A Simple Algorithm for Superimposing Sets of NMR Derived Structures: Its Application to the Conformational Study of Cephalomannine in Lipophobic and Lipophilic Solution

Guillermo Moyna, Sanjai Mediwala, Howard J. Williams, and A. I. Scott*

Center for Biological NMR, Department of Chemistry, Texas A&M University, College Station, Texas 77843-3255

Received July 17, 1996[⊗]

A simple iterative method for superimposing sets of NMR derived structures and calculation of the root mean square deviation (RMSD) of the sets is described. It was compared to the commonly used algorithm involving pairwise best fitting in the conformational study of the taxoid anticancer drug cephalomannine in lipophobic and lipophilic solvents. Lower RMSD values were obtained, indicating a better superposition of the structures in the sets. The conformations of cephalomannine in the two solvent systems reported are in good agreement with earlier conformational studies on other active taxoids.

INTRODUCTION

The 3D structure determination of proteins and natural products by multidimensional NMR techniques and molecular modeling has become an important field in the last decade. The method is well established, and several volumes have been written on the subject.[1−3]

The final step of the process involves generation of sets of conformers of the molecule that comply with distance constraints from NOE experiments as well as dihedral angle constraints from ^{1}H coupling constants. This is usually accomplished by a combination of simulated annealing followed by energy minimization of the resulting structures using an appropriate energy force field. In order to estimate the precision of the structures thus obtained, those with the lowest energy are superimposed as tightly as possible in space, and the RMSD for all corresponding atoms, or a group of selected corresponding atoms in the structures, is calculated. The global RMSD, which represents the precision of the set, is calculated as the average of the RMSDs for all the atoms included in the calculation. This is illustrated in the following equations

$$\mathrm{RMSD}_i = \left(\sum_j \left(X_{i,j}^{\ 2} - \mathbf{X}_i^{\ 2}\right)^{-1/2}\right) \div n_{\mathrm{structures}}$$

$$\text{global RMSD} = \left(\sum_i (\mathrm{RMSD}_i)\right) \div n_{\mathrm{atoms}}$$

where $X_{i,j}$ represents the *x y z* position of atom *i* in structure *j*, $\mathbf{X}_i$ is the mean position of atom *i*, $n_{\mathrm{structures}}$ is the total number of structures superimposed, and n_{atoms} is the number of atoms considered in the calculation.

The superposition of conformers is usually done by best fitting pairs of molecules in the set, repeating the process until all the possible pairs have been best fitted. The precision of the set is then calculated as the average of the RMSD from each pair of conformers evaluated.[2−6] This method gives a good estimation of the precision of the set and usually suffices for the elimination of undesired conformers. However, since the fitting process is done between pairs of conformers, the final set of superimposed structures does not represent the best overall fit of all the structures in the ensemble. This problem can be solved by applying a modified best fit algorithm to sets of more than two structures that considers all the structures of the set simultaneously. Although this is the best solution to the problem, the number of calculations involved make it unsuitable for cases where 10 or more structures of molecules with more than 200 atoms have to be best fitted.

We wish to report a simple iterative method that can be used for the superposition and global RMSD calculation of a set of conformers that takes into account all the molecules in the set simultaneously and requires less computational power than a multiple best fit superposition of all the structures of the ensemble. The algorithm is compared to pairwise best fitting in the conformational study of the natural product cephalomannine (Figure 1), an anticancer agent closely related to taxol, in lipophilic ($CDCl_3$) and lipophobic (DMSO−H_2O) solvents.

MATERIALS AND METHODS

NMR Experiments and Sample Preparation. 1D and 2D experiments were recorded on a Bruker ARX 500 spectrometer at 500 MHz, using either a ^{1}H inverse or a triple resonance HCN probe. For 1D ^{1}H experiments 32K data points were used, and 2K × 0.5 K data point matrices were employed for 2D experiments. A 600 ms hard pulse spin lock was used for ROESY experiments.[7] Solvent suppression in the H_2O containing samples was by presaturation. Cephalomannine samples were dissolved in $CDCl_3$ or DMSO−H_2O (1:1) to a final concentration of 10 mM. NMR data for both solvents is reported in Table 1. The residual solvent peak (7.24 ppm for $CHCl_3$ and 2.49 ppm for DMSO) was used for spectral calibration. Cephalomannine was obtained from a crude taxol−cephalomannine mixture by reported methods.[8]

Generation of Cephalomannine Conformer Sets. Lipophobic and lipophilic solution conformer sets were created with Sybyl 6.2 (Tripos Associates Inc.) running on a Silicon Graphics Iris Indigo R4000. Starting structures for both sets were built from the X-ray crystal structure of Taxotere[9] and had eight distance constraints and one dihedral angle

[⊗] Abstract published in *Advance ACS Abstracts,* November 1, 1996.

S0095-2338(96)00118-7 CCC: $12.00

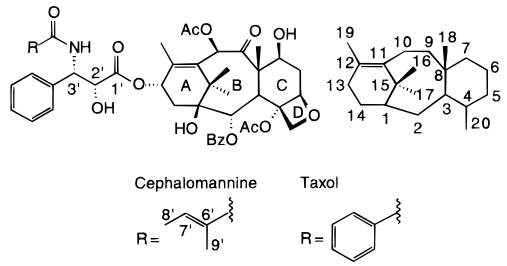

Figure 1. Chemical structure and numbering of cephalomannine and taxol.

Table 1. NMR Data for Cephalomannine in $CDCl_3$ and DMSO−H_2O Solution[a]

^{1}H on	$CDCl_3$	DMSO−H_2O
C-2	5.65 (d, 7.0)	5.37 (d, 7.3)
C-3	3.77 (d, 7.0)	3.55 (d, 7.3)
C-5	4.92 (dd, 2.3, 9.7)	4.88 (dd, 1.9, 9.7)
C-6	2.53 (ddd, 6.7, 9.7, 14.8) α	2.33 (ddd, 6.6, 9.7, 14.6) α
	1.86 (ddd, 2.3, 11.0, 14.8) *β*	1.74 (ddd, 1.9, 10.9, 14.6) *β*
C-7	4.38 (ddd, 4.0, 6.7, 11.0)	4.04 (dd, 6.6, 10.9)
C-10	6.25 (s)	6.22 (s)
C-13	6.19 (qdd, 1.2, 9.1, 9.1)	5.84 (qdd, 0.9, 9.4, 9.4)
C-14	2.28 (dd, 9.1, 15.5) α	1.84 (dd, 9.4, 15.5) α
	2.21 (dd, 9.1, 15.5) *β*	1.63 (dd, 9.4, 15.5) *β*
C-16	1.24 (s)	0.98 (s)
C-17	1.13 (s)	0.97 (s)
C-18	1.77 (d, 1.2)	1.71 (d, 0.9)
C-19	1.66 (s)	1.45 (s)
C-20	4.27 (d, 8.4) α	4.00 (m) α and *β*
	4.16 (d, 8.4) *β*	
C-2′	4.69 (dd, 2.8, 5.3)	4.47 (d, 7.0)
C-3′	5.59 (dd, 2.8, 8.8)	5.16 (d, 7.0)
C-7′	6.41 (qq, 1.3, 6.8)	6.32 (qq, 1.4, 6.9)
C-8′	1.70 (qd, 1.0, 6.8)	1.64 (qd, 0.9, 6.9)
C-9′	1.79 (dq, 1.0, 1.3)	1.69 (dq, 0.9, 1.4)
NH	6.48 (d, 8.8)	
OAc-10	2.23 (s)	2.06 (s)
OAc-4	2.34 (s)	2.15 (s)
OBz-*o*	8.10 (dd, 1.1, 7.8)	7.91 (dd, 1.4, 7.2)
OBz-*m*	7.49 (dd, 7.2, 7.8)	7.56 (dd, 7.2, 7.9)
OBz-*p*	7.60 (tt, 1.1, 7.2)	7.66 (tt, 1.4, 7.9)
Ph-*o*	7.40 (m)	7.25 (dd, 1.5, 7.8)
Ph-*m*	7.38 (m)	7.32 (dd, 7.5, 7.8)
Ph-*p*	7.32 (m)	7.15 (tt, 1.5, 7.5)

[a] Chemical shifts are indicated in ppm and coupling constants in Hz.

constraint. Conformer sets were generated from the starting structures by 100 cycles of simulated annealing. Each cycle consisted of heating at 800 K for 1000 fs followed by exponential annealing to 200 K for 1000 fs. Energy minimization of the annealed structures was performed using Tripos force field with Gasteiger−Hückel charges.[10,11] The ten lowest energy conformers obtained by this process were used for structure comparison and superposition.

Algorithm Description. The target of the algorithm is the minimization of the RMSD between the average molecule of the set and all the molecules in the set. The average molecule is calculated in the first iteration, all the conformers are pairwise best fitted to it, and a global RMSD is calculated. The process is then repeated, a new average molecule is calculated, and all the molecules of the set are pairwise best fitted to it, obtaining a second global RMSD. The iterative process is continued until a convergence criteria is reached. A difference between the global RMSD from two consecutive iterations smaller than 5 ppm was used as termination criteria. An example of the code needed for the algorithm, written in Sybyl Programming Language (SPL), is outlined below.

```
uims define macro iterative_superposition sybylbasic yes

# initialize variables and calculate the number of molecules in the set

setvar loop true
setvar rmsd 0
setvar rmsd_1 0
setvar number_of_molecules %count(%mols(*))

# repeat until convergence reached

while loop

# prepare a concatenated list of the molecules in the set

  setvar molecule_list %mols(*)
  setvar molecule_list %set_create($molecule_list)
  setvar mollecule_list %cat("$molecule_list" "m100")

# calculate the average molecule

  average_mol $molecule_list |> $nulldev

# superimpose all to the average and calculate the RMSD of the fitting

  for each_molecule in %mols(*-m100)
    match m100({atom_selection}) $each_molecule({{atom_selection}}) |> $nulldev
    setvar rmsd %math( $rmsd + $match_rms )
  endfor

# calculate the global RMSD for the iteration and the tolerance

  setvar rmsd %math( $rmsd / $number_of_molecules )
  setvar conconvergence %math( $rmsd / $rmsd_1 )
  setvar conconvergence %math( $convergence - 1 )
  setvar conconvergence %abs( $convergence )

# if convergence is reached stop iterations

  if %strgt("0.000005" "$convergence")
    setvar loop ""
  endif

# display the global RMSD from the last iteration

  echo $rmsd

# save the RMSD value in rmsd_1 and reset variables for next loop

  setvar rmsd_1 $rmsd
  setvar rmsd 0
  zap m100

endwhile
```

Copies of the program in SPL format can be obtained by e-mail from the authors at gmoyna@cbnmr.chem.tamu.edu.

RESULTS AND DISCUSSION

Algorithm Evaluation. Two sets of ten low energy cephalomannine conformers superimposed with the iterative algorithm described above are shown in Figure 2. They represent the preferred conformations of cephalomannine in lipophilic and lipophobic solvents. In both cases the atoms of the amide residue were not considered in the superposition due to the lack of significant NMR constraints. The global RMSD obtained was 0.42 ± 0.02 Å for the lipophilic set and 0.39 ± 0.02 Å for the lipophobic set. The global RMSDs calculated by the pairwise best fit method using the same set of conformers were, respectively, 0.46 ± 0.08 and 0.42 ± 0.08 Å. The amide residue was not considered in the superposition for this case either. The decrease in the global RMSD indicates that the iterative method does a better job in overlapping the conformers of the set than the pairwise best fit. By considering the average molecule as the reference for the fitting, an equally weighted component of each molecule in the set is taken into account simultaneously in the process. This removes the bias of the pairwise best fit and gives a better superposition of the molecules in space.

Description of Cephalomannine Conformers. The conformations of cephalomannine in the two solvent systems correlate well with the conformation of taxol and other active taxoid compounds previously reported by us and other groups.[12−15] For the lipophobic solution conformers, the 4-acetyl group is in close contact to protons H2′ and H3′, as well as to the *o*-protons of the 3′-phenyl ring, indicating a clustering of the A-ring side chain and the 2-benzoate under the diterpene moiety. A <H2′-C-C-H3′> dihedral angle of

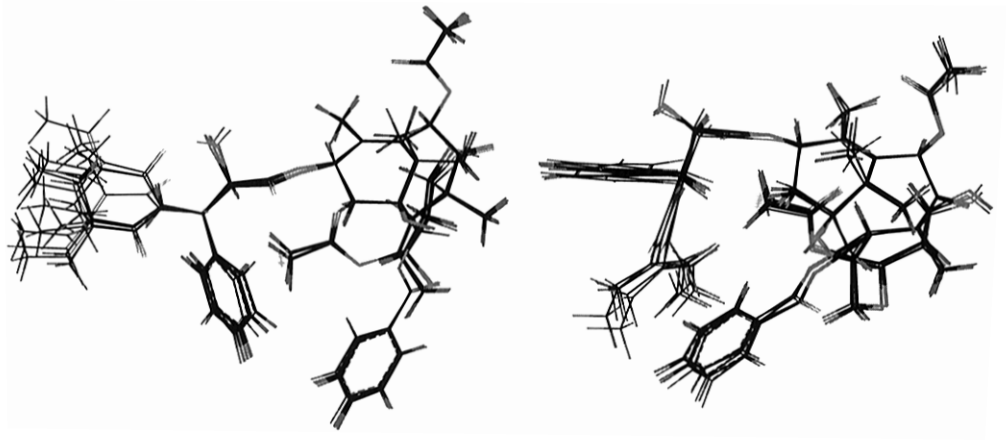

Figure 2. Superposition of ten NMR derived structures of cephalomannine in lipophobic and lipophilic solvent systems.

152 ± 5 degrees indicates an anti arrangement of the side chain in lipophobic media. For the lipophilic solution conformers, the 3′-phenyl points away from the 2-benzoate and the diterpene moiety, indicating no appreciable clustering of lipophilic groups. The <H2′-C2-C3-H3′> dihedral angle was 52 ± 5 degrees in this case, indicative of a gauche configuration of the side chain in lipophilic media.

Taxol binds and stabilizes tubulin in its polymeric state preventing cell division,[16] and it is presently used for treatment of breast and ovarian cancer. However, little information exists about the structure of the taxol−tubulin complex or the structure of the drug binding site in the protein, which is crucial for the development of taxol-like synthetic drugs. This makes the determination of the conformation of taxol and active taxol analogs in different solvent systems that can mimic different cellular environments a matter of great interest. All active taxol analogs previously studied have the same general lipophobic and lipophilic conformations, and a correlation of activity with the lipophilic conformation has been proposed.[13] In the case of cephalomannine this correlation could also be established, but other binding conformations cannot be ruled out until the 3D structure of a representative number of active and inactive taxol analogs is analyzed.

Application to Conformational Studies of Small Peptides. The method is not limited to small organic molecules. It has been used successfully in conformational studies of neuropeptide analogs from the allatostatins family. These peptides regulate insect growth and represent good leads for the development of pseudopeptide pest management agents.[17] A family of ten low energy conformers of analog 396-1 [Ala-Arg-Pro-Tyr-Asn-Aic-Gly-Leu-NH_2] generated from NMR constraints by simulated annealing and energy minimization was superimposed using the two methods discussed avobe. A global RMSD of 0.6 ± 0.2 Å was obtained when the superposition was done with the iterative algorithm reported here. When done by the pairwise best fit method, the global RMSD for the ten structures was 0.8 ± 0.4 Å. Only backbone atoms were considered for the calculations in both cases. Although these molecules possess a higher degree of freedom than cephalomannine, better structure superposition was also obtained with the iterative algorithm. A detailed report of these studies will be published elswhere.

CONCLUSIONS

Although not as rigorous as a multiple best fit in which all the conformers are considered simultaneously, the algorithm reported here has several advantages over previous molecular superposition methods. These are ease of implementation, increased speed compared to more computer intensive algorithms, and, most importantly, removal of bias in the superposition calculations. Its application to the conformational study of cephalomannine in lipophobic and lipophilic solvents helped us to conclude that this active taxoid has the same general conformation as other active taxoids previously reported. The iterative algorithm works better than the pairwise best fit method for superimposing conformer sets of small peptides as well, making it a useful tool in conformational studies of a wide variety of molecules.

ACKNOWLEDGMENT

The authors wish to acknowledge Dr. Ricardo Rodriguez Iglesias and Dr. Joseph Reibenspies for insightful discussions. Financial aid from the NIH is also acknowledged.

REFERENCES AND NOTES

(1) Wütrich, K. *NMR of Proteins and Nucleic Acids*; Wiley & Sons: New York, 1986.
(2) Evans, J. N. S. *Biomolecular NMR Spectroscopy*; Oxford University Press: New York, 1995.
(3) Williamson, M. P. NMR of Proteins. *Nat. Prod. Rep.* **1993**, *10*, 207−232.
(4) Nigles, M.; Clore, M. G.; Gronenborn, A. M. A simple method for delineating well-defined and variable regions in protein structures determined from interproton distance data. *FEBS Lett.* **1987**, *219*, 11−16.
(5) Nyburg, S. C. Some uses of a best molecular fit routine. *Acta Crystallogr.* **1974**, *B30*, 251−253.
(6) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1976**, *A32*, 922−923.
(7) Kessler, H.; Griesinger, C.; Kerssebaum, R.; Wagner, K.; Ernst, R. R. Separation of cross- relaxation and *J* cross peaks in 2D rotating-frame NMR spectroscopy. *J. Am. Chem. Soc.* **1987**, *109*, 607−609.
(8) Rimoldi, J. M.; Molinero, A. A.; Chordia, M. D.; Gharpure, M. M.; Kingston, D. G. I. An improved procedure for the separation of paclitaxel and cephalomannine. *J. Nat. Prod.* **1996**, *59*, 167−168.
(9) Guéritte-Voegelein, F.; Guénard, D.; Mangatal, L.; Potier, P.; Guilhem, J.; Cesario, M.; Pascard, C. Structure of a synthetic taxol precursor: *N-tert*-butoxycarbonyl-10-deacetyl-*N*- debenzoyltaxol. *Acta Crystallogr.* **1990**, *C46*, 781−784.
(10) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(11) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; ACS Monograph 177, 1982.
(12) Williams, H. J.; Scott, A. I.; Dieden, R. A.; Swindell, C. S.; Chirlian, L. E.; Francl, M. M.; Heerding, J. M.; Krauss, N. E. NMR and molecular modeling study of the conformations of taxol and of its side chain methyl ester in aqueous and non-aqueous solution. *Tetrahedron* **1993**, *49*, 6545−6560.
(13) Williams, H. J.; Scott, A. I.; Dieden, R. A.; Swindell, C. S.; Chirlian, L. E.; Francl, M. M.; Heerding, J. M.; Krauss, N. E. NMR and molecular modeling study of active and inactive taxol analogs in aqueous and non-aqueous solution. *Can. J. Chem.* **1994**, *72*, 252-260.
(14) Williams, H. J.; Moyna, G.; Scott, A. I.; Swindell, C. S.; Chirlian, L. E.; Heerding, J. M.; Williams, D. K. NMR and molecular modeling study of the conformation of taxol-2′-acetate in chloroform and aqueous dimethyl sulfoxide solutions. *J. Med. Chem.* **1996**, *39*, 1555-1559.
(15) Vander Velde, D. G.; Georg, G. I.; Grunewald, G. L.; Gunn, C. W.; Mitscher, L. A. "Hydrophobic collapse" of taxol and Taxotere solution conformations in mixtures of water and organic solvents. *J. Am. Chem. Soc.* **1993**, *115*, 11650−11651.
(16) Schiff, P. B.; Fant, J.; Horowitz, S. B. Promotion of microtubule assembly *in vitro* by taxol. *Nature (London)* **1979**, *227*, 665.
(17) Nachman, R. J.; Tilley, J. W.; Hayes, T. K.; Holman, G. M.; Beier, R. C. Pseudopeptide mimetic analogs of insect Neuropeptides. In *Natural and Engineered Pest Management Agents*; Hedin, P. A., Menn, J. J., Hollingworth, R. M., Eds.; ACS Symposium Series 551, American Chemical Society: Washington, DC, 1994; Chapter 15.

CI960118S