BUTTERWORTH HEINEMANN

MOLGEN: Personal computer-based modeling system

P. Baričič and M. Mackov

Drug Research Institute, Modra, Slovak Republic

MOLGEN is a comprehensive molecular modeling package that runs on personal computers and allows building, drawing, storing, and comparison of molecular structures. The system contains modules for geometry optimization and conformational analysis, and modules for calculation and description of lipophilicity and electrostatic potentials and for their three-dimensional matching based on gnomonic projection. Described here are a systematic conformation search running on PCs and a new approach to 3D similarity analysis of lipophilic potentials. The drug design applicability of the program is further enhanced by database facilities for retrieving and storing any structure data sets. The program features and examples of its applications are presented.

Keywords: molecular modeling, personal computer, lipophilicity potential, conformational analysis, database, three-dimensional similarity

INTRODUCTION

Molecular modeling has become an effective method for facilitating the invention of new bioactive compounds. Until recently the packages for molecular design ran only on expensive microcomputers. The expansion of low-cost but powerful personal computers (e.g., 486 CPU PCs, Apple Macintoshes) has prompted chemists to use their computers for work other than word processing or spreadsheet calculations. In chemistry and biochemistry many useful packages designed to run on PCs exist for use in molecular modeling.[1] Thus, although excellent achievements have been recorded in the field of PC molecular modeling, there is still room for further improvements in molecular modeling algorithms.

The objective in designing the program presented here was to produce a low-cost, user-friendly system that incorporates, in addition to routinely used molecular modeling functions, new and important options not found in PC packages. Because precise structural knowledge about binding sites is not available in most drug design studies, the three-dimensional quantitative structure-activity relationship (3D-QSAR) based on molecular similarity remains an important technique in the design of novel biologically active compounds. Therefore the field of small- and medium-sized molecule similarity search was the focus of our program.

To achieve a universal package for molecular similarity studies, the following objectives were implemented into the system presented here.

1. Simple and quick molecular builder with the ability to edit chiral centers.
2. Communications with quantum chemistry methods and interpretation of their outputs.
3. User-friendly manipulation and organization of directories with the files containing molecular structures of differing formats.
4. Geometry optimization and conformational analysis.
5. Calculation and description of electrostatic and lipophilicity potentials.
6. Similarity search by 3D matching of molecular shape and electrostatic and lipophilicity potential.
7. Database storage of any three-dimensional structures with the available substructure search and other retrieval features.
8. Quality printing of molecular models or calculated potentials on the most widespread printers and plotters.

The program presented here was written for scientists, students, and teachers who are interested in molecular modeling and structure–activity relationships and who use personal computers in their everyday activities. Clear menus and context-sensitive help functions are available for all options, which are selected by mouse; MOLGEN is therefore well suited also for beginners and for occasional users. The main features are discussed in detail, with examples, in the following sections.

METHODS

Geometry optimization and conformational analysis

MOLGEN provides users with the possibility of geometry optimization and conformational analysis, using the molec-

Color plates for this article are on p. 198.

Address reprint requests to Dr. Baričič at the Drug Research Institute, Horná 36, 900 01 Modra, Slovak Republic.

Received 29 November 1994; revised 20 March 1995; accepted 29 March 1995

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00028-5

ular mechanics with the MM2 force field.[2] This molecular mechanics force field was chosen because of its wide distribution among the scientific community and the frequent updates of its force field parameters. Missing values of the applied force field are automatically fulfilled by the method of Schnur et al.,[3] or they can be added directly to the text file with force field parameters. To include hydrogen bond contributions we have implemented the Allinger et al.[4] modification of the applied MM2 force field.[4]

A grid search[5] of up to five rotation bonds was implemented to generate all conformational possibilities. All geometries represented by values of torsional angles and related energies generated by grid search are stored in files. The geometry to be sought for a low-energy conformer must fulfill the condition that its energy be smaller then the selected cut-off energy and represents local minima on the generated energy hypersurface. The retrieving of local minima is based on the following condition: for any minimum on the energy hypersurface the following must be in force:

$$\delta E(\mathrm{tor}_1, \ldots, \mathrm{tor}_k, \ldots, \mathrm{tor}_{N\mathrm{tor}})/\delta\, \mathrm{tor}_k = 0, \quad k = 1 \text{ to } N\mathrm{tor} \tag{1}$$

where E is an internal energy, tor is the ith torsion, and Ntor is the number of torsions. For finite change of torsional angle these conditions can be rewritten as

$$E(\mathrm{tor}_1, \ldots, \mathrm{tor}_k + \Delta\mathrm{tor}, \ldots, \mathrm{tor}_{N\mathrm{tor}}) > E(\mathrm{tor}_1, \ldots, \mathrm{tor}_k, \ldots, \mathrm{tor}_{N\mathrm{tor}}) \tag{2a}$$

and

$$E(\mathrm{tor}_1, \ldots, \mathrm{tor}_k - \Delta\mathrm{tor}, \ldots, \mathrm{tor}_{N\mathrm{tor}}) > E(\mathrm{tor}_1, \ldots, \mathrm{tor}_k, \ldots, \mathrm{tor}_{N\mathrm{tor}}), \quad k = 1 \text{ to } N\mathrm{tor} \tag{2b}$$

where Δtor is a rotational step. The algorithm automatically searches the file of conformers to find all the local minima fulfilling the above-mentioned conditions. The retrieved conformers are saved into a separate file. To cover all minima it is necessary to choose an appropriate size of rotational step. The size of 30 is a suitable choice for most conformational analyses. Further refinement of the selected conformers can be carried out by internal molecular mechanics.

Examples of the CPU times and numbers of generated conformers for the conformation analysis of a molecule of *m*-guanidinophenylpropionic acid phenyl ester, an enzyme inhibitor, are presented in Table 1. The molecule contains 38 atoms and the exhaustive conformational searches were carried out on a PC 386/387 (40 MHz) computer with a rotational step of 30°. The bonds in the chain connecting the phenyl groups were used as the rotational torsions. Color Plate 1 presents examples of graphic outputs from the conformational analysis. Because, as follows from Table 1, the conformational analysis is a time-consuming process, MOLGEN includes an option for treating the conformational analysis of a set of molecules in batches.

Table 1. CPU times for conformation analysis of *m*-guanidinophenylpropionic acid phenyl ester

Number of torsions	Number of geometries	Number of retrieved conformers	CPU time (min)
3	1 728	3	2.3
4	20 736	8	28.0
5	248 832	23	336.0

Lipophilicity potential: Three-dimensional molecular descriptor

While the steric or electrostatic properties of a molecular structure are dealt with through their natural three-dimensional descriptors, molecular shape and electrostatic potential, the hydrophobic properties, are much more difficult to handle in three dimensions. The currently used logarithm of the partition coefficient of *n*-octanol/water (log *P*) gives only one-dimensional information about the global molecular hydrophobicity. Such a representation of the hydrophobicity becomes insufficient when three-dimensional aspects of the molecules under study are included into QSAR. To solve this problem the concept of lipophilicity potential was introduced and applied in 3D QSAR. In MOLGEN two approaches have been implemented to express lipophilicity. In the first approach the molecular lipophilicity potential (MLP) in a point is expressed by[6]

$$\mathrm{MLP} = \Sigma\, C_i/(1 + d_i) \tag{3}$$

where C is a fragment or atomic contribution to the total log P and d is the distance between the point and atom i. In the presented program the Ghose and Crippen fragmentation scheme[7] was used.

In the second approach, to introduce the dependence lipophilicity vs. molecular geometry, for MLP Eq. (4) was suggested:

$$\mathrm{MLP} = \Sigma\, A_i C_i (f_i + 1)/A(\mathrm{max})_i/(d_i + 1) \tag{4}$$

where A_i is the actual solvent-accessible van der Waals surface of the fragment or atom i, $A(\mathrm{max})_i$ is a maximum solvent-accessible van der Waals surface of the fragment or atom i, and f_i is a stepwise function[8,9] that relates the number of possible external hydrogen bonds to the accessible area of the fragment or atom i. Summation goes through all available fragments. In that approach the same fragmentation scheme[7] was used but contrary to the former approach the following modifications have been used.

1. To increase the effect of solvent accessible surfaces of fragments, they were constituted from heavy atom and connected hydrogen atoms.
2. The reparametrization was done because the solvent accessibilities of some fragments, even though the hydrogen atoms were taken into account, were insignificant and there was no reason to include them in the account. This led to a reduction in the size of the fragment set used.

Solvent-accessible surfaces were calculated using the algorithm in which 642 symmetrically placed points on the atom van der Waals surface were tested for solvent accessibility. The solvent-water was approximated by a sphere of 1.4 Å. The calculation was repeated for eight random molecule orientations and an average value of the accessible surfaces

was used. The refined values of fragment contributions were calculated by least-squares fitting of Eq. (5):

$$\text{Log } P = \Sigma A_i C_i (f_i + 1)/A(\text{max})_i \quad (5)$$

The 507 molecules[10] consisting exclusively of C, H, N, and O atoms with geometries optimized with the MM2 force field were included in the regression. For 85 chosen fragments the method gives a standard error of 0.37 log P unit. Equation (5) can be used in calculating lipophilicity changes resulting from molecular geometry changes.

For illustration, in Color Plate 2 the lipophilicity potentials of all conformers described in Color Plate 1c are shown. The lipophilicity potential, together with electrostatic potential and molecular shape, give a sufficient basis for comparison of different molecules and quantification of their three-dimensional similarity.

Three-dimensional molecular similarity search

It has often been said that the interactions between a bioactive molecule and its binding site are governed by different intermolecular forces, which can be classified as steric, electrostatic, and hydrophobic.[11] One can conclude that two molecules (or only their parts) with similar steric, electrostatic, and hydrophobic properties would dispose themselves with similar activity. It makes sense, therefore, to perceive molecular similarity throughout the three-dimensional overlapping of the appropriate representations of these properties.

To estimate molecular similarity based on overlapping 3D properties we used the method originally described by Chau and Dean.[12,13] In this method, an arbitrary property of the molecular surface is mapped onto the sphere surface, using gnomonic projection.

The surface of a sphere can be represented by the vertices of an icosahedron or dodecahedron. As a point from which the projection will be made, the center of mass or center of particular interest can be used. If the projections for two different molecules are carried out in a consistent manner, the measure of their similarity can be quickly calculated as the sum of the differences of the studied property at the sphere points used in projections.[14] To find the mutual orientation of compared molecules that gives the best overlap of molecular properties, it is necessary to minimize the sum of the differences. This leads to an optimization in six-dimensional space consisting of three translational degrees and three rotational degrees. Because the rotational degrees have a significantly larger effect on the final results than the translational degrees, only the rotations were used in the best overlap searching. Several techniques can be used to find the maximal overlap.[13,15,16] A module that searches for the best match in two steps was built into the program presented here.

1. Estimation of the orientations of matched molecules by comparison of gnomonic projections.[17]
2. Final optimization of the object function by the direct search method of Hooke and Jeeves[18] to obtain an exact minimum.

Generally, any property available for evaluation in 3D space can be treated by the previously described approach. In MOLGEN the following species can be chosen for 3D matching: van der Waals surface, and electrostatic and lipophilicity potentials calculated on the van der Waals surface. Color Plate 3a–c show the best matchings of the molecular shapes, electrostatic potentials, and lipophilicity potentials of the lowest energy conformers of *m*-guanidinophenylacetic acid phenyl ester and *m*-guanidinophenylpropionic acid phenyl ester. The 3D matching of molecules depicted in Color Plate 3 took about 100 s of CPU time on a PC 386 (40 MHz) with mathematical coprocessor.

Database features

Concerning the problem of molecular building and parametrization of the fragmentation schemes used in an estimation of physicochemical properties, an internal MOLGEN database was established. The database contains about 1500 entries of quantum chemistry-optimized structures, 2500 experimental partition coefficients (*n*-octanol/water, log P), and 1500 experimental molar refractivities. Because most problems are caused by the construction of molecular structures consisting of polycondensed cyclic subsystems, the database was first filled with cyclic molecules. The needed structures can be retrieved by the database searching options and used in the molecular builder.

The data contained in the MOLGEN database are available by various search criteria:

1. Atom-by-atom substructure search.
2. Exact match searching.
3. Numeric field range search.
4. Text field substring search.
5. Summary formula search.

These search criteria may be combined; for example, all substituted nitrobenzenes with log P in interval $(-0.5, 2.0)$ can be retrieved simply by one query. The form of output from this searching is presented in Color Plate 4. Because the validity of the data included in the MOLGEN database is guaranteed by the authors, the data are not available for changing or deleting. MOLGEN allows the user to create his own in-house databases to meet his needs.

Integration with other systems

MOLGEN provides the user with interfaces to a wide range of external software to broaden the capabilities for molecular building. The interfaces can be divided into six groups:

1. Molecular mechanics (MM2/MM2PI [QCPE 318])
2. Quantum chemistry (AMPAC 1.0, MOPAC [QCPE 455], GEOMO [QCPE 485], PCILO [QCPE 220])
3. Molecular modeling (SYBYL,[19] ALCHEMY[1])
4. Databases (Cambridge Structural Database,[20] ISIS[21])
5. Molecular 2D editor (CHEMWINDOW[22])
6. General (x, y, z Cartesian coordinates, crystallographic coordinates)

For any interface the geometry of an exported molecule is checked; if any incorrect geometric parameters are found the molecule can be returned to the molecular editor for modification. For molecules exported into quantum chemistry methods MNDO, MINDO/3, AM1, and PM3 (in-

cluded in AMPAC and MOPAC) the following options are available.

1. Interactive definition of geometry parameters to be optimized (bond length, bond angle, torsions, and complete cyclic subsystems): In cases in which the optimization of the whole molecular geometry is not required, this option helps save CPU time.
2. Choice of key words that control the self consistent field (SCF) calculation and output options: The following choice of key words is available:

   Define the electronic state (OPEN, CLOSED, SINGLET, DOUBLET, etc.)
   Control the choice of the method (MNDO, MINDO/3, AM1, PM3)
   Control the calculation (PRECISE, NLLSQ, UHF)
   Modify the output (BONDS, DENSITY, LOCALIZE, VECTORS)

Other key words can be easily added by the internal text editor.

3. Reading the previously prepared data input files for the chosen method: The geometry parameters to be optimized can be highlighted.

In the case of using appropriate key words to produce the quantum chemistry output file, the MOLGEN allows the user to describe the calculated canonical or localized molecular orbitals, bond orders, and electron density. The optimized geometry parameters can be highlighted.

The PCILO program generates its input data file in an extremely cumbersome manner. MOLGEN can create the PCILO input data file and import PCILO results for molecules with up to 1000 atoms. The lone electron pairs required in PCILO calculations are added automatically.

The Cambridge Structural Database (CSD) is one of the most powerful sources of three-dimensional templates usable in molecular building. MOLGEN interfaces to the CSD search program QUEST88[20] and higher through the CSD standard format. The typical data file from a search session using QUEST88 consists of many crystallographic hits (usually several hundred). MOLGEN automatically decomposes such a file into MOLGEN data files containing single molecules or, more often, the molecular system including molecules of solvent. The internal MOLGEN atom types and bond types are automatically reconstructed from crystallographic atomic coordinates. The survey of all MOLGEN interfaces is presented in Table 2.

Output from MOLGEN

A plethora of different printers and plotters are available, mostly incompatible with each other. This may cause unpleasant trouble when printing on any special type of printer. We strove to surmount this problem by incorporating the external printer and plotter drivers into MOLGEN. For that purpose we used the GRAF/DRIVE PLUS package.[23] which provides the drivers for more than 900 different printers or plotters from more 140 producers. Table 3 lists the main types of available printers and plotters. MOLGEN can also export its pictures in most industry-standard graphics formats. A list of them is given in Table

Table 2. Survey of MOLGEN interfaces

Program	Option
Tripos—SYBYL MOL/2 File	Import/export
CHEMWINDOW Molfile	Import
ISIS/Base	Import
CSD (Cambridge Structural Database)	Import
x, *y*, *z* Cartesian atom coordinates	Import/export
Crystallographic atom coordinates	Import
AMPAC	Import/export
MOPAC	Import/export
PCILO	Import/export
GEOMO	Import/export
MM2	Import/export
MM2PI	Import/export

4. The MOLGEN menu system provides the user with comprehensive options for setting up printer type, quality of printing, and type of output file.

Hardware and software required for MOLGEN

MOLGEN runs on all IBM PC/AT/PS/2 or compatible computers with MS-DOS 5.0 or higher. It requires about 20MB of free space on hard disk. The mathematical coprocessor is optional but recommended. The basic version requires 640K of RAM. The extended version running in protected mode requires a minimum of 4MB of RAM. MOLGEN can draw the graphical outputs using the following graphic modes.

(640 × 350) 16 colors for graphics cards with minimum of 256Kb
(640 × 480) 256 colors for graphics cards with minimum of 512Kb (Logix, ATI Prism Elite, Maxxon, SEFCO TVGA, Imtec Zymos Poach, Trident TVGA 8800, or compatible)
(800 × 600) 256 colors for graphics cards with 1024Kb (Genoa 6400, Zymos Poach, Trident 8900, or compatible)

Table 3. Printers and plotters that can be used by MOLGEN system

1. Nine-pin dot matrix printers—including Epson FX and MX, IBM Graphics Printer and Proprinter, and Panasonic, OkiData, and other printers with Epson or IBM emulation—at up to 240 × 216 dpi
2. 24-pin dot matrix printers—including Epson LQ, NEC Pinwriter, Proprinter X24, and other printers with Epson LQ or Proprinter X24 emulation—at up to 360 × 180 dpi
3. Laser and ink jet printers—including LaserJet, DeskJet, DeskJet 500C, DeskJet 550C, PaintJet, and Canon LBP 8
4. PostScript printers
5. Hewlett-Packard pen plotters (HPGL)

Table 4. Programs with applicable import formats that are available in MOLGEN

Program	Import format
PC Paintbrush	.PCX
Bitmap	.BMP
Gem IMG	.IMG
Tag image format	.TIF
Computer graphics metafile	.CGM
AutoCad	.DXF
Windows metafile	.WMF
Word Perfect	.WPG
Color QuickDraw (PICT)	.PCT
Video Show	.VSH
Adobe Illustrator PostScript	.AIF

(1024 × 768) 256 colors for graphics cards with 1024Kb (Trident TVGA 8900, Zymos)

Coding details

MOLGEN has been written in Borland Pascal 7.0 with objects and includes about 150 000 lines of source code. The dynamic memory allocation available in Borland Pascal 7.0 with objects allows the size of molecules that may be displayed to be limited only by the amount of available memory.

Two versions of the software have been produced. The basic version will run on any PC having at least 640K of memory. This memory allows the user to display 3 000 atoms. The second version that works in protected mode requires at least 4MB of memory. Compilation in protected mode of Borland Pascal 7 with objects allows the user to utilize all memory. The 4MB memory allows the user to display a maximum of 16 000 atoms. In addition, this version includes quantum chemistry package MOPAC 6 configured for 25 heavy atoms and 25 hydrogen atoms.

CONCLUSION

The MOLGEN program provides a wide range of representation and manipulation functions for the investigation of molecules and molecular systems on personal computers, allowing various molecular mechanics and quantum chemistry calculations. Its special options for studying molecular similarity using 3D overlapping of molecular shapes and electrostatic and lipophilic potentials, options for conformational analysis, and database features make it an efficient took for QSAR-oriented studies. Work in progress includes an automatic 2D-to-3D converter and functions with which to search the database for similarities between database structures and the query structure in terms of shape or potential overlap.

REFERENCES

1 HYPERCHEM for Windows: Release 4. Hypercube, Inc., Waterloo, Ontario, Canada, 1994; ALCHEMY III. Tripos Associates, St. Louis, Missouri 63117; MOBY 1.4. Springer New Media, New York; NEMESIS V2.1 (Mac), Nemesis V2.0(PC). Oxford Molecular, Ltd., Magdalen Centre, Oxford Science Park, Oxford OX4 4GA, 1994; CHEM-X, Chemical Design, Ltd., Roundway House, Cromwell Park, Chipping Norton, Oxon OX7 5SR, UK; PCMODEL, Serena Software, Bloomington, Indiana 47402-3076

2 Burkert, U. and Allinger, N.L. *Molecular Mechanics*. American Chemical Society, Washington, D.C., 1982

3 Schnur, D.M., Grieshaber, M.V., and Bowen, J.P. Development of a internal searching algorithm for parametrization of the MM2/MM3 force field. *J. Comput. Chem.* 1991, **12,** 844–849

4 Allinger, N.L., Kok, R.A., and Imam, M.R. Hydrogen bonding in MM2. *J. Comput. Chem.* 1988, **9,** 591–595

5 Taylor, R., Mullier, G.W., and Sexton, G.J. Automation of conformational analysis and other molecular modeling calculations. *J. Mol. Graphics* 1992, **10,** 152–160

6 Furet, P., Sele, A., and Cohen, N.C. 3D molecular lipophilicity potential profiles a new tool in molecular modeling. *J. Mol. Graphics* 1988, **6,** 182–189

7 Ghose, A.K., Pritchett, A., and Crippen, G.M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. III. Modeling hydrophobic interactions. *J. Comput. Chem.* 1988, **9,** 80–90

8 Richards, N.G.J., Williams, P.B., and Tute, M.S. Empirical methods for computing molecular partition coefficients. I. Upon the need to model the specific hydration of polar group in fragment-based approaches. *Int. J. Quant. Chem. Quant. Biol. Symp.* 1991, **18,** 299–316

9 Richards, N.G.J., Williams, P.B., and Tute, M.S. Empirical methods for computing molecular partition coefficients. II. Inclusion of conformational flexibility within fragment-based approaches. *Int. J. Quant. Chem.* 1992, **18,** 219–233

10 Hansch, C. and Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Wiley, New York, 1979

11 Cohen, N.C. *Molecular Mimics and Drug Design: Trends in Medicinal Chemistry '88*. Elsevier Science, Amsterdam, 1989

12 Chau, P.L. and Dean, P.M. Molecular recognition: 3D surface structure comparison by gnomonic projection. *J. Mol. Graphics* 1987, **5,** 97–100

13 Dean, P.M. and Chau, P.-L. Molecular recognition optimized searching through rotational 3-space for pattern matches on molecular surfaces. *J. Mol. Graphics* 1987, **5,** 152–158

14 Namasivayam, S. and Dean, P.M. Statistical method for surface pattern-matching between dissimilar molecules: Electrostatic potentials and accessible surfaces. *J. Mol. Graphics* 1986, **4,** 46–50

15 Blaney, F., Edge, C., and Phippen, R.W. "Real-time" similarity calculations: A joint study between Smith-Kline Beecham and IBM. *Chem. Design Automat. News* 1993, **8,** 27–36; Blaney, F., Finn, P., Phippen, R.W., and Wyatt, M. Molecular surface comparison: Application to drug design. *J. Mol. Graphics* 1993, **11,** 98–105

16 Bladon, P. A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects. *J. Mol. Graphics* 1989, **7,** 130–135
17 Perry, N.C. and van Geerestein, V.J. Database searching on the basis of three dimensional molecular similarity using the SPERM program. *J. Chem. Inf. Comput. Sci.* 1992, **32,** 607–616
18 Kowalik, J. and Osborne, M.R. *Methods for Unconstrained Optimization Problems*. American Elsevier, New York, 1968, p. 21
19 SYBYL. Tripos Associates, St. Louis, Missouri 63117.
20 User's Manual Part I. QUEST88. Cambridge Crystallographic Data Centre, 1988
21 ISIS[TM]/Base 1.0. Molecular Design, Ltd., 1991
22 ChemWindow 2.1. SoftShell 1992
23 GRAF/DRIVE PLUS 4.0 M.K. Fleming Software, Oakton, Virginia 22121