

Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA

P. Güntert, C. Mumenthaler and K. Wüthrich*

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

The new program DYANA (DYnamics Algorithm for Nmr Applications) for efficient calculation of three-dimensional protein and nucleic acid structures from distance constraints and torsion angle constraints collected by nuclear magnetic resonance (NMR) experiments performs simulated annealing by molecular dynamics in torsion angle space and uses a fast recursive algorithm to integrate the equations of motions. Torsion angle dynamics can be more efficient than molecular dynamics in Cartesian coordinate space because of the reduced number of degrees of freedom and the concomitant absence of high-frequency bond and angle vibrations, which allows for the use of longer time-steps and/or higher temperatures in the structure calculation. It also represents a significant advance over the variable target function method in torsion angle space with the REDAC strategy used by the predecessor program DIANA. DYANA computation times per accepted conformer in the "bundle" used to represent the NMR structure compare favorably with those of other presently available structure calculation algorithms, and are of the order of 160 seconds for a protein of 165 amino acid residues when using a DEC Alpha 8400 5/300 computer. Test calculations starting from conformers with random torsion angle values further showed that DYANA is capable of efficient calculation of high-quality protein structures with up to 400 amino acid residues, and of nucleic acid structures.

© 1997 Academic Press Limited

*Corresponding author

Keywords: NMR structure determination; molecular dynamics in torsion angle space; simulated annealing; protein structure; nucleic acid structure

Introduction

Here, we present a novel implementation of torsion angle dynamics (TAD; Bae & Haug, 1987; Güntert *et al.*, 1996; Jain *et al.*, 1993; Katz *et al.*,

1979; Kneller & Hinsen, 1994; Mathiowetz *et al.*, 1994; Mazur & Abagyan, 1989; Mazur *et al.*, 1991; Rice & Brünger, 1994; Stein *et al.*, 1997) in the program DYANA for the calculation of three-dimensional protein and nucleic acid structures from sets of experimental distance constraints and torsion angle constraints collected by NMR spectroscopy (Wüthrich, 1986). In common with theoretical approaches to the protein folding problem, NMR structure calculation must afford adequate sampling of conformation space. In light of the fact that the energy landscape of a protein is complex and studded with many local minima (e.g. see Wolynes *et al.*, 1995), new approaches for structure calculation from NMR data can therefore be of potential interest also with regard to gaining novel insights for studies on *in vivo* protein folding. Although we concentrate here on the aspect of NMR structure calculation, a detailed presentation of the algorithms underlying the high efficiency of DYANA for sampling of the conformation space of

Present address: C. Mumenthaler, Boston Consulting Group AG, Zollikerstr. 226, CH-8008 Zürich, Switzerland.

Abbreviations used: ADB, activation domain of porcine procarboxypeptidase B; *Antp*(C39S), *Antp*(C39S) homeodomain; APK, RNA pseudoknot APK; BPTI, basic pancreatic trypsin inhibitor; CPU, central processing unit; CypA, human cyclophilin A; DIANA, distance geometry algorithm for NMR applications; DYANA, dynamics algorithm for NMR applications; INCLAN, interactive command language; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; NOESY, NOE spectroscopy; PDB, Protein Data Bank; PGK, 3-phosphoglycerate kinase; PrP(121-231), mouse prion protein domain PrP(121-231); REDAC, use of redundant dihedral angle constraints; rmsd, root-mean-square deviation; TAD, torsion angle dynamics.

biological macromolecules is justified in view of potential alternative applications.

Currently, most NMR structures of proteins are calculated either by simulated annealing with molecular dynamics in Cartesian space (Brünger, 1992; Brünger & Nilges, 1993), starting from conformers generated by metric matrix distance geometry (Havel & Wüthrich, 1984; Nilges *et al.*, 1988) or from extended conformers (Nilges *et al.*, 1991), or by conjugate gradient minimization of a variable target function in torsion angle space (Braun & Gö, 1985; Güntert *et al.*, 1991a) using redundant dihedral angle constraints (REDAC; Güntert & Wüthrich, 1991). For larger molecules all these approaches tend to use long computation times, in particular when only relatively poor input data are available. The calculation of each conformer can be time-consuming, and the percentage of structures that reach small residual constraint violations can be so low that an excessively large number of computations need to be started. Usually, a low yield of converged structure calculations is due to the fact that the objective function to be minimized has many local minima, into which a minimizer may be trapped as long as it takes exclusively downhill steps. Although simulated annealing using molecular dynamics has in principle the possibility to "escape" from unfavorable local minima, Cartesian space molecular dynamics often fails to calculate successfully the structures of large proteins or of nucleic acids, unless a well-defined starting conformation is available (Stein *et al.*, 1997).

When compared with other structure calculation algorithms based on simulated annealing (Brünger, 1992), the principal difference of TAD is that it works with internal coordinates rather than Cartesian coordinates, which it has in common with the variable target function approach of the predecessor program to DYANA, DIANA (Güntert *et al.*, 1991a). Thereby the number of degrees of freedom is decreased about tenfold, since the covalent structure parameters (bond lengths, bond angles, chiralities and planarities) are kept fixed at their optimal values during the calculation. The strong potentials required to preserve the covalent structure in conventional Cartesian space molecular dynamics and the concomitant high-frequency motions are absent in torsion angle dynamics. This results in a simpler potential energy function and the use of longer time-steps for the numerical integration of the equations of motion, and hence in higher efficiency of the structure calculation. An additional, practical advantage of working in torsion angle space is that this approach excludes inadvertent changes of chirality during the course of a structure calculation, which has apparently resulted in errors in structures calculated with Cartesian space molecular dynamics (Schultze & Feigon, 1997). To fully exploit the potential advantages of TAD for macromolecular structure calculation, its implementation

has to take account of the fact that the Lagrange equations of motion with torsion angles as degrees of freedom are more complex than Newton's equations in Cartesian coordinates. A "naive" implementation of torsion angle dynamics (Mazur & Abagyan, 1989; Mazur *et al.*, 1991) would entail the solution of a system of n linear equations in each time-step (n being the number of degrees of freedom) and thus require a computational effort proportional to n^3 , which would inevitably render the algorithm unattractive for larger systems. DYANA gets around this potential drawback by using a fast recursive implementation of the equations of motion that was originally developed for spacecraft dynamics and robotics (Jain *et al.*, 1993), which entails a computational effort proportional to n .

As the successor of the program DIANA (Güntert *et al.*, 1991a), DYANA contains, in addition to the new features described here, all functions of the program DIANA and its supporting programs. This includes in particular protein structure calculation with the variable target function method (Braun & Gö, 1985) supported by the REDAC strategy (Güntert & Wüthrich, 1991), conversion of NOESY cross-peak volumes into upper distance bounds, and determination of stereospecific assignments (Güntert *et al.*, 1989, 1991a,b). Two separate supporting programs of DIANA, CALIBA and GLOMSA, and an automatic assignment method for NOESY spectra, NOAH (Mumenthaler & Braun, 1995; Mumenthaler *et al.*, 1997), have been incorporated directly into DYANA. The program package is available from the authors for use on a variety of Unix systems. For further details, please consult the DYANA home page at <http://www.mol.biol.ethz.ch/dyana/>.

Theory and Algorithms

For TAD calculations with DYANA the molecule is represented as a tree structure consisting of a base rigid body that is fixed in space and n rigid bodies, which are connected by n rotatable bonds (Abe *et al.*, 1983). The degrees of freedom are exclusively torsion angles, i.e. rotations about single bonds. Each rigid body is made up of one or several mass points (atoms) for which the relative positions are invariable. The tree structure starts from a "base", typically at the N terminus of the polypeptide chain, and terminates with "leaves" at the ends of the side-chains and at the C terminus. The rigid bodies are numbered from 0 to n . The base has the number 0. Each other rigid body, with a number $k \geq 1$, has a single nearest neighbor in the direction toward the base, which has a number $p(k) < k$. The torsion angle between the rigid bodies $p(k)$ and k is denoted by θ_k . The conformation of the molecule is uniquely specified by the values of all torsion angles, $\theta = (\theta_1, \dots, \theta_n)$. All three-dimensional vectors are in an inertial frame of reference that is fixed in

space. For each rotatable bond, \vec{e}_k denotes a unit vector in the direction of the bond, and \vec{r}_k is the position vector of its end point, which is subsequently used as the "reference point" of the rigid body k in the inertial reference frame.

Potential energy

For TAD in DYANA the DIANA target function (Güntert *et al.*, 1991a), V , has the role of the potential energy, E_{pot} , i.e. $E_{\text{pot}} = w_0 V$, with an overall weighting factor $w_0 = 10 \text{ kJ mol}^{-1} \text{ \AA}^{-2}$ ($1 \text{ \AA} = 10^{-10} \text{ m}$). The target function, V , with $V \geq 0$, is defined such that $V = 0$ if and only if all experimental distance constraints and torsion angle constraints are fulfilled and all non-bonded atom pairs satisfy a check for the absence of steric overlap. We then have that $V(\theta) < V(\theta')$ whenever the conformation satisfies the constraints more closely than the conformation θ' . The exact definition of the DYANA target function is:

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta) \in I_c} f_c(d_{\alpha\beta}, b_{\alpha\beta}) + w_d \sum_{k \in I_d} \left(1 - \frac{1}{2} \left(\frac{\Delta_k}{\Gamma_k} \right)^2 \right) \Delta_k^2 \quad (1)$$

Upper and lower bounds, $b_{\alpha\beta}$, on distances between two atoms α and β , $d_{\alpha\beta}$, and constraints on individual torsion angles, θ_k , in the form of allowed intervals, $[\theta_k^{\min}, \theta_k^{\max}]$, are considered. I_u , I_l and I_v are the sets of atom pairs (α, β) with upper, lower or van der Waals distance bounds, respectively, and I_d is the set of restrained torsion angles. w_u , w_l , w_v and w_d are weighting factors for the different types of constraints. $\Gamma_k = \pi - (\theta_k^{\max} - \theta_k^{\min})/2$ denotes the half-width of the forbidden range of torsion angle values, and Δ_k is the size of the torsion angle constraint violation.

The function $f_c(d, b)$, which measures the contribution of a violated distance constraint to the target function, may have any of the three forms given in equations (2) to (4). Equation (2):

$$f_c(d, b) = \left(\frac{d^2 - b^2}{2b} \right)^2 \quad (2)$$

corresponds to the same definition that was used in DIANA (Güntert *et al.*, 1991a). Equation (3):

$$f_c(d, b) = (d - b)^2 \quad (3)$$

is a simple square potential. Equation (4):

$$f_c(d, b) = 2\beta^2 b^2 \left[\sqrt{1 + \left(\frac{d-b}{\beta b} \right)^2} - 1 \right] \quad (4)$$

is a function with a linear asymptote for large constraint violations, where β is a dimensionless parameter that weighs large violations relative to small ones. For small constraint violations the

value of $f_c(d, b)$ is always given by equation (3). To ensure that target function values are directly comparable with those from DIANA, equation (2) was used for all calculations with proteins in this work. Equation (3) was used for work with nucleic acids.

The torques about the rotatable bonds, i.e. the negative gradients of the potential energy with respect to torsion angles, $-\nabla E_{\text{pot}}$, are calculated by the fast recursive algorithm of Abe *et al.* (1983).

Kinetic energy

For all rigid bodies with $k = 1, \dots, n$ (see above), the angular velocity vector, $\vec{\omega}_k$, and the linear velocity of the reference point, $\vec{v}_k = \dot{\vec{r}}_k$, are calculated recursively (Jain *et al.*, 1993):

$$\vec{\omega}_k = \vec{\omega}_{p(k)} + \vec{e}_k \dot{\theta}_k \quad (5)$$

$$\vec{v}_k = \vec{v}_{p(k)} - (\vec{r}_k - \vec{r}_{p(k)}) \wedge \vec{\omega}_{p(k)} \quad (6)$$

For the rigid body k , the vector from its reference point to the center of mass is denoted by \vec{Y}_k , the total mass by m_k , and the inertia tensor (see below) by I_k . The kinetic energy, E_{kin} , is then given by:

$$E_{\text{kin}} = \frac{1}{2} \sum_{k=1}^n [m_k \vec{v}_k^2 + \vec{\omega}_k \cdot I_k \vec{\omega}_k + 2\vec{v}_k \cdot (\vec{\omega}_k \wedge m_k \vec{Y}_k)] \quad (7)$$

Inertia tensors

The inertia tensor of the rigid body k with respect to its reference point, I_k , is a symmetric 3×3 matrix with elements given by (Arnold, 1978):

$$(I_k)_{ij} = \sum_{\alpha} m_{\alpha} (\vec{y}_{\alpha}^2 \delta_{ij} - y_{\alpha i} y_{\alpha j}) \quad (8)$$

The sum runs over all atoms, α , with mass m_{α} , in the rigid body k . \vec{y}_{α} is the vector from the reference point to the atom α , and δ_{ij} is the Kronecker symbol. In DYANA, the center of mass vectors and the inertia tensors are calculated only once by summing over all atoms, with each rigid body in a standard orientation. The resulting standard values are stored as $\vec{Y}_k^{(0)}$ and $I_k^{(0)}$, respectively. If these quantities are required later for a different conformation, for which the orientation of the rigid body k can be generated from its standard orientation by a rotation, R_k , they can be computed by the following transformations (a superscript T denotes the transposed matrix):

$$\vec{Y}_k = R_k \vec{Y}_k^{(0)} \quad (9)$$

$$I_k = R_k I_k^{(0)} R_k^T \quad (10)$$

Since the shape of a rigid body enters the

equations of motion only *via* the inertia tensor and the center of mass vector, it is not essential to derive these quantities from the masses and relative positions of the individual atoms that constitute the rigid body, as in equation (8). Therefore the efficiency of the TAD algorithm can be improved by different choices of inertia tensors and center of mass vectors. By treating, with the use of equations (11) and (12):

$$\vec{Y}_k = \vec{0} \quad (11)$$

$$\mathbf{I}_k = \frac{2}{5} m_k \rho^2 \mathbf{1}_3 \quad (12)$$

all rigid bodies, $k = 1, \dots, n$, as solid spheres of mass m_k and radius ρ centered at their reference points, \vec{r}_k , evaluation of equation (7) is simplified and the transformation of equations (9) and (10) is avoided entirely. In equations (11) and (12), $\vec{0}$ denotes the zero vector and $\mathbf{1}_3$ is the 3×3 unit matrix. Further improvement is obtained by setting the masses of the rigid bodies, m_k , equal to $10\sqrt{n_k} m_0$, where n_k denotes the number of atoms in the rigid body k (pseudoatoms are not counted), and $m_0 = 1.66 \times 10^{-27}$ kg is the atomic mass unit. In this way, fast motions of light rigid bodies, for example hydroxyl protons, are slowed, thereby permitting longer integration time-steps. For all TAD calculations in this work, equations (11) and (12) were used with this expression for m_k , and $\rho = 5.0$ Å. Note that the use of equations (11) and (12) does not imply an approximation of the van der Waals interaction: the steric repulsion is still calculated for each individual atom pair.

Torsional accelerations

The equations of motion for a classical mechanical system with generalized coordinates are the Lagrange equations (Arnold, 1978):

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\theta}_k} \right) - \frac{\partial L}{\partial \theta_k} = 0 \quad (k = 1, \dots, n) \quad (13)$$

with $L = E_{\text{kin}} - E_{\text{pot}}$. They lead to equations of motion of the form:

$$M(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) = 0 \quad (14)$$

Using torsion angles as degrees of freedom, the $n \times n$ mass matrix $M(\theta)$ and $C(\theta, \dot{\theta})$ have been derived explicitly by Mazur & Abagyan (1989) and Mazur *et al.* (1991). However, to integrate the equations of motion, equation (14) would have to be solved in each time-step for the torsional accelerations, $\ddot{\theta}$. This requires the solution of a system of n linear equations and hence entails a computational effort proportional to n^3 , which would become prohibitively expensive for larger systems. Therefore, in DYANA the fast recursive algorithm of Jain *et al.* (1993) is implemented to compute the torsional accelerations, which makes explicit use of the aforementioned tree structure of the molecule

in order to obtain $\ddot{\theta}$ with a computational effort that is proportional to n .

The algorithm of Jain *et al.* (1993) is initialized by calculating for all rigid bodies, $k = 1, \dots, n$, the six-dimensional vectors a_k , e_k and z_k :

$$a_k = \begin{bmatrix} (\vec{\omega}_k \wedge \vec{e}_k) \dot{\theta}_k \\ \vec{\omega}_{p(k)} \wedge (\vec{v}_k - \vec{v}_{p(k)}) \end{bmatrix} \quad (15)$$

$$e_k = \begin{bmatrix} \vec{e}_k \\ \vec{0} \end{bmatrix} \quad (16)$$

$$z_k = \begin{bmatrix} \vec{\omega}_k \wedge \mathbf{I}_k \vec{\omega}_k \\ (\vec{\omega}_k \cdot m_k \vec{Y}_k) \vec{\omega}_k - \omega_k^2 m_k \vec{Y}_k \end{bmatrix} \quad (17)$$

Furthermore, the 6×6 matrices P_k and ϕ_k need to be calculated:

$$P_k = \begin{bmatrix} \mathbf{I}_k & m_k \mathbf{A}(\vec{Y}_k) \\ -m_k \mathbf{A}(\vec{Y}_k) & m_k \mathbf{1}_3 \end{bmatrix} \quad (18)$$

$$\phi_k = \begin{bmatrix} \mathbf{1}_3 & \mathbf{A}(\vec{r}_k - \vec{r}_{p(k)}) \\ \mathbf{0}_3 & \mathbf{1}_3 \end{bmatrix} \quad (19)$$

$\mathbf{0}_3$ is the 3×3 zero matrix, and $\mathbf{A}(\vec{x})$ denotes the antisymmetric 3×3 matrix associated with the cross product, i.e. $\mathbf{A}(\vec{x})\vec{y} = \vec{x} \wedge \vec{y}$ for all vectors \vec{y} .

Next, a number of auxiliary quantities is calculated by executing a recursive loop over all rigid bodies in the backward direction, $k = n, n-1, \dots, 1$:

$$D_k = e_k \cdot P_k e_k$$

$$G_k = P_k e_k / D_k$$

$$\varepsilon_k = -e_k \cdot (z_k + P_k a_k) - \frac{\partial V}{\partial \theta_k}$$

$$P_{p(k)} \leftarrow P_{p(k)} + \phi_k (P_k - G_k e_k^T P_k) \phi_k^T$$

$$z_{p(k)} \leftarrow z_{p(k)} + \phi_k (z_k + P_k a_k + G_k \varepsilon_k) \quad (20)$$

D_k and ε_k are scalars, G_k is a six-dimensional vector, and \leftarrow means: assign the result of the expression on the right-hand side to the variable on the left-hand side. Finally, the torsional accelerations are obtained by executing another recursive loop over all rigid bodies in the forward direction, $k = 1, \dots, n$:

$$\alpha_k = \phi_k^T \alpha_{p(k)}$$

$$\ddot{\theta}_k = \varepsilon_k / D_k - G_k \cdot \alpha_k \quad (21)$$

$$\alpha_k \leftarrow \alpha_k + e_k \ddot{\theta}_k + a_k$$

The auxiliary quantities α_k are six-dimensional vectors, with α_0 being equal to the zero vector.

Integration of the equations of motion

The integration scheme for the equations of motion in TAD (Mathiowetz *et al.*, 1994) is a variant of the leap-frog algorithm used in Cartesian dynamics (Allen & Tildesley, 1987). The temperature is controlled by weak coupling to an external bath (Berendsen *et al.*, 1984). A time step, $t \rightarrow t + \Delta t$, that follows a preceding time step, $t - \Delta t' \rightarrow t$, consists of the following parts.

- (1) On the basis of the torsional positions, $\theta(t)$, calculate the Cartesian coordinates of all atoms (Güntert, 1993).
- (2) Using the Cartesian coordinates, calculate the potential energy function of equation (1), $E_{\text{pot}}(t) = E_{\text{pot}}(\theta(t))$, and its gradient $\nabla E_{\text{pot}}(t)$.
- (3) Determine the time-step, $\Delta t = \lambda_\varepsilon \Delta t'$, using the time-step scaling factor:

$$\lambda_\varepsilon = \min\left(\lambda_\varepsilon^{\text{max}}, \sqrt{1 + \frac{\varepsilon^{\text{ref}} - \varepsilon(t)}{\tau \varepsilon(t)}}\right) \quad (22)$$

λ_ε is based on the reference value for the relative accuracy of energy conservation, ε^{ref} , and on the relative change of the total energy, $E = E_{\text{kin}} + E_{\text{pot}}$ in the preceding time-step, $\varepsilon(t)$, as given by:

$$\varepsilon(t) = \left| \frac{E(t) - E(t - \Delta t')}{E(t)} \right| \quad (23)$$

$\lambda_\varepsilon^{\text{max}}$ is the maximally allowed value of the scaling factor. $\lambda_\varepsilon^{\text{max}} = 1.025$ was used for all calculations in this work. The time constant, $\tau \gg 1$, is a user-defined parameter that is measured in units of the time-step. A value of $\tau = 20$ was used for all calculations in this work. To calculate $\varepsilon(t)$ in equation (23), $E(t)$ is evaluated before velocity scaling is applied (step 4 below), whereas for $E(t - \Delta t')$ the value after velocity scaling in the preceding time-step is used. Thus, the measurement of the accuracy of energy conservation in equation (23) is not affected by the scaling of velocities. An exact algorithm would yield $E(t) = E(t - \Delta t')$ and consequently $\varepsilon(t) = 0$.

- (4) Adapt the temperature by scaling of the torsional velocities, i.e. replace $\dot{\theta}(t - \Delta t'/2)$ by $\lambda_T \dot{\theta}(t - \Delta t'/2)$ and $\dot{\theta}_e(t)$ by $\lambda_T \dot{\theta}_e(t)$ (see equation (27) below for the definition of $\dot{\theta}_e(t)$). The velocity scaling factor, λ_T , is given by (Berendsen *et al.*, 1984):

$$\lambda_T = \sqrt{1 + \frac{T^{\text{ref}} - T(t)}{\tau T(t)}} \quad (24)$$

where T^{ref} is the reference value of the temperature. The instantaneous temperature, $T(t)$, is given by:

$$T(t) = \frac{2E_{\text{kin}}(t)}{nk_B} \quad (25)$$

where n denotes the number of torsion angles and $k_B = 1.38 \times 10^{-23} \text{ J K}^{-1}$ is the Boltzmann constant. The kinetic energy, $E_{\text{kin}}(t) = E_{\text{kin}}(\theta(t), \dot{\theta}_e(t))$, is taken from equation (7). Temperature control by coupling to an external heat bath (and time-step control) can be turned off by setting $\tau = \infty$.

- (5) Calculate the torsional accelerations, $\ddot{\theta}(t) = \ddot{\theta}(\theta(t), \dot{\theta}_e(t))$, using equations (15) to (21).
- (6) Calculate the new velocities at half time-step:

$$\dot{\theta}(t + \Delta t/2) = \dot{\theta}(t - \Delta t'/2) + \frac{\Delta t + \Delta t'}{2} \ddot{\theta}(t) \quad (26)$$

- (7) Calculate the new estimated velocities at full time step:

$$\begin{aligned} \dot{\theta}_e(t + \Delta t) = & \left(1 + \frac{\Delta t}{\Delta t + \Delta t'}\right) \dot{\theta}_e(t + \Delta t/2) \\ & - \frac{\Delta t}{\Delta t + \Delta t'} \dot{\theta}_e(t - \Delta t'/2) \end{aligned} \quad (27)$$

- (8) Calculate the new torsional positions:

$$\theta(t + \Delta t) = \theta(t) + \Delta t \dot{\theta}(t + \Delta t/2) \quad (28)$$

The algorithm is initialized by setting $t = 0$, $\Delta t' = \Delta t$ and $\theta_e(0) = \dot{\theta}(-\Delta t/2)$. Furthermore, the initial torsional velocities, $\dot{\theta}(-\Delta t/2)$, are chosen randomly from a normal distribution with zero mean value and a standard deviation that ensures that the initial temperature has a predefined value, $T(0)$. Once the time step $t \rightarrow t + \Delta t$ is completed by going through the operations (1) to (8), t is replaced by $t + \Delta t$, and $\Delta t'$ by Δt .

Unlike the situation in Cartesian dynamics, where the accelerations are a function of the positions only, the torsional accelerations depend also on the velocities. These, however, are only known at half time-steps, whereas the positions and accelerations are required at full time-steps. In equation (27) the presently used algorithm employs linear extrapolation to obtain an estimate of the velocity after the full time-step, $\dot{\theta}_e(t)$, which is used in the next integration step to calculate the torsional accelerations (step (5)). They are therefore beset with an additional error that could be eliminated by iterating the steps (5) to (7). In general, no significant improvement can be observed after one iteration (Mathiowetz *et al.*, 1994). Using $\dot{\theta}_e(t)$ instead of $\dot{\theta}(t)$ introduces an additional error of order $O(\Delta t^3)$ into the velocity calculation of the leap-frog algorithm in equation (26). However, the intrinsic accuracy of the velocity step is also $O(\Delta t^3)$. Thus, using the estimated velocities, $\dot{\theta}_e(t)$, does not change the order of accuracy of the integration algorithm. Each iteration of steps (5) to (7) requires the calculation of the torsional accelerations which, in general, takes as long as the calculation of the target function and its gradient. It is

therefore more efficient to slightly decrease the time-step, Δt , than to perform steps (5) to (7) twice in each integration step. (The situation would be different if a full physical force field were used, since then the calculation of torsional accelerations would require only a negligible fraction of the total computation time.)

Since in structure calculations with TAD the time-steps are made as long as possible, a safeguard against occasional strong violations of energy conservation was introduced. If the total energy, E , has changed by more than 10% in a single time-step, this time-step is rejected and replaced by two time-steps of half length.

Energy conservation

Energy conservation is a key feature of proper functioning of any MD algorithm (Allen & Tildesley, 1987). To verify that TAD was properly implemented in DYANA and to determine the admissible size range for the integration time-step, Δt , we performed a series of TAD runs under conditions where the total energy, E , should be conserved. The experimental NMR data set for cyclophilin A (Ottiger *et al.*, 1997) was used with $\tau = \infty$ in equations (22) and (24), i.e. without coupling to an external heat bath and without automatic time-step control. The accuracy of energy conservation was monitored by the standard deviation, $\sigma(E) = \sqrt{\langle (E - \langle E \rangle)^2 \rangle}$, and the rms change of the total energy between successive integration steps, $\sqrt{\langle \Delta E^2 \rangle}$, where $\langle \dots \rangle$ denotes the average over all steps of a TAD run (Beeman, 1976; van Gunsteren & Berendsen, 1977). The latter parameter, which is closely related to $\varepsilon(t)$ in equation (23), probes the local error in one time-step of the integration algorithm, whereas the standard deviation is sensitive both to local errors and to slow drifts of the total energy over many time-steps and is thus dependent on the length of TAD run. The results (Figure 1) show that long time-steps of up to 100 fs are tolerated by the algorithm, that $\sigma(E)$ is proportional to Δt^2 , as expected for Verlet-type integration algorithms (Allen & Tildesley, 1987), and that $\sqrt{\langle \Delta E^2 \rangle}$ is proportional to Δt^3 . It shows also that the use of "exact" torsional accelerations leads to smaller violations of energy conservation than calculations with accelerations based on approximate velocities obtained by linear extrapolation from preceding time-steps. However, the resulting slight improvement does not warrant the additional computational effort needed to obtain "exact" accelerations, in particular since their use does not permit an increase of the maximally allowed length of the time-step. All calculations in the remainder of this work have therefore been performed with accelerations based on approximate velocities.

For practical applications, the most relevant result from the test calculations of Figure 1 is that long time-steps, about 100, 30 and 7 fs at low, medium and high temperatures, respectively, can

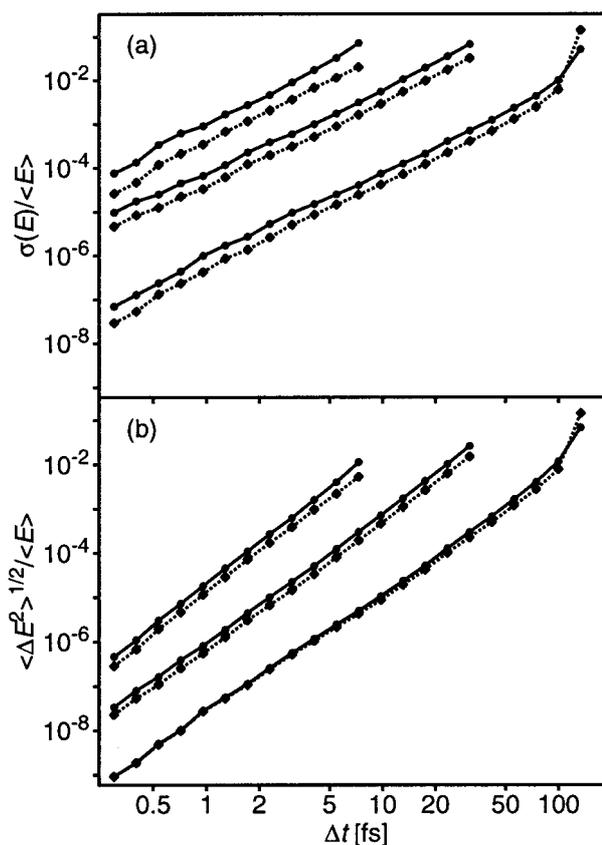


Figure 1. Plots versus integration time-step length, Δt of (a) the standard deviation, $\sigma(E) = \sqrt{\langle (E - \langle E \rangle)^2 \rangle}$, and (b) the rms change of the total energy between successive integration steps, $\sqrt{\langle \Delta E^2 \rangle}$, for TAD calculations using the experimental NMR data set for cyclophilin A. The values given are relative to the average total energy, $\langle E \rangle$. Circles and continuous lines correspond, from top to bottom, to initial temperatures of 10,600, 390, and 1.18 K, respectively, and were obtained using torsional accelerations calculated on the basis of linearly extrapolated torsional velocities. Squares and dotted lines resulted from identical calculations with exact torsional accelerations. The duration of each TAD run was 0.9 ps, and no coupling to an external heat bath was applied.

be used in TAD calculations with DYANA. The concomitant fast exploration of conformation space provides the basis for efficient structure calculation protocols. These time-steps should not be compared directly with the steps of 1 to 2 fs duration used in conventional Cartesian space MD (Allen & Tildesley, 1987), since DYANA uses more uniform masses and a much simpler potential energy function than standard molecular dynamics programs (Brooks *et al.*, 1983; Cornell *et al.*, 1995; van Gunsteren *et al.*, 1996).

Computation times with the standard DYANA simulated annealing protocol

CPU times per starting conformer for protein structure calculations with the standard DYANA

Table 1. CPU times (seconds) for DYANA structure calculations of the proteins BPTI and CypA on different computers

Computer	BPTI	CypA
NEC SX-4	13	36
DEC Alpha 8400 5/300	20	86
SGI Indigo ² R10000 (175 MHz)	23	127
IBM RS/6000-590	35	141
Cray J-90	44	141
Convex Exemplar SPP1600	44	177
Hewlett-Packard 735	47	209

CPU times are for the calculation of one conformer, using the experimental NMR data sets (Berndt *et al.*, 1992; Ottiger *et al.*, 1997) and the standard DYANA simulated annealing protocol (see Methods).

simulated annealing protocol (see Methods) on a variety of different computer systems are in the range of 13 to 47 seconds for the 58 residue protein BPTI, and of 36 to 209 seconds for the 165 residue protein cyclophilin A (Table 1).

The dependence of the calculation times on the protein size is illustrated in Figure 2 for a scalar computer (DEC Alpha 8400 5/300) and for a vector supercomputer (NEC SX-4). CPU time as a function of protein size increases linearly on the vector machine and slightly faster than linear on the scalar machine. This moderate increase of computational effort with protein size results from the efficient algorithm used to calculate the torsional accelerations (Jain *et al.*, 1993), which scales linearly with the number of torsion angles. It allows handling of macromolecules of any size for which data sets can be collected with present state-of-the-art NMR techniques. Note that the calculation of torsional accelerations in equations (15) to (21) scales linearly with protein size, and that the steeper than

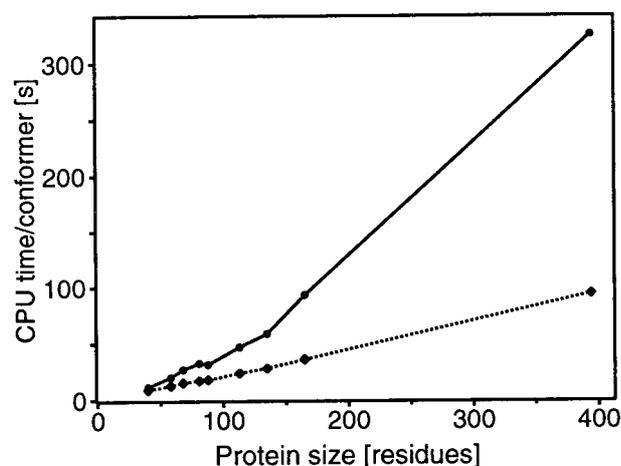


Figure 2. Dependence on protein size of the CPU time for the calculation of one conformer using the standard DYANA simulated annealing protocol (see Methods) with 4000 TAD steps. CPU times were measured on DEC Alpha 8400 5/300 computers (circles and continuous line) and on a NEC SX-4 computer (squares and dotted line).

linear increase of the overall computation time on scalar computers is related to the evaluation of the target function of equation (1).

Since an NMR structure calculation always involves the computation of a group of conformers (Wüthrich, 1986), it is highly efficient to run calculations of multiple conformers in parallel. In DYANA, parallel computing is implemented on the level of the command language, INCLAN, and achieves nearly ideal speedup. Using a dedicated Cray J-90 computer with eight processors, the elapsed time for the calculation of 100 BPTI conformers on $n = 1, \dots, 8$ processors was closely proportional to $1/n$ (Figure 3), which is the theoretically achievable optimum.

Results and Discussion

Structure calculations with DYANA using TAD

Structure calculations were performed with experimental and simulated NMR data sets of six different proteins and an RNA 32-mer (Table 2). The experimental data sets include a small globular protein (BPTI), an all- α -helical protein (*Antp*(C39S)), a predominantly β -sheet protein (ADB), a medium-size protein with a moderately dense network of constraints (PrP(121-231)), and a larger β -barrel protein with an outstandingly complete set of constraints (CypA). The simulated NMR data sets include a 394 residue protein (PGK) and an RNA molecule (APK), which were selected to illustrate the application of DYANA with protein sizes at the limit of present-day NMR technology, and with nucleic acids.

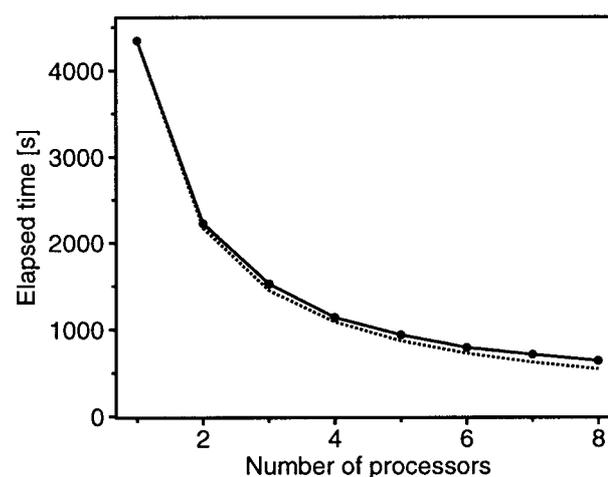


Figure 3. Elapsed time for the calculation of 100 BPTI conformers as a function of the number of processors used. Calculations were run on a dedicated Cray J-90 computer with eight processors. Dots represent measured elapsed times. The dotted line indicates the theoretically achievable optimum. The standard DYANA simulated annealing protocol (see Methods) was used with the experimental NMR data set from Berndt *et al.* (1992).

Table 2. Experimental and simulated NMR data sets used for the DYANA structure calculations discussed in the text

Molecule	Residues	Torsion angles	Upper distance bounds	Torsion angle constraints	Reference
BPTI ^a	58	241	651	115	Berndt <i>et al.</i> (1992)
Antp(C39S) ^a	68	351	894	171	Güntert <i>et al.</i> 1991b)
ADB ^a	81	379	795	168	Vendrell <i>et al.</i> (1991)
PrP(121-231) ^a	113	521	1508	256	Riek <i>et al.</i> (1996)
CypA ^a	165	722	4093	371	Ottiger <i>et al.</i> (1997)
PGK ^b	394	1778	14,161	1117	Davies <i>et al.</i> (1994)
APK (RNA) ^b	32	318	929	284	Kang <i>et al.</i> (1996)

^a For these proteins, the previously published sets of experimental NMR constraints were used.
^b Conformational constraints were simulated from the atomic coordinates of the crystal or NMR structure (see Methods).

Figure 4 presents key characteristics of structure calculations with the standard simulated annealing protocol of DYANA (see Methods), using CypA as an illustration. One fifth of the TAD steps are performed at the initial high reference temperature of 9600 K, and during the remainder of the TAD calculation the reference temperature slowly

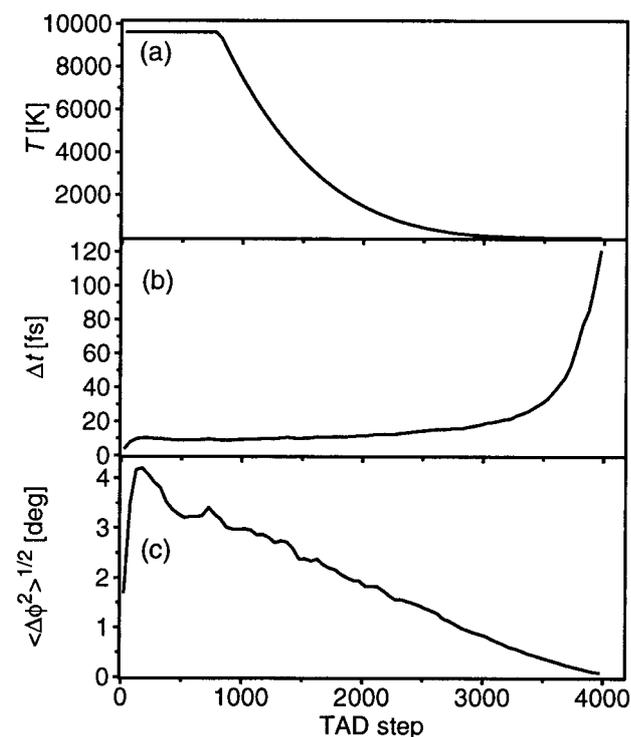


Figure 4. Plots versus the number of TAD steps for a structure calculation using the standard DYANA simulated annealing protocol (see Methods) with the experimental NMR data set for CypA. (a) Temperature of the heat bath to which the system is weakly coupled. (b) Integration time-step length, Δt . (c) The rms torsion angle change between successive TAD steps, $\sqrt{\langle \Delta\phi^2 \rangle}$. The plot (b) shows data representing averages over both 50 TAD steps and the 80 conformers. The plot (c) shows data representing averages over both 50 TAD steps and all rotatable torsion angles of the 80 independently calculated conformers.

approaches 0 K (Figure 4(a)). The length of the integration time-step, Δt , is adapted so as to attain a user-defined accuracy of energy conservation, with short Δt values at the outset and an increase to above 100 fs toward the end of the calculation (Figure 4(b)). The rms torsion angle change between successive integration steps is up to 4° during the high-temperature phase of the annealing schedule and then decreases linearly to small values (Figure 4(c)).

To assess the influence on the quality of the resulting structures of the parameter that affects the computation time most strongly, we performed structure calculations with different numbers of TAD steps for each of the data sets in Table 2. As expected, increasing the number of TAD steps generally leads to a reduction of the residual target function values for a given data set, and provided that enough TAD steps were performed the percentage of conformers with “acceptable” low target function values approached 90 to 100% in all proteins except CypA (Figure 5). However, for most data sets, little further improvement resulted from increasing the number of TAD steps above 6000. In the other limit, for most data sets a decrease of the number of TAD steps to 2000 resulted in significant deterioration of the quality of the resulting structures (Figure 5). The dependence of the quality of the structures on the number of TAD steps is most pronounced if the calculation is either based on incomplete input sets of constraints, such as for PrP(121-231), or when there is intrinsic scarcity of constraints, such as for APK (RNA). The size of the protein appears not to be a decisive factor, since the target function distributions for the 394 residue protein PGK are quite similar to those for the small protein BPTI.

The relevant measure for the overall efficiency of a structure calculation program is the computation time per accepted conformer, i.e. the total computation time for a group of starting conformers divided by the number of acceptable conformers. Here, conformers with target function values below the cutoffs indicated by horizontal broken lines in Figure 5 are considered to be acceptable (see Methods), and in Figure 6 the

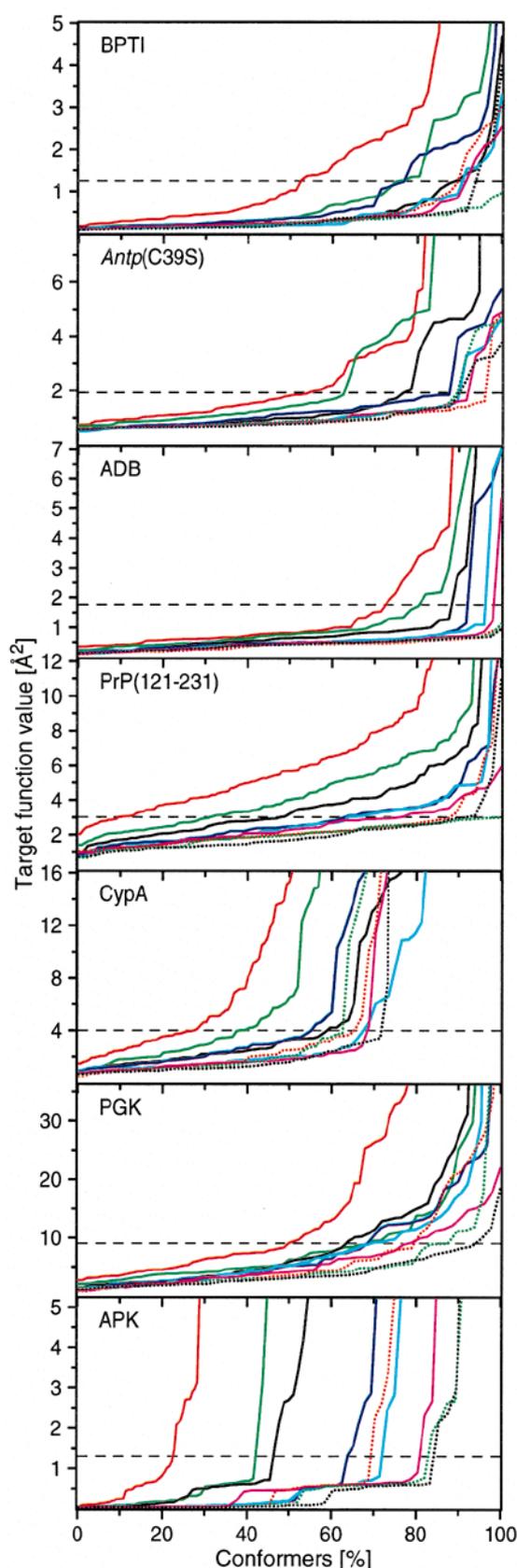


Figure 5. Final target function values of structure calculations for the data sets of Table 2 using the standard DYANA simulated annealing protocol with different numbers of TAD steps, N : $N = 2000$, red; $N = 3000$, green; $N = 4000$, black; $N = 6000$, blue; $N = 8000$,

resulting computation times per accepted conformer are plotted as a function of the number of TAD steps. Optimal efficiency for all data sets is attained at about 4000 TAD steps with the given acceptance criteria (see Methods). Possible increase of the percentage of acceptable conformers by going to higher numbers of TAD steps is always overcompensated by the concomitant rise of the computation time per conformer, which is proportional to the number of TAD steps. On the basis of the data in Figure 6, individual optimization of this parameter for different systems seems unnecessary if good input data sets are available. Therefore, 4000 TAD steps are used in the standard DYANA simulated annealing protocol (Table 3). The numbers in Table 3 show that a group of 20 acceptable conformers can be obtained in about eight minutes for a small protein (BPTI), in less than one hour for a medium-size protein (CypA), and in less than three hours for a large protein (PGK). These data are for the use of a single processor and can of course be further reduced if multi-processor parallel computing is available (Figure 3).

The definition used here for acceptable conformers (see Methods) excludes conformers with large residual constraint violations (Table 4). The average maximal violations (Table 4) are below 0.5 Å for upper distance bounds, below 0.4 Å for steric lower distance bounds, and below 7° for torsion angle constraints. The sums of all violations and hence the violation per constraint are very small. Since these results were obtained in torsion angle space, the program had, in contrast to calculations in Cartesian space, no possibility to reduce constraint violations by introducing distortions of bond lengths and bond angles. Finally, the rmsd values for the well-defined parts of the polypeptide backbone are indicative of high-quality structures for all the proteins investigated (Table 3).

Efficiency of structure calculations with TAD: comparison with DIANA/REDAC

Since DYANA includes all the functions of the predecessor program DIANA as well as the novel TAD approach, it enabled a direct comparison of the TAD routine with the DIANA/REDAC method (Güntert & Wüthrich, 1991). The previously pub-

cyan; $N = 10,000$, magenta; $N = 12,000$, red, dotted; $N = 16,000$, green, dotted; $N = 20,000$, black, dotted. For each value of N the results obtained with groups of calculations using randomized starting conformers are sorted by increasing residual target function values. The horizontal axis is calibrated by the percentage of all starting conformers; to obtain at least 40 acceptable conformers in each case, different numbers of starting conformers were calculated for the different curves. The horizontal broken line indicates the target function cutoff value, V_{cut} , for acceptable conformers (see Methods).

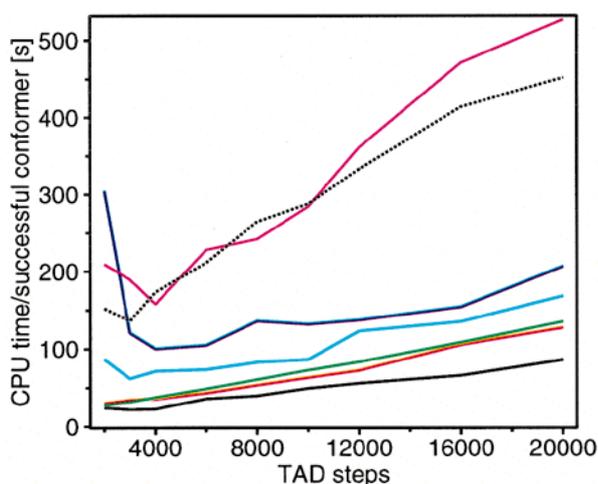


Figure 6. Plots of the CPU times per acceptable conformer as a function of the number of TAD steps in structure calculations with the standard DYANA simulated annealing protocol for the data sets of Table 2, measured on DEC Alpha 8400 5/300 computers: BPTI, black; *Antp*(C39S), red; ADB, green; PrP(121-231), blue; CypA, magenta; PGK, black, dotted; APK (RNA), cyan. The CPU times for the largest protein, PGK, are scaled down by a factor of 3.

lished structure calculations for BPTI, *Antp*(C39S), and ADB (Güntert & Wüthrich, 1991), for PrP(121-231) (Riek *et al.*, 1996) and for CypA (Ottiger *et al.*, 1997) were repeated using DYANA, after the REDAC strategy had been implemented as an INCLAN macro. The same parameters and acceptance criteria were used as in the original calculations. Table 3 shows that TAD is between two and ten times faster than the REDAC strategy, depending primarily on the size and topology of the structure,

† Stein *et al.* (1997) recently reported on the efficiency of TAD and Cartesian space molecular dynamics protocols implemented in the program XPLOR. According to this reference the CPU time for the calculation of BPTI on a Hewlett-Packard 735 computer with the TAD protocol in XPLOR is 921 s (Table 6 of Stein *et al.*, 1997). This compares with a CPU time of 47 seconds when using the TAD protocol of DYANA for BPTI on the same computer (Table 1). Although small contributions to this difference of more than an order of magnitude may be due to the fact that slightly different input data sets were used, the high performance of DYANA can be traced primarily to differences in the implementation of TAD and the force calculation: DYANA uses the TAD algorithm of Jain *et al.* (1993), whereas XPLOR employs the algorithm of Bae & Haug (1987) and follows a fourth-order Runge-Kutta scheme for integration of the equations of motion (Rice & Brünger, 1994). The latter requires four force evaluations per time-step, whereas a single force evaluation per time-step is performed in the leap-frog algorithm used by DYANA. Overall, both methods have similar convergence rates and yield structures of comparable quality (Tables 3 and 4; Tables 5 to 7 of Stein *et al.*, 1997).

and yields structures of comparable or higher quality†. TAD turns out to compare particularly favorably with DIANA/REDAC in the calculation of the predominantly β -sheet protein ADB, and for the larger proteins PrP(121-231) and CypA. Although for BPTI and *Antp*(C39S) all 20 accepted REDAC conformers would also be acceptable following the criterion used for TAD conformers (see Methods), only four, one and six out of the 20 accepted REDAC conformers for ADB, PrP(121-231) and CypA, respectively, would be accepted by the same criterion. In parallel with the lower residual target function values, the groups of TAD conformers have slightly lower rmsd values for the well-defined parts of the polypeptide backbone than the bundles of REDAC conformers for all proteins except CypA (Table 3).

Structure calculations for larger proteins

The structure calculations with a simulated NMR data set for the 394 residue protein PGK were included to investigate the potential of the program DYANA for use with larger molecules, for which experimental NMR data sets may become available in the future. The results in Tables 3 and 4, and in Figures 5 and 6 show that with this data set DYANA can be used with the same simulated annealing schedule as for smaller proteins, and that the yield of acceptable conformers is similar to that for the smaller molecules.

A comparison of the PGK structure calculated using DYANA with the X-ray structure from which the simulated NMR data set was derived (Davies *et al.*, 1994) is afforded by Figure 7. The two structures coincide closely, with an rmsd between the mean DYANA conformer and the X-ray structure calculated for all backbone atoms N, C α and C' of 1.05 Å (Figure 7(a)). This is actually significantly smaller than the corresponding average of the rmsd values between the individual accepted DYANA conformers and their mean (Table 3). Figure 7(b) reveals that the spread among the DYANA conformers is predominantly caused by slightly different relative orientations of the two domains. Within each domain the precision is much higher, with average backbone rmsd values to the mean of 0.30 Å for residues 1 to 170 and 0.36 Å for residues 183 to 381. The larger variations observed for the complete protein are directly related to the scarcity of constraints that characterize the orientation of the two domains relative to each other and do not indicate a limitation of the DYANA program.

Structure calculations for RNA

NMR structure determination of DNA or RNA molecules is more difficult than for proteins because the network of NOE distance constraints in nucleic acids is less dense than in proteins (Wüthrich, 1986). For instance, the number of distance constraints per degree of freedom is 2.7 times

Table 3. Results of TAD structure calculations and comparison with corresponding calculations performed with REDAC in DYANA

Molecule	TAD				REDAC ^a		
	Accepted conformers (%)	CPU time/accepted conformer (s) ^b	Target function (Å ²)	rmsd (Å) ^c	CPU time/accepted conformer (s) ^b	Target function (Å ²)	rmsd (Å) ^c
BPTI	88	23	0.33	0.37	56	0.36	0.47
<i>Antp</i> (C39S)	75	36	0.94	0.38	74	1.00	0.49
ADB	88	37	0.60	0.92	162	2.06	1.18
PrP(121-231)	47	101	2.19	1.11	960	5.19	1.28
CypA	58	159	2.19	0.68	1627	4.59	0.54
PGK	63	521	3.99	1.62			
APK (RNA)	45	72	0.32	3.08			

^a REDAC calculations with DYANA were performed with the experimental data sets using the same parameters and acceptance criteria as for the DIANA calculations described in the original publications.

^b CPU times were measured on DEC Alpha 8400 5/300 computers.

^c rmsd values are given for the backbone atoms N, C^α, C' of the following residues: BPTI, 3 to 55; *Antp*(C39S), 8 to 60; ADB, 11 to 76; PrP(121-231), 124 to 168 and 173 to 227; CypA, PGK and APK (RNA), all residues.

lower in the simulated data set for the pseudoknot RNA APK than for the protein PGK (Table 2), even though the same criteria were applied to derive NOE distance constraints from the published structures. For this reason, structure calculation programs based on Cartesian space molecular dynamics and metric-matrix distance geometry often fail to find nucleic acid conformers that satisfy the experimental data (Stein *et al.*, 1997) unless *ad hoc* assumptions about the three-dimensional structure are made, usually by starting the calculations from standard A or B-type duplex conformations. Stein *et al.* (1997) showed for a 12 base-pair DNA duplex that torsion angle dynamics was successful in finding structures that satisfy the experimental constraints in a situation where metric-matrix distance geometry and Cartesian space molecular dynamics did not lead to acceptable results.

Here, we used simulated data for a 32 nucleotide pseudoknot RNA (Kang *et al.*, 1996), to investigate whether TAD as implemented in DYANA is able to calculate RNA structures starting from conformers with random torsion angle values. In these calculations the sugar rings remained flexible, which was achieved by "cutting" the bond between C4' and O4' and imposing distance constraints to fix

the length of the C4'–O4' bond and the corresponding bond angles. The results of the structure calculations (Figure 5; Tables 3 and 4) show that the standard DYANA simulated annealing protocol for proteins (see Methods) is also adequate for nucleic acid structure calculations, indicating that with the TAD approach in DYANA there is no fundamental difference between the two classes of molecules from the point of view of structure calculation. In particular, the DYANA calculations do not need to start from well-defined structures, which could introduce a bias into the final result of the structure calculation.

Conclusions

The results of this work show that the implementation of torsion angle dynamics in the program DYANA is a powerful method for the calculation of protein and nucleic acid structures from NMR data. It represents a significant advance over the previously used implementations of the variable target function method (Braun & Gö, 1985) with the REDAC strategy (Güntert & Wüthrich, 1991) in DIANA (Güntert *et al.*, 1991a,b), the currently widely used simulated annealing by Cartesian space molecular dynamics (Brünger, 1992)

Table 4. Constraint violations of structures calculated using TAD in DYANA

Molecule	Upper distance limits		Steric distance limits		Torsion angle constraints	
	Average (Å)	Max. (Å)	Sum (Å)	Max. (Å)	Average (°)	Max. (°)
BPTI	0.0040	0.19	1.2	0.09	0.04	1.6
<i>Antp</i> (C39S)	0.0050	0.28	2.5	0.17	0.11	4.5
ADB	0.0025	0.25	1.9	0.20	0.04	2.7
PrP(121-231)	0.0062	0.41	5.2	0.25	0.10	5.3
CypA	0.0028	0.44	5.1	0.23	0.03	3.7
PGK	0.0009	0.47	11.0	0.39	0.01	6.9
APK (RNA)	0.0014	0.22	0.8	0.13	0.03	1.4

Average, Sum and Max. denote the average violation, the sum of all violations, and the maximal violation per conformer, respectively. Values are averaged over all accepted conformers. The average violation equals the sum of violations divided by the number of constraints and was for practical reasons not evaluated for the steric distance limits.

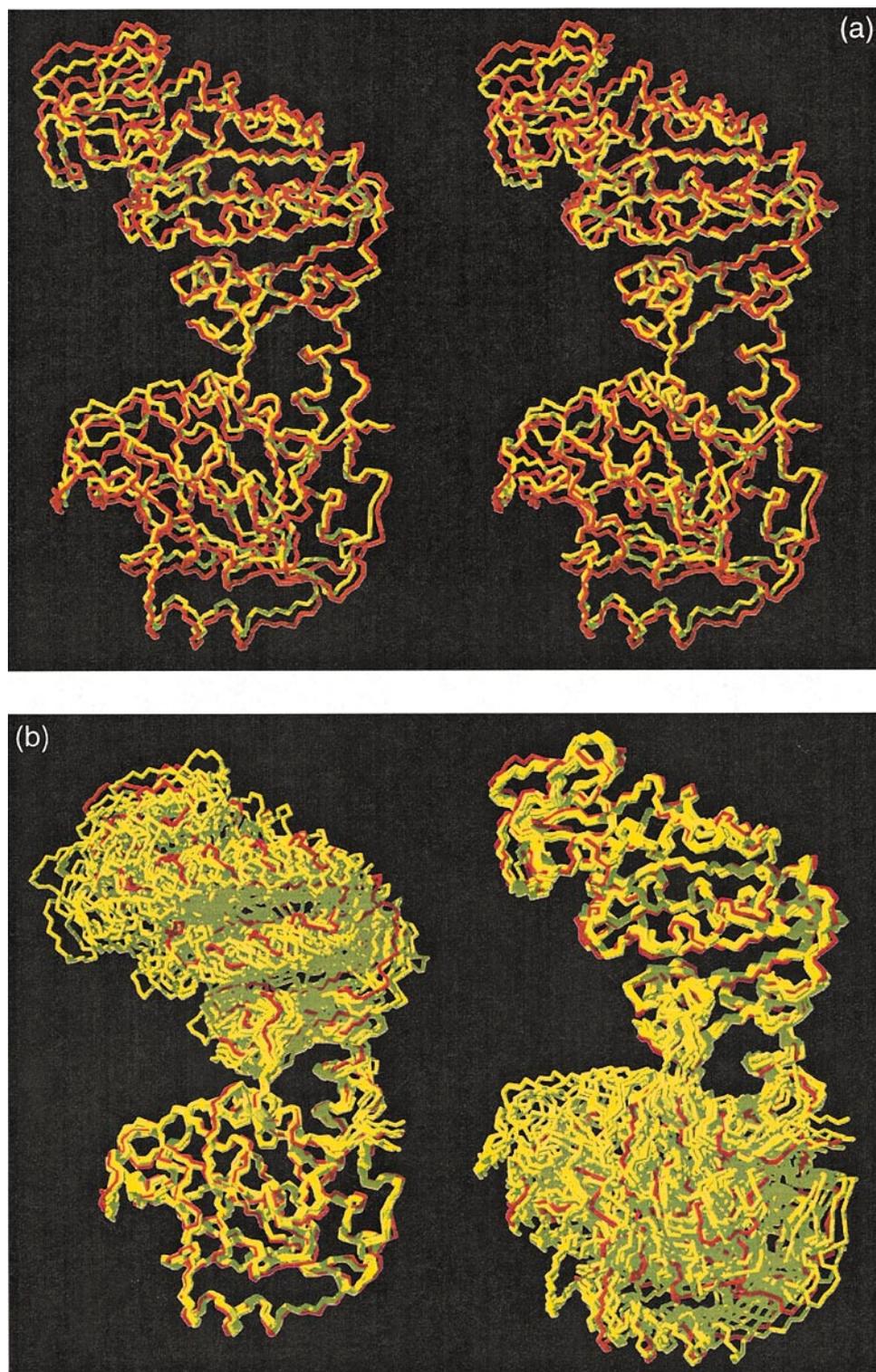


Figure 7. Superpositions of the X-ray structure of PGK (Davies *et al.*, 1994; red) and the structure calculated with the standard DYANA simulated annealing protocol using a simulated NMR data set derived from the crystal coordinates (yellow). (a) Stereo view of a superposition for minimal rmsd of all backbone atoms N, C α and C' of the X-ray structure and the mean of the ten DYANA conformers with the lowest residual target function values. (b) Mono views of superpositions for minimal backbone rmsd of the X-ray structure and the ten DYANA conformers with lowest final target function values; left: superposition for best fit of residues 1 to 170; right: superposition for best fit of residues 183 to 381. These Figures were generated with the program MOLMOL (Koradi *et al.*, 1996).

and, as judged by comparison with the available literature data (Table 6 of Stein *et al.*, 1997), other presently available implementations of torsion angle dynamics algorithms for structure calculation of biological macromolecules from NMR data. With DYANA, the calculation of three-dimensional structures will not be a bottleneck for NMR structure determination of proteins up to 400 residues. Considering its high efficiency for exploring the conformation space of biological macromolecules, future applications of DYANA might well include problems in molecular modeling, tertiary structure prediction and protein folding.

Methods

The program DYANA

DYANA is written in FORTRAN-77 and has been implemented and optimized for a variety of serial, parallel and vector computers. It uses a new user-interface based on the interactive command language INCLAN (see below). Data file formats are compatible with the programs XEASY (Bartels *et al.*, 1995), GARANT (Bartels *et al.*, 1997), DIANA (Güntert *et al.*, 1991a), MOLMOL (Koradi *et al.*, 1996), OPAL (Luginbühl *et al.*, 1996), XPLOR (Brünger, 1992), and with the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). Furthermore, DYANA has a mechanism for the transparent use of foreign notations for atoms and angles, and new library entries for non-standard residues can readily be created with the molecular graphics program MOLMOL. The program package includes standard residue libraries for the ECEPP/2 (Momany *et al.*, 1975; Némethy *et al.*, 1983) and AMBER (Cornell *et al.*, 1995) force fields.

INCLAN

The interactive command language INCLAN (P.G., unpublished), which is used also by the NMR data processing program PROSA (Güntert *et al.*, 1992), the automatic assignment program GARANT (Bartels *et al.*, 1997), and the molecular dynamics program OPAL (Luginbühl *et al.*, 1996), provides DYANA with a highly flexible user-interface. INCLAN is an interpreted programming language that combines features of Unix shells with those of conventional programming languages such as FORTRAN or C, and produces high-quality two-dimensional graphics. INCLAN gives the user the possibility to create macros, new commands that are built from existing ones with the help of conditional statements, loops, jumps, variables, functions, FORTRAN-77 mathematical, logical and character expressions, etc. Many standard commands of DYANA are implemented as macros, and can therefore be modified by the user without changing the source code of the program.

Parallelization

A structure calculation that consists of the independent generation of multiple conformers from the same input data has a high degree of inherent parallelism. For shared-memory parallel computers, almost optimal parallelization is achieved by DYANA on the level of the command language INCLAN, without any modifications in the specialized core parts of the program. In INCLAN, parallel execution of a loop is accomplished by creating several copies of the program in its current state, using the Unix system call *fork*. These copies are identical except for the value of the loop counter, and run in parallel. Except for one instance of the program, the "main process", all copies terminate after the execution of the last statement in the loop body.

Standard DYANA simulated annealing protocol

The DYANA standard simulated annealing protocol used for all structure calculations in this work consists of the following steps.

- (1) Create a start conformation by selecting all torsion angles as independent, uniformly distributed random variables.
- (2) Exclude all hydrogen atoms from the check for steric overlap, and increase the repulsive core radii of those heavy atoms that are covalently bound to hydrogen atoms by 0.15 Å with respect to their standard values (Güntert *et al.*, 1991a). Set the weighting factor for user-defined upper and lower distance bounds to 1, for steric lower distance bounds to 0.5, and for torsion angle constraints to 5 Å².
- (3) Perform 100 conjugate gradient minimization steps at level 3, i.e. including only distance constraints between atoms up to three residues apart along the sequence, followed by further 100 minimization steps including all constraints. Optionally, for large systems, 100 minimization steps can be performed at each of the intermediate levels 1, 3, 10, 20, 50, 100, 150, 200, 300, 600, etc.
- (4) Perform $N/5$ TAD steps at a constant high reference temperature, T_{high} , where N is the user-defined total number of TAD steps. The initial time-step, $\Delta t = 2$ fs, is adapted such as to achieve a relative accuracy of energy conservation of $\epsilon^{\text{ref}} = 0.005$.
- (5) Perform $4N/5$ TAD steps with reference values for the temperature and the relative accuracy of energy conservation of:

$$T^{\text{ref}}(s) = (1 - s)^4 T_{\text{high}}$$

$$\epsilon^{\text{ref}}(s) = 0.005 \cdot 0.02^s \quad (29)$$

The parameter s varies linearly from 0 in the first time-step to 1 in the last time-step.

- (6) Include all atoms into the check for steric overlap, reset the repulsive core radii to their stan-

dard values, increase the weighting factor for steric constraints to 2.0, and perform 100 conjugate gradient minimization steps with inclusion of all constraints.

- (7) Perform 200 TAD steps at zero reference temperature and $\varepsilon^{\text{ref}} = 0.0001$. Initial velocities are those from the last time-step of step (5) above, and the length of the initial time-step is set to one fourth of the last time-step.
- (8) Perform 1000 conjugate gradient minimization steps, including all constraints.

Throughout the calculation the list of the van der Waals lower distance bounds is updated every 50 TAD steps, or, during minimization, whenever a torsion angle has changed by more than 10° since the last update, or after 100 minimization steps. Unless otherwise noted, the default values for the initial temperature, $T_{\text{high}} = 9600$ K, and the number of torsion angle dynamics steps, $N = 4000$, were used.

In DYANA the standard simulated annealing protocol is implemented as an INCLAN macro and can be modified or replaced by the user without changing the source code of the program.

Input data sets for structure calculations

For the present evaluation of the performance of DYANA, we mainly used the experimental NMR input data sets of NOE upper distance limits and torsion angle constraints for BPTI (Berndt *et al.*, 1992), *Antp*(C39S) (Güntert *et al.*, 1991b), ADB (Vendrell *et al.*, 1991), PrP(121-231) (Riek *et al.*, 1996), and CypA (Ottiger *et al.*, 1997). The data sets for BPTI, *Antp*(C39S) and ADB are identical with those used previously by Güntert & Wüthrich (1991) to evaluate the performance of the REDAC strategy. In addition, experimental NMR input data sets for the pheromone Er-2 from *Euplotes raikovi* (Ottiger *et al.*, 1994), the killer toxin from *Williopsis mrakii* (Antuch *et al.*, 1996) and the pathogenesis-related protein P14a from tomato (Fernández *et al.*, 1997) were used to measure the data in Figure 2.

For structure calculations with the 394 residue protein 3-phosphoglycerate kinase (PGK), we created a simulated set of NMR conformational constraints on the basis of the high-resolution X-ray crystal structure (Davies *et al.*, 1994; PDB code 1PHP) as follows: Hydrogen atoms were attached to the crystal structure with the program MOLMOL (Koradi *et al.*, 1996). NOE distance constraints were generated for all proton-proton pairs (excluding OH and SH) that are closer than 4.5 \AA , whereby the distance bound was set to the actual distance plus 0.5 \AA . Constraints with methyl groups were referred to pseudo atoms, and the appropriate correction was added to the distance limit (Wüthrich *et al.*, 1983). Constraints were then processed by DYANA in order to account for the absence of stereospecific assignments (these were included only for the isopropyl methyl groups of Val and

Leu) and to remove irrelevant constraints (Güntert *et al.*, 1991a): 14,161 relevant upper distance limits and 1117 ϕ , ψ and χ^1 torsion angle constraints were thus generated, where the torsion angle constraints were taken to be $\pm 60^\circ$ about the value in the crystal structure.

A second simulated NMR data set was used for structure calculations with a 32 nucleotide pseudoknot RNA, APK. Constraints were derived from the NMR solution structure (Kang *et al.*, 1996; PDB code 1KAJ) in the same way as for PGK, except that angle constraints were created for all torsion angles, and that the allowed range for torsion angles in sugar rings was reduced to $\pm 5^\circ$ about the value in the reference structure. The resulting data set consisted of 769 relevant upper distance limits and 284 torsion angle constraints. In addition, the following five upper and five lower distance limits were used to "close" the sugar rings: $1.39 \text{ \AA} \leq d(\text{C4}', \text{O4}') \leq 1.41 \text{ \AA}$, $2.38 \text{ \AA} \leq d(\text{C4}', \text{C1}') \leq 2.40 \text{ \AA}$, $2.39 \text{ \AA} \leq d(\text{C5}', \text{O4}') \leq 2.39 \text{ \AA}$, $2.11 \text{ \AA} \leq d(\text{H4}', \text{O4}') \leq 2.12 \text{ \AA}$, and $2.19 \text{ \AA} \leq d(\text{C3}', \text{O4}') \leq 2.28 \text{ \AA}$.

Structure calculations

For each of the input data sets in Table 2, structure calculations were performed using the standard simulated annealing protocol with $N = 2000$, 3000, 4000, 6000, 8000, 10,000, 12,000, 16,000 or 20,000 TAD steps. The default parameters were used, except that additional relaxation was included for the large protein PGK, as outlined in step (3) of the standard simulated annealing protocol, and that equation (3) was used in place of the default term of equation (2) to calculate the target function for the RNA. Calculations were performed on DEC Alpha 8400 5/300 computers, using up to eight processors in parallel. For each system and each value of N , at least 40 acceptable (see the next section) conformers were computed. The exact number of accepted conformers may vary between 40 and 47, since the calculations were started in parallel in groups of up to eight conformers.

For comparison, structure calculations for the experimental data sets of Table 2 were also accomplished with DYANA using the REDAC strategy (Güntert & Wüthrich, 1991), whereby the parameters were set as given by Güntert & Wüthrich (1991) for BPTI, *Antp*(C39S) and ADB, by Riek *et al.* (1996) for PrP(121-231), and by Ottiger *et al.* (1997) for CypA.

Acceptance criteria for conformers

Two factors determine the overall efficiency of an NMR structure calculation: the time needed to calculate one conformer, and the "success rate", i.e. the percentage of conformers that reach small residual constraint violations as manifested by low residual target function values. Criteria for identification of acceptable conformers are necessarily somewhat arbitrary. For a given molecule and con-

straint data set, the residual target function value can be seen as resulting in part from inconsistencies in the data set manifested in the target function value at the global minimum, and in part from contributions due to local minima. The former depends only on the data set, whereas the latter generally depends also on the algorithm used. Along these lines, we define an acceptable conformer as one with a target function value, V , below a cutoff, $V_{\text{cut}} = V_{\text{min}} + \kappa N_r$, where V_{min} is an approximation of the global minimum value of the target function, N_r denotes the number of residues in the molecule, and $\kappa = 0.02 \text{ \AA}^2$ for proteins and 0.04 \AA^2 for nucleic acids is the tolerated contribution per residue from local minima. V_{min} is the lowest residual target function value found for the system in a calculation of 40 to 47 acceptable conformers, using the standard simulated annealing protocol, except that $N = 20,000$. The resulting cutoff values, V_{cut} were: 1.25 \AA^2 for BPTI, 1.93 \AA^2 for *Antp*(C39S), 1.77 \AA^2 for ADB, 3.02 \AA^2 for PrP(121-231), 4.02 \AA^2 for CypA, 9.03 \AA^2 for PGK, and 1.30 \AA^2 for APK.

Acknowledgments

Financial support by the Schweizerischer Nationalfonds (project 31.32033.91), the use of the C4 and Cray J-90 clusters of the ETH Zürich, and the use of the NEC SX-4 supercomputer of the Centro Svizzero di Calcolo Scientifico are gratefully acknowledged. We thank Mrs R. Hug for the careful processing of the typescript.

References

- Abe, H., Braun, W., Noguti, T. & Gō, N. (1983). Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles in proteins. General recurrent equations. *Comput. Chem.* **8**, 239–247.
- Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids*, Clarendon Press, Oxford.
- Antuch, W., Güntert, P. & Wüthrich, K. (1996). β Ancestral $\beta\gamma$ -crystallin precursor structure in a yeast killer toxin. *Nature Struct. Biol.* **3**, 662–665.
- Arnold, V. I. (1978). *Mathematical Methods of Classical Mechanics*, Springer, New York.
- Bae, D. S. & Haug, E. J. (1987). A recursive formulation for constrained mechanical system dynamics: Part I. Open loop systems. *Mech. Struct. Mech.* **15**, 359–382.
- Bartels, C., Xia, T. H., Billeter, M., Güntert, P. & Wüthrich, K. (1995). The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR*, **5**, 1–10.
- Bartels, C., Güntert, P., Billeter, M. & Wüthrich, K. (1997). GARANT: a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comput. Chem.* **18**, 139–149.
- Beeman, D. (1976). Some multistep methods for use in molecular dynamics calculations. *J. Comput. Phys.* **20**, 130–139.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
- Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.
- Berndt, K. D., Güntert, P., Orbons, L. P. M. & Wüthrich, K. (1992). Determination of a high-quality NMR solution structure of the bovine pancreatic trypsin inhibitor (BPTI) and comparison with three crystal structures. *J. Mol. Biol.* **227**, 757–775.
- Braun, W. & Gō, N. (1985). Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* **186**, 611–626.
- Brooks, B. R., Brucoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
- Brünger, A. T. (1992). *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR*, Yale University Press, New Haven, USA.
- Brünger, A. T. & Nilges, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Quart. Rev. Biophys.* **26**, 49–125.
- Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Jr, Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.
- Davies, G. J., Gamblin, S. J., Littlechild, J. A., Dauter, Z., Wilson, K. S. & Watson, H. C. (1994). Structure of the ADP complex of the 3-phosphoglycerate kinase from *Bacillus sterothermophilus* at 1.65 Å. *Acta Crystallog. sect. D*, **50**, 202–209.
- Fernández, C., Szyperski, T., Bruyère, T., Ramage, P., Mössinger, E. & Wüthrich, K. (1997). NMR solution structure of the pathogenesis-related protein P14a. *J. Mol. Biol.* **266**, 576–593.
- Güntert, P. (1993). Neue Rechenverfahren für die Proteinstrukturbestimmung mit Hilfe der magnetischen Kernspinresonanz, PhD thesis, 10135 ETH, Zürich.
- Güntert, P. & Wüthrich, K. (1991). Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *J. Biomol. NMR*, **1**, 446–456.
- Güntert, P., Braun, W., Billeter, M. & Wüthrich, K. (1989). Automated stereospecific ^1H NMR assignments and their impact on the precision of protein structure determinations in solution. *J. Am. Chem. Soc.* **111**, 3997–4004.
- Güntert, P., Braun, W. & Wüthrich, K. (1991a). 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. *J. Mol. Biol.* **217**, 517–530.
- Güntert, P., Qian, Y. Q., Otting, G., Müller, M., Gehring, W. J. & Wüthrich, K. (1991b). Structure determination of the *Antp*(C39 \rightarrow S) homeodomain from nuclear magnetic resonance data in solution using a novel strategy for the structure calculation with the programs DIANA, CALIBA, HABAS and GLOMSA. *J. Mol. Biol.* **217**, 531–540.
- Güntert, P., Dötsch, V., Wider, G. & Wüthrich, K. (1992). Processing of multi-dimensional NMR data

- with the new software Prosa. *J. Biomol. NMR*, **2**, 619–629.
- Güntert, P., Mumenthaler, C. & Wüthrich, K. (1996). DYANA: Torsion angle dynamics for structure calculation and automatic assignment of NOESY spectra. *Keystone/Colorado, USA, August 18–23, 1996, Abstracts XVIIth International Conference on Magnetic Resonance in Biological Systems*.
- Havel, T. F. & Wüthrich, K. (1984). A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular ^1H - ^1H proximities in solution. *Bull. Math. Biol.* **46**, 673–698.
- Jain, A., Vaidehi, N. & Rodriguez, G. (1993). A fast recursive algorithm for molecular dynamics simulation. *J. Comput. Phys.* **106**, 258–268.
- Kang, H., Hines, J. V. & Tinoco, I., Jr (1996). Conformation of a non-frameshifting RNA pseudoknot from mouse mammary tumor virus. *J. Mol. Biol.* **259**, 135–147.
- Katz, H., Walter, R. & Somorjay, R. L. (1979). Rotational dynamics of large molecules. *Comput. Chem.* **3**, 25–32.
- Kneller, G. R. & Hinsen, K. (1994). Generalized Euler equations for linked rigid bodies. *Phys. Rev. ser. E*, **50**, 1559–1564.
- Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
- Luginbühl, P., Güntert, P., Billeter, M. & Wüthrich, K. (1996). The new program OPAL for molecular dynamics simulation of biological macromolecules. *J. Biomol. NMR*, **8**, 136–146.
- Mathiowetz, A. M., Jain, A., Karasawa, N. & Goddard, W. A., III (1994). Protein simulations using techniques suitable for large systems: the cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. *Proteins: Struct. Funct. Genet.* **20**, 227–247.
- Mazur, A. K. & Abagyan, R. A. (1989). New methodology for computer-aided modelling of biomolecular structure and dynamics. (I) Non-cyclic structures. *J. Biomol. Struct. Dynam.* **4**, 815–832.
- Mazur, A. K., Dorofeev, V. E. & Abagyan, R. A. (1991). Derivation and testing of explicit equations of motion for polymers described by internal coordinates. *J. Comput. Phys.* **92**, 261–272.
- Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.* **79**, 2361–2381.
- Mumenthaler, C. & Braun, W. (1995). Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J. Mol. Biol.* **254**, 465–480.
- Mumenthaler, C., Güntert, P., Braun, W. & Wüthrich, K. (1997). Automated procedure for combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR*, in the press.
- Némethy, G., Pottle, M. S. & Scheraga, H. A. (1983). Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.* **87**, 1883–1887.
- Nilges, M., Clore, G. M. & Gronenborn, A. M. (1988). Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Letters*, **229**, 317–324.
- Nilges, M., Kuszewski, J. & Brünger, A. T. (1991). Sampling properties of simulated annealing and distance geometry. In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (Hoch, J. C., Poulsen, F. M. & Redfield, C., eds), pp. 451–455, Plenum Press, New York.
- Ottiger, M., Szyperski, T., Luginbühl, P., Ortenzi, C., Luporini, P., Bradshaw, R. A. & Wüthrich, K. (1994). The NMR solution structure of the pheromone Er-2 from the ciliated protozoan *Euplotes raikovi*. *Protein Sci.* **3**, 1515–1526.
- Ottiger, M., Zerbe, O., Güntert, P. & Wüthrich, K. (1997). The NMR solution conformation of unligated human Cyclophilin A. *J. Mol. Biol.* **272**, 64–81.
- Riek, R., Hornemann, S., Wider, G., Billeter, M., Glockshuber, R. & Wüthrich, K. (1996). NMR structure of the mouse prion protein domain PrP(121–231). *Nature*, **382**, 180–182.
- Rice, L. M. & Brünger, A. T. (1994). Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins: Struct. Funct. Genet.* **19**, 277–290.
- Schultze, P. & Feigon, J. (1997). Chirality errors in nucleic acid structures. *Nature*, **387**, 668.
- Stein, E. G., Rice, L. M. & Brünger, A. T. (1997). Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J. Magn. Reson.* **124**, 154–164.
- van Gunsteren, W. F. & Berendsen, H. J. C. (1977). Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.* **34**, 1311–1327.
- van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P. & Tironi, I. G. (1996). *Biomolecular Simulation: The GROMOS96 Manual and User Guide*, vdf Hochschulverlag, Zürich, Switzerland.
- Vendrell, J., Billeter, M., Wider, G., Avilés, F. X. & Wüthrich, K. (1991). The NMR structure of the activation domain isolated from porcine procarboxypeptidase B. *EMBO J.* **10**, 11–15.
- Wolynes, P., Onuchic, J. N. & Thirumalai, D. (1995). Navigating the folding routes. *Science*, **267**, 1619–1620.
- Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, Wiley, New York.
- Wüthrich, K., Billeter, M. & Braun, W. (1983). Pseudostructures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with nuclear magnetic resonance. *J. Mol. Biol.* **169**, 949–961.

Edited by P. E. Wright

(Received 15 April 1997; received in revised form 11 July 1997; accepted 14 July 1997)