Comprehensive molecular modelling system

David N J White[†] and John E Pearson*

[†]Chemistry Department, The University, Glasgow G12 8QQ, UK
*Ciba-Geigy AG, R1046.108, CH-4002 Basel, Switzerland

A molecular modelling system, for small and macromolecules, which incorporates a wide range of functionality has been developed. The system is 'user friendly' and is controlled almost exclusively by a puck (mouse), in a manner akin to the Apple Macintosh. The system is written in FORTRAN 77 and the graphics adheres to the CORE standard, so that a reasonable degree of portability is assured.

Keywords: molecular graphics, molecular modelling, Chemically Oriented Graphics System

received 2 September 1985, revised 25 November 1985

Chemically Oriented Graphics System (COGS) is a molecular modelling program designed to be useful in four major areas: small molecule stereochemical design, model building, and comparison; investigation of the interaction between organic or inorganic crystal lattices with small and medium sized organic molecules; enzyme–substrate or enzyme–inhibitor docking studies; and protein engineering. The system is designed to be as comprehensive as possible, whilst remaining easy to use. COGS is almost entirely driven by a hand-held puck or mouse which is used to select items from a series of menus. (A cursor on the graphics screen tracks movement of the mouse and items are selected by positioning the cursor over them and pressing any button on the mouse. 'Pointing to', 'selecting', 'picking', and 'hitting' are synonyms for this process. 'Selecting a null atom' involves pointing to a blank area on the screen.) In other instances a different peripheral may be more appropriate (e.g. rotations are controlled by dials). This means that absolutely no knowledge of the host computer operating system is necessary in order to use COGS (the only requirement is to log-on and type COGS). Most of the facilities in COGS perform their operations immediately, with the obvious exception of molecular mechanics and molecular dynamics calculations, and conformational search procedures. COGS incorporates extensive error checking so that it is very difficult, but not yet impossible, to crash the program by doing something nonsensical. When an error is trapped the user is always given an opportunity to correct it (interactively) without losing place in a sequence of commands. COGS has been implemented on DEC VAX and Norsk Data 570 minicomputers driving Megatek 7200 or Sigmex S5600 series displays. The system and some of its facilities are illustrated in Colour Plates 1–6.

Table 1. Standard features available in COGS

Function	Purpose	Comments
Fragment build	Joins two molecules into one with the elimination of H_2	Used for making complex molecule from building blocks
Add hydrogens	Adds hydrogen atoms at all unfilled valencies	Uses an internal table of lengths and angles
Delete items	Deletes single or multiple atoms and/or molecules	
Cut/Join	Makes or breaks bonds	Splits one molecule into two or more if enough bonds are broken
Measure geometry	Measures lengths, angles and torsion angles	Atoms concerned need not be bonded
New projection	Draws Newman projections and/or mean plane views	
Solid models	Draws a colour filled CPK model of the molecules in the workspace	
Save to File	Saves workspace in Brookhaven[7] or COGS format	All atoms or only the visible ones may be saved
Superimpose	Superimposes two or more molecules	
Steric congest	Calculates steric accessibilities of potential reaction centres	Uses modified Wipke and Gund algorithm[8]
Delre charge	Calculates partial atomic charges	Uses the Delre method[9]
Areas/Volumes	Calculates molecular volumes and surface areas	Uses the Lee and Richards procedure[10]
Quit	The only way to exit from COGS	Prevents exit with unsaved workspace

COGS SYSTEM

There are a number of main menu entries in COGS, each of which is the head of a tree of a number of submenus. Because of the richness of the facilities offered it is not feasible to discuss all of them in detail. COGS is still growing and evolving both in terms of new facilities, and improvements to existing procedures; these will be reported at a later date. What follows is a condensed description of the major COGS menu trees. Descriptions of standard features are shown in Table 1, while options containing novel features are discussed more completely.

0263-7855/86/030134-09 $03.00 © 1986 Butterworth & Co (Publishers) Ltd

Display files

This option allows the reading of existing files using either COGS or Brookhaven formats. The COGS format is primarily intended for small molecules of any kind while the Brookhaven format is reserved for polypeptides, proteins, and nucleic acids. At present the maximum number of atoms handled in either format is 10 000. There is no limit to the number of independent molecules that may be accessed by this option; if the system is in Single mode (see later) the new molecule overwrites those currently in the workspace; in Multiple mode the new molecule(s) is (are) added to those already in the workspace.

When display files is selected a directory of all of the user's data files appears on the graphics screen. Selecting one of these by pointing to it with the puck, enters the file into a COGS workspace (format determination is automatic) and displays the molecule(s) on the screen.

Peptide build

Selection of successive amino-acids, from a screen menu with the mouse, results in a growing polypeptide chain appearing on the screen in an extended conformation. The polypeptide may be composed of any of the twenty 'protein building' amino-acids, or aminobutyric acid, sarcosine, ornithine, statine in any combination and in any number subject to a maximum of 10 000 atoms. The geometry of each amino-acid and the atomic charges come from COGS internal tables. Hydrogen atoms may be added automatically, or omitted, at will after the CNOS skeleton is complete. Models assembled with peptide build will overwrite any other molecule(s) currently in the workspace.

3D Quick build

This is another model building routine similar to peptide build but the building blocks are atoms rather than amino-acids[1]. Picking a succession of atom types with the puck results in the construction and simultaneous display of the molecule in 3D. If a branch is desired in the growing chain then the atom vertex corresponding to the branch point is picked before the branch atom type. The branch atom then becomes the head of a new chain to which new atoms are added (until a new branch is chosen). Bond lengths and angles for all combinations of atom types are stored in internal tables and are not the concern of the user. The torsion angles used are arbitrary. If the ring option is selected and between five and ten atoms marked with the puck, the system will present a code for all low-energy conformations of that ring size and type across the top of the screen. Pointing to the cross next to the desired code will cause the selected atoms to be folded up into the selected ring conformation. Hydrogen atoms may be added to the model prior to exit from this option. 3D quick build may be used to build a whole new molecule from scratch or may be used to add to an existing molecule.

Draw molecules

The 'draw molecules' option allows models to be constructed from a 2D structural chemical formula of the desired molecule drawn with the puck. This occurs in exactly the same way as it would be drawn by a chemist using pencil and paper. This 2D diagram, of connected atoms of different types, is converted automatically by COGS into a full 3D molecular representation which is then displayed on the screen[2]. Models constructed with draw molecules option will overwrite any other molecule(s) currently in the workspace.

Geometry build

This is the option which allows the addition of one atom at a time onto a molecule (containing a minimum of three atoms) by picking four atoms (usually three real atoms and a null atom), a bond length (*l*34), a bond angle (*a*234), and a torsion angle (*t*1234) for the group of atoms (1-2-3-4). Atom 4 is always the new atom, and its type and charge may also be specified. However, if all four atoms picked already exist in the molecule then geometry build will alter the molecular geometry without adding a new atom. The geometric parameters, atom type, and atomic charge may be changed individually or in any combination using geometry build.

Input coordinates

Crystallographic or orthogonal Cartesian coordinates for a molecule may be input via the input coordinates option. The main use of this option is the input of crystal structure coordinates which are sufficiently new as to have not appeared in the Cambridge Crystallographic Data base. A molecule input in this way will overwrite any other molecule already in the workspace.

Protein build

This option is intended to assist the building of new protein structure models starting from the already known 3D structure of a related (homologous?) protein[3]. There are therefore facilities available to substitute, delete or insert amino-acids into a preexisting protein structure.

Substitute

Selection of any atom in the residue to be substituted causes the system to respond by drawing a menu of amino-acids down the left hand side of the screen. Pointing to the desired amino-acid will cause it to be substituted for the original residue pointed to in the protein. Before performing the operation COGS will check to see that the φ and ψ torsion angles at the point of substitution in the protein are allowed values for the new residue, that helix breakers are not substituted into helical regions of the protein, that charged residues are not inserted into the hydrophobic core etc. In the event that something unusual is attempted a warning and opportunity to retract will be given. The checking is done by making use of internal tables in COGS and the topological analysis performed if secondary structure is not defined in the input (Brookhaven format) file. If everything is correct or it is desired to proceed in any event, the substitution will be made with the side chain torsion angles as close as possible to those of the original residue. If the new side chain contains a potential hydrogen donor or acceptor then a search is made for complementary atoms in spatially adjoining residues. Two options to optimize the position of the side chain atoms offered. The first is to perform a Sitar (see later) scan on the X torsion angles to find the most favourable values, while keeping the side chain bond lengths and angles fixed; and the second is to perform a full relaxation Newton–Raphson optimization of the positions of the side chain atoms only, under the influence of the forces arising from atoms less than 4.0Å distant from the side chain atoms.

Delete

The residue to be deleted is selected with the puck. Again the system will check insofar as is possible to prevent peculiar operations from being executed (e.g. attempts to delete residues not in surface loops will be queried). After the atoms comprising the residue to be deleted have been removed from the workspace the polypetide chain is rejoined with a long bond (spanning the deleted residue) and a Newton–Raphson optimization calculation initiated to close the gap. This is achieved by designating a 'molten zone' for a few residues on either side of the gap and allowing the positions of the atoms in these residues to refine in the force field defined by their <4.0Å neighbours, in a similar manner to the substitute operation.

Insert

Selection of the protein residue after which the insertion is to be made and the residue to be inserted, from the menu of amino-acids, initiates the insertion option. A range of checks similar to those described under substitute will be performed before any operation is commenced and the usual chance to retract given in the event of ambiguities. If everything is alright then the chain will be broken at the point of insertion and the new residue inserted (at this stage directly on top of the residue after the point of insertion). The φ, ψ torsion angles of the inserted residue are set to values corresponding to the A, C, D, E, F, G, A^* regions of a Ramachandran map[4] in turn, and the φ, ψ torsion angles of a nine residue molton zone (four on each side of the inserted residue) adjusted to give the best geometry in the region of the chain break, consistent with minimum change in the position of the existing residues. The results of the seven trials are then compared and the molten zone torsion angles set to correspond to the most favourable result. After insertion an energy minimization is performed in order to idealize the side chain geometries of the new residues. (Notice that these optimizations are all very fast because only a few iterations are performed on a few tens of atoms including only short nonbonded interactions — if required a minimization over the whole, or large pieces, of the protein could be performed at the end of a sequence of substitutions, insertions, and deletions but this would take several hours for a large protein).

Alter geometry

This option has two suboptions: alter geometry or close ring. Alter geometry allows bond lengths, angles or torsion angles to be varied. If the default picture type is a dot surface, alter lengths and alter angles would cause the surface to 'tear' as it moved; so the picture temporarily reverts to a coloured stick model while part of the molecular is in motion. There is no problem with dot surfaces when alter torsion angle is being used.

Alter lengths

Bond lengths may be adjusted interactively by first selecting the atom at each end of the bond with the puck. COGS will then display the current length of the selected bond and a scale/pointer in the dialogue area at the top of the screen. Pressing the left hand button on the puck will move the pointer up the scale in the positive direction and the bond length indicated will increase. Simultaneously, the bond in the on-screen molecular model will begin to stretch. Pressing the same button again will accelerate all of the above processes; while pressing the right hand button will cause deceleration and ultimately reversal of direction (i.e. the bond will contract) if pressed often enough. The stretching/contraction may be stopped at any point either by moving the pointer to the zero on the scale and pressing the yellow button on the puck, or simply by pressing the 'up' button. The user may then alter another bond length by choosing another pair of atoms and repeating the above procedure, or exit from this option by pointing to + exit with the puck. On exit from alter length the bond distance will remain at its new value.

Alter angles

This option is entered from alter length by selecting a null atom, when the indicator at the top centre of the molecule area will change from distances to angles. Valency angles may be altered in an analogous fashion to bond lengths, but by selecting three rather than two atoms. The chosen angle will be seen to bend in real-time on the graphics screen and may be stopped when the desired value is reached.

Alter torsion angles

This option adjusts torsion angles, defined by connected groups of four atoms, in a similar way to the working of alter angle with valency angles. Again, moving pictures are drawn in real-time to illustrate the motion. The alter torsion angle is entered from alter angles in an identical manner to swapping from alter lengths to alter angles, except that the indicator changes from 'angles' to 'torsion angles'. Selecting a null atom in alter torsion angles will cycle to alter length. An additional feature of this suboption is that the steric energy may be monitored during real-time bond rotation and/or short nonbonded distances arising from the rotation may be highlighted by flashing dotted lines colour coded according to length; again in real-time. The monitoring/highlighting works in real-time for all molecules up to the 10 000 atom system limit.

Close ring

It frequently happens that a ring structure is built by folding up a linear chain of atoms into a cyclic conformation using a predetermined set of torsion angles (obtained, say, from an NMR experiment). What usually happens then is that the ends of the chain are not sufficiently close to each other to make a reasonable bond. This could be corrected by a couple of iterations of energy minimization, but this is a bit wasteful. Instead, close ring will make the smallest conformational adjustment necessary to close the ring subject to a maximum ±20° change in any torsion angle. If larger adjustments are required close ring can be used repeatedly.

The close ring option is invoked by selecting the atoms defining the chain ends with the puck. COGS will then automatically work out which torsion angles to adjust and write the post-optimization end-to-end distance on the dialogue area of the screen.

Transforms

Interactive picture transforms including *x, y, z* rotation; *x, y, z* translation; and scaling may be performed on one or more molecules in real time. For single molecules turning one of seven dials affects the appropriate transformation and moves the pointer on the corresponding scale in the dialogue area. Motion is halted by turning the dial until the pointer is zeroed or exiting by depressing the puck 'up' button. When multiple molecules are displayed in single mode (i.e. the mode indicator in the dialogue area reads 'single') 'Transforms' will operate on all molecules simultaneously. Transforms works with all picture types.

If multiple molecules are displayed in the 'multiple' mode then transforms will operate on the selected molecule. If energy monitoring is requested the intermolecular energy (nonbonded, electrostatic and hydrogen bonding) between the fixed molecule(s) and the moving molecule is continuously displayed on the graphics screen in real-time (i.e. no pauses while the energy is calculated). If distance monitoring is requested then as the moving molecule comes into proximity with a fixed molecule dashed, flashing, coloured lines are drawn between atoms in the different molecules closer together than a specified upper bound. The colour of the line fixes the distance within a specific range, the key to which is displayed on the graphics screen. Requesting both energy and distance monitoring gives rise to both indications being active. When the dials are turned, only the selected molecule is transformed and the others remain stationary.

On exit from transforms the molecule remains in its new orientation. If the operation of transforms has resulted in atoms being clipped from the molecule, these atoms will remain invisible on exit from transforms and the visible atoms only may be saved with save to file if desired. In order to restore the hidden atoms the visible fragment may be translated and/or scaled down in such a manner that the hidden atoms would reappear were they not still marked as invisible. Alternatively, the suboption of new defaults which restores invisible atoms may be used.

Chain editor

Chain editor is used for interactive editing of amino-acid residues from polypeptides and proteins. The picture is redrawn to show the α-carbon atoms coloured in blue no matter what the current default picture type (see later). One residue will be drawn in its entirety in a colour corresponding to its type (i.e. basic, hydrophobic etc.) and will be shown flashing. As the appropriate dial is rotated the sequence number will increase or decrease incrementally according to the direction of rotation, and the corresponding residue together with its name (e.g. Asp 215) will be displayed. Pushing the designated 'keep' button on the dial box while rotating the dial will cause all residues traversed to remain on the screen. Pushing 'skip' button while turning the dial will erase all residues traversed. When the desired residues are displayed chain editor may be terminated by pushing any button on the puck. COGS continues with only those residues displayed by side chains and the partial sequence may be saved with save to file. Notice that until save to file is invoked the nondisplayed atoms have not been deleted but only marked as invisible. Provided that save to file has not been used, a view of the entire polypeptide may be restored by entering transforms and scaling down the partial display by an appropriate amount, or preferably by using the new defaults option which restores invisible atoms.

Clip molecule

This routine serves the same purpose as chain editor but in a much less selective manner. In this option six dials are linked interactively to clip planes (1, $+x$; 2, $-x$; 3, $+y$; 4, $-y$; 5, $+z$; 6, $-z$) which form an 'elastic sided parallelopiped'. Atoms within the volumes are visible, those without are invisible, and this distinction remains upon exit from clip molecule. Any portion of a molecule or group of molecules may be 'boxed' and the visible atoms only saved with save to file if desired.

Dot surfaces

The first implementation (+ Clip van der Waals surface) is due to Pearson, and modelled on an approach by Pearl and Honegger[5]. It is much faster than earlier algorithms. The pixels are illuminated on the van der Waals surface of the molecule and each one may be colour coded according to the atom type to which it belongs, the molecule to which it belongs, or the electrostatic potential at the pixel *xyz* coordinates. Dot surfaces must be used to generate a dot surface before the default picture type is changed to a dot surface by new defaults. After the surface has been drawn it may be interactively *z*-axis clipped, rotated, etc. The Clip van der Waals surface suboption will not allow selective dot surfacing of parts of a molecule; a surface is always generated over the whole surface of all of the molecules present.

An alternative and much faster procedure (+ Fast van der Waals surface) is also available. This works by placing precalculated dot spheres at the atom positions and scooping out the dots internal to the overlapping volume of spheres[6]. The fast procedure will calculate and display the dot-surface of a 1000 atom molecule in around 1 s and also, as a bonus, allows rotation of partial dot surfaces around bonds in real-time without 'tearing'. The surface may be coloured as before plus a few additional variations. The fast dot surface suboption allows selective surfacing of individual atoms, amino-acids, or molecules; and lists of atoms, amino-acids, or molecules by pointing to the appropriate atoms in the molecule or molecules with the puck. There are no restrictions on the colours in which the various (mixed, if desired) partial surfaces may be drawn, and these are completely under the control of the user via puck selectable menus.

Toggle flags

In this option, hydrogen atoms, atom and amino-acid labels, and the spheres with radius proportional to Z coordinates ('z-discs') are drawn in different segments from the rest of the molecule and so may be turned on and off instantaneously at will by selecting the appropriate menu item. The choice remains, in effect, after exit from toggle flags and may only be modified by using toggle flags again.

The mode of operation of COGS may also be toggled between Single and Multiple using this option. The significance of Single and Multiple is as follows. If files are read in Single mode then the molecule(s) described therein, overwrite any other molecule(s) in the current workspace and only the new molecules are drawn. If the file describes more than one molecule and these were saved in Multiple mode, then the mode is automatically switched to Multiple after reading in and display are complete. If files are read in via display files in Multiple mode then the new molecule(s) are added to the current workspace and all of the old as well as the new molecules are displayed. If multiple molecules are saved via save to file in Multiple mode then each molecule retains a separate identity and may be manipulated independently of the others. If multiple molecules are saved in Single mode, then they form an indivisible entity when recalled by display files and cannot be manipulated separately. Attempts to save single molecules in Multiple mode, and multiple molecules in Single mode are queried by COGS and an opportunity to retract given before executing the save.

The mode also affects the operation of transforms. In Single mode anything in the workspace and on the screen at the time is operated upon as a single entity. In Multiple mode with multiple displayed molecules the user is offered the opportunity to monitor the intermolecular energy or intermolecular short nonbonded contacts in real time if a 'docking' experiment is required, or simply to rotate/translate/scale one molecule with no monitoring if the purpose of the manipulation is merely aesthetic.

In other instances attempts to perform an operation in the incorrect mode will produce warning messages together with a suggested course of action. For example entering superimpose in Single mode with one molecule in the workspace will result in the user being advised to get another molecule and change to Multiple mode before COGS returns to the main menu level.

New defaults

New defaults allows the user to change various parameters affecting molecules and the display of molecules. The title of a workspace may be changed and displayed so that when save to file is used the new title is written into the file.

Atoms may be changed from one type to another and the picture redrawn to reflect this change. As many atom types as desired may be changed.

Atoms which have been rendered invisible (but not erased from the workspace) by options such as transforms, chain editor, etc. may be restored to visibility by pointing to the + Restore invisible atoms subsuboption.

The default picture type may also be changed using new defaults. When COGS is first started the default picture type is a coloured stick model, but this may be changed to red–green stereo, dot surface, or α-carbon atom backbone models. Pictures are then always drawn in the chosen type until the default is changed again with new defaults.

Sometimes the quality of a model derived from crystallographic experiments is so poor that the algorithm used for calculating connectivity does not work. Atoms are connected together if the distance between them lies within ±0.1Å of the reference bond length for the atom types concerned, which is stored internally in COGS. In order to allow correct connectivities to be calculated for poor structures the default tolerance of ±0.1Å may be relaxed to any user defined value, and COGS will continue to use this value up to the point of exit from the program or specification of a new default tolerance.

Show features

The show features option can be used to highlight various features of molecules including atomic properties, bond/group related features, amino-acid dependent features, and secondary structure in polypeptides and proteins. Each of these four options has a number of suboptions. The highlighting consists of, for example, flashing the desired feature on and off in blue for 5 s; drawing it permanently in a different colour, colour coding etc., as appropriate. Normal default picture style is reverted to on exit from show features.

Named atoms may be highlighted by supplying a list of atom names via the graphics screen pseudo keyboard; atomic properties such as type, charge, electronegativity, or coordination number may be written against all atoms displayed on the screen by selecting from the appropriate submenu with the puck. All atoms proximate to a chosen atom within nominated upper and lower bounds may be highlighted in a colour coded fashion. In the first instance the choice between intra-, inter-, or all distances must be made by menu selection, and the lower and upper bounds on distance chosen with the puck and screen keypad. The computer then calculates all of the distances concerned and divides them up into five equal ranges between the bounds. The distances are highlighted by drawing a dashed line between the atoms concerned, the colour of the line depending on which of the five ranges encloses the distance in question. All short nonbonded distances involving all atoms may be colour code highlighted in an identical manner to the highlighting of atoms proximate to a chosen atom.

Bonds and groups may be located as follows. The bond type or multiplicity (i.e. single, double, single conjugated, amide) may be written against all bonds displayed on the screen, all bonds between atoms of given types (e.g. Nsp2-Csp3) are highlighted by picking the atom types from a menu with the puck, all potential hydrogen bonds may be highlighted by dashed blue lines, and all groups of a given kind (e.g. CH3, NO2 etc.) can be shown by picking the group from a menu with the puck.

Specific amino-acid residues are located by supplying the residue names with the pseudo keyboard (e.g. Lys45 Arg167), all residues of a particular class (e.g. acidic, hydrophobic, etc.) are located by picking the class from a menu with the puck, and all residues of a specific kind (e.g. Ala, Pro, Gly) are located similarly. All residues lying within a sphere centred upon a target residue may be highlighted. The target residue is selected by pointing to any of its constituent atoms with the puck, the sphere radius is then selected with the mouse and screen keypad. The picture on the graphics screen is then redrawn showing only those residues within the chosen sphere (it must be noted that the atoms rendered invisible in this way remain invisible on exit from this option; allowing work with a specific residue and its immediate environment; frequently used in conjunction with protein build). The other atoms have not been deleted from the workspace but merely rendered invisible. All amino-acids of a specific kind such as hydrophobic, hydrophilic, positively charged, negatively charged, neutral, etc. may be highlighted individually or in groups in user chosen colours by making the appropriate menu selections with the puck.

Helices, β-sheet and bends in polypeptides are located by selecting the appropriate menu item from the dialogue area (+ helices, + β-sheet, + bends).

If multiple molecules are displayed then selecting + molecule numbers will display the molecule numbers assigned by COGS.

Compose slides

Pictures may be annotated before photographs are taken from the graphics screen. On entry to this option the molecule/surface etc. is rescaled to fill the entire graphics screen and the menu and picture frame turned off (i.e. made invisible). The picture may then be annotated with text, ikons, or graphs; before the photograph is taken.

Text may be line/stroke or outline, of any colour; outline text may be filled with any colour, not necessarily the same as the outline; and the text may be scaled to any size. After the text has been written to the screen in the desired font, colour, and size it may be 'dragged' around by moving the puck. When the text is in the desired position it may be fixed by pressing any button on the puck. The user is then returned to the text/ikons/graphs menu.

Ikons are dealt with in much the same way as text except that the ikon (square circle, sphere, arrow etc.) is chosen from a menu of ikons.

In the case of graphs a question and answer session is used to get the axes, scales and annotation. The graph is drawn from a user supplied function or set of values, and the whole picture may be scaled up or down. When everything is correct the graph may be dragged to its final position.

The screen is cleared of all extraneous dialogue etc. when the slide is ready to be taken.

Molecular mechanics

Molecular mechanics calculations of three flavours may be performed namely; pattern search energy minimization using a valence force field, Newton–Raphson energy minimization or maximization (for locating transition state structures) again with a valence force field, and molecular dynamics simulation calculations[11]. One very important and unique feature of these calculations is the fact that the valence force field employed is completely orthogonal (i.e. there is a complete set of force constants for all 29 atom types recognized by the system, in any chemically sensible combination). The force field is implemented as a 'mini' knowledge-based expert system[11]. Contributors to the steric energy are bond length stretching, valency angle bending, bond torsion, nonbonded contacts, out-of-plane bending at trigonal atoms, coulombic interactions, and hydrogen bonding. Every coordinate of every atom is allowed full relaxation.

Pattern search

In this energy minimization procedure the Cartesian coordinates of each atom are moved, one at a time, by positive and negative values of a user supplied sampling offset (the default value of which is 0.1Å) and the steric energy evaluated after every move. If at any point the energy decreases then the coordinate is allowed to remain at its offset value, otherwise it is reset to its initial value. The moves which result in an energy reduction are noted. After one pass through the coordinates of all of the atoms the energy has been lowered and a pattern of successful moves stored. It is then assumed that what worked once will work again, and the whole stored pattern of moves is applied to the model structure again and the steric energy checked. If it goes down then the new structure is retained otherwise it is reset. When a pattern is no longer successful in lowering the energy, the magnitude of each pattern offset is halved and the energy sampled again with lowering giving rise to structure retention. When the lower limit on pattern offset is reached, or no further lowering in energy can be obtained, a new pattern is established using an offset equal to half of the value at the beginning of the calculation (0.1Å). The whole calculation is cycled until no energy lowering can be obtained or a user specified limit on the number of iterations is reached. This procedure is very useful for 'three dimensionalizing' 2D structures built with the draw molecules option (in this case use 10 iterations, the default van der Waals cutoff, an initial step size or offset of 0.5Å, short or long output, and no constraints. In this context note that selecting zero for any numeric input value, by 'hitting' Enter on the 'screen keypad', will give the default value not 0). Constraints may be applied during minimization to fix various atomic positions, bond lengths, bond angles, and torsion angles or to drive these quantities to user specified values. The constraints are specified by pointing to the atoms comprising the feature to be constrained with the puck, and the desired value of the constrained quantity is again entered on the 'screen keypad' with the puck, and the severity of the constraint selected from a menu in the same way.

One very important caveat is that search procedures should not be used to perform calculations on ensembles of molecules (e.g. enzyme and substrate) because the nature of the algorithm prevents one molecule from moving with respect to the other. This problem may be circumvented by explicitly including rotation and translation of the molecules in the minimization process; however this cure is rather messy and it is better to use a Newton–Raphson optimizer (which does not have the same difficulty) under these circumstances. Furthermore, search procedures should not be used to idealize protein structures during model building because spatially adjacent strands of sequentially remote amino-acids will not move relative to each other during minimization. In this case there is no solution to the problem (separate rotation and translation of secondary structural features ?!) and search procedures should be avoided.

Newton–Raphson

As far as usage is concerned this option is almost identical to pattern search. The user may select the number of iterations required,

- an upper bound on nonbonded distances (beyond which they will not contribute to the van der Waals energy),
- a threshold on the contribution of an interaction to the steric energy (below which the interaction will still contribute to the energy but will not be listed on the long printed output),
- the maximum allowable shift in a single Cartesian coordinate during an iteration of optimization,
- whether short or long printed output is required,
- whether the second derivatives of the steric energy with respect to the coordinates should be calculated for every iteration or the same derivatives used for a few iterations (this is possible because the second derivatives change much more slowly than the first during the process of optimization),
- and whether the minimization should be constrained or unconstrained.

The Newton–Raphson option allows molecules to be fixed in place if required (the option of fixing molecules appears in pattern search but is redundant as the centroids of molecules remain fixed during pattern search). Furthermore, the minimization may be confined to a number of consecutive amino-acid residues within a larger (fixed) polypeptide chain.

There are no restrictions on the use of the Newton–Raphson optimizer but the user should be aware of some points. In the first place the usual Newton–Raphson iteration is rather sensitive to the quality of the starting geometry of the model, and may not converge from really poor starting geometries. However, the procedure used here will trap situations that it cannot deal with and temporarily resort to a 'Steepest Descents' algorithm for the duration of the problem, before reverting to the Newton–Raphson algorithm. The user sees only uninterrupted convergence.

The Block Diagonal (BDNR) approximation to the Full Matrix Newton–Raphson (FMNR) Iteration is used here[12] so that COGS can energy-minimize protein structures adequately. The FMNR procedure has storage requirements $O(9N**3)$ floating point words for the matrix alone and this means 144 Mbyte in the R*16 precision necessary to invert the matrix resulting from, say, a 3 000 atom problem. By contrast the BDNR iteration needs only 6 floating point words for matrix storage as each atom is dealt with individually. The price paid for using BDNR as opposed to FMNR is poorer convergence and seriously incomplete optimization in around 3% of molecules tackled — a small price to pay when BDNR calculations are practicable and FMNR are not (even on most 'virtual memory' minicomputers where 'virtual' means OK if the program is under 20 Mbyte in size).

Finally, it should be remembered that Newton–Raphson procedures are not strictly energy minimizers but force minimizers. This makes the Newton–Raphson procedures suitable for calculations on ensembles of molecules and for idealizing protein structures, but, the Newton–Raphson procedure is also suitable for calculating the energy maxima corresponding to transition state structures where the forces on the atoms are also minimized, as they are at an energy minimum. This makes the Newton–Raphson procedure very useful for investigating transition state structures but means that the user has to take care that the starting model employed is closer to an energy minimum than to a maximum if minimization is desired (this is the purpose of the conformational space searching options and one of the uses of pattern searching which is always guaranteed to energy minimize).

Molecular dynamics

The molecular dynamics option uses the same expert system force-field as pattern search and Newton–Raphson and also the same subroutines to calculate the forces on the atoms (which are related to $\mathrm{d}E/\mathrm{d}x$). The integration of the equations of motion may be accomplished by means of Gear, Verlet, or Beeman (or modified versions thereof)[13] numerical procedures where the timestep is specified by the user in femtoseconds. The user may also specify the viscosity of the solvent in centipoise. The equilibration of the structure and the increase up to the required temperature for the simulation run may be handled by the user, or the program will attempt to do it automatically. In either case the sequence of events will be similar but the user will have little control over the automatic procedure. The automatic procedure works as follows: first the input structure is examined to see if it has been energy minimized, by checking the forces on the atoms, if it hasn't then a Newton–Raphson energy minimization run is performed on the molecule. The temperature is then gradually raised in a series of steps (5–10°) to the required temperature by increasing the atomic velocities and consequently the kinetic energy of the system. The velocities are increased in a manner that conserves the linear and angular momentum of the system. When the desired temperature has been reached the system is equilibrated until the total energy (potential + kinetic) is as close to constant as is practicable. The system is then checked for 'hot spots' or pooling of energy in particular regions of the molecule; if this is detected then the velocities are randomly redistributed and the simulation run until the desired distribution is achieved. Running up to temperature is difficult and becomes easier with experience. At this point the simulation can begin, and run for as long as the user has access to the computer. The configuration of the system may be saved to a disc file at intervals for later analysis. A facility is provided for 'playing back' the simulation on the graphics screen; turning a potentiometer dial governs the speed of playback.

Conformational search

This option has three major suboptions which allow conformational space searches, or global energy minimization, to be executed in different ways according to circumstances and the molecule concerned.

Sitar

Sequential Iterative Torsion Angle Refinement is used to locate energy minima of open chain molecules, or the linear side-chains of cyclic molecules. The user selects as many variable torsion angles as required and a search interval in degrees. The Sitar routine will then spin around the first torsion angle and locate the minimum energy position; the torsion angle is then set to its value at the minimum. This procedure is then repeated for the second and subsequent torsion angles until all have been set to the most energetically favourable position. This process will in all probability have altered the minimum energy position of the first torsion angle, and so this is reoptimized, as are the second and subsequent torsion angles. This whole process is iterated until the conformational energy can no longer be lowered any further. Sitar may be rerun a specified number of times with randomly chosen sets of torsion angles if required. In this case the sets of torsion angle values and the conformational energy at each of the minima are recorded on a specified disc file and the molecular conformation set to correspond with the lowest energy located.

Cyclo-glomin

This suboption is used for the location of all low energy minima of cyclic molecules. The subroutine accepts a linear chain structure, corresponding to the opened ring, as input. The linear structure is then folded into a number of conformations by permuting all of the local potential energy minima around each of the isolated four atom torsion angle units composing the ring. If cyclohexane is taken as a simple example; the linear chain is -CH2-CH2-(*)CH2-(*)CH2-(*)CH2-CH2 and there are three variable torsion angles (*) where the local potential energy minima of the R1CH2-CH2R2 fragments are at −60°, 60° and 180°, giving $3 \times 3 \times 3 = 27$ generated conformations. These conformations are then adjusted in an attempt to effect ring closure; subject to a maximum variation of ±20° from the torsion angles ('generators') representing the local potential energy minima, and values of the three torsion angles which become defined on ring closure (the three dependent torsion angles in the cyclohexane example) being within ±20° of the appropriate generator value. Conformations which pass this preliminary test are passed through a battery of additional tests (which must be computationally fast) and the surviving structures are saved on a disc file for subsequent refinement by energy minimization calculations. If cyclohexene is chosen as an example -CH2-CH2-CH=CH-CH2-CH2 there are still three sets of generators but their values are (−60, 60, 180), (0, 180), and (−60, 60, 180) giving $3 \times 2 \times 3 = 18$ generated conformations. In the case of polypeptides the generators are (φ, ψ) pairs of torsion angles, corresponding to energy minima in Ramachandran maps for the isolated *N*-methyl,C′-methyl amino-acids. Generators are stored within Cyclo-glomin and do not have to be supplied by the user.

On entering Cyclo-glomin[14] an output disc file must be nominated and the two atoms defining the ring junction picked; the ring is then defined by the system to be the shortest distance between the two atoms on the connectivity tree. Once the ring is defined the system automatically locates the variable torsion angles and their associated generators. A maximum of 15 variable torsion angles or torsion angle pairs is allowed. The next step is to select a tolerance of between 0.05 and 0.40Å for ring closure and the maximum number of short nonbonded contacts to be allowed (1, 2, 5, 10) after the ring has been folded and crudely geometry optimized. Each generator for a variable torsion angle is allocated a serial number from 1 to *n* so that a ring conformation can be described in terms of a string of digits. Short combinations of generators exist which give rise to extended chain conformations so that no matter what the value of the remaining generators it will be impossible to close the ring. These combinations of generators are known as suicide sequences and may be input to Cyclo-glomin as digit strings using the puck. A few exclusion strings reduce the run time by a considerable amount. The calculations are also simplified if the ring structure is declared to be a homopolymer. The proximity of the dependent torsion angles over the ring junction to their generator values may be used to filter conformations. The calculations may be started or restarted at any point.

These calculations can take some time for a large ring structure (about 2 h for a 15 residue cyclic polypeptide). Although the calculations will run with substituents attached to the ring atoms they will take much longer and may exclude perfectly good conformations because of substituent clashes which Glomin cannot fix, but which could easily be relieved by pattern search energy minimization on return to COGS. The best way to run Glomin is with the ring atoms only, the substituents may be replaced with 'fragment build' after Glomin has generated the basic conformations.

Monte-Carlo

The Monte-Carlo[15] suboption is a conformational space searching algorithm for linear molecules or the linear side-chains of cyclic molecules. The subroutine uses the Metropolis algorithm and operates as follows. The variable torsion angles are specified in the same way as for Sitar and Monte-Carlo begins by calculating the steric energy of the initial conformation of the molecule; the torsion angles are then set to random values between −180 and +180° using the random number generator supplied in the FORTRAN run-time library and the steric energy is reevaluated. If the new steric energy is lower than the old value then the molecule is left in the conformation defined by the random number generator. If the new steric energy is higher than the old then the fraction $\exp(-\mathrm{d}E/RT)$ is evaluated and compared with a random number between 0.0 and 1.0. If the fraction is higher than the random number then the current conformation is retained, otherwise the conformation is reset to that prevailing before the torsion angles were randomly perturbed. The torsion angles corresponding to conformations in which the molecule spends a significant amount of time before escaping may be saved to a user specified disc file. The routine terminates after a predetermined number of energy evaluations. The strength of the Monte-Carlo method is that it allows the molecule to climb energy barriers in order to sample the hypersurface on the other side.

Energy contour

The molecular potential energy, or steric energy, may be calculated (without minimization at each point) and plotted on the graphics screen as a function of one or two torsion angles.

When a single torsion angle is specified the only other information required is the angular interval (+30°, +20°, +10° or +5°) between steric energy evaluations. After the calculation is complete a graph of steric energy *versus* torsion angle will be drawn and annotated on the graphics screen. The option may be exited with the molecular conformation unchanged, or set automatically by COGS to correspond to the lowest energy point on the graph.

The procedure is similar if the energy is to be plotted as a function of two torsion angles except that two atom quartets must be chosen. The results are presented as contours coloured on the graphics screen to reflect relative energy (white: 0 kcal; bright blue: 3 kcal; dark green: 6 kcal; yellow: 9 kcal; bright red: 12 kcal; brown: 14 kcal mol^{-1}) as a function of the two torsion angles. The axes are annotated at 30° intervals. As an alternative to contouring, the area between contours may be blocked out in solid colour. The molecular conformation may also be set to correspond to the lowest energy point on the map before exiting energy contour. These calculations may take some minutes on the minicomputer if a 5° interval is chosen for a large molecule.

Fold predictor

Given an amino-acid sequence either from a Brookhaven file or a manual input via the screen pseudo keyboard this module will attempt to predict the locations of helices, β-sheet, and β-bends. A number of different prediction schemes are available (e.g. Robson[16], Chou and Fasman[17], etc.).

Homology comparisons

Given the amino-acid sequences of two polypeptides, either from file or keyboard as in fold predictor, this option will attempt to align the chain sequences, so that the homology is maximized, by conservative substitutions and a minimum of insertions and deletions. The procedure is basically that of Needleman and Wunsch[18] together with subsequent improvements by Smith and Waterman[19]; the substitution probabilities are due to McLachlan[20].

Automatic dock

Automatic dock is the option that positions a substrate in a local minimum energy orientation in the active site of a receptor. Auto docker first asks for complementary pairs of atoms in the same way as superimpose, although in this case the goal is to get the atoms in each pair in van der Waals contact, not superimposed. Torsion angles in the substrate which will be allowed to vary in order to improve the receptor–substrate fit may also be chosen, or the substrate may remain rigid. Docking is achieved by a rigid body rotations and translations of the substrate, plus torsion angle rotations if required, which minimize the energy of nonbonded, electrostatic, and hydrogen bonding interactions between receptor and substrate. The routine then locates the local energy minimum 'dock' closest to the starting arrangement in hyperspace. In order to locate the best 'dock' (i.e. the global minimum) it may be necessary to try several different starting points, or to manually perturb an existing 'dock' and restart the calculation from the new position.

CONCLUSIONS

The molecular modelling system described in this paper provides a range of facilities in a single integrated package which are normally implemented as several separate programs.

Work is in hand to extend the facilities offered by COGS and this is rendered very straightforward by the modular construction of the program. COGS consists of a family of subroutines (around 150), each with a single well defined and logically distinct function, which communicate with each other via FORTRAN common blocks (arguments to subroutines are avoided where sensibly possible). Extensive use is made of structured programming techniques and long variable names (with words separated by underscores) so that the intent of each subroutine is clear and may almost be read as English text. In addition each subroutine has a multiline text header describing its function, entry points, references to the algorithm used, etc.

REFERENCES

1 **Liljefors, T** *J. Mol. Graph.* Vol 1 (1983) p 111
2 **Wipke, W T et al.** in **Wipke, W T and Howe, W J** (Eds) *Computer assisted organic synthesis* ACS Symposium Series 61 (1977) p 97
3 **Greer, J** *J. Mol. Biol.* Vol 153 (1981) p 1027
4 **Zimmerman, S S et al.** *Macromolecules* Vol 10 (1977) p 1
5 **Pearl, L H and Honegger, A** *J. Mol. Graph.* Vol 1 (1983) p 9
6 **Bash, P A et al.** *Science* Vol 222 (1983) p 1325
7 **Bernstein, F C et al.** *J. Mol. Biol.* Vol 112 (1977) p 535
8 **Wipke, W T and Gund, P** *J. Am. Chem. Soc.* Vol 98 (1976) p 8107
9 **Del Re, G** *Biochem. Biophys. Acta* Vol 75 (1963) p 153
10 **Lee, B and Richards, F M** *J. Mol. Biol.* Vol 55 (1971) p 379
11 **White, D N J** in *Computer aided molecular design* Oyez, London (1984) p 73
12 **White, D N J** *Comput. & Chem.* Vol 1 (1977) p 225
13 **Levitt, M and Meirovitch, H** *J. Mol. Biol.* Vol 168 (1983) p 595
14 **White, D N J and Morrow, C** *Comput. & Chem.* Vol 3 (1979) p 33
15 **Go, N and Scheraga, H A** *Macromolecules* Vol 11 (1978) p 552
16 **Garnier, J et al.** *J. Mol. Biol.* Vol 120 (1978) p 97
17 **Chou, P Y and Fasman, G D** *Ann. Rev. Biochem* Vol 47 (1978) p 251
18 **Needleman, S B and Wunsch, C D** *J. Mol. Biol.* Vol 48 (1979) p 443
19 **Smith, T F and Waterman, M S** *J. Mol. Biol.* Vol 147 (1981) p 195
20 **McLachlan, A D** *J. Mol. Biol.* Vol 61 (1971) p 409