angle constraints (REDAC) derived from preliminary calculations of the complete structure. The REDAC approach reduces the computation time for obtaining a group of acceptable conformers with the program DIANA 5–100-fold, depending on the complexity of the protein structure, and retains good sampling of conformation space.

INTRODUCTION

NMR structures of proteins in solution are commonly presented as a group of conformers, each of which has been calculated individually from the same experimental input data. In a high-quality structure determination each individual conformer has small residual violations of the experimental conformational constraints, and the root-mean-square deviations (RMSDs) among all conformers in the group are small (Braun, 1987; Wüthrich, 1986,1989). In this paper we describe

Abbreviations: RMSD, root-mean-square deviation; REDAC, use of redundant dihedral angle constraints; HD, mutant *Antennapedia* homeodomain with Cys^{39} replaced by Ser; BPTI, basic pancreatic trypsin inhibitor; ADB, activation domain from porcine procarboxypeptidase B.

0925-2738/$ 5.00

a new strategy for the use of the variable target function program DIANA (Güntert et al., 1991a) that enables efficient calculation of such groups of conformers.

The basic idea of the variable target function algorithm (Braun and Gō, 1985) is to *gradually* fit an initially randomized starting structure to the conformational constraints collected with the use of NMR experiments, starting with intraresidual constraints only, and increasing the 'target size' stepwise up to the length of the complete polypeptide chain. Since 1986 different implementations of the variable target function algorithm in the programs DISMAN (Braun and Gō, 1985), DADAS (Kohda et al., 1988) and DIANA (Güntert et al., 1991a) have been used for numerous structure determinations (e.g., Wagner et al., 1987; Arseniev et al., 1988; Kline et al., 1988; Qian et al., 1989; Widmer et al., 1989; Güntert et al., 1991b; Ikura et al., 1991), so that its performance in practice can be quite reliably evaluated. Advantages of the method are its conceptual simplicity and the fact that it works in dihedral angle space, so that the covalent geometry is preserved during the entire calculation. A drawback is that for all but the most simple molecular topologies (see below) only a small percentage of the calculations converge with small residual constraint violations, which is a typical local minimum problem (Li and Scheraga, 1987). Because of the low yield of acceptable conformers, calculations have typically been started with a large number of randomized starting conformers in order to obtain a group of good solutions, and sometimes a compromise had to be made between the requirements of small residual violations, the availability of approximately 10–20 'good' conformers to represent the solution conformation, and the available computing time (Kline et al., 1988; Widmer et al., 1989). With the introduction of the highly optimized program DIANA, which significantly reduced the computation time needed for the calculation of a single conformer, a workable situation was achieved for α-proteins (Güntert et al., 1991b), but for β-proteins with more complex topology the situation remained unsatisfactory. With the use of redundant dihedral angle constraints (REDAC) described in this paper, a greatly improved yield of converged conformers is now obtained also for β-proteins.

METHOD

In Fig. 1 the new strategy for the use of DIANA with REDAC is outlined and placed in perspective with the 'direct' variable target function method as proposed originally by Braun and Gō (1985) and used here as a reference for evaluating the merits of the new approach. In the direct approach, n start conformers with randomized dihedral angles are selected, and the program HABAS (Güntert et al., 1989) is applied for an initial analysis of the intraresidual and sequential NMR constraints (A in Fig. 1). The n conformers are then subjected to DIANA minimization against the experimental NMR constraints ($B^{(0)}$). Experience has shown that for well-converged solutions, the target function can be further reduced by repeating the DIANA refinement at L_{max} with variable weights for the van der Waals constraints. A limited number of k conformers ($m \leqslant k \leqslant n$) is subjected to this refinement in step D. Among the resulting solutions, m conformers with the smallest final target function values are selected to represent the solution structure. In practice, n is adjusted so as to obtain $m = 10$–20 acceptable conformers.

To use REDAC, one or several cycles $C^{(i)}$–$B^{(i)}$ are added to the calculation, providing a partial feedback of structural information from all conformers that were calculated up to the maximal level L_{max} (for a definition of L, see Güntert et al., 1991a) in the step $B^{(i-1)}$. In the step $C^{(i)}$, a particular amino acid residue is considered to have an acceptably well-defined conformation if the tar-

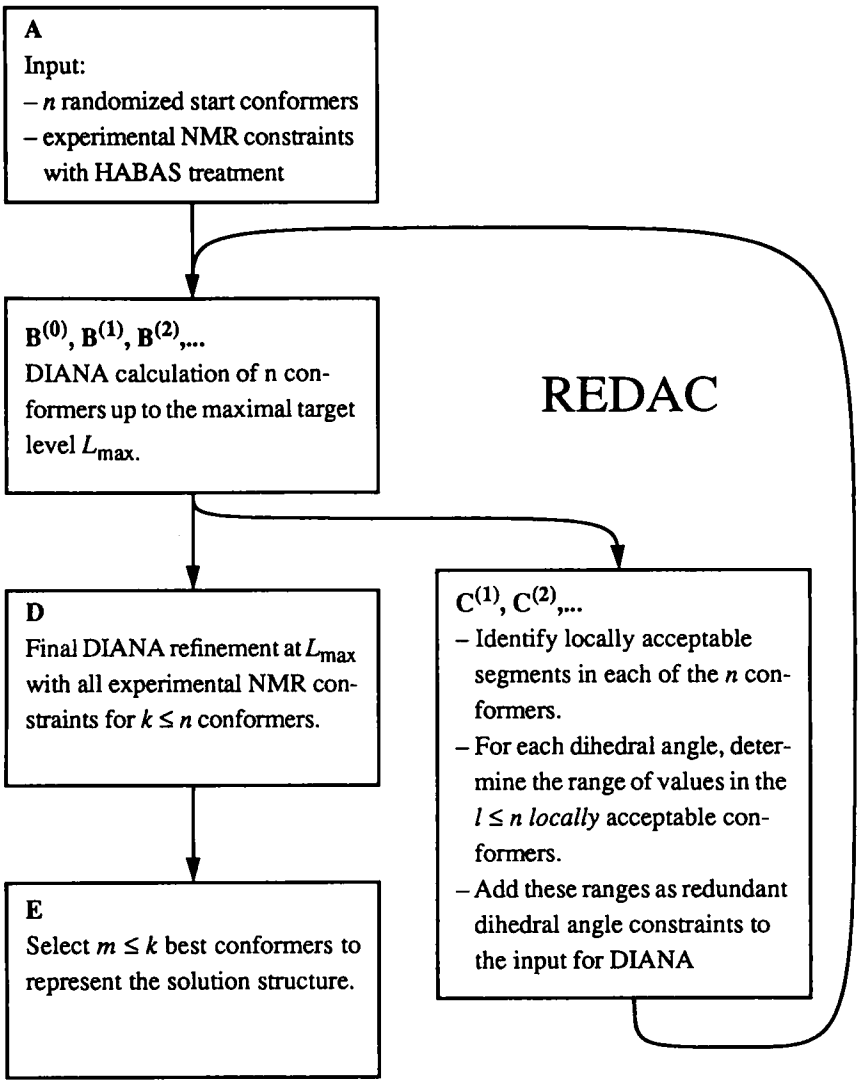


Fig. 1. Flowchart outlining the course of a protein structure calculation with the program DIANA using either the 'direct' way (A-$B^{(0)}$-D-E) or REDAC (A-$B^{(0)}$-[$C^{(1)}$-$B^{(1)}$-...]-D-E). Typically, the number of REDAC-cycles is 1 or 2.

get function value due to constraint violations that involve atoms or dihedral angles of this residue is less than a predefined value, typically 0.4 $Å^2$, and if the same condition holds for the two sequentially neighboring residues. Redundant dihedral angle constraints are generated for all those residues that were found to be acceptable in at least a predefined minimal number of conformers, typically 10 if the calculation is started with $n = 50$ randomized conformations (see Fig. 1), by taking the two extreme dihedral angle values in the group of acceptable conformers as upper and lower bounds. If the dihedral angle interval defined by these bounds is larger than a predefined maximal width, typically 270°, the redundant dihedral angle constraint is discarded, otherwise it is added to the input for the DIANA structure calculation in step $B^{(i)}$. The automated use of REDAC is implemented in version 1.14 of the program DIANA.

RESULTS

To compare the efficiency of DIANA calculations with and without use of REDAC, we calculated structures from high-quality experimental NMR data sets for the *Antp*(C39→S) homeodomain (HD; Güntert et al., 1991b), the basic pancreatic trypsin inhibitor (BPTI; K.D. Berndt, P. Güntert, L. Orbons and K. Wüthrich, to be published), and the activation domain from porcine procarboxypeptidase B (ADB; Vendrell et al., 1991). The HD is a typical α-protein, BPTI contains α and β secondary structure, and ADB has a more complex topology including a four-stranded β-sheet, two α-helices, and three loops that are only poorly determined by the NMR data (Vendrell et al., 1990). For each protein we performed a structure calculation starting with $n = 50$ randomized conformations and using REDAC, and selected the $m = 20$ conformers with the smallest final target function values for further analysis. In step $B^{(i)}$ we used the standard selection of minimization levels and parameters of DIANA, i.e., a maximal number of 150 conjugate gradient iterations and a weighting factor w_v of 0.2 for the van der Waals constraints (for a definition, see Güntert et al., 1991b) at all but the final level L_{max}, and three times 400 iterations with van der Waals weights of 0.2, 0.6, and 2.0 at L_{max}. The other weights had the same values throughout the entire calculation, with $w_u = w_l = 1$, and $w_a = 5$ Å^2. In step D we used $k = 50$, and we allowed for a maximal number of three times 1000 iterations at L_{max}, using the van der Waals weights $w_v = 0.2$, 0.6, and 2.0, respectively. For the HD and BPTI more than 40 conformers with final target function values at L_{max} below 2.1 Å^2 and 1.3 Å^2, respectively, were obtained after one REDAC cycle (Fig. 2). For ADB two REDAC cycles were needed to yield a group of 20 conformers with target function values below 2.9 Å^2 (Fig. 2).

For the HD and BPTI the DIANA calculations were repeated with the direct approach (Fig. 1), with the aim of producing a group of 20 conformers of equal quality, i.e., with final target function values in the same range as the 20 best conformers obtained from 50 starting conformers with the use of REDAC. For the HD this was achieved with $n = 400$ starting conformers, for BPTI with $n = 2000$ (Table 1). To obtain a fair comparison, the maximally allowed number of iterations for each target level in step $B^{(0)}$ was doubled when compared with the aforementioned parameters for the calculations with REDAC, and only the $k = 50$ conformers with lowest target function values at the end of step $B^{(0)}$ were further refined in step D. For the ADB it was found that calcula-

TABLE 1
EFFICIENCY OF DIANA CALCULATIONS WITH AND WITHOUT USE OF REDAC

	HD		BPTI		ADB	
	Direct	REDAC	Direct	REDAC	Direct[a]	REDAC
n[b]	400[b]	50	2000[b]	50	≈8000[b]	50
CPU time (h)[c]	3.8	0.66	17.7	0.61	≈140	1.48

[a] Estimated (see text).

[b] n is the number of randomized starting conformers (see Fig. 1). For the direct approach, n values were chosen so as to obtain the same number of acceptable conformers as with $n = 50$ and use of REDAC.

[c] Measured on a Cray Y/MP using one processor.

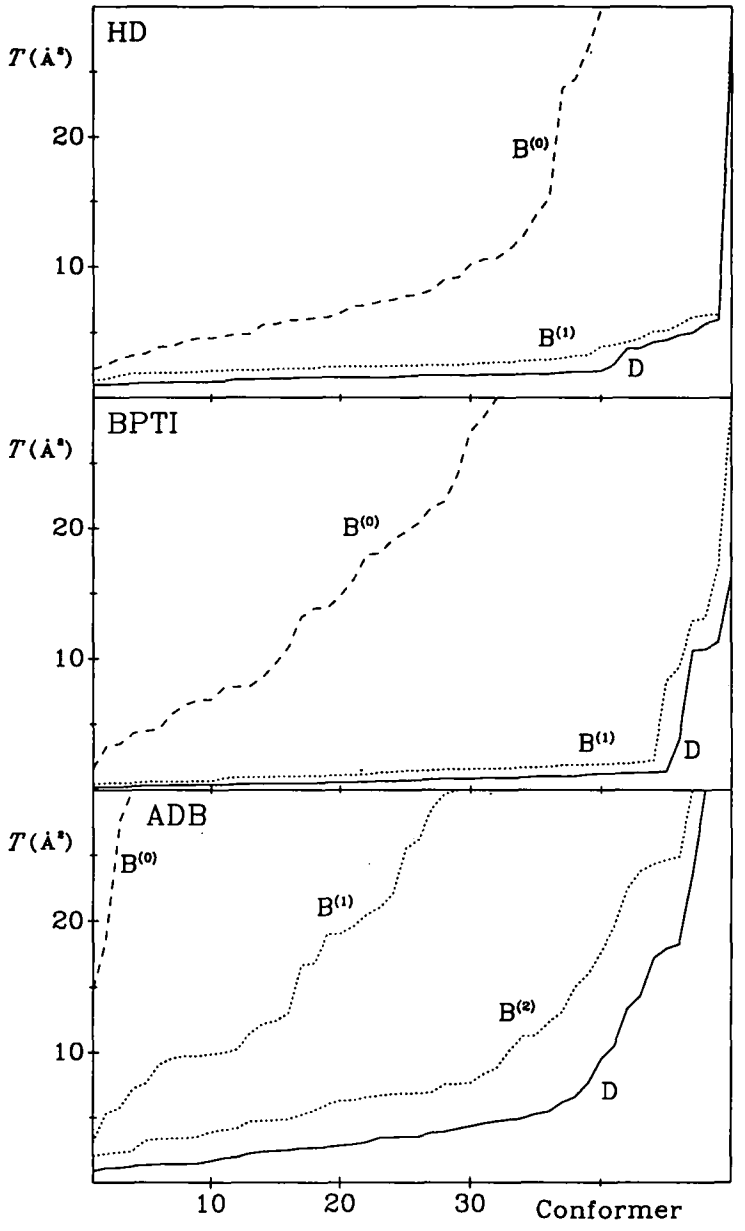
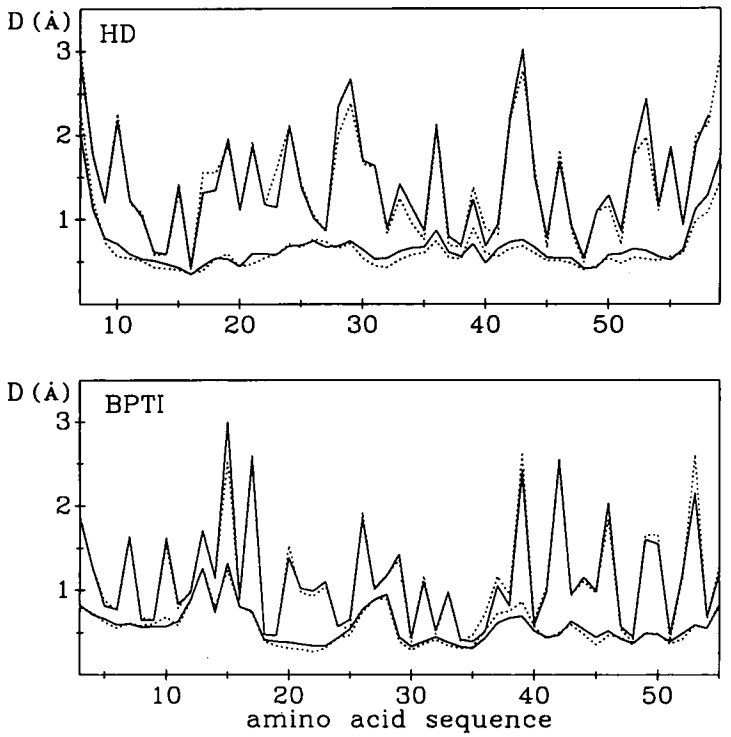


Fig. 2. Distribution of the DIANA target function values, T, among the 50 conformers calculated using REDAC for each of the three proteins *Antp*($C39 \rightarrow S$) homeodomain (HD), BPTI, and activation domain B (ADB). Along the horizontal axis the 50 conformers are ordered according to increasing target function value. For the HD and for BPTI the calculation consisted of the sequence of steps A-$B^{(0)}$-$C^{(1)}$-$B^{(1)}$-D; for ADB the sequence of steps was A-$B^{(0)}$-$C^{(1)}$-$B^{(1)}$-$C^{(2)}$-$B^{(2)}$-D. The letters identify the results at the end of the respective step in Fig. 1.



Fig, 3. Mean displacements (Billeter et al., 1989), i.e., averages of the pairwise RMSD values for the backbone atoms N, C^{α}, and C′ (lower curves) and for all heavy atoms (upper curves) of the individual amino acid residues after global superposition of the 20 best DIANA conformers obtained with (solid lines) or without (dashed lines) use of REDAC. The global superposition was made for the regions of the polypeptide chain that are well defined in the solution structure, i.e. residues 7-59 for the HD and residues 3-55 for BPTI, and displacements are presented only for the residues located within these regions.

tions without use of REDAC produced only one acceptable converged conformer from 400 starting conformers, so that we would have had to compute of the order of 8000 conformers in order to obtain a comparable result to that shown in Fig. 2 after the steps $B^{(2)}$ and D. Table 1 shows that for the three proteins the effective overall CPU time was reduced through the use of REDAC by factors of 5.7, 29, and about 100, respectively. The improved efficiency when using REDAC is particularly pronounced for the β-proteins, i.e., those proteins where the direct approach gives the lowest yields of acceptable structures.

To further evaluate the relevancy of the data in Table 1, we compared the quality of the DIANA structure calculations with and without REDAC on the basis of the parameters in Table 2 and Fig. 3. For both the HD and BPTI both types of calculations gave nearly identical values for the final target function, the different types of residual constraint violations, and the global pairwise RMSDs (McLachlan, 1979) among the 20 best conformers. Nearly identical values were also obtained for the displacements (Billeter et al., 1989) of the individual amino acid residues (Fig. 3). In addition, for both proteins the global pairwise RMSDs between conformers calculated with and without use of REDAC are for all practical purposes identical to the RMSDs between different conformers calculated using the same protocol (Tables 2 and 3).

TABLE 2
ANALYSIS OF THE 20 BEST CONFORMERS OBTAINED WITH THE DIANA CALCULATIONS OF TABLE 1

Quantity[a]	HD		BPTI	
	Direct[b]	REDAC[b]	Direct[b]	REDAC[b]
Final target function values (Å^2)	1.31 ± 0.19	1.29 ± 0.19	0.40 ± 0.12	0.39 ± 0.11
Distance constraint violations[c]:				
Number > 0.2 Å	3 ± 2	3 ± 2	0	0
Maximum (Å)	0.28 ± 0.06	0.28 ± 0.06	0.18 ± 0.02	0.19 ± 0.03
Sum (Å)	9.8 ± 0.9	9.8 ± 0.8	3.1 ± 0.6	3.1 ± 0.6
Dihedral angle constraint violations:				
Number > 5°	0	1 ± 1	0	0
Maximum (°)	4.5 ± 1.5	5.0 ± 1.4	1.9 ± 1.0	1.8 ± 1.0
Sum (°)	25.8 ± 4.8	24.7 ± 4.7	5.9 ± 2.3	5.7 ± 2.0
Average pairwise RMSDs (Å)[d]:				
Backbone atoms N, C$^\alpha$, C′	0.76 ± 0.14	0.80 ± 0.16	0.67 ± 0.12	0.67 ± 0.13
All heavy atoms	1.70 ± 0.13	1.76 ± 0.16	1.49 ± 0.12	1.48 ± 0.14

[a] The average value and the standard deviation are given.
[b] Without (direct) or with use of REDAC.
[c] These include both violations of the distance constraints in the NMR input to DIANA and violations of the van der Waals lower distance limits imposed by DIANA.
[d] Only the well-defined parts of the protein structures were used for the superposition and the RMSD calculation, i.e., the residues 7-59 in the HD, and 3-55 in BPTI.

DISCUSSION

The empirically found higher yield of good conformers with the use of REDAC can be rationalized as follows: In many regions of a protein structure, in particular in β-strands, the local conformation is determined not only by the local conformational constraints derived from intraresidual, sequential and medium-range NOEs (Wüthrich, 1986), but also by longer-range constraints, e.g., interstrand distance constraints in β-sheets. Therefore, the local constraints alone may allow for

TABLE 3
COMPARISON OF THE 20 BEST CONFORMERS OBTAINED USING THE DIANA CALCULATIONS OF FIG. 1 WITH AND WITHOUT REDAC

	Average pairwise RMSD ± standard deviation (Å)[a]	
	HD	BPTI
Backbone atoms N, C$^\alpha$, C′	0.77 ± 0.15	0.67 ± 0.14
All heavy atoms	1.74 ± 0.15	1.47 ± 0.16

[a] Each of the 20 conformers calculated with REDAC was compared with each of the 20 conformers obtained with the direct approach (see text). The numbers given are the average and the standard deviation for the resulting 400 pairwise RMSDs, calculated for residues 7-59 in the HD and 3-55 in BPTI.

multiple different local conformations at low target levels in a DIANA calculation, of which some may be incompatible with the longer-range constraints taken into account at higher minimization levels. Obviously, incorrect local conformations that satisfy the experimentally available local constraints are potential local minima, which could only be ruled out from the beginning if the information contained in the long-range constraints were already available at low levels of the mini-

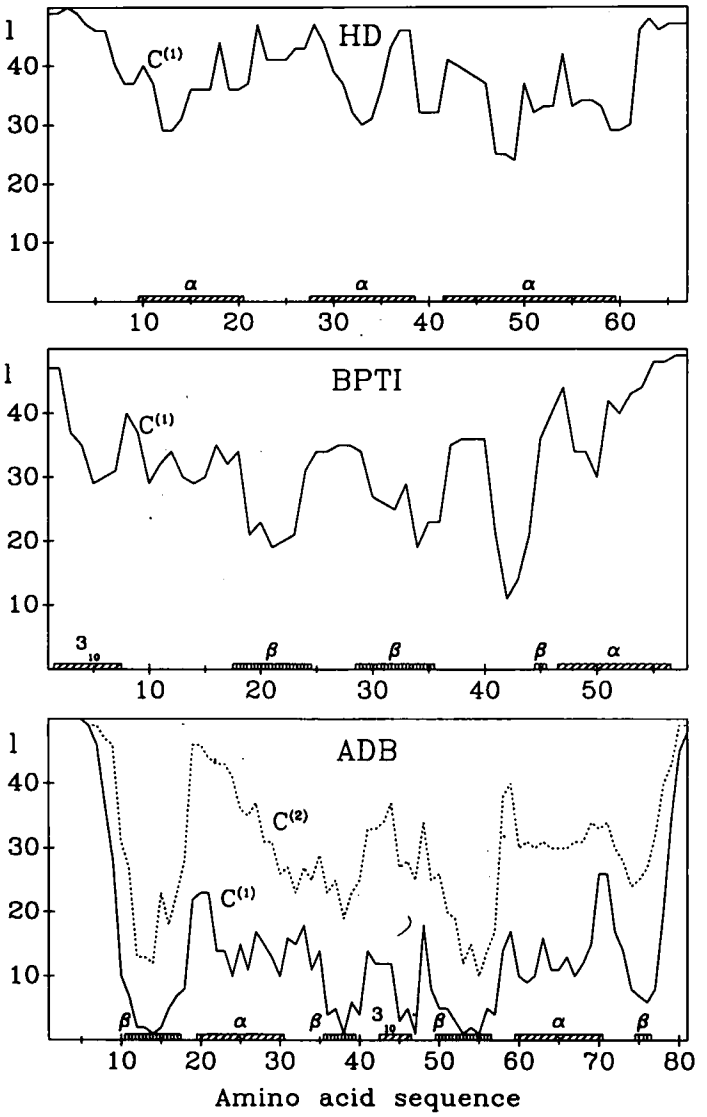


Fig. 4. Plot versus the amino acid sequence of the number of locally acceptable conformers, *l*, found in step $C^{(i)}$ of the DIANA structure calculations using REDAC. An amino acid residue is acceptable if the target function value due to constraint violations that involve atoms or dihedral angles of this residue is less than 0.4 Å^2, and if the same condition holds for the immediately preceding and following residues in the sequence. Helices (α,3_{10}) and β-strands (β) are indicated by hatched bars.

mization. The use of REDAC achieves this: information contained in the complete data set is translated into (by definition intraresidual) dihedral angle constraints. The same argument also explains why earlier attempts to use redundant dihedral angle constraints taken from calculations with L up to about 5 (Kline et al., 1988; Widmer et al., 1989; Billeter et al., 1990) had only limited success. It further makes clear why the yield of good solutions with the direct strategy was in general higher for α-proteins than for β-proteins, since the conformation of an α-helix is particularly well-determined by sequential and medium-range constraints (Wüthrich et al., 1984; Wüthrich, 1986). The plots of the number of locally acceptable conformers versus the amino acid sequence in Fig. 4 show that on the average the highest number of locally acceptable conformers for the generation of REDAC was obtained for the HD, even though the *final* target function values for BPTI were lower. Both in BPTI and in ADB particularly low numbers of locally acceptable conformers are observed in the β-strands, whereas loops, helices, and (often poorly determined) segments with non-regular secondary structure gave, in general, a higher number of locally acceptable conformers. This finding strongly supports the aforementioned explanation of the higher yield of good solutions with the use of REDAC.

In conclusion, the success of DIANA structure calculations using REDAC is primarily due to the feedback of useful structural information derived from conformers calculated up to the maximal level L_{max} into a subsequent round of structure calculations, which starts with local constraints only. In this way information gathered during the entire duration of the structure calculation is used in obtaining the final result, whereas most of this information (up to 95%) is discarded in the direct approach. Figure 3 and Tables 2 and 3 show that groups of conformers of equal quality are obtained with and without use of REDAC, and that the only significant effect of the use of REDAC is a large reduction of the overall computation time (Table 1). The use of REDAC should therefore become the standard strategy for protein structure calculations with the program DIANA and, more generally, with all implementations of the variable target function algorithm.

ACKNOWLEDGEMENTS

We thank Drs. Kurt Berndt and Leonard Orbons for the use of unpublished data of BPTI, and Drs. Martin Billeter and Werner Braun for helpful discussions. Financial support by the Schweizerischer Nationalfonds (project 31.25174.88) and the use of the Cray Y/MP of the ETH Zürich are gratefully acknowledged.

REFERENCES

Arseniev, A., Schultze, P., Wörgötter, E., Braun, W., Wagner, G., Vašák, M., Kägi, J.H.R. and Wüthrich, K. (1988) *J. Mol. Biol.*, **201**, 637–657.

Billeter, M., Kline, A.D., Braun, W., Huber, R. and Wüthrich, K. (1989) *J. Mol. Biol.*, **206**, 677–687.

Billeter, M., Qian, Y.Q., Otting, G., Müller, M., Gehring, W.J. and Wüthrich, K. (1990) *J. Mol. Biol.*, **214**, 183–197.

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

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

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

Güntert, P., Braun, W. and Wüthrich, K. (1991a) *J. Mol. Biol.*, **217**, 517–530.

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

Ikura, T., Gō, N. and Inagaki, F. (1991) *Proteins*, **9**, 81–89.

Kline, A.D., Braun, W. and Wüthrich, K. (1988) *J. Mol. Biol.*, **204**, 675–724.

Kohda, D., Gō, N., Hayashi, K. and Inagaki, F. (1988) *J. Biochem.*, **103**, 741–743.

Li, Z. and Scheraga, H.A. (1987) *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 6611–6615.
McLachlan, A.D. (1979) *J. Mol. Biol.*, **128**, 49–79.
Qian, Y.Q., Billeter, M., Otting, G., Müller, M., Gehring, W.J. and Wüthrich, K. (1989) *Cell*, **59**, 573–580.
Vendrell, J., Wider, G., Avilés, F.X. and Wüthrich, K. (1990) *Biochemistry*, **29**, 7515–7522.
Vendrell, J., Billeter, M., Wider, G., Avilés, F.X. and Wüthrich, K. (1991) *EMBO J.*, **10**, 11–15.
Wagner, G., Braun, W., Havel, T.F., Schaumann, T., Gō, N. and Wüthrich, K. (1987) *J. Mol. Biol.*, **196**, 611–639.
Widmer, H., Billeter, M. and Wüthrich, K. (1989) *Proteins*, **6**, 357–371.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.
Wüthrich, K. (1989) *Science*, **243**, 45–50.
Wüthrich, K., Billeter, M. and Braun, W. (1984) *J. Mol. Biol.*, **180**, 715–740.