Journal of Biomolecular NMR, 3 (1993) 307–324
ESCOM



J-Bio NMR 121

# Extensive distance geometry calculations with different NOE calibrations: New criteria for structure selection applied to Sandostatin and BPTI

Hans Widmer[a,*], Armin Widmer[a] and Werner Braun[b]

[a]*Preclinical Research, Sandoz Pharma Ltd., CH-4002 Basel, Switzerland*
[b]*Institut für Molekularbiologie und Biophysik, ETH-Hönggerberg, CH-8093 Zürich, Switzerland*

Received 4 January 1993
Accepted 2 March 1993

*Keywords:* NMR structure determination; Structure selection; Distance geometry; NOE calibration; Sandostatin

## SUMMARY

To generate structures efficiently, a version of the distance geometry program DIANA for a parallel computer was developed, new objective criteria for the selection of NMR solution structures are presented, and the influence of using different calibrations of NOE intensities on the final structures are described. The methods are applied to the structure determination of Sandostatin, a disulfide-bridged octapeptide, and to model calculations of BPTI. On an Alliant FX2800 computer using 10 processors in parallel, the calculations were done 9.2 times faster than with a single processor. Up to 7000 Sandostatin structures were calculated with distance and angular constraints. The procedure for selecting acceptable structures is based on the maximum values of pairwise RMSDs between structures. Suitable target function cut-offs are defined independent of the number of starting structures. The method allowed for an objective comparison of three groups of Sandostatin structures that were calculated from different sets of upper distance constraints which were derived from the same NOE intensity data using three empirical calibration curves. The number of converged structures and the target function values differed significantly among the three groups, but the structures were qualitatively and quantitatively very similar. The conformation is well determined in the cyclic region Cys$^2$–Cys$^7$ and adopts a β-turn centered at D-Trp$^4$–Lys$^5$. The criteria for structure selection were further tested with BPTI. Results obtained from sets of structures calculated with and without using the REDAC strategy are consistent and suggest that the structure selection method is objective and generally applicable.

## INTRODUCTION

The standard procedure for NMR structure determination of biopolymers (Braun et al., 1981, 1983; Clore et al., 1985; Havel and Wüthrich, 1985; Kaptein et al., 1985; Braun et al., 1986) consists of several steps. After the assignments of resonances, NOESY cross-peak intensities and J-coupling constants are measured. These NMR data are then used as constraints in structure calculations by distance geometry (Havel et al., 1983; Havel and Wüthrich, 1984; Braun and Go,

*To whom correspondence should be addressed.

1985; Braun, 1987) or restrained molecular dynamics (Clore et al., 1985; Kaptein et al., 1985). A crucial step in this procedure, the conversion of the NOE intensities into distance bounds, can be done in several ways. The simplest method is to deduce a uniform upper bound, regardless of the intensity, for the distance between the two spins to which the NOE is assigned. Another, by now a classic, strategy is to classify the NOEs according to intensities, e.g., strong, medium, and weak, and to associate them with different constraints (Wüthrich, 1986). Alternatively the effect of spin diffusion can be accounted for by relaxation matrix calculations (Keepers and James, 1984; Borgias and James, 1988, 1990; Boelens et al., 1989). Analogous considerations lead to lower distance bounds, which can either be taken to be the sum of van der Waals radii, or larger minimum distances can be inferred from weak or missing NOEs (Borgias and James, 1990). The quantification of distance constraints and the criteria for structure selection influence the final 3D structure.

A second unresolved problem concerns the selection of structures which are considered to represent a reliable sampling of the allowed conformation space. Often an arbitrary number of structures (e.g., 20) with lowest target function values (TF) or structures with a residual TF less than a cut-off value are selected. These criteria need arbitrary cut-off numbers which cannot be applied in cases where different calibration schemes are used for the same experimental NOE cross-peak intensities since they lead to quite different residual target function values.

Both of these problems are addressed here for determining the structure of Sandostatin, an octapeptide analogue of somatostatin, with NMR. Sandostatin is a potent inhibitor of growth hormone secretion and is used in the treatment of acromegaly and gastrointestinal diseases. The structure determination is based on a set of assigned and integrated NOE cross peaks from which three lists of distance constraints are generated using different NOE calibration parameters. To efficiently compute a large number of structures, the program DIANA was ported to a parallel computer. Calculations were performed to probe the number of structures needed for sufficient sampling. Then the groups of Sandostatin structures calculated from the three constraints lists were compared. Our results suggest a new and more objective criterion for selection of structures which is based on examination of the maximum pairwise RMSD. The general usefulness of this procedure is shown with calculations of a protein, BPTI.

## MATERIALS AND METHODS

### *Collection of NMR data of Sandostatin*

Sandostatin (D-Phe-Cys-Phe-D-Trp-Lys-Thr-Cys-Thr(ol), SMS 201-995) was synthesized by Bauer et al. (1982). A 10-mM peptide solution was prepared in $H_2O$ buffered with 100 mM deuterated acetate, pH 4.6, and 15% methanol-$d_4$ was added to avoid sample freezing at the measurement temperature of −13 °C. NMR spectra were recorded on a Bruker AM500 spectrometer. 2QF-COSY (Rance et al., 1983) and NOESY (Kumar et al., 1980) spectra using water presaturation were recorded with 800 increments, corresponding to 49.6 ms in $t_1$; the acquisition time in $t_2$ was 164 ms with a recycle time per scan of 1.33 s. The NOESY spectrum was measured with a 40-ms mixing time. A sample of Sandostatin in a fully deuterated solvent mixture was used

*Abbreviations:* BPTI, bovine pancreatic trypsin inhibitor; NOE, nuclear Overhauser effect; NOESY, two-dimensional NOE spectroscopy; RMSD, root-mean-square deviation; $RMSD_{max}$ and $RMSD_{ave}$, maximum and average RMSD, respectively; TF, target function value.

to measure scalar H,H-coupling constants in 1D experiments at various temperatures between −13 and 24 °C.

The data were processed on an Aspect X32 computer using Bruker UXNMR software. Cosine window functions were applied in both dimensions of the NOESY data, and the frequency domain spectrum was baseline corrected with a polynomial of third degree. It was analyzed on a Sun SparcStation 1+ using the EASY program package (Eccles et al., 1991) for assignments and NOE peak integration.

The NOE intensities, V, were converted to upper distance bounds, b, by CALIBA (Güntert et al., 1991) using the relation $V = c/b^n$. Three sets of constraints were generated using different combinations of the proportionality constant c and the power n. In the first and second set, A and B, n was equal to 6; in the third set, C, n was equal to 4, which provides an empirical relation suitable for NOEs involving peripheral side-chain protons (Güntert et al., 1991). For set B the constant c was chosen such that the intraresidual distances between $H^\alpha$ and NH were 2.8 Å or slightly longer; for set A the constant c was 1.5 times smaller, leading to tighter constraints; the curve for group C intersects the curve of group B at b = 2.9 Å. In all three groups the distance limits were further restricted to the range 2.4–5.0 Å. No explicit lower bounds were applied.

*Adaptations of DIANA for parallel processing*

The program DIANA 2.0 (Güntert et al., 1991) was adapted for parallel processing. The computer used was an Alliant FX-2800 configured with 29 Intel i860 RISC processors; the hierarchical memory system consisted of 512 MB global memory, 16 MB global cache, and a small cache (8 KB data, 4 KB instruction) on each chip. DIANA was parallelized on the level of the outmost loop over the structures. Each structure is calculated by an independent process, and several processes are running in parallel. The creation of the overview file, which is updated during the calculation and provides a list of structures sorted according to their final TFs, requires special consideration.

The calculation is divided into three stages. First, the starting conformation is set up. Second, the variable target function is minimized. Third, the output is made of the resulting structure and of an overview file which may be monitored during a long calculation. Using a master/slave model the calculation is distributed over several simultaneous processes. The master iterates over the main loop in which the first stage of a structure calculation is performed and a slave process is forked off (UNIX system service call 'fork'). The independent slaves proceed to the second and third stages of their current structure generation and then exit. If the number of working slaves has reached a user-specified threshold, the master waits for termination of a slave before starting a new process. The master also synchronizes the output by the slaves, since only one process may write into the overview file at any given time.

The implementation of this scheme in DIANA requires a few additional subroutines handling the process control between the master and the slaves. Other changes in the serial version are minimal. All variables that must be shared between processes are collected into a shared memory region. These are some new variables needed for process control and all those which are updated by the subroutine LISVIO during the generation of the overview file. In addition, all output is deferred to the third stage. Since none of these changes makes use of Alliant FX-2800 specific features the resulting parallel code is machine independent within the class of moderately parallel-shared memory machines.

*Structure calculations of Sandostatin and BPTI*

The three lists of upper bounds of Sandostatin were used separately in structure calculations following the same protocol. First, HABAS (Güntert et al., 1989) was used with the NOE distance constraints and with the J-coupling constants to obtain dihedral angle constraints. Structures were calculated with the parallel version of DIANA in 10 minimization steps. The target function evaluation in the final step used weighting factors of 1 for NOE-derived distance constraints and for dihedral constraints, and 100 for the van der Waals terms. The disulfide bridge between $Cys^2$ and $Cys^7$ was imposed by lower and upper distances with a relative weight of 10; the upper limits were 0.1 Å longer than the following lower limits: S-S, 2.0 Å; S-$C^\beta$ and $C^\beta$-S, 3.0 Å; $C^\beta$-$C^\beta$, 3.8 Å. Stereospecific assignments were obtained from analyses of preliminary structures with GLOMSA (Güntert et al., 1991). Instead of the reduced C-terminus, $-CH_2OH$, of Sandostatin, a carboxylate was used in the calculations.

For the BPTI calculations the set of distance and angular constraints was provided by Berndt et al. (1992). The DIANA calculations were performed with and without the use of REDAC and applying the same strategies and parameters as described by Güntert and Wüthrich (1991), i.e., the standard minimization protocol and subsequent DIANA refinement at the maximum level of the target function. In the calculations evaluating CPU times for parallel DIANA, the REDAC cycle and the refinement step were omitted. The relative weights in the final target function evaluation were 1.0 for lower and upper limits, 5.0 for angular and 2.0 for van der Waals constraints.

Violations and RMSDs of the resulting conformers were calculated with the routines built in DIANA. The structures were sorted in ascending order of the final target function value. For each structure the pairwise RMSDs to all structures with lower target function values were computed, and the average and maximum of these RMSDs as a function of the target function cut-off was plotted.

The occurrence of very similar Sandostatin structures was checked. Two structures were regarded as belonging to the same class if none of the $\varphi$- and $\psi$-angles intervening the $C^\alpha$s of $Cys^2$–$Cys^7$ differed by more than 2.5°.

## RESULTS

*Parallelized version of DIANA*

To evaluate the efficiency of the parallel version of DIANA, a series of calculations with an increasing number of slaves was performed. Each calculation generated 200 structures of BPTI from the same random seed following the standard minimization protocol. The measured elapsed and CPU times are listed in Table 1. 'Speed-up' is the ratio between the elapsed time measured for the serial calculation and the elapsed time measured for the calculation using N slaves and the master. For example, a speed-up of 9.2 is obtained for N = 10. For larger numbers of slaves the speed-up levels off (Table 1). The non-linearity of the performance is caused by the part of the code running serially, i.e. the set-up and output stages. Also, an increase in the number of processes leads to increasing competition for resources, such as cache and memory bandwidth. This competition effect was estimated by comparing the CPU times of a serial (i.e. 1 process) calculation when run in an otherwise idle system, with the CPU times of N simultaneous copies of the same serial calculation. For N = 16 in a calculation of 10 BPTI structures the average

elapsed time was 13.7% longer than for a single copy of this calculation. This effect is also manifested in the rising CPU times of parallel calculations when the number of slaves is increased (Table 1).

*Generation of constraints used for Sandostatin structure calculations*

All Sandostatin resonances were assigned unambiguously to individual protons. There was one set of resonances, indicating that on the time scale of chemical shifts there is only one resolved conformation. Most NOEs were integrated on both sides of the diagonal. The volumes were within 20% of each other for peaks with good signal-to-noise ratios, or within 30% for weak peaks: The more intense peak was used. Stereospecific assignments of the β-protons of $Trp^4$, $Lys^5$, and $Cys^7$, and the γ-methylene group of $Lys^5$ could be made after initial structure calculations by statistical analyses of the structures and comparison with NOEs and J-couplings (programs HABAS and GLOMSA). NOEs were referred to pseudoatoms (Wüthrich et al., 1983) if stereospecific assignments could not be obtained, e.g., for the β-methylene protons of $Cys^2$ which overlap at all temperatures examined. About 100 NOEs with unique assignments were integrated. The intensities were then translated into distances using three calibrations as outlined in the Methods section; distances not constraining the conformation were eliminated by a routine of DIANA. Three intraresidual NOEs of D-$Trp^4$ were excluded because they were mutually incompatible (control calculations showed that inclusion of these constraints did not alter the structures to any significant extent). The final lists consisted of 67, 57, and 51 NOEs for the three sets A, B, or C, respectively, and are given explicitly in Table 2. Set A is the most restrictive one; 20 distance bounds are shorter by 0.3 Å than in set B, and there are differences up to 1.0 Å as compared to C. In data set B 23 constraints are tighter by ≥0.3 Å and 5 are less restrictive than in C. The scalar J-coupling constants did not change measurably in the temperature range −13 to 24 °C and the values at 24 °C are given in Table 3.

TABLE 1
PERFORMANCE OF THE PARALLEL VERSION OF DIANA IN CALCULATIONS OF 200 BPTI STRUCTURES[a]

| N slaves[b] | Elapsed time (min) | CPU time (min) | Speed-up[c] |
|---|---|---|---|
| 0[d] | 440.6 | 439.0 | |
| 2 | 227.4 | 458.2 | 1.94 |
| 4 | 116.0 | 459.2 | 3.80 |
| 8 | 58.1 | 465.6 | 7.58 |
| 10 | 47.9 | 468.9 | 9.20 |
| 16 | 33.5 | 510.5 | 13.15 |
| 20 | 27.6 | 529.7 | 15.96 |
| 28 | 24.6 | 623.0 | 17.91 |

[a] Standard minimization protocol: the average number of iterations per structure is 4898.
[b] Total number of processes is N+1 (master and N slaves).
[c] Ratio between the elapsed time measured for the serial calculation and the elapsed time measured for the calculation using N slaves and the master.
[d] Serial calculation performed by one process.

TABLE 2
THREE SETS OF UPPER DISTANCE CONSTRAINTS USED FOR STRUCTURE CALCULATIONS OF SANDOSTATIN

| Atom 1 | Atom 2 | Constraints sets[a] | | |
|---|---|---|---|---|
| | | A | B | C |
| D-Phe[1] | | | | |
| HA | Cys[2] HN | 2.40 | 2.60 | 2.40 |
| QB | Cys[2] HN | 4.18 | 4.38 | – |
| Cys[2] | | | | |
| HA | Phe[3] HN | 2.60 | 2.80 | 2.70 |
| HB2 | Phe[3] HN | 3.30 | 3.50 | 3.80 |
| HB3 | Phe[3] HN | 3.30 | 3.50 | 3.80 |
| QB | Phe[3] HN | 4.30 | 4.50 | 4.80 |
| HA | Cys[7] HA | 3.70 | 4.00 | 4.60 |
| QB | Cys[7] HN | 5.90 | 6.00 | 6.00 |
| QB | Cys[7] HA | 4.08 | 4.28 | 4.48 |
| Phe[3] | | | | |
| HN | Phe[3] HB2 | 3.50 | 3.70 | 4.10 |
| HN | Phe[3] HB3 | 3.50 | 3.70 | 4.10 |
| HN | Phe[3] QB | 3.29 | 3.49 | 3.85 |
| HA | Phe[3] HB2 | 3.00 | – | – |
| HA | Phe[3] HB3 | 3.00 | – | – |
| HA | Phe[3] QB | 2.77 | – | – |
| HA | D-Trp[4] HN | 2.40 | 2.60 | 2.40 |
| CG | D-Trp[4] HA | 6 30 | 6.60 | 7.00 |
| HN | Thr[4] HN | 4.00 | 4.30 | 5.00 |
| HN | Thr[6] HB | 4.60 | 4.90 | 5.00 |
| HN | Cys[7] HA | 4.30 | 4.60 | 5.00 |
| D-Trp[4] | | | | |
| HN | D-Trp[4] HB2 | 3.40 | 3.60 | 4.00 |
| HN | D-Trp[4] HB3 | 2.70 | 2.90 | 2.80 |
| HA | D-Trp[4] HB2 | 3.00 | – | – |
| HA | D-Trp[4] HE3 | 3.00 | 3.20 | 3.40 |
| HB2 | D-Trp[4] HD1 | 3.50 | 3.70 | – |
| HB3 | D-Trp[4] HD1 | 3.20 | 3.40 | 3.60 |
| HA | Lys[5] HN | 2.40 | 2.50 | 2.40 |
| HA | Lys[5] HG3 | 4.40 | 4.70 | 5.00 |
| HB2 | Lys[5] HN | 4.60 | 4.90 | 5.00 |
| HB3 | Lys[5] HN | 4.50 | 4.80 | 5.00 |
| HD1 | Lys[5] HA | 4.80 | 5.00 | 5.00 |
| HE1 | Lys[5] HG2 | 4.50 | 4.80 | 5.00 |
| HE1 | Lys[5] HG3 | 4.10 | 4.40 | 5.00 |
| HE1 | Lys[5] QE | 6.00 | 6.00 | 6.00 |
| HE3 | Lys[5] HA | 4.60 | 4.90 | 5.00 |
| HZ2 | Lys[5] QE | 5.20 | 5.40 | 6.00 |
| HA | Thr[6] HN | 4.30 | 4.60 | 5.00 |

TABLE 2 (continued)

| Atom 1 | Atom 2 | Constraints sets[a] | | |
|---|---|---|---|---|
| | | A | B | C |
| $Lys^5$ | | | | |
| HN | $Lys^5$ HB2 | 2.70 | 2.90 | 2.90 |
| HN | $Lys^5$ HB3 | 3.50 | 3.70 | 4.20 |
| HN | $Lys^5$ HG2 | 3.40 | 3.60 | 4.00 |
| HN | $Lys^5$ HG3 | 2.90 | 3.10 | 3.10 |
| HN | $Lys^5$ QE | 6.00 | 6.00 | 6.00 |
| HA | $Lys^5$ HB2 | 3.00 | – | – |
| HA | $Lys^5$ HB3 | 2.90 | – | – |
| HA | $Lys^5$ HG2 | 2.90 | 3.10 | 3.10 |
| HA | $Lys^5$ HG3 | 3.20 | 3.50 | 3.70 |
| HB2 | $Lys^5$ HG2 | 2.80 | 3.00 | – |
| HB3 | $Lys^5$ HG2 | 2.90 | – | – |
| HB3 | $Lys^5$ HG3 | 3.00 | – | – |
| HG3 | $Lys^5$ QE | 3.82 | 3.91 | – |
| HN | $Thr^6$ HN | 2.90 | 3.10 | 3.20 |
| HB3 | $Thr^6$ HN | 4.20 | 4.50 | 5.00 |
| HA | $Cys^7$ HN | 4.60 | 4.90 | 5.00 |
| $Thr^6$ | | | | |
| HN | $Thr^6$ HA | 2.80 | – | – |
| HN | $Thr^6$ HB | 3.40 | 3.60 | 4.00 |
| HN | $Cys^7$ HN | 3.10 | 3.30 | 3.50 |
| HA | $Cys^7$ HN | 2.90 | 3.10 | 3.30 |
| HB | $Cys^7$ HN | 4.40 | 4.70 | 5.00 |
| HA | $Thr^8$ HN | 4.40 | 4.70 | 5.00 |
| $Cys^7$ | | | | |
| HN | $Cys^7$ HB2 | 2.90 | 3.10 | 3.10 |
| HN | $Cys^7$ HB3 | 3.80 | 4.10 | – |
| HA | $Cys^7$ HB3 | 2.90 | – | – |
| HN | $Thr^8$ HN | 3.70 | 4.00 | 4.70 |
| HA | $Thr^8$ HN | 2.80 | 3.00 | 3.10 |
| HB2 | $Thr^8$ HN | 4.60 | 4.90 | 5.00 |
| HB3 | $Thr^8$ HN | 3.80 | 4.10 | 4.80 |
| $Thr^8$ | | | | |
| HN | $Thr^8$ HB | 3.70 | 3.90 | – |

[a] Only conformationally restricting constraints are listed (Å); nonrestricting constraints are indicated by a dash (–). In case of missing stereospecific assignments, the distances as modified by DIANA are given.

*Calculations with different numbers of starting structures*

A first series of calculations probed the number of starting structures needed for sufficient sampling of the conformational space. The constraints set B was used in a series of DIANA

calculations starting from 100, 1000, or 7000 random structures. The conformational sampling was assessed by calculating pairwise RMSDs of backbone atoms of residues 2–7. The maximum and the average RMSDs, $RMSD_{max}$ and $RMSD_{ave}$, among all pairs of structures with TFs smaller than a cut-off value were determined. The dependences of $RMSD_{max}$ on the TF cut-off is plotted in Fig. 1 (cf. Table 4). This graph is characterized by steps, which indicates that additional conformational space becomes accessible at increasing target function values, i.e. when more constraint violations are tolerated. The improved sampling by using larger numbers of starting conformations manifests itself by higher values of $RMSD_{max}$ at the same TF cut-offs (Fig. 1; Table 4). For example at the cut-off value 0.2 $Å^2$ the maximum RMSD is 1.28 Å or 1.91 Å, respectively, when 100 or 1000 conformations are sampled, and it increases to 2.02 Å with 7000 starting structures. However, the latter increase obtained from 6000 additional structures is small and so 1000 starting structures were considered to be sufficient for further calculations of Sandostatin. Also note that all three curves show the same plateau at around 1.2 Å RMSD. For completeness the average RMSD is also listed in Table 4; it shows no general trend when the number of starting structures is changed. With higher numbers of conformers there is also a higher chance of finding the same structures more than once. For example, 219 structures are found with TFs up to 0.2 $Å^2$ when starting from 1000 structures, but 31 (14%) of them are redundant; with 7000 starting structures only ca 50% of the conformations are unique (Table 4).

*Calculations with constraints sets obtained from different NOE distance calibrations*

In addition to the structures obtained from set B (above), 1000 DIANA structures were calculated with each of the constraints sets A and C. In the three calculations the final TFs differ by several orders of magnitude (Table 5; Fig. 2). In set C there is an exact solution with TF equal

TABLE 3
COUPLING CONSTANTS OF SANDOSTATIN

| Residue | Proton pair | J-coupling (Hz)[a] |
|---|---|---|
| $Cys^2$ | HN-HA | 7.3 |
| $Cys^2$ | HA-HB2 / 3[b] | 4.5 / 6.0 |
| $Phe^3$ | HN-HA | 8.0 |
| $Phe^3$ | HA-HB2 / 3[b] | 6.1/9.2 |
| D-$Trp^4$ | HN-HA | 4.2 |
| D-$Trp^4$ | HA-HB2 / 3 | 5.6 / 11.3 |
| $Lys^5$ | HN-HA | 6.5 |
| $Lys^5$ | HA-HB2 / 3 | 11.2 / 3.5 |
| $Thr^6$ | HN-HA | 8.0 |
| $Thr^6$ | HA-HB | 4.9 |
| $Cys^7$ | HN-HA | 8.0 |
| $Cys^7$ | HA-HB2 / 3 | 11.7 / 3.4 |
| $Thr^8$ | HN-HA | 9.5 |
| $Thr^8$ | HA-HB | 3.6 |

A solution of 10 mM Sandostatin in $H_2O$+15% methanol at 24 °C was used.
[a] Measured from multiplet splittings in 1D spectra.
[b] HB2 and HB3 are not stereospecifically assigned.

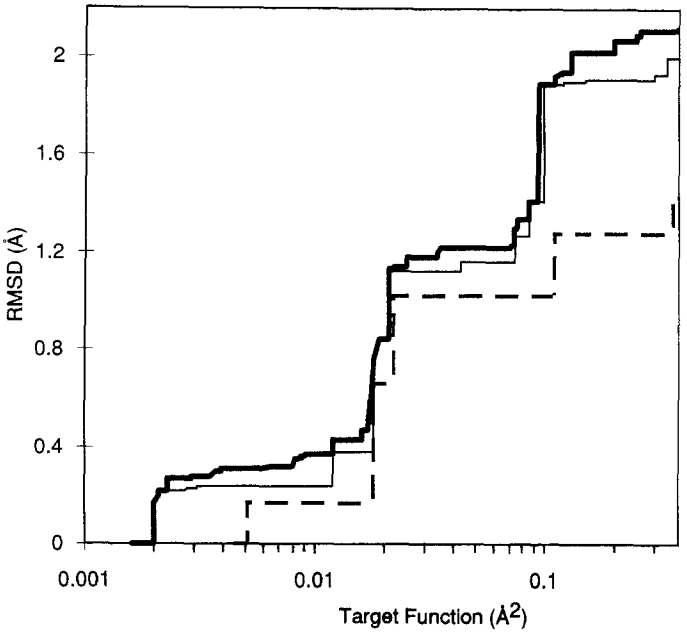


Fig. 1. Maximum of the RMSD between pairs of Sandostatin structures with a target function smaller then a cut-off value. The sample of structures for RMSD calculations gets progressively larger with increasing cut-off value which is plotted on the abscissa in a logarithmic scale. The numbers of starting conformations are 100 (broken line), 1000 (solid thin line), and 7000 (solid thick line); the constraints set B (Table 2) was used in the DIANA calculations. The RMSDs are calculated for the backbone atoms N, $C^{\alpha}$, and C' of residues 2–7.

to zero, whereas in the most restrictive set A the best TF is 0.9 Å$^2$. The dependence of the maximum and average pairwise RMSD on the TF cut-off are plotted in Fig. 2 (upper panel). In all three sets of calculations the graph of $RMSD_{max}$ levels off before it rises sharply to >1.8 Å. In this region with stable $RMSD_{max}$ values the TF cut-off was chosen (indicated by arrows, Fig. 2) to select structures from each group, namely 304 structures from A, 87 from B, and 14 from C. A histogram showing the distribution of structures along the logarithmic scale of the TF is plotted in the lower panel of Fig. 2; both the total number of structures and the number of unique structures (for definition, see the Methods section) are indicated. For set A a much more even distribution is obtained than for groups B and especially C, for which structures with relatively high TFs predominate. In set A, and to a lesser extent in set B, many conformations are found several times in different runs, whereas set C contains almost no redundant conformations. Among the structures selected from each set, only the unique conformers were subjected to further analyses and comparisons; these are 107 structures from A, 78 from B, and all 14 from C (Table 5) and they are plotted in stereo (Fig. 3).

Almost all structures completely fulfill the steric and the disulfide bridge constraints; the maximum violation is 0.04 Å in a structure of group A. In group C, the distance and dihedral angle constraints are also fulfilled equally well (Table 5). In group A, however, the sum of NOE violations is 2.9 Å in the best structure; while no individual NOE is consistently violated, each structure has 9 or more violations >0.1 Å and at least one of 0.4–0.5 Å. The sum of dihedral angle violations ranges from 4.5 to 76 degrees (Table 5). In group B, moderate violations occur (Table

5). The RMSD comparisons for the segment from $Cys^2$ to $Cys^7$ are listed in Table 6. For the backbone, the maximum pairwise RMSDs within the individual groups (A, B, or C) of structures are 1.0 Å, 1.2 Å, or 1.4 Å, respectively, and the minima are 0.1 Å or less. In the cross comparisons between the groups the minimum RMSDs are very low (<0.2 Å, backbone), and the maxima do not exceed the larger of the two maxima found within the groups which are compared.

The structures of Sandostatin (Fig. 3) are characterized by two opposing strands of residues 1–4 and 5–8. The strands are held together by the disulfide bridge between $Cys^2$ and $Cys^7$; the exocyclic, terminal residues are disordered. The turn centered at D-$Trp^4$-$Lys^5$ has the following dihedral angles (all groups, n = 199; average ± standard deviation): $\varphi(4) = 88° \pm 16°$; $\psi(4) = -127° \pm 3°$; $\varphi(5) = -89° \pm 3°$; $\psi(5) = 20° \pm 14°$. It thus falls into the category of a type II' β-turn. There are no consistent hydrogen bonds. In individual structures, hydrogen bonds occur between $Thr^6$(NH) and $Phe^3$(CO) or D-$Trp^4$(CO), or between $Lys^5$(NH) and $Cys^7$(CO).

TABLE 4
OVERVIEW OF STRUCTURES CALCULATED FROM CONSTRAINTS SET B[a]

| | Number of starting structures | | |
|---|---|---|---|
| | 100 | 1000 | 7000 |
| Minimum TF (Å$^2$) | 0.0045 | 0.0016 | 0.0015 |
| TF < 0.010 Å$^2$ | | | |
| Number of structures[b] | 2 (2) | 35 (27) | 336 (156) |
| Maximum RMSD | 0.17 | 0.24 | 0.38 |
| Average RMSD | 0.17 | 0.13 | 0.14 |
| TF < 0.027 Å$^2$ | | | |
| Number of structures[b] | 6 (6) | 72 (63) | 548 (286) |
| Maximum RMSD | 1.02 | 1.12 | 1.15 |
| Average RMSD | 0.63 | 0.43 | 0.38 |
| TF < 0.200 Å$^2$ | | | |
| Number of structures[b] | 26 (24) | 219 (188) | 1573 (901) |
| Maximum RMSD | 1.28 | 1.91 | 2.02[c] |
| Average RMSD | 0.64 | 0.65 | 0.64[c] |
| TF < 3.000 Å$^2$ | | | |
| Number of structures[b] | 74 (71) | 777 (627) | 5430 |
| Maximum RMSD | 2.19 | 2.40 | –[d] |
| Average RMSD | 1.07 | 1.21 | –[d] |

[a] cf. Table 2.

[b] Total number of structures with a target function value (TF) smaller than the indicated cut-off. In parentheses the number of unique structures is given which differ from each other in at least one of the backbone dihedral angles intervening the $C^\alpha$s of $Cys^2$–$Cys^7$ by > 2.5°.

[c] Only unique structures as defined in [b] are considered.

[d] Not determined.

Experimentally, all amide protons exchange within one minute (not shown). The temperature dependence of the chemical shift is small for the amide proton of Thr[6] (Wynants et al., 1985a), supporting a transient involvement in a hydrogen bond.

*Calculations with BPTI data*

The usefulness of the procedure for structure selection was tested by application to BPTI. First, the calculations described by Güntert et al. (1991) were repeated with the recently published BPTI data set (Berndt et al., 1992). In short, 2000 structures were calculated without the use of REDAC and a second set of 50 structures was calculated with one REDAC cycle. Then the plots of the maximum and the average RMSD versus the TF cut-off were generated for both sets of calculations (Fig. 4A). The distribution of structures is shown in Figs. 4B (no REDAC) and 4C (one REDAC cycle). $RMSD_{max}$ adopts a constant value of 1.0 Å in the range of TFs from 0.37 to 0.67 $Å^2$ in the calculations without REDAC. Another stable $RMSD_{max}$ of 1.35 Å is observed at TFs from 1.0 to 1.5 $Å^2$. These flat regions coincide with those found in the calculations with REDAC, although the second region has a lower $RMSD_{max}$ of 1.18 Å and extends to TFs >5 $Å^2$ (Fig. 4A). In contrast to the calculations without REDAC, the peak in the distribution of structures calculated with REDAC is at TF 0.7 $Å^2$ and very few structures are sampled with TFs >1.5 $Å^2$ (Fig. 4B,C).

TABLE 5
OVERVIEW OF STRUCTURES CALCULATED FROM CONSTRAINTS SETS A, B, AND C[a]

| | A | B | C |
|---|---|---|---|
| Minimum TF ($Å^2$) | 0.90 | 0.0016 | 0.00 |
| TF cut-off ($Å^2$) | 1.40 | 0.043 | 0.00025 |
| Number of structures | | | |
| Total[b] | 304 | 87 | 14 |
| Unique[c] | 107 | 78 | 14 |
| NOE constraints violations[d] | | | |
| Number > 0.1 Å | 11.6 (9–15) | 0.4 (0–2) | 0 (0–0) |
| Sum (Å) | 3.2 (2.9–3.8) | 0.3 (0.1–0.6) | 0.0 (0.0–0.02) |
| Maximum (Å) | 0.5 (0.4–0.5) | 0.1 (0.0–0.2) | 0.0 (0.00–0.01) |
| Angle constraints violations[d] | | | |
| Number > 5° | 1.8 (0–5) | 0.2 (0–1) | 0 (0–0) |
| Sum (°) | 35.8 (4.5–76.1) | 1.4 (0.0–6.3) | 0.1 (0.0–0.3) |
| Maximum (°) | 23.7 (4.0–32.7) | 1.4 (0.0–6.3) | 0.1 (0.0–0.3) |

[a] cf. Table 2.
[b] Total number of structures with TF lower than the cut-off.
[c] For definition, cf. footnote [b] in Table 4.
[d] Average and range (in parentheses) considering only unique structures.

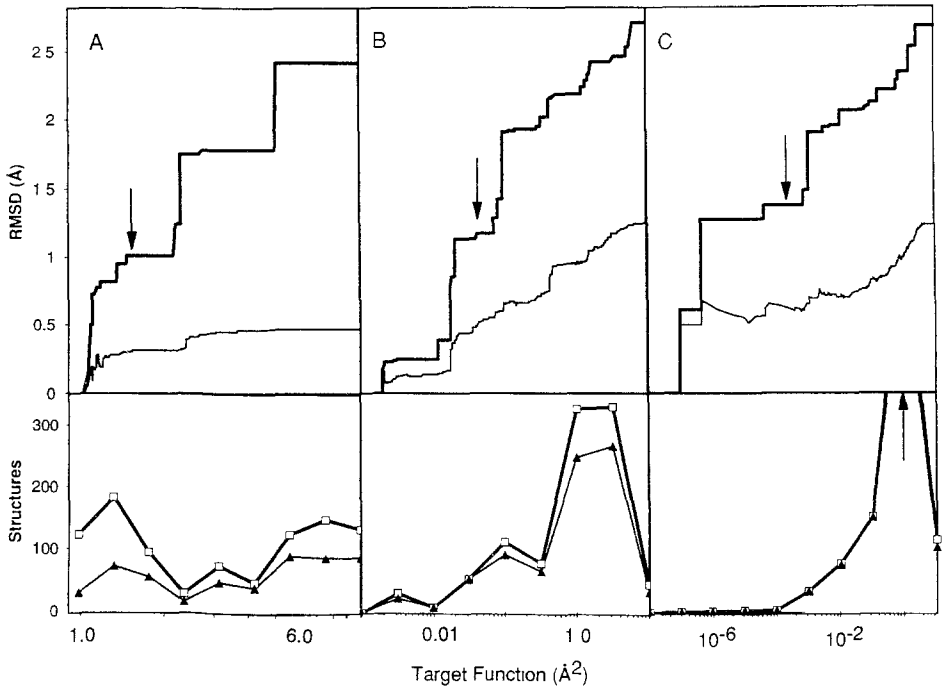


Fig. 2. Pairwise backbone RMSDs and numbers of Sandostatin structures as a function of the TF cut-off. The upper panel shows the maximum (thick line) and the average (thin line) of the RMSDs obtained in all structures with target functions smaller than the cut-off value which is plotted on the abscissa in a logarithmic scale. The RMSDs are calculated for the backbone atoms N, $C^{\alpha}$, and C' of residues 2–7. In the lower panel, a histogram shows the number of all (open squares) or only of the unique (filled triangles) structures with target function values within the interval between two data points (in C, the arrow points to an off-scale value of 599 structures); the areas under the curves represent the integrated number of structures. Structures were calculated separately using constraints sets A (A), B (B), and C (C), each starting from 1000 random conformations. Data from > 900 structures are included in each graph. The arrows in the upper panel indicate the target function cut-off used to select the final set of structures representing each group.

## DISCUSSION

*Parallel processing*

The efficiency of DIANA calculations is improved significantly by parallelization. The speed-up increases almost linearly up to ca. 10 processes. This brings the performance of parallel DIANA on the FX-2800 into the performance range of serial DIANA on currently available supercomputers. Further increase in speed achieved by using higher numbers of processes is underproportional.

The higher throughput of parallel DIANA calculations has valuable practical implications. In the course of an NMR structure determination, the calculations are repeated several times with modifications in the input, such as the NOE list, stereospecific assignments, or a REDAC cycle. A fast feedback from a calculation (e.g. 200 BPTI structures in half an hour) therefore facilitates the careful evaluation of parameters. Furthermore, extensive calculations (e.g. 5000 structures) can be run overnight, allowing statistical analyses of the resulting structures.

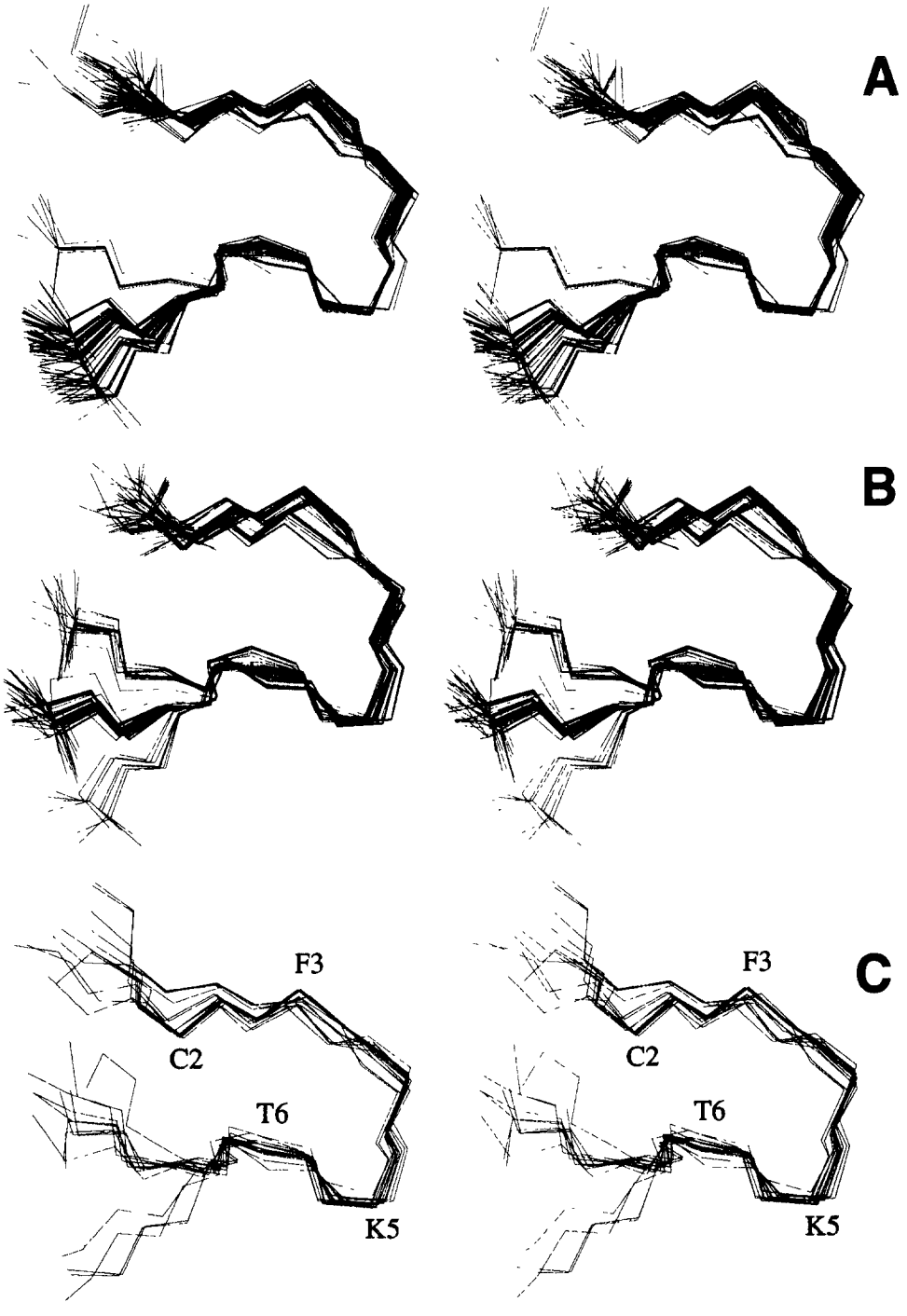


Fig. 3. Stereoplots of the backbone (residues 1–8) of Sandostatin after least squares superposition of the backbone atoms of $Cys^2$–$Cys^7$. Shown are the groups of unique conformers obtained from constraints set A (107 structures), set B (78 structures), and set C (14 structures). The orientations are identical. Selected residues are labeled in (C).

TABLE 6
COMPARISON OF THREE GROUPS OF SANDOSTATIN STRUCTURES[a]

| Groups | Pairwise backbone RMSD (Å)[b] | | | Pairwise heavy atom RMSD (Å)[c] | | |
|---|---|---|---|---|---|---|
| | Min | Max | Ave | Min | Max | Ave |
| A vs. A | 0.00 | 1.00 | 0.38 | 0.01 | 1.39 | 0.73 |
| B vs. B | 0.00 | 1.18 | 0.50 | 0.03 | 1.53 | 0.85 |
| C vs. C | 0.10 | 1.36 | 0.61 | 0.32 | 1.74 | 0.98 |
| A vs. B | 0.05 | 1.16 | 0.62 | 0.19 | 1.53 | 0.96 |
| A vs. C | 0.14 | 1.23 | 0.70 | 0.36 | 1.56 | 1.03 |
| B vs. C | 0.18 | 1.36 | 0.80 | 0.44 | 1.73 | 1.11 |

[a] The unique structures as defined in Table 5 are analyzed.
[b] Backbone atoms $C^{\alpha}$, N, and C' of residues 2–7 are considered.
[c] Backbone atoms as defined in footnote [b] and heavy atoms of the side chains of $Cys^2$, D-$Trp^4$, $Lys^5$, $Thr^6$, and $Cys^7$ are considered.

*Selection of structures*

In order to provide a meaningful comparison of Sandostatin structures, common selection criteria applicable to all three sets of structures were required. As the structures were calculated from the same experimental NOESY spectrum with different calibrations, the groups of structures differ significantly in distance constraints violations and concomitantly in the target function values. In group C an exact solution (TF = 0.0 $Å^2$) was obtained, but the best structure in group A has a TF of 0.9 $Å^2$ with an individual NOE violation of 0.49 Å and all violations summing to 3.3 Å. As this example shows, the absolute TF alone cannot provide a sufficient criterion for selection of structures, as it depends on the calibration. Furthermore, in these and in the BPTI calculations we did not find a definite TF step which would suggest a distinction between acceptable and inacceptable solutions. Selection of a predefined number of structures also seemed inappropriate as they would represent a rather arbitrary sample of conformational space. Instead, we plotted the maximum pairwise backbone RMSD as a function of the TF cut-off. A general discussion of these graphs is provided by schematic drawings for three typical cases (Fig. 5). For a single family of conformations, the $RMSD_{max}$ rises continuously as a function of the TF cut-off (Fig. 5A). Each increment of TF expands the conformational space by a small amount. The $RMSD_{max}$ takes a step (Fig. 5B) when a particular conformation becomes accessible which differs significantly from all structures with lower target function values, and then levels off again. In actual calculations, the conformational space is sampled at discrete points, i.e. as a limited number of conformational states in local minima. In this case the maximum RMSD rises discontinuously until a steady state is reached when the most distant subconformations are sampled (Fig. 5C). The selection of NMR structures aims at defining a group of widely sampled structures with low residual violations, whereby the selected conformation space should be relatively stable with respect to the TF cut-off. These criteria are met in the flat regions in the plot of $RMSD_{max}$ vs. TF cut-off. This method should be equally applicable to select structures calculated with other programs, e.g., DGII (Havel, 1991) or X-PLOR (Kuszewski et al., 1992). The parameter to rank the structures would be the target function defined in the metric matrix approach, or the total of the NOE restraint energy defined in simulated annealing.

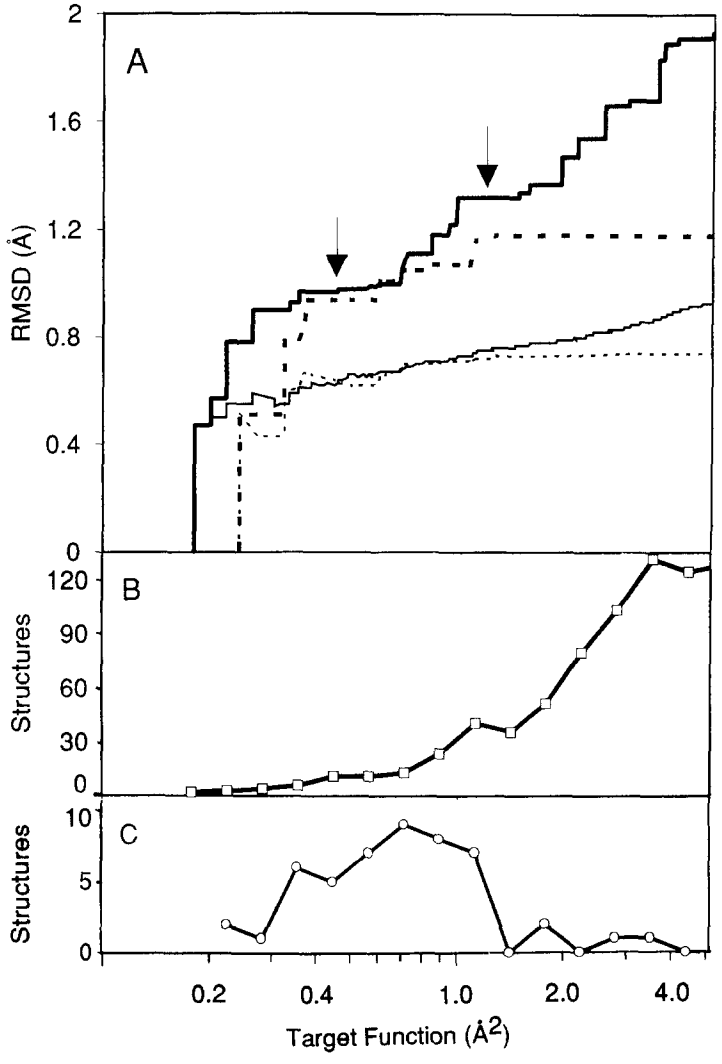


Fig. 4. Pairwise backbone RMSDs and yield of BPTI structures as a function of the TF cut-off. The data result from DIANA calculations without use of the REDAC approach (2000 starting structures) or with use of REDAC (50 starting structures). In (A), the maximum RMSD (thick solid line, with REDAC; thick broken line, without REDAC) and the average RMSD (thin solid line, without REDAC; thin broken line, with REDAC) is shown as obtained for the structures with TFs smaller than the cut-off value which is plotted on the abscissa in a logarithmic scale. The RMSDs are calculated for the backbone atoms N, $C^{\alpha}$, and C' of residues 3–55. The arrows point to possible cut-offs which can be used to select the final structures. (B) and (C), graphs of the number of the structures obtained with TFs within the interval between two data points; (B), without REDAC; (C), with REDAC. The vertical scale in B and C is different. The areas under the curves represent the integrated numbers of structures. The range of TFs displayed includes 450 structures without use of REDAC and 48 structures with use of REDAC.

In an independent study recently presented by Kohda and Inagaki (1992), the selection of structures is based on the dependence of the average RMSD on the target function cut-off. However, as Fig. 2 shows, the maximum RMSD is much more sensitive to expansion of confor-

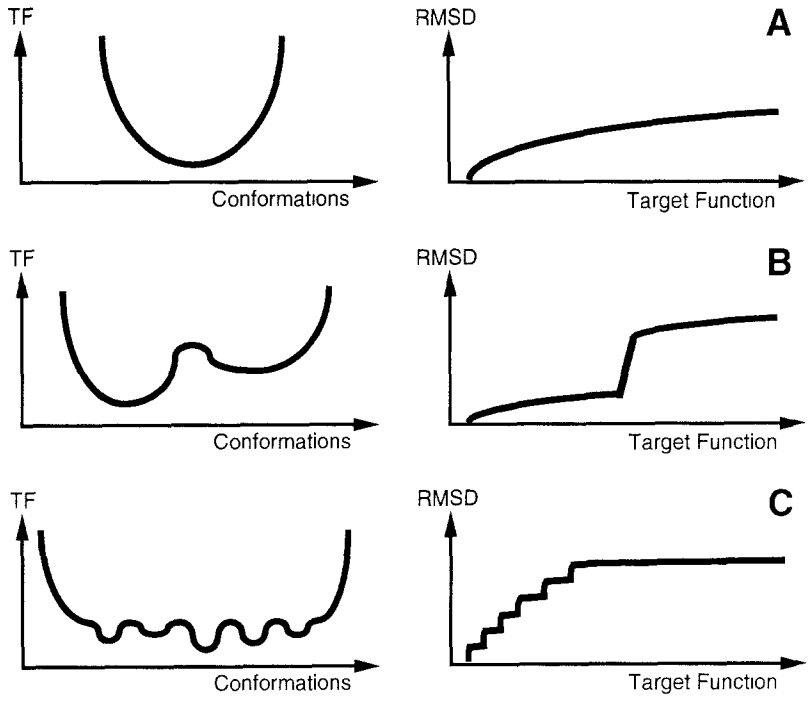


Fig. 5. Schematic drawings of the relations between conformations, $RMSD_{max}$, and target function values for three typical cases. (A), single conformation; (B) two conformations; (C) set of substates. See text (Discussion).

mational space because the corresponding step of the average RMSD is strongly attenuated by all previous pairs of structures. The average RMSD is also influenced by the distribution of the different structures and thus reflects a property of the search algorithm, rather than the size of the conformational space.

The curves obtained for Sandostatin (Fig. 2) and BPTI (Fig. 4A) can be interpreted as superpositions of the idealized graphs of Fig. 5. They basically represent the case of two or more states (Fig. 5B), where each individual state is composed of a large number of substates (Fig. 5C). The TF cut-offs used for Sandostatin were based on these plots (Fig. 2, Table 5) and differed by factors of >5000 between the most (A) and the least (C) restrictive sets. This indicates that this method provides criteria for structure selection also for inconsistent data sets. In particular, despite the exact solution for data set C, the plot suggests that a higher target function cut-off, where the dependence of the $RMSD_{max}$ is low, is more advantageous. The yield of structures also depends strongly on the data set. From constraints set A 304 structures fulfilled the defined criterion, and 107 (35%) of these are unique; the respective numbers are 87 (78 unique) structures for set B, and only 14 structures for set C. This indicates that the restrictive data set leads to a higher convergence rate and to multiple definition of structures. Much wider sampling is obtained with the looser constraints sets B and particularly with C.

Experience shows that restricting the dihedral angles to certain ranges can considerably improve the convergence rate of distance geometry calculations in dihedral angle space (Kline et

al., 1988; Widmer et al., 1989; Güntert and Wüthrich, 1991) or of Monte Carlo calculations (von Freyberg and Braun, 1991). The selection of structures is thereby quite critical as the calculations are done already in a restricted sample space. The test calculation with the BPTI NMR data set shows that the REDAC procedure finds the same $RMSD_{max}$ value in the first plateau (Fig. 4A). However, the second plateau value is underestimated, because very few structures with higher TFs are sampled (Fig. 4A,C). Whether or not it is desirable to obtain these structures depends on the judgement of acceptable error and on how restrictively the NOEs were calibrated. In both cases the plot $RMSD_{max}$ vs. TF provides a basis for the selection of NMR structures which is independent of an arbitrary choice of the numbers of starting or final structures.

*Comparison of structures obtained with different distance calibrations of NOE intensities*

The final Sandostatin structures obtained with the three constraints sets are very similar, as the visual impression suggests (Fig. 3). The maximum pairwise RMSD within individual groups increases in the order A to C (Table 6) and reflects the degree of restrictions which are applied by the corresponding constraints set. In the group-to-group comparisons (Table 6), the minimum RMSDs of backbone atoms of residues 2–7 are as low as 0.05 Å, and the maximum RMSDs indicate that the more restricted group of structures is also contained in the less restricted group. Therefore the NOE calibration influenced the size of the spanned conformational space but did not change the structures systematically. The existence of a βII' turn has previously been proposed for Sandostatin in DMSO (Wynants et al., 1985a) but could not be confirmed in aqueous solution at 24 °C (Wynants et al., 1985b). In this investigation, a much larger number of NOEs was observable due to the slower molecular tumbling rate at −13 °C in the solvent mixture of water and 15% methanol-$d_4$, and a complete 3D structure was determined.

## CONCLUSION

We have shown that the different parametrizations of the NOE calibration curves used in this study have little effect on the resulting 3D structures of Sandostatin. The method is therefore quite robust. However, the value of the target function and the sampling properties are strongly influenced by the different NOE calibrations. Therefore the criteria for selection of structures (i.e. the tolerance for constraints violations) are very critical and have to be adjusted to the data set. The need for an unbiased selection of representative structures is further emphasized by the higher efficiency of structure calculations obtained by parallel computing or by use of REDAC. In this study a plot of the maximum RMSD as a function of the TF cut-off was shown to be useful. Clearly this one-parameter descriptor of conformational sampling is no substitute for cluster analyses, or identification of subsets of NOEs which are compatible with distinct conformations (e.g., MEDUSA algorithm; Brüschweiler et al., 1991). However, it provides a very simple criterion for more objective selection of structures.

## ACKNOWLEDGEMENTS

We thank Dr. W. Bauer for providing Sandostatin, Dr. Ph. Flörsheim for help with the structure analyses, L. Oberer for acquiring and assigning the NMR spectra, and Dr. C. Schein for a careful reading of the text.

## REFERENCES


Bauer, W., Briner, U., Dopfner, W., Haller, R., Huguenin, R., Marbach, P., Petcher, T.J. and Pless, J. (1982) *Life Sci.*, **31**, 1133–1140.

Berndt, K.D., Güntert, P., Orbons, L.P.M. and Wüthrich, K. (1992) *J. Mol. Biol*, **227**, 757–775.

Boelens, R., Koning, T.M.G., van der Marel, G.A., van Boom, J.H. and Kaptein, R. (1989) *J. Magn. Reson.*, **82**, 290–308.

Borgias, B.A. and James, T.L. (1988) *J. Magn. Reson*, **79**, 493–512.

Borgias, B.A. and James, T.L. (1990) *J. Magn Reson.*, **87**, 475–487.

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

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

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

Braun, W., Wagner, G., Wörgötter, E., Vašák, M., Kägi, J.H.R and Wüthrich, K. (1986) *J. Mol. Biol.*, **187**, 125–129.

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

Brüschweiler, R., Blackledge, M. and Ernst, R.R. (1991) *J. Biomol. NMR*, **1**, 3–11.

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

Eccles, C., Güntert, P., Billeter, M. and Wüthrich, K. (1991) *J. Biomol. NMR*, **1**, 111–130.

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

Güntert, P. and Wüthrich, K. (1991) *J Biomol NMR*, **1**, 447–456.

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

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

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

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

Havel, T.F. (1991) *Prog. Biophys. Mol. Biol.*, **56**, 43–78.

Kaptein, R., Zuiderweg, E.R.P., Scheek, R.M., Boelens, R. and van Gunsteren, W.F. (1985) *J Mol. Biol.*, **182**, 179–182.

Keepers, J.W. and James, T.L. (1984) *J. Magn. Reson.*, **57**, 404–426.

Kline, A.D., Braun, W. and Wüthrich, K. (1988) *J. Mol. Biol*, **189**, 377–382.

Kohda, D. and Inagaki, F. (1992) *15th International Conference on Magnetic Resonance in Biological Systems, Jerusalem, Israel*, abstr. P201.

Kumar, A., Ernst, R.R. and Wüthrich, K. (1980) *Biochem. Biophys. Res. Commun.*, **95**, 1–6.

Kuszewski, J., Nilges, M. and Brünger, A.T. (1992) *J. Biomol. NMR*, **2**, 33–56.

Rance, M., Sørensen, O.W., Bodenhausen, G., Wagner, G., Ernst, R.R. and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun.*, **117**, 479–485.

von Freyberg, B. and Braun, W. (1991) *J. Comput. Chem.*, **12**, 1065–1076.

Widmer, H., Billeter, M. and Wüthrich, K. (1989) *Proteins*, **6**, 357–371.

Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol Biol.*, **169**, 949–961.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.

Wynants, C., van Binst, G. and Loosli, H.R. (1985a) *Int. J. Pept. Protein Res.*, **25**, 608–614.

Wynants, C., van Binst, G. and Loosli, H.R. (1985b) *Int. J. Pept Protein Res.*, **25**, 615–621.