Interactive flexible molecular fitting program to be integrated into computer-aided molecular modelling systems

J Lejeune, A G Michel* and D P Vercauteren

Laboratoire de Chimie Moléculaire Structurale, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles, 61, B-5000 Namur, Belgium
*Laboratoire de Chimie Structurale, Université de Sherbrooke, Sherbrooke, J1K 2R1, Canada

Barino's flexible molecular fitting program FMFIT is significantly revised by the use of the distance matrix algorithm. The implementation of a new program, Improved or Interactive Flexible Molecule Fit (IFMFIT), into an in-house molecular modelling system is detailed. An application is presented relevant to the field of computer-aided drug design studies. The possibility of chemical and spatial equivalences between functional groups of rigid opiates and those of conformationally constrained cyclic analogues of enkephalin is demonstrated.

Keywords: computer-aided molecular modelling, molecular fitting program, distance matrix algorithm, IFMFIT, rigid opiates, enkephalin

Received 3 December 1984
Revised version accepted 28 July 1986

MOLECULAR FITTING CONCEPTS AND DISTANCE MATRIX ALGORITHM

Three-dimensional models have been used for many years as routine tools in the general understanding and analysis of structural properties of molecules, or as a visual aid for investigating spatial molecular relationships.

The estimations given by the classical mechanical models (such as the Dreiding or CPK models) were too crude, and computer generated molecular models and computer-assisted molecular modelling systems are commonly used. One of the definite advantages of these computer models and systems lies in their ability to help structural chemists to obtain a better understanding via more reliable quantification and systematization of the molecular structural parameters.

Among various techniques developed in computer-aided molecular modelling systems, superposition and comparison methods, often called 'molecular fitting', are now widely used for the investigation of structural data as an aid in the more general approach of structure–activity relationships; it must be kept in mind that each piece of geometrical information is part of a complex puzzle.

The essential purpose of these approaches is to compare series of molecules on a pair wise basis — one molecule being considered as the reference, the other as a comparison — in order to discover structural and geometrical similarities or differences. This is achieved by geometrical least-squares minimization on the 3D atomic coordinates.

From a computer graphics point of view, molecular superposition does not differ from any other methods used to display molecular structures, and most of the computer graphics chemical design systems are now providing such facilities[1]. Moreover, several groups have attempted to resolve typical superposition problems, especially for large molecules[2].

Molecular representations may be considered as being very dependent on the way users interact with pictorial data. The use of wire frame representations of molecules is widespread due to the ease of handling highly interactive manipulations; it is now common to find representations based on sophisticated molecular surfaces and shapes[3] and more complex comparison concepts such as excluded volumes[4].

For smaller molecules — less than 100 atoms — the most widely used fitting technique was introduced in 1974 by Nyburg[5] in a computer program called BMFIT (*B*est *M*olecular *Fit*)[6]. In Nyburg's scheme, the user needs only to define two rigid sets of three stereochemically comparable atoms. One set is taken as fixed for the reference molecule; the other rotates rigidly during the fitting process. Essentially, the procedure consists of computing the best rotational angle around the best rotational axis. Since direction cosines for this axis are unknown, the best molecular fit will be obtained only after successive minimizations; this means that the rotated atomic set is iteratively compared with the fixed (or reference) set until the sum of squares of interatomic distances between the two sets, obtained during two successive cycles, becomes smaller than a given threshold.

The main drawback of BMFIT comes from its inability

 0263-7855/86/040194-06 $03.00

to take into account the internal rotational freedom of the flexible molecules. If this were possible, the result would be a much better and more complete matching of molecular fragments during the fitting calculation. This shortcoming was eliminated in 1981 by Barino with the introduction of the FMFIT program (*F*lexible *M*olecular *Fit*)[7]. Based on the same algorithm as BMFIT, FMFIT's subdivides the molecules into groups of atoms and 'branches'; it also matches two molecules 'branch to branch'. A 'branch' is described as an internally rigid set of atoms.

For interactive computer graphics, the least-squares fitting procedure used for flexible structures represents an adequate interactive graphic application. This is because the user may take into account all the internal degrees of freedom in order to adjust the geometries of the molecules with respect to each other. This would be a meaningless purpose using a simple rigid fit. A user first defines the rotatable bonds, then the 'branches' that will rotate orthonormally around one or more bonds (each rotation can in fact involve one or more groups of atoms), and finally the atoms to be matched in each molecule.

Despite its great theoretical and practical interest, the FMFIT program has not been widely adopted, essentially because its full exploitation entails a tedious data input. In other words, in addition to the atomic Cartesian coordinates, it required many formatted data lists (rotatable bonds, composition of groups and 'branches', and atoms to be matched) without having available an easy checking procedure or visual feedback. The original FMFIT program was, in fact, a batch-processing oriented application.

The authors' first attempt to improve the FMFIT input mode, in order to interface it to the interactive molecular modelling system developed at the Chemistry Department of the University of Namur[8], involved the implementation of a conversational input mode connected to a continuous checking on the consistency of the user options with the intramolecular connectivity. Furthermore, a natural way, from the user's point of view, to introduce FMFIT data, was to take into account all the intrinsic information derived from the molecular structure definition itself, freeing the user to provide all the groups and 'branches' composition. All this data is, indeed, automatically defined by selecting the flexible fitting degrees of freedom (i.e. the possible internal rotations). The only thing needed, therefore, is the intramolecular connectivity matrix for both structures.

Regarding the molecule as a connected and labelled graph of points and lines, the atoms and the bonds, respectively (the atoms being the nodes of the graph), the intramolecular connectivity of a target structure may be expressed mathematically by the adjacency matrix **A** of that graph. The matrix **A** is an $N*N$ square symmetric matrix, where N is the total number of atoms in the molecule. The matrix elements $\mathbf{A}(i,j)$ are 1 or 0 if the atoms i and j are, or are not, bonded to each other; the diagonal elements **A**(i,j) are all equal to 0[9]. This matrix contains, in fact, all the connectivity information about the structure. Therefore, for this reason and due to its ideal form for computer storage as a Boolean array, the adjacency matrix is now extensively used in chemical structure codification and cataloguing, canonical numbering of atoms, partitioning into classes, discovery of functional groups, substructure sorting and retrieval, etc...[10].

Another matrix of interest is the distance matrix $\mathbf{D}(N,N)$ in which each array element **D**(i,j) has the value of the number of edges on the shortest path from node i to node j in the graph. For any molecular structure, **D**(i,j) would be the number of bonds which separates atom i and j on the shortest path from i to j.

Considering the molecular superposition problem, the above defined distance matrix $\mathbf{D}(N,N)$ will be an essential aid for discovering which atoms of the molecule form a rotatable subgroup just by computing differences between appropriate matrix elements. Let us suppose that we want to rotate around a given bond J–K. The corresponding torsional angle is defined by the atomic indices I, J, K and L, where I is the first of several possible atoms connected to J located on the part of the molecule which is kept fixed during the rotation process, and L is the first neighbour of K which is to be moved. Every atom X involved in the rotation around J–K must satisfy the following relation:

$$\mathbf{D}(J,X) - \mathbf{D}(K,X) = +1$$

The atoms of both molecules are consequently partitioned into three groups. For the molecule being fitted to the reference molecule: group 1 contains the atoms to be matched; group 2 contains the set of atoms involved in the rotation around the given bond, excluding the one to be matched; group 3 is the group of all remaining atoms (groups 1 and 2 are the actual 'branches' of the original flexible fitting program FMFIT). For the reference molecule: group 1 contains the atoms to be matched; group 2 is a 'null' group for consistency; and group 3 all remaining atoms.

Implementation of the distance matrix algorithm, originally wirtten in PASCAL[11], allowed the complete revision of the input mode of the program. The resulting program has been called *I*mproved or *I*nteractive *F*lexible *M*olecular *Fit* (IFMFIT).

MOLECULAR MODELLING SYSTEM ARCHITECTURE

Hardware architecture

The graphics workstation used comprises a Megatek Whizzard 7210 graphics processor with 16 kbyte of 32 bit-words memory. The graphics processor is connected to a CPU processor — a Digital PDP 11/60 with 128 kbyte of 16 bit-words core memory and 15Mbyte of disc memory — using a direct memory access connection. Picture display is performed by a full refresh monochromic calligraphic monitor with 4096 × 4096 addressable points resolution. The workstation is fully equipped with interactive input devices, namely a Summagraphics 2000 data tablet, a keyboard, a joystick, and a function box (with 16 buttons and 8 dials). The 3D appearance of the objects displayed on the screen is achieved through continuous real-time rotations performed by the Megatek hardware HCRST (*H*ardware *C*lipping, *R*otating, *S*caling, and *T*ranslating) module. Computer graphics output is obtained using a MEGATEK hardware 'Rasterizor' unit connected to a Versatec V-80 printer plotter.

The PDP 11/60 and the University mainframe Digital

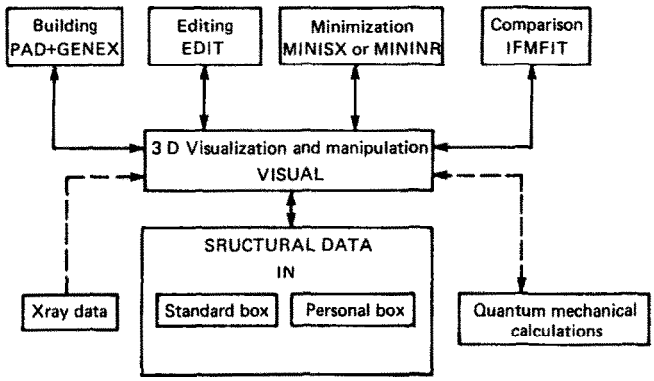

Figure 1. Schematic representation of the software architecture of the molecular modelling system developed at the University of Namur

DEC 20/60 are connected by a serial line for information interchange.

Software architecture

All the molecular modelling software developed at the University of Namur has been written in FORTRAN IV (F4P) under PDP11 RSX-11M OS, using the graphics primitives from the Megatek Wand 7200 Graphics Package. Wand 7200 respects the GSPC Siggraph Core Graphics Standard guidelines[12], which insures software portability.

Owing to the lack of core memory, the software has been divided into different modules and overlays, each overlay being responsible for a specific process. The actual system is shown schematically in Figure 1; its functions may be assumed as:

- an assembling module divided into two overlays, PAD, a 3D model building system utilizing the data tablet input[13], and Genex, a 3D interactive building system allowing links to existing molecular fragments;
- an editing module EDIT allowing minor modifications (atom label, atom type, bond length, valence and torsion angles, . . .) to existing structures;
- an energy minimization module MINISX or MININR (using the *simplex* or *N*ewton-*R*aphson method) based on molecular mechanical concepts;
- a comparison module IFMFIT based on the previously described molecular fitting concepts;
- all modules are linked together by VISUAL, a 3D visualization (with scaling, translation, and rotation) and manipulation (rotation around freely rotating bonds) module which also allows geometrical calculations and analyses.

Structural data obtained by monocrystal X-ray diffraction measurements, search and retrieval from the Cambridge Crystallographic Database, or quantum mechanical (semiempirical or *ab initio*) calculations are transferred from the mainframe DEC to the PDP with the Kermit software. Computer graphics utilities[14] are also provided for displaying individual structures on medium-resolution graphics terminals or on a multipen plotter connected to the mainframe.

Data organization

Structural data needed for and produced by the authors molecular modelling system are organized in two basic different and related ways, called 'MOL' and 'BOX' type organization.

A 'MOL' data file is described as the set of formatted structural data needed to characterize a single molecular structure. A typical 'MOL' file content would be the following:

- a structure index name and header,
- a list of atom identifiers (atom type and atom numbering indices),
- a list of orthonormalized x, y, and z atomic coordinates,
- molecular bonds information (internal connectivity setting and bond type indices).

The 'BOX' data file is defined as a collection file of molecules, or fragments (the simplest case being a single atom, such as a sp^3 carbon atom); each fragment data set having the same format as in the 'MOL' case. Each 'BOX' file carries its own directory header. Further retrievals, using the structure index names as keyword, are therefore easily achieved.

Utilities are provided to manipulate these 'BOX' files. The main possible operations are creation, updating, insertion, removal, extraction, and directory[15].

We have assumed, until now, that all structural data — especially the 3D atomic coordinates — were easily available. Trying to set a data pipeline, from raw data to structural analysis results, implies strong coherence within the data paths, and good program interfaces between all data processing steps. We have tried to provide the user with useful tools to overcome these obstacles. A typical implemented aid[16] provides an interface between the Shelx X-ray structure determination programs[17], the Cambridge Crystallographic Database, the main quantum mechanical programs and the graphics system internal data organization. Using X-ray or theoretical data, 'MOL' files, containing the orthogonalized atomic coordinates but also the atom type and the intramolecular connectivity, are automatically produced for further use.

The intramolecular connectivity setting is based upon the assumption that, if atom i and atom j are connected one to each other, the distance d(i,j) must lie in the interval defined by:

$$C(i,j) - T < d(i,j) < C(i,j) + T$$

where C(i,j) is the half sum of covalent radii for atoms i and j and T is a given tolerance factor.

The atom type is automatically induced by the internal connectivity, e.g. a carbon atom with 4 neighbouring atoms will be an sp^3 carbon. If hydrogen atoms were a feature not provided by the Shelx program, corrected connectivities and the setting of proper atom type indices could be improved using an 'hydrogen positions findings' algorithm, or after user modifications upon request.

The atom type indices used are:

Carbon	sp^3, sp^2, sp, and aromatic
Nitrogen	sp^3, sp^2, sp, and aromatic
Oxygen	sp^3, and sp^2
Sulphur	sp^3, and sp^2
Phosphorus	sp^3
Hydrogen	
Fluorine	
Chlorine	
Iodine	
Bromine	

IFMFIT PROGRAM

Having decided to compare two molecules, one for reference (molecule A) and the other for comparison (molecule B) with a respective total number of atoms N(A) and N(B), the user is prompted to specify the input files for both molecules. These files may be 'MOL' or 'BOX' files; in the latter case the user scans the entire BOX to find the desired one.

Using the internal molecular connectivity (list of pair wise atom indices), distance matrices are computed for structures A and B. Due to the square symmetrical character of the distance matrix, they are stored as the upper and lower triangle of an $N \times N$ matrix whose size N is the maximum value of N(A) and N(B).

Using the concept of computer graphics segments, each molecular structure will exist in a particular segment. Atom labelling will be considered to be independent of the graphics segments. This allows the user to easily manipulate the molecular graphics representations using the transformation module of the HCRST hardware (particular dials are set for rotation, translation, and scaling). The user is also able to interactively modify the graphics attributes for each segment (with a dial for changing the intensity level, and buttons for turning on/off the visibility, line type drawing, blinking, etc...).

The 3D Cartesian coordinates contained in the input data files do not always correspond to the best orientation when displaying the molecule on the screen. The user therefore has the ability to set the best orientation for viewing using the function box rotation dials; their coordinates are updated according to the 3D transformation matrix and subsequent molecular pictures will respect the given orientation.

Starting the superposition process itself, the user is prompted to specify the first fitting step type. Two alternatives are offered here:

- a rigid fitting step
- a flexible fitting step

The rigid fitting step corresponds to the original BMFIT scheme[6]. The user is prompted to enter two sets of atoms whose position are to be matched. At least three atoms of the two sets are assumed to be stereochemically comparable. The order used to enter atom indices is essential, since it implies the geometrical setting of the origin of the reference system. Two possibilities are offered: to choose two particular comparable atomic positions or to choose the centroids of two corresponding sets of atoms. The fitting process translates the two chosen points to the origin and keeps them coincident during the whole of the rigid calculation. The process is determined by the minimization of the sum of the squares of the distances between corresponding atoms, using the global translational changes as variables.

Alternatively or subsequently, the user has the opportunity to perform the flexible fitting process. This part of the application starts with the freely rotatable bond and pairs of atoms to be matched being chosen; the minimization process will only use the dihedral rotation around the chosen bonds as variable. The part of the molecule affected by the dihedral rotation and including the atoms to be fitted is reoriented to achieve the best matching. Consecutive steps allow successive rotations around bonds interactively selected by the user.

The computer graphics representation for the molecule under study is updated after each step, providing the user with three superposed molecular pictures: the unmodified reference molecule, the molecule under study in the initial state (before any fitting steps), and the actual state of the molecule. Input graphics devices may be used to manipulate all of the representations in order to clarify conformational changes and geometrical modifications. At each step, the 3D Cartesian coordinates of the compared molecule are also updated and converted into the coordinate system of the reference molecule. The process ends by providing a root mean square (RMS) value, considered as a typical geometrical 'agreement' value of the pair wise analysis.

Optionally, it is possible to obtain the overlap percent between nonbonded atoms in the rotated molecule for each 'branch'-to'branch' comparison. Overlaps are computed by considering atoms as rigid spheres having Van der Waals radii. The overlap sum value gives a rough estimate of the conformational stability for the flexible molecule, and may be used as a fitting efficiency indicator during successive steps. Nonbonded atoms interactions (such as 1–4 interactions) are easily found using the distance matrix since, in this case, the distance matrix element **D**(i,j) for two nonbonded atoms must have a value equal or superior to three.

Recently, Morffew has proposed another way of comparing two molecular conformations so that the distance matrices are represented as a net, with the height of the intersections above the base proportional to the distance between corresponding atoms[18].

APPLICATION TO DRUG DESIGN

Introduction

Studies with synthetic analogues of endogenous pentapeptides have established the relative importance and contribution to opiate activity of the structural elements in the peptide backbone and side chains of enkephalins[19].

Conformationally constrained cyclic analogues of enkephalin which have been prototyped by the cyclic peptide Tyr-cyclo (N^{δ}-D-Orn-Gly-Phe-Leu-) (see Figure 2(b)) bind to brain opiate receptors with high affinity. The *in vitro* opiate activity is subject to similar structural and stereochemical factors as linear enkephalins[20].

Using IFMFIT the authors have investigated the possibility of topological congruency of pharmacophoric groups (particularly the tyrosine residue and phenylalanine side chain) in the cyclic peptide (see Figure 2(b)) with corresponding functional groups in the model rigid opiate PEO (see Figure 2(a)).

Fitting procedure

The conformational preferences of the aromatic rings for the 'rigid' opioid PEO were determined by molecular mechanics calculations[21]. The PEO conformer was retained as the 'rigid' reference model for the fitting process (see Figure 3(a)). The backbone conformation of the cyclic peptide Tyr-cyclo (N^{δ}-D-Orn-Gly-Phe-Leu-) (see Figure 2(b)) was invoked as described by Kessler *et al.*[22].

The procedure respects the following guidelines:

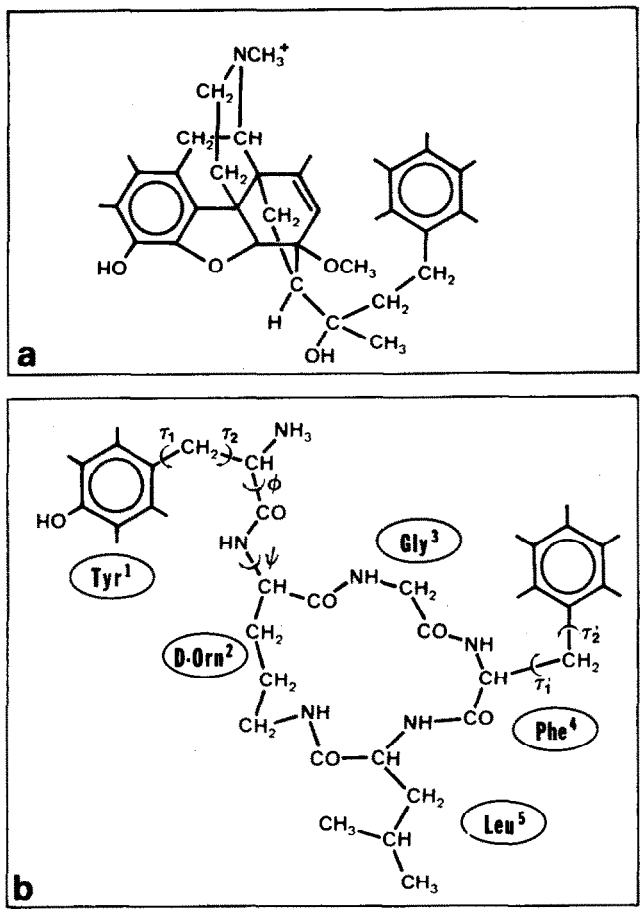

Figure 2. Planar molecular structure of (a) the rigid opiate PEO; and (b) the cyclic peptide Tyr-cyclo (N$^{\delta}$-D-Orn-Gly-Phe-Leu-). The freely rotating bonds used in the flexible fitting step are labelled: τ_1, τ_2, τ'_1, τ'_2, φ and ψ

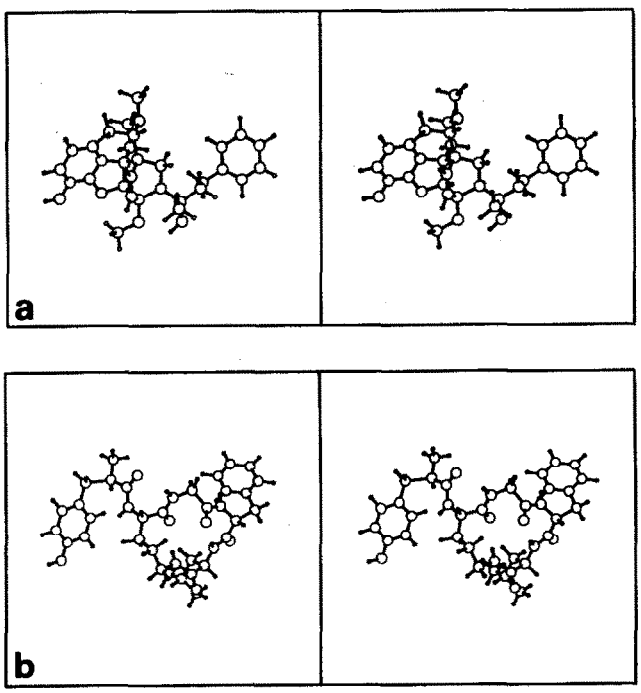

Figure 3. Stereoscopic view of (a) the rigid opiate PEO and (b) the cyclic peptide Tyr-cyclo (N$^{\delta}$-D-Orn-Gly-Phe-Leu-) after flexible fitting around the freely rotating bonds labelled τ_1, τ_2, τ'_1, τ'_2, φ, and ψ in Figure 2

- molecules are compared by pairs;
- the first molecule PEO is kept fixed;
- the second molecule is allowed to modify its conformation by rotations around specified torsional angles to achieve the best fit.

First step: a rigid fit. The ends of the vectors normal to the phenyl and phenol rings of the peptide, the phenolic oxygen, and tyramine nitrogen were superimposed on the corresponding atoms of PEO for the rigid fit.

Next step: a flexible fit (varying τ_1, τ_2, τ'_1, τ'_2, φ, ψ). Torsional angles used in this analysis are defined by clockwise rotations around the appropriate bonds according to the convention of Klyne and Prelog[23].

The ends of the vectors normal to the phenyl and phenol rings of the peptide were superimposed on the corresponding atoms of PEO. The resulting conformer has been shown to be energetically reliable. Only minor changes were observed after energy minimization by molecular mechanics techniques.

RESULTS

The ability of the cyclic peptide to adopt a conformation whose phenyl and phenol rings coincided with the PEO corresponding moieties is demonstrated. Figure 3(b) shows a stereoscopic view of the cyclic peptide fitted on PEO as presented in Figure 3(a). Examination of the stereo structures reveals that:

- We find that the conformation reported for the cyclic opioid peptide Tyr-cyclo (*N*$^{\delta}$-D-Orn-Gly-Phe-Leu-) provides the possibility of topographical congruency between the aromatic ring of PEO and that of the phenylalanine residue in the peptide.
- The tyramine moiety does not assume an identical geometrical arrangement as the same fragment found in the alkaloid.
- The basic nitrogens of the peptide (Figure 3(b)) and PEO conformer (Figure 3(a)) are 0.9 Å apart. This finding suggests that the N locus for receptor activation is variable; this agrees with previous reports[24].

CONCLUSIONS

Using the properties of the distance matrix, a molecular superposition problem for flexible structures can be significantly improved; this ends with a program called IFMFIT which presents a data input mode suitable for a highly interactive molecular modelling system.

The comparison between the results obtained with IFMFIT and those of similar applications performed on more sophisticated systems[21] demonstrate the reliability of this program.

The portability of this interactive version of the flexible fit justifies the efforts needed to implement this version of the flexible fitting program into local molecular modelling systems.

ACKNOWLEDGEMENTS

The authors thank Dr L Barino for providing the FMFIT program source code and for valuable contacts

on the molecular fitting problems. They also wish to thank UCB (Braine-l'Alleud, Belgium) for financial support, and Dr L Rodriguez (UCB), Dr B Burtin (UCB) and Prof F Durant (Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium) for fruitful discussions.

REFERENCES

1 **Barry, C D and MacAlister, J P** 'High performance molecular graphics, a hardware review' in **Sayre, D (Ed)** *Computational crystallography* Oxford Science Publications, Clarendon Press, UK (1982) p 274
2 **MacLachlan, A D** *Acta Crystallogr. A* Vol 38 (1982) p 871
3 **Morffew, A J** *J. Mol. Graphics* Vol 1 No 1 (March 1983) p 17
4 **Marshall, G R et al.** 'The conformational parameter in drug design: the active analogue approach' in **Olson, E C and Christoffersen, R E (Eds)** *Computer assisted drug design* Symposium Series 112, American Chemical Society (1979) p 205
5 **Nyburg, S C** *Acta Crystallogr. B.* Vol 30 (1974) p 251
6 **Yuen, P S and Nyburg, S C** *J. Appl. Cryst.* Vol 12 (1979) p 258
7 **Barino, L** *Comp. Chem.* Vol 5 (1981) p 85
8 **Lejeune, J and Michel, A** 'An interactive graphics system for 3-D molecular analysis' in **Frigerio, A and Milon, H (Eds)** *Chromatography and mass spectrometry in nutrition science and food safety* Elsevier Science, Amsterdam (1984) p 249
9 **Spialter, L** *J. Amer. Chem. Soc.* Vol 85 (1963) p 2012
10 **Colbourn, C J** *Bibliography of the graph isomorphism problem, Technical Report 123/78* Computer Science Department, University of Toronto, Toronto (1978). **Hendrickson, J B and Toczko, A G** *J. Chem. Inform. Comput. Sci.* Vol 23 (1983) p 171. **Hendrickson, J B et al.** *J. Chem. Inform. Comput. Sci.* Vol 24 (1984) p 195
11 **Bersohn, M J** *Comput. Chem.* Vol 4 (1983) p 110
12 *Status report of the graphics standard planning committee* SIGGRAPH-ACM Vol 13 (1979)
13 **White, W G et al.** *J. Chem. Educ.* Vol 59 (1982) p 515
14 **Michel, A** *Program VISUAL (Fortran 77 DEC 20/60) using the multi-purpose Graphics Package DI-3000* Facultés Universitaires Notre-Dame de la Paix, Belgium (1983)
15 **Lejeune, J** *Program GENBOX (Fortran 77 DEC 20/60)* Facultés Universitaires Notre-Dame de la Paix, Belgium (1983)
16 **Lejeune, J and Vercauteren, D P** *Program GENMOL (Fortran 77 DEC 20/60)* Facultés Universitaires Notre-Dame de la Paix, Belgium (1984)
17 **Sheldrick, G** *SHELX76, a system of computing programs* Cambridge University, UK (1976)
18 **Morffew, A J** *J. Mol. Graphics* Vol 1 No 2 (June 1983) p 43
19 **Morley, J S** *Ann. Rev. Pharmacol. Toxicol.* Vol 20 (1980) p 81
20 **Di Maio, J et al.** *Life Sci.* Vol 31 (1982) p 2253
21 **Bayly, C I et al.** *Int. Narcotics Res. Conf.* Vol P130 (1985) p 42
22 **Kessler, H et al.** *Int. J. Pep. Prot. Res.* Vol 25 (1985) p 267
23 **Klyne, W and Prelog, V** *Experientia* Vol 16 (1960) p 521
23 **Shiotani, S et al.** *J. Med. Chem.* Vol 19 (1976) p 803