WHAT IF: A molecular modeling and drug design program

G. Vriend

BIOSON Research Institute, Laboratory of Chemical Physics, Department of Chemistry, University of Groningen, Groningen, The Netherlands

A FORTRAN 77 computer program has been written to aid with macromolecular modeling and drug design. Called WHAT IF, it provides an intelligent and flexible environment for displaying, manipulating, and analyzing small molecules, proteins, nucleic acids, and their interactions. A relational protein structure database is incorporated to be queried. The program is suitable for most common crystallographic work. The menu-driven operation of WHAT IF, combined with the use of default values wherever user input is required, makes it very easy to use for a novice user while keeping full flexibility for more sophisticated studies. Although there are not too many unique features in WHAT IF, the fact that everything is integrated in one program makes it a unique tool for many purposes.

Keywords: molecular modeling, drug design, molecular graphics

INTRODUCTION

Molecular modeling is becoming ever more important in the fields of protein engineering and drug design. X-ray diffraction and nuclear magnetic resonance techniques are improving rapidly, but nevertheless it is unlikely that the three-dimensional structures of all molecules of interest will be elucidated in the foreseeable future. If no experimental structural data is available one needs to resort to modeled data if one wants to understand a molecule, or predict mutants with certain, altered characteristics.

Another line of work is drug design. Here often the main goal is to design a small molecule that binds tightly to a crucial target protein in a pathogen, while leaving the analogous human isoenzyme, if existing, unaffected. In many cases such a target protein is an unknown protein that needs to be modeled before the actual drug design process can start.

In all molecular modeling applications high quality computer graphics plays an important role, because visualization of results is often a most important road to new ideas. We have set out to design a computer program that meets the criteria described below. These criteria can be divided into two classes: usage criteria and design criteria. The following usage criteria have been considered:

(1) easy to use for a novice user;
(2) flexible enough to allow an expert user to perform even the most sophisticated studies;
(3) allow for all kinds of high quality graphics in two or three dimensions;
(4) incorporate sequence and structure databases;
(5) highly modular—a user should only need to know very few global commands plus the commands of the module of interest to be able to work;
(6) transparent interfaces must be available to popular programs used in the field, e.g., GROMOS,[1] MODEL,[4] PIR/PSQ,[5] etc.;
(7) online HELP must be available at several levels at every stage of user interaction;
(8) the program should be useful to molecular modelers, drug designers, and crystallographers;
(9) one command should only invoke one action—no subcommands or modes should be used. (An experienced user, however, should not be limited by this requirement.)

Design criteria taken into account:

(1) written in FORTRAN 77, therefore easy to port to other computers. Even the HELP facility should be fully FORTRAN 77;
(2) highly modular—if somebody wants to make a change or extension, only a very limited number of subroutines need to be studied;
(3) options that are not CPU intensive should be an integral part of the program; options that are CPU intensive should be submittable as batch jobs;

Dr. Vriend's permanent address is the BIOSON Research Institute, Laboratory of Chemical Physics, Department of Chemistry, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.

Address reprint requests to Dr. Vriend at his present address: EMBL, Postfach 10.2209, 6900 Heidelberg, FRG.
Received 10 July 1989; accepted 25 July 1989

(4) the idea that one command performs one option is also extended to the source code—the routines needed for the options in one menu are all stored in one library and one subroutine should always only perform one task.

WHAT IF options fall into several categories: graphics, databases, molecular comparisons, crystallographic tools, protein mutations, drug docking, structure analysis, parameter correlation analysis,[2] atom-, residue-, and molecule operations, etc.

GENERAL FEATURES

WHAT IF is written in FORTRAN 77 and uses an Evans and Sutherland PS300 series high quality graphics system. The host program contains about 125.000 lines of code. The associated "loosely-coupled" programs comprise another few hundred thousand lines. Figure 1 shows the general layout of WHAT IF. Its core is the so-called "soup." This name has been chosen because of the homology with culinary soup; both consist of water with other molecules floating around in it. In the WHAT IF soup, however, water does not need to be present for proper program operation. Also, in this soup one finds soup-related parameters, e.g., matrices, file names, pointers to databases, standard data sets, etc. Around the soup one finds the soup-interface routines. All operations that the user has available from any of the roughly fifty menus operate on the soup by means of these interface routines.

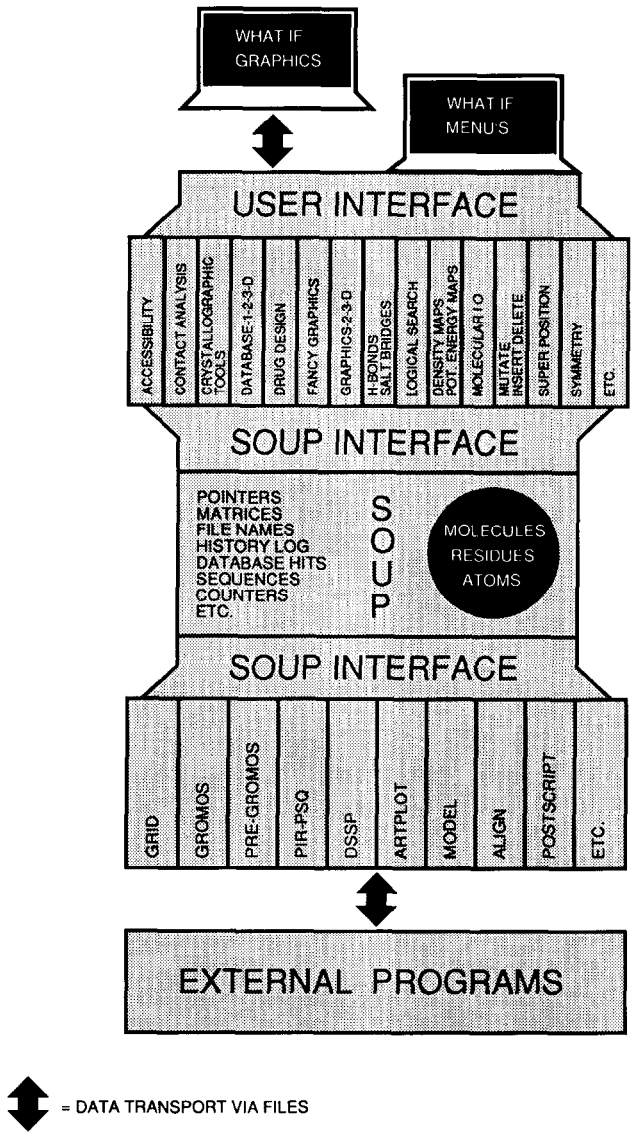

Figure 1. Organization of WHAT IF. Arrows indicate communication via files

Another set of interfaces is available for fast and flexible use of external programs. Examples are GROMOS,[1] (via PRE-GROMOS[3]) for molecular dynamics and energy minimizations, MODEL[4] for design and analysis of small molecules, and PSQ/PIR[5] for access to protein sequence databases.

Because there does not exist a really widely accepted computer graphics standard, great care has been taken to assure that all graphics options are implemented in a very general way. WHAT IF uses an object oriented graphics management system. The objects are called MOL-items. Although MOL-items are primarily meant to hold pictures of molecules or parts of molecules, they are also used for all other kinds of information. Examples of possible MOL-items are: contoured electron density maps, database hits, arrows, text strings, external general line- or dot-files, or, if the user defines them, even more exotic items like a race car or next year's calendar. MOL-items can, regardless of their contents, be rotated, translated, and scaled together with, or in the context of, the rest of the MOL-items present in the screen. Dials can be used to apply overall transformations or to operate on one or more MOL-items.

Most operations take place via a separate terminal. Only those operations that are specific for usage of the graphics system, such as picking atoms or requesting atom positions, distances, or bond angles, are performed via the graphics data tablet.

The user can, at every moment, obtain help with the requested user input by just hitting the return key. The command SHORT gives, at every moment, a one-line description for all options in the active menu. Extensive help can be obtained by typing HELP ⟨option⟩.

On a VAX computer all one line VMS commands can be executed directly from the WHAT IF menus. It is also possible to spawn subprocesses.

COORDINATES

In its present setup the program can manipulate one hundred molecules at the same time. Molecules can be of the type protein, nucleic acid, (group of) waters, ion, or "drug." As yet, no provisions have been made for polysaccharides or other polymeric molecules. All multi-atomic molecules not known to WHAT IF are treated as "drugs." Very large drugs (>1000 atoms) are not yet allowed. WHAT IF uses IUPAC nomenclature wherever possible. The standard coordinate input and output files are in the Brookhaven protein data bank format.[6] Because these files are normally not completely correct (e.g., see Ref. 7), some artificial intelligence is applied upon reading these files. WHAT IF can also read GROMOS[1] and MODEL[4] output files.

The program works internally with orthogonal coordi-

nates. Fractional coordinates (as are often used by crystallographers) will automatically be converted.

GRAPHICS OPTIONS

The main purpose of the graphics part of WHAT IF is the visualization of (parts of) molecules. This can be done in several ways (see Color Plate 1). Line drawings, ball and stick models, and smooth secondary structure dependent ribbons can be rotated in real time. Space-filling models (CPK-models) can only be produced as static pictures. Atoms can be colored individually, per residue, or per molecule as function of their type or property. Properties can be, among others, charge, B-factor, accessibility, logical combination of parameters,[2] energy terms, van der Waals radii, one-dimensional property moment,[8] or any combination of these. Every time an atom is colored, all parts of MOL-items generated later that belong with this atom will obtain that atom's color. For example, if the atoms are colored as a function of the temperature factor, then a generated van der Waals dot surface will have every dot in the same color as the atom it belongs to.

Several possiblities are available to plot screen images. Fully general formatted files that contain all vectors, together with all viewing parameters, can be written out. These files have the same format as the files generated by FRODO.[9] A program to plot these files on an HP-plotter is available. It is also possible to make a direct plot on a laser writer using a postscript format intermediate file. Solid rendered (CPK) plots can only be made using the postscript option.

DATABASES

Since information is the key to knowledge, WHAT IF has access to several databases. The EMBL, Swissprot, and NBRF protein sequence databases can be inspected using the PIR/PSQ software.[5]

A three-dimensional fragment database as described by T. A. Jones[10] is available. (The data itself will not be redistributed, but the program to reformat the PDB files to fragments is part of WHAT IF.) A relational protein structure database is part of WHAT IF. There is no SQL-like query language, but the user can ask one question at a time and combine them by means of logical operations. Questions can involve topics such as sequence characteristics, secondary structure, torsion angles, and accessibilities. One-hundred twenty-four proteins are present in this database, and routines are available to increase this to any desired extent; the availability of solved structures is the only limiting factor. The proteins that went into this database were carefully selected in such a way that no pair of them had more than 50% sequence homology.[11]

COMPARING MOLECULES

One of the WHAT IF menus allows all kinds of three-dimensional superposition and comparison options. The main principle is that the user first selects ranges of atoms or residues to be compared. WHAT IF will then generate the 3×4 matrix needed to superimpose these ranges with minimal RMS deviation. Many matrices can be generated and stored. Every matrix can then be applied to any desired range of atoms, residues or molecules. A nice feature is the automatic comparison of proteins. WHAT IF can make a pairwise comparison of every stretch of every number of residues in one molecule with every stretch of the same length in the same or in another molecule much like described by Matthews *et al.*[12] However, the application of a new algorithm makes this comparison very fast in WHAT IF. This is shown by the comparison of glutathione reductase[13] with parahydroxybenzoate hydroxylase.[14] The program found the similarity of the $\beta\alpha\beta$-fold in parahydroxybenzoate hydroxylase with the two $\beta\alpha\beta$-folds in glutathione reductase in only seventeen seconds CPU time on a VAX WS3200. The equivalence of millions of three-dimensional superpositionings had to be considered in this short time.

SYMMETRY

The symmetry menu in WHAT IF allows the user to read, type, edit, delete or apply crystallographic, noncrystallographic or, if wanted, even nonorthogonal matrices. The program provides all matrices for every crystallographic space group by merely typing the name of that space group. Integrity checks of matrices or groups of matrices can be performed. Matrices can be used to generate multiple copies of the same molecule in memory or in files. The special PS300 hardware, however, also allows for the very fast application of matrices in the graphics device. Color Plate 2 shows the sixty copies of the ~800 amino acids of one protomer of the human common cold virus R14.[15] The almost one million vectors were generated in only a few seconds.

CRYSTALLOGRAPHIC TOOLS

Almost all tools that crystallographers use in the program FRODO[9] are available in WHAT IF too. Many options, however, have been made much faster and more user-friendly.

The user is allowed to work with ten electron density maps (contoured at five levels) at a time. These maps can be brought up at the screen in their entirety if desired. The maps need no conversion prior to usage. However, maps do not necessarily need to represent electron density. Any three-dimensional data structure that can be represented on a grid can be used.

An interface to the electron density skeletonization BONES program[16] has been incorporated. Amino acid regularization is available via the algorithm of Dodson *et al.*[17]

AMINO ACID MUTATIONS

There are two ways to mutate amino acids in WHAT IF. The first is by using the fragment database in combination with the sequence data base (see Color Plate 3). The second is by an algorithm designed by R. J. Read. This algorithm uses statistically significant correlations between observed frequencies of occurrence of side chain torsion angles.[18] Mutations are visualized immediately at the graphics screen. A local conformational search can be performed to minimize spatial overlap with neighboring residues. The fragment database can also be used to search for potential deletions, insertions or loop transplantations.

COMBINING INFORMATION

The program gives the user the possibility to combine information to answer complicated questions such as: "Where are the fully buried hydrogen bond donors or acceptors that are not involved in a hydrogen bond?" The possibility exists to automatically repeat such questions for any number of PDB files. The full potential of the combination of so-called parameter correlation rows will be described in a separate article.[2]

TWO-DIMENSIONAL GRAPHICS

Often three-dimensional data is represented clearer in only two dimensions. Examples include Ramachandran plots, crystallographic B-factor plots and amino acid contact analysis plots. In WHAT IF, all these two-dimensional plots can be made. Since two-dimensional plots are sent to the graphics device as MOL-items, the graphics object manager can keep track of what is in them. Therefore, you can pick points of the plots and the part of the molecule that gave rise to them will automatically be displayed.

ENERGY CALCULATIONS

WHAT IF can use the pre-GROMOS program written by J. P. M. Postma[3] to prepare the files and input decks needed to run the molecular dynamics and energy minimization program GROMOS.[1] There are also several facilities to evaluate the results of molecular dynamics runs. A slightly modified subset of the GROMOS package can be used to obtain the contribution of every individual atom to the energy terms as calculated by GROMOS.[19]

OTHER OPTIONS

It is beyond the scope of this article to describe all of the more than five hundred options that WHAT IF makes available to the user. The program has many options to work with sequences. It can create all kinds of two-dimensional graphics; it can even create text slides. It has been the aim to give here a short overview of some of the main options of the program, and a few of the "loosely-coupled" programs. A fast drug docking option has not been mentioned, but will be the subject of a separate article.[20] Some of the other options that have not been mentioned include:

- inspection and comparison of crystallographic reflection data sets;
- manual inspection and editing of crystallographic envelop maps, needed for solvent flattening;
- accessibility calculations;
- evaluation of contacts, bond distances, bond angles, torsion angles;
- cavity searches;
- keeping a partly automatic notebook;
- three-dimensional sequence alignment;
- analysis and prediction of water positions.

DISCUSSION

As is usual with large computer programs, the work on WHAT IF is on-going. Part of the effort is used to improve the general scheme of options, and part is used to implement new algorithms and options that are the result of interaction with users. At present, the program is being used by several groups of colleagues who are active in the fields of protein engineering, molecular dynamics and pharmacology.

The program is available for a very minor fee. This includes all source codes that are produced by us and the full user manual and programmer's manual. The loosely coupled programs like GROMOS, MODEL, ARTPLOT, GRID, ALIGN, etc., that are used by WHAT IF will obviously not be redistributed. The routines needed to use these programs, and a listing of every line of code we changed in them, will be distributed upon request.

ACKNOWLEDGEMENTS

I would like to thank all my colleagues in and around the Groningen protein crystallography group and the biocomputing group at the EMBL for using and testing the program while it was developed. Their remarks, ideas and sometimes their actual help with coding has been a decisive factor for the current shape of the program. I thank Evans and Sutherland for help with the coding of some of the graphics parts and for continuous support.

REFERENCES

1 GROMOS is written by W. F. van Gunsteren, Dept. of Physical Chemistry, University of Groningen, The Netherlands
2 Vriend, G. Parameter correlation rows, a fast way to answer molecular questions. Manuscript in preparation
3 PRE-GROMOS is written by Johan Postma at the EMBL in Heidelberg, FRG
4 MODEL is an update version of C. Still's MODEL program, written by K. Steliou, University of Montreal, Canada
5 PIR/PSQ is public domain software produced by the NBRF, Georgetown University Medical Center, 3900 Reservoir Road, N.W. Washington, D.C. (PIR is a registered mark of NBRF)
6 Bernstein, F. C., *et al. J. Mol. Biol.* 1977, **112**, 535–542
7 Cherfils, J., Vaney, M. C., Morize, I., Surcouf, E., Colloc's, N. and Mornon, J. P. *J. Mol. Graphics* 1988, **3**, 155–160
8 Eisenberg, D., Weiss, R. M. and Terwillinger, T. C. *Proc. Nat'l. Acad. Sci.* 1984, **81**, 140–144
9 Jones, T. A. *J. Appl. Crystallogr.* 1978, **11**, 268–272
10 Jones, T. A. and Thirup, S. *EMBO J.* 1986, **5**, 819–822
11 Argos, P., personal communication
12 Remmington, S. J. and Matthews, B. W. *Proc. Nat'l. Acad. Sci.* 1978, **75**, 2180–2184
13 Karplus, P. A. and Schulz, G. E. *J. Mol. Biol.* 1987, **195**, 701
14 Schreuder, H. A., Laan, J. M. van der, Hol, W. G. J. and Drenth, J. *J. Mol. Biol.* 1988, **199**, 637
15 Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H.-J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Ruecjert, R. R., Sherry, B. and Vriend, G. *Nature* 1985, **317**, 145–153

16 The BONES electron density skeletonization program is written by T. A. Jones
17 Dodson, E. J., Isaacs, N. W. and Rollett, J. S. *Acta Crystallogr.* 1976, **A32**, 311–315
18 The MUTATE program is written by R. J. Read, University of Edmonton, Canada
19 The program to determine energy terms per atom is a modification of GROMOS, made by J. P. M. Postma at the EMBL, Heidelberg, FRG
20 Huckriede, D., Vriend, G. and Hol, W. G. J., manuscript in preparation
21 The very fast surface calculation program is written by R. Voorinthold, Dept. Comp. Sci., University of Groningen, The Netherlands