Journal of Computer-Aided Molecular Design, 8 (1994) 347–366
ESCOM

J-CAMD 253

GREEN: A program package for docking studies in rational drug design

Nobuo Tomioka and Akiko Itai*

Faculty of Pharmaceutical Sciences, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan

Received 19 October 1993
Accepted 18 March 1994

Key words: Docking study; Protein–ligand interaction; Drug design; Trypsin–inhibitor complex

SUMMARY

A program package, GREEN, has been developed that enables docking studies between ligand molecules and a protein molecule. Based on the structure of the protein molecule, the physical and chemical environment of the ligand-binding site is expressed as three-dimensional grid-point data. The grid-point data are used for the real-time evaluation of the protein–ligand interaction energy, as well as for the graphical representation of the binding-site environment. The interactive docking operation is facilitated by various built-in functions, such as energy minimization, energy contribution analysis and logging of the manipulation trajectory. Interactive modeling functions are incorporated for designing new ligand molecules while considering the binding-site environment and the protein–ligand interaction. As an example of the application of GREEN, a docking study is presented on the complex between trypsin and a synthetic trypsin inhibitor. The program package will be useful for rational drug design, based on the 3D structure of the target protein.

INTRODUCTION

Computer simulation of protein–ligand interactions, so-called docking, has become one of the most important methods in computer-aided drug design. On the basis of the 3D structure of the target protein (hereafter called the receptor molecule) elucidated by X-ray crystallography, we can simulate the interaction between small bound molecules (hereafter called ligand molecules) and the protein by means of visual inspection and empirical energy calculations. Docking is also useful for elucidating the mechanisms of biochemical processes, such as enzyme reactions. Once a rational model of the protein–drug complex has been constructed by docking simulations, the model can be used for explaining relationships between structures and activities of drug molecules. Furthermore, the docked drug molecule can be modified by an interactive modeling method, so as to attain higher affinity and specificity for the target protein.

*To whom correspondence should be addressed.

As in many applications to structural studies of molecules, molecular mechanics calculations are often used for docking studies. The position and conformation of the ligand molecule can be optimized by using an empirical potential function. Since the radius of convergence of molecular mechanics calculations is rather narrow, the obtained structure is often one of the local minimum structures that is dependent on the initial model adopted for the calculation. In order to avoid this 'local-minimum problem', an appropriate initial geometry of the ligand molecule should be determined beforehand by some means, such as visual inspection with 3D computer graphics. However, an enormous number of local minima occur, because docking modes (position and orientation of the ligand relative to the protein) and ligand conformations are completely interlinked. Therefore, we need a rational method that can select likely stable docked models from all the possible models as efficiently as possible. For such a purpose, a simple interactive method to manipulate molecules on the graphics display is not sufficient. We need both proper graphics to assist the docking operation and rational scores to guide the operation in real time.

Pattabiraman et al. reported a method to calculate an approximate intermolecular interaction energy between protein and ligand in real time by using a 3D grid [1]. They decomposed calculation of the interaction energy (van der Waals and electrostatic) into two phases. On every 3D grid point in the binding-site cavity, they calculated and stored intermediate results of the energy calculation. When the ligand molecule is placed in the grid region, the interaction energy can be calculated in real time, using the precalculated data. However, these grid-point data cannot be used for other purposes, such as graphic expression, because they do not directly represent physical properties of the binding site.

On the other hand, Goodford developed a method to represent favorable sites for some functional groups (such as $-NH_3^+$, $-COO^-$, -OH and H_2O) in the binding pocket of a protein [2]. This program calculates the interaction energy of the functional groups at each 3D grid point and displays favorable sites for these groups by contouring with the energy values. Although the method seems to be useful for modifying the docked ligand molecule, it cannot be used for calculating the interaction energy of the ligand molecule in the binding pocket.

Almost simultaneously with Pattabiraman and Goodford, we also developed a method to describe the physical and chemical environment of the binding pocket by using a 3D grid [3,4]. At each grid point, the van der Waals potential of several probe atoms (C, H, N, O), the electrostatic potential and the hydrogen-bonding characteristics are calculated and stored. The stored grid-point data enable real-time calculation of intermolecular interaction energy (van der Waals, electrostatic and hydrogen bond) during the docking operation, as in Pattabiraman's method. The energy values are updated at every manipulation step of the ligand molecule and are shown both numerically and graphically. This facilitates finding stable local minima in docking studies.

As well as for the energy calculation, we use the grid-point data for showing the 'atom-acceptable region', which is defined by contouring at a certain level of the van der Waals potential for each probe atom, as in Goodford's method. The physical meaning of this region is that the center of a ligand atom can exist inside the region without contact of the atom with the protein atoms. Furthermore, the electrostatic potential on the surface grid points can be shown by color-coding the atom-acceptable region according to the potential values. The hydrogen-bonding flag in the grid-point data represents the positions and characters of hydrogen-bonding partners expected from the hydrogen-bonding functional groups in the protein.

We previously reported a program package GREEN (Grid for Receptor Environment and

ENergy calculation) for docking studies, which was based on the above ideas [5]. At that time, the package had basic functionalities implementing these ideas, but it was not sufficiently developed to efficiently perform a docking study. Since then, the program package has been thoroughly revised and improved, with addition of various functions. The major improvements and added functionalities so far are the following:

(1) Plotting of the energy profile along the trajectory of the ligand manipulation.
(2) Real-time analysis of the intermolecular energy contribution of each ligand atom.
(3) A built-in energy minimization function for refining the docking model.
(4) A 'precise' energy calculation, employing the conventional atom pair-type equation, which can be used as an alternative to the grid point-based energy calculation for fine discrimination.
(5) Calculation of the intermolecular energy of the ligand, using the conventional force field.
(6) Model building functions for modifying the ligand structure.
(7) Improvements of the user interface by use of the functionalities of the 3D graphics workstation.

In this paper, we describe these new functionalities of the GREEN package, as well as the details of its basic ideas and algorithms. In order to demonstrate the usefulness of the package, an example of the docking study of a synthetic inhibitor to the enzyme trypsin is also presented.

BASIC ALGORITHMS

Description of binding-site environment

To describe the physical and chemical environment of the ligand-binding site of the receptor molecule, we use a 3D grid and calculate several quantities (grid-point data) that express the environment at each grid point. The grid-point data consist of van der Waals and electrostatic interactions, and hydrogen-bonding properties. The data are used for the graphical representation of the binding-site environment, as well as for the real-time calculation of the protein–ligand interaction energy.

A 3D grid with a regular interval (typically 0.2–0.5 Å) is generated inside the binding site of the target protein molecule. For calculating the van der Waals and electrostatic interactions, a probe atom is placed on each grid point and the interaction between the probe and the protein is calculated and stored as the grid-point data (Fig. 1A). If the distance between a grid point and one of the protein atoms is much shorter than the sum of the van der Waals radii of the probe and protein atoms, the grid point is treated as 'out of bounds' and the data are not calculated on that point.

The van der Waals interaction term G_{vdW} is calculated by using the Lennard-Jones potential, as shown in Eq. 1. r_{ij} is the distance between the *i*th grid point and the *j*th protein atom. k and l are the atom types of the probe atom and of the *j*th protein atom, respectively. A_{kl} and B_{kl} are the potential-function parameters between the atom types k and l. This term is calculated and stored separately for several types of probe atoms, such as carbon, hydrogen and oxygen. All atom species that exist in the ligand molecule to be docked need to be used as probe atoms. As for the parameters A_{kl} and B_{kl}, we currently use the values published by Weiner et al. [6,7], but any other parameter set and function form can be adopted easily.

$$G_{vdW}(i,k) = \sum (A_{kl}\, r_{ij}^{-12} - B_{kl}\, r_{ij}^{-6}) \qquad (1)$$

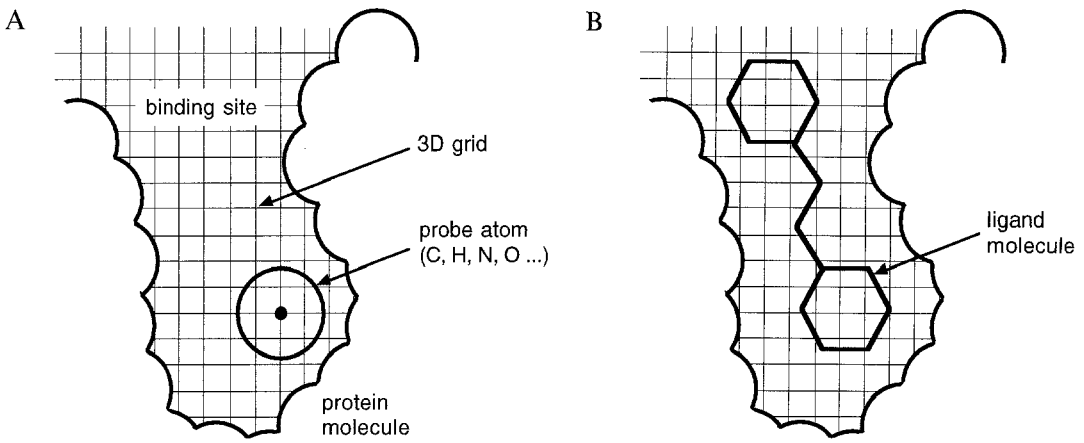


Fig. 1. Grid-point approximation of the protein–ligand interaction energy. (A) Calculation of the grid-point data. (B) Calculation of the interaction energy, using the grid-point data.

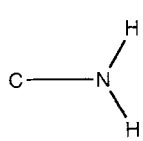
The electrostatic energy term G_{elc} is calculated by using the Coulomb potential, as shown in Eq. 2. Parameter q_j is the atomic charge of the *j*th protein atom, ε is the dielectric constant and K is the constant to convert the energy value into kcal/mol. The probe atom is assumed to bear a positive charge of one electron unit. By default, a distance-dependent treatment of the dielectric constant ($\varepsilon = 1\ r_{ij}$) is used.

$$G_{elc}(i) = K \sum q_j / \varepsilon r_{ij} \tag{2}$$

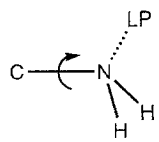
Hydrogen bonds play an important role in the specific recognition of molecules in biological systems. Introduction of hydrogen-bonding functional groups into a lead compound is often an effective method for designing a drug molecule with higher specificity and binding affinity for the target protein. In order to facilitate the investigation of hydrogen-bonding patterns between protein and ligand molecules, the grid-point data contain flags to indicate favorable regions for hydrogen bonding (hereafter called 'hydrogen-bonding regions'). These are determined based on the positions and directions of hydrogen-bonding functional groups in the protein. The region is calculated based on empirical geometrical criteria that have been determined from published statistical data of hydrogen-bonding patterns, searched from crystallographic databases [8–11]. The criteria are defined for typical functional groups found in protein structures. Categories of the functional groups and angle criteria are shown in Fig. 2. Criteria of minimum and maximum distances between the grid point and the heteroatom in the functional group are 2.5 and 3.1 Å, respectively. Grid points within the hydrogen-bonding region are marked with the properties (donor, acceptor, or both) of the functional group that is affecting the region. The information is stored as flags in the grid-point data and is used for visualizing the hydrogen-bonding region.

Approximation of protein–ligand interaction energy

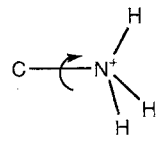
In the conventional methods of molecular mechanics and molecular dynamics, the intermolecular interaction energy is calculated by using the empirical potential function given in Eq. 3. The first summation term is the Lennard-Jones 6–12 potential, that accounts for the van



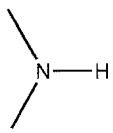
#1 primary amine N (sp2)
D: θ(N–H···G) > 150°



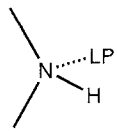
#2 primary amine N
(sp3, free rotating)
D, A: |θ(C–N···G)–109.5°| < 30°



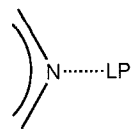
#3 ammonium ion N
(free rotating)
D: |θ(C–N···G)–109.5°| < 30°



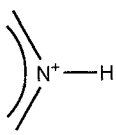
#4 amide N
D: θ(N–H···G) > 150°



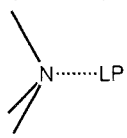
#5 secondary amine N
(H and LP interchangeable)
D, A: θ(N–H···G) > 150°
D, A: θ(N–LP···G) > 150°



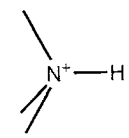
#6 aromatic N
A: θ(N–LP···G) > 150°



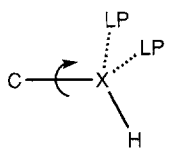
#7 aromatic N (with H)
D: θ(N–H···G) > 150°



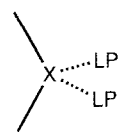
#8 tertiary amine N
A: θ(N–LP···G) > 150°



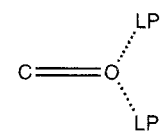
#9 tertiary amine N (with H)
D: θ(N–H···G) > 150°



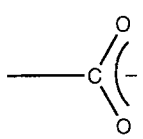
#10 hydroxyl O (free rotating)
#17 thiol S (free rotating)
D, A:|θ(C–X···G)–109.5°| < 30°



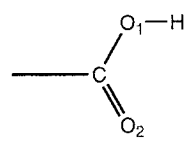
#11 ether O
#18 thioether S
A: θ(X–LP···G) > 150°



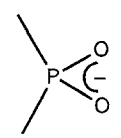
#12 carbonyl O
A: θ(O–LP···G) > 120°



#13 carboxylate anion O
A: θ(C–O···G) > 90°

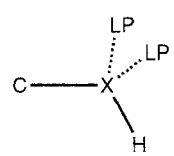


#14 carboxylic acid O
D: θ(C–O_1···G) > 140°
A: θ(C–O_2···G) > 120°

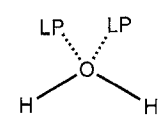


#15 phosphate O
A: θ(P–O···G) > 150°

#16 water O (without explicit H)
D, A: no angle condition



#19 hydroxyl O (with fixed H)
#20 thiol S (with fixed H)
D: θ(X–H···G) > 150°
A: θ(X–LP···G) > 150°



#21 water O (with fixed H)
D: θ(O–H···G) > 150°
A: θ(O–LP···G) > 150°

Fig. 2. Categories of hydrogen-bonding functional groups and angle criteria to determine hydrogen-bonding regions. D: donor; A: acceptor; G: grid point; LP: lone pair.

der Waals interaction between pairs of atoms. A and B are the potential parameters and r_{ij} is the distance between atoms i and j. The second summation term accounts for the electrostatic interaction. q_i and q_j are the atomic charges of atoms i and j, and ε is the dielectric constant. The third term (10–12 potential) is used in the AMBER force field to treat the interaction of hydrogen-bonding atoms separately from the other interacting pairs. The summations in Eq. 3 are taken for all interacting pairs of atoms in the system. For the calculation of protein–ligand interactions, the number of pairs to be calculated becomes large, resulting in rather long computation times (in practice, a distance cutoff is usually applied in creating the atom-pair list, but the number of pairs is still very large).

$$E_{inter} = \sum\sum (Ar_{ij}^{-12} - Br_{ij}^{-6}) + \sum\sum q_i q_j / \varepsilon r_{ij} + \sum\sum (Cr_{ij}^{-12} - Dr_{ij}^{-10}) \quad (3)$$

We have developed a method to accelerate the calculation of the protein–ligand interaction energy by using the van der Waals and electrostatic terms of the grid-point data [3–5]. In this method, the interaction energy between protein and ligand is approximated by Eq. 4. For each atom of the ligand molecule placed in the grid region, the grid-point values $G_{vdW}(m,k_m)$ and $G_{elc}(m)$ are determined by interpolating from the values on the eight grid points surrounding the atom (Fig. 1B). Here, m is the sequential number of the ligand atom and k_m is its atom type. For the van der Waals energy term $G_{vdW}(m,k_m)$, an appropriate probe atom is selected from the tables for several types of probe atoms. The obtained value of $G_{vdW}(m,k_m)$ is taken as the value of the van der Waals interaction energy. For the electrostatic energy term, the interaction energy is calculated by scaling the value of $G_{elc}(m)$ with the atomic charge of the ligand atom q_m.

$$E_{inter} = \sum (G_{vdW}(m,k_m) + G_{elc}(m)q_m) \quad (4)$$

The advantage of this approximation method is that the amount of computation required for the second stage depends only on the number of atoms in the ligand molecule. Therefore, even for a very large protein molecule, the protein–ligand interaction energy can be calculated quite rapidly. Furthermore, once the grid-point data have been collected, the second phase can be performed repeatedly for various ligand molecules by using the same grid-point data.

PROGRAMS AND FUNCTIONS

Overview of the program package

Figure 3 shows the programs and data flow of the program package. The package consists of a program for preprocessing the protein coordinate data (PDBFIL), one for calculating the grid-point data (CALGRID), a main program for the docking operations (MMGRD) and several utility programs. The package is implemented for the IRIS 4D graphics workstation, running under UNIX. The program MMGRD employs the 3D graphics capability of the IRIS workstation.

PDBFIL: Preprocessing of the protein coordinate file

The program PDBFIL reads atomic coordinates of the target protein from a file in the Protein Data Bank [12] format and assigns force-field atom types, atomic charges, connectivity and types of hydrogen-bonding heteroatoms. Definitions of the force-field atom types and atomic charges

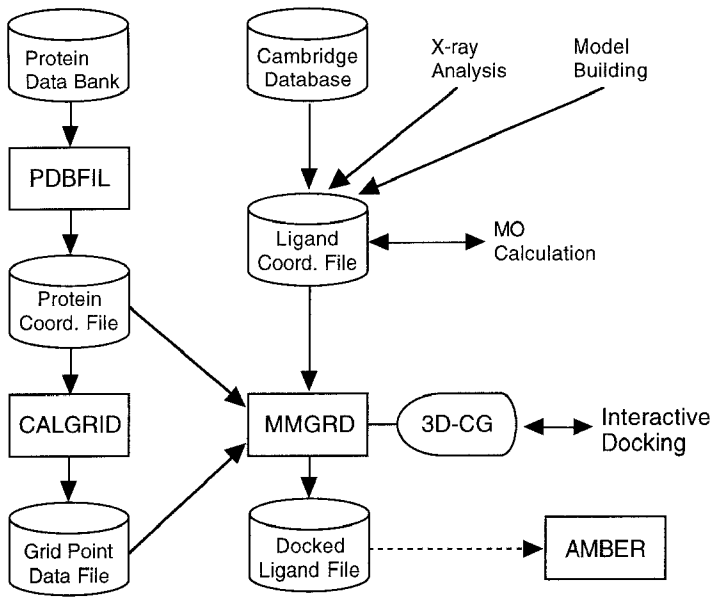


Fig. 3. Programs and data flow of the GREEN package.

are the same as those in the AMBER program package. The type of hydrogen-bonding hetero-atom (Fig. 2) is used to select the geometrical criteria for calculating the hydrogen-bonding region. The data for atoms in standard amino acids and nucleic acids are stored within the program. The program also calculates coordinates of hydrogen atoms that are usually missing from PDB data. An extended conformation is assumed for the hydrogen atom with rotational freedom such as that of a serine side chain. By default, only hydrogen atoms attached to hetero-atoms (oxygen, nitrogen and sulfur) are explicitly generated, and hydrogen atoms attached to a carbon atom are treated as a 'united atom' that is placed at the position of the carbon atom. Optionally, 'all-atom' treatment can be used; all hydrogen atoms are then explicitly generated. The coordinates and attributes of atoms are stored in a standard format file (GREEN standard format). This file consists of several header records and atom data records, one record per atom.

CALGRID: Calculation of grid-point data

The program CALGRID calculates the grid-point data that represent the physical and chemical environment of the ligand-binding site. The program reads the protein coordinate file generated by PDBFIL and a control file for the calculation. Grid-point data are written to a binary file that is used by the program MMGRD during the interactive docking operation. The control file contains specifications for the types of probe atoms to be used, selection of dielectric models (constant or distance dependent) in the electrostatic calculation, grid spacing, and the range of the rectangular box region within which grid-point data are to be calculated. Atom types that exist in ligand molecules to be docked must be selected as probe atoms. The range of grid points needs to be determined such that it surrounds the whole binding pocket where protein–ligand interaction may take place. The default grid spacing is 0.4 Å. If a finer grid spacing is used, the approximation of the interaction energy becomes more precise. However, at the same time, the

size of the grid-point data file and the amount of memory used by the MMGRD program also increase. The calculation of the grid-point data takes relatively much CPU time and is usually done as a 'background job'. The program MMGRD contains a function to graphically specify the region of the grid calculation and other control parameters of the CALGRID program.

Preparation of the ligand coordinate file

A ligand molecule file is prepared as a GREEN standard coordinates file. The data necessary for the ligand molecule are the atomic coordinates, charge distribution, force-field atom types and bond tables. The atomic coordinates are prepared from various sources, such as the Cambridge Crystallographic Database [13], X-ray crystal analysis, and model building. The conformation of the ligand is arbitrary, as it can be changed during the docking operation. The atomic charges on the ligand atoms are calculated by semiempirical or ab initio molecular orbital calculation programs. For the calculation of the intramolecular energy, a force-field atom type must be specified for every ligand atom with a two-letter symbol, as in the AMBER package.

MMGRD: Program for interactive docking

The program MMGRD is an interactive program for 3D graphics display. The main features of the program are the following:

(1) Various representations of the physical and chemical environment of the ligand-binding site.
(2) Real-time calculation of protein–ligand interaction energy during the interactive docking operation.
(3) Energy minimization, including the conformational degrees of freedom of the ligand molecule.
(4) Interactive model-building functions for designing new ligand structures.

The program has a hierarchical menu structure. The main menu (Table 1) is always displayed on the upper part of the screen. From the main menu, one can select various groups of commands. When a main-menu item is selected by pressing a mouse button, a corresponding submenu appears as a pull-down menu. Most of the commands of the program are invoked by selecting a submenu item.

Representation of molecules

By default, molecules are displayed as wire-frame skeleton models. In the docking study of a large protein molecule, it is often unnecessary to display the whole protein structure, especially when one is concentrating on the binding site. In such cases, the skeletal representation can be

TABLE 1
MAIN-MENU ITEMS OF THE PROGRAM MMGRD

INPUT	Read molecule data, grid-point data, etc. from files
OUTPUT	Store coordinates of the docked drug molecule in a file
ENERGY	Energy calculation, energy minimization, energy decomposition analysis, etc.
BUILD	Model building commands
CALC	Monitoring of geometries, close contacts, etc.
DISPLAY	Change representations of molecule models and grid-point data
INFO	Set up optional parameters, show summary of grid-point data
TORSION	Set bond rotation of the drug molecule
VIEW	Set pivot for rotation of drug or protein molecules

restricted to a part of the whole protein structure. The residues to be displayed can be specified by residue numbers, by a spherical cutoff from a specified atom, by a distance cutoff from the grid region, or by a distance cutoff from the docked ligand molecule. For the purpose of examining the folding pattern, the protein model can be simplified to a 'pseudo-skeleton' model in which successive α-carbon atoms of the protein are connected, or to a schematic 'ribbon-and-cylinder' picture in which α-helices are displayed as shaded cylinders and other backbone atoms as ribbons. Atoms of the skeleton model can be color-coded according to their atom type (default), amino acid type (acidic, basic, hydrophilic or hydrophobic), crystallographic temperature factor, or chain number (useful option for multi-subunit proteins). The color-coding options can be specified separately from the ligand skeleton model. For highlighting the ligand molecule or a part of the protein molecule, a ball-and-stick model or a spacefilling model can be displayed as a smoothly shaded picture.

Representation of binding-site environment

Using the van der Waals energy term of the grid-point data, the 'atom-acceptable region' can be displayed. This region is defined as a group of grid points where the van der Waals interaction energy between the protein and the probe atom is less than a certain level. The atom-acceptable region is displayed as a bird-cage representation, which is drawn by contouring with the van der Waals energy term of the grid-point data. By default, the contouring is done with the carbon probe atom at a potential-energy level of 0 kcal/mol. Optionally, different energy levels and probe atoms can be used for the contouring. The cage is color-coded according to the value of the electrostatic potential of the grid-point data (blue: most negative; green: neutral; red: most positive). A ligand atom can exist energetically favorably if its center is within the atom-acceptable region. Therefore, the representation of the atom-acceptable region is a convenient guide for docking a ligand molecule that is displayed as a skeleton model.

The 'hydrogen-bonding region' is displayed by using the hydrogen-bonding flags stored in the grid-point data. The region is shown either as a bird-cage picture, which is drawn by surrounding the grid points with the hydrogen-bonding flags turned on, or as a group of markers drawn at the hydrogen-bonding grid points. They are color-coded according to the hydrogen-bonding properties of the protein functional group affecting the region (blue: hydrogen acceptor; red: hydrogen donor; yellow: ambivalent hydrogen-bonding group such as a hydroxyl). In other words, the color coding indicates the expected hydrogen-bonding character of the ligand atom

TABLE 2
OPERATION-MENU ITEMS OF THE PROGRAM MMGRD

WHOLE	Switch to the whole-model manipulation mode
LIGAND	Switch to the ligand manipulation mode
X-ROT	Pseudo-dial bar for X rotation
Y-ROT	Pseudo-dial bar for Y rotation
Z-ROT	Pseudo-dial bar for Z rotation
X-TRAN	Pseudo-dial bar for X translation
Y-TRAN	Pseudo-dial bar for Y translation
Z-TRAN	Pseudo-dial bar for Z translation
STEREO	Display stereopair pictures
CLIP	Set depth clipping of the picture

(blue: hydrogen donor; red: hydrogen acceptor; yellow: either donor or acceptor). Therefore, this representation is useful for the initial placement of the ligand molecule on the basis of the hydrogen-bonding patterns and also for designing a ligand structure with specific hydrogen-bonding functional groups.

Manipulation of molecules

Basic manipulations of the displayed molecules are controlled by the operation-menu items (Table 2). The operation menu is always displayed on the right side of the screen. By selecting the 'WHOLE' or 'LIGAND' menu, the program switches between two manipulation modes for 3D rotation and translation of the molecules, i.e., a whole-model manipulation mode and a ligand manipulation mode. In the whole-model manipulation mode, which is used for changing the viewing direction of the molecules, protein and ligand molecules are rotated and translated together. The rotation and translation are controlled either by a dial device attached to the workstation or by pressing a mouse button while the cursor is on one of the pseudo-dial menus. In the ligand manipulation mode, which is repeatedly used during the docking operation, the ligand molecule is rotated and translated relative to the protein molecule. The conformation of the ligand molecule can be changed interactively by specifying rotatable bonds (the maximum number of bonds allowed is 20) and by controlling the torsional angles by pressing keys on the keyboard assigned for this function. All degrees of freedom of the ligand molecule (i.e., rotation, translation and bond rotations) are treated as variables in the energy minimization function which is described below.

Docking operation

In the ligand manipulation mode, a ligand molecule is fitted into the binding-site pocket by rotation and translation relative to the protein molecule and by rotations of the rotatable bonds. When the 'MANUAL FIT' command in the 'ENERGY' menu is turned on, the interaction energy between ligand and protein is calculated and displayed on the screen while the ligand molecule is manipulated. By the use of the grid-point approximation, the energy calculation is very fast, so that the energy value is updated almost in real time, following each manipulation step. By turning on the 'ATOM ENERGY' command, the contribution of each ligand atom to the total interaction energy is displayed as a label near the ligand atom. The display of the energy contribution is also updated in real time, together with the total interaction energy, during the interactive manipulation of the ligand molecule. If a part of the ligand molecule has an unfavorable interaction with the protein, that part can be easily identified by the energy contribution monitor. As well as in the docking operation, this function is also useful for modifying the docked ligand molecule by model building, so as to attain a more favorable interaction with the protein.

When the conformation of the ligand molecule is varied by bond rotations during the docking operation, it is important to evaluate the intramolecular energy of the ligand molecule as well as the intermolecular energy. If the 'EINTRA' option is turned on, the intramolecular energy of the ligand molecule is calculated by using the AMBER force field, together with the intermolecular energy. Since the calculation of the intramolecular energy is applied to the ligand molecule only, it can be done almost in real time for typical ligand molecules with less than 100 atoms. As in the AMBER package, the force-field parameters that are necessary for the calculation are read

from an external file. Users have to supply appropriate force-field parameters for ligand molecules other than peptides or nucleotides.

The energy values that are continuously calculated during the docking operation are sampled and stored as 'trajectory data'. The trajectory data also contain values of rotation angles, the translation vector, and bond rotation angles of the ligand molecule at every sample point. If the 'ENERGY GRAPH' option is turned on, the energy values of the trajectory data are plotted against the sample points. The energy graph is redrawn and grows to the right in real time while the ligand molecule is manipulated. From the energy graph, a low-energy geometry of the ligand molecule can be easily located. By picking one of the points on the energy graph, the corresponding geometry of the ligand molecule can be reproduced at any stage of the docking operation.

Optimization of the docked model

The low-energy geometry, determined roughly by manual docking, can be further optimized by the built-in simplex minimizer. The minimizer is invoked by selecting the 'SIMPLEX MIN' command of the 'ENERGY' menu. The sum of the protein–ligand intermolecular energy and the intramolecular energy of the ligand (if the 'EINTRA' option is turned on) is minimized by varying the translation vector, rotation angles and bond rotation angles of the ligand molecule, starting from the current docked structure. Initial step widths for searching the minimum can be specified separately for translation, rotation and bond rotation.

The optimized coordinates of the ligand molecule are stored into the standard format file by the 'WRITE STD' command of the 'OUTPUT' menu. Further optimization of the model, including the flexibility of the protein molecule, is done by other external programs, such as the AMBER package.

The grid-point approximation versus precise energy calculation

Although the grid-point approximation is advantageous for the interactive docking operation because of its rapidity, errors in the calculated energy value become a problem, especially when we want to consider subtle differences in the energy values. As the curve of the Lennard-Jones potential changes sharply near the interatomic distance of van der Waals contact, the error of the grid-point approximation becomes especially large when the ligand atom is close to the

TABLE 3
MODEL-BUILDING COMMANDS

ADD ATOM	Add new atoms
ADD BOND	Add new bonds
ADD FRAG	Add substructural fragment such as functional group
DEL ATOM	Delete atoms
DEL BOND	Delete bonds
DEL FRAG	Delete group of atoms
CNG BONLEN	Change bond length
CNG BONANG	Change bond angle
CNG TORSON	Change dihedral angle
CNG ATOM	Change atomic number
CNG CHARGE	Change atomic charge, invoke MOPAC calculation
CNG TYPE	Change AMBER force field atom type
CNG HBTYPE	Change hydrogen-bond group type

binding-site pocket. The error is reduced to some extent by the interpolation method, as described before, but still remains too large for a quantitative discussion of the energy values. In order to avoid this problem, one can turn on the 'PRECISE MODE' option and calculate the intermolecular energy exactly with the AMBER force field, without using the grid-point approximation. This option applies to all energy-related functions, such as manual fitting and energy minimization. In the example of the docking to trypsin, the 'PRECISE MODE' requires a factor of three longer computation time compared to the grid-approximation method.

Model-building commands

For the purpose of designing new ligand molecules, the structure of the docked ligand molecule can be modified interactively by using the model-building commands listed in Table 3.

With the 'ADD ATOM' command, new atoms are added to the specified position in the molecule in standard geometry (sp^3, sp^2 or sp) or by explicitly specifying bond length, bond angle and dihedral angle with a Z-matrix. With the 'ADD BOND' command, new bonds are formed between specified pairs of atoms. With the 'ADD FRAG' command, substructural fragments such as rings or functional groups are added to the ligand structure by overlapping 'joint bonds' that must be specified on both the ligand and the fragment. Fragment structures have to be prepared as standard coordinate files beforehand.

Geometrical parameters such as bond length, bond angle and dihedral angle are modified by picking pairs, triplets or quartets of atoms, respectively, and then specifying new parameters from the keyboard (commands 'CNG BONLEN', 'CNG BONANG' and 'CNG TORSON', respectively). There are also commands for changing attributes of atoms such as atomic number ('CNG ATOM'), atomic charge ('CNG CHARGE'), atom type of the AMBER force field ('CNG TYPE') and type of hydrogen-bonding heteroatom ('CNG HBTYPE'). The 'CNG CHARGE' command has an option to invoke a semiempirical molecular orbital calculation by the MOPAC [14] program. The MMGRD program awaits termination of the MOPAC calculation and the resulting atomic charges are automatically assigned to the ligand atoms.

When a model-building operation is finished by selecting other menus, the program reassigns force-field parameters according to the modified structure and a docking operation can be readily restarted.

DOCKING STUDY OF A TRYPSIN INHIBITOR TO TRYPSIN

Trypsin is a serine protease that hydrolyzes an amide bond at the C-terminal side of lysine or arginine residues of a polypeptide chain. Trypsin is a good target for testing a new computational

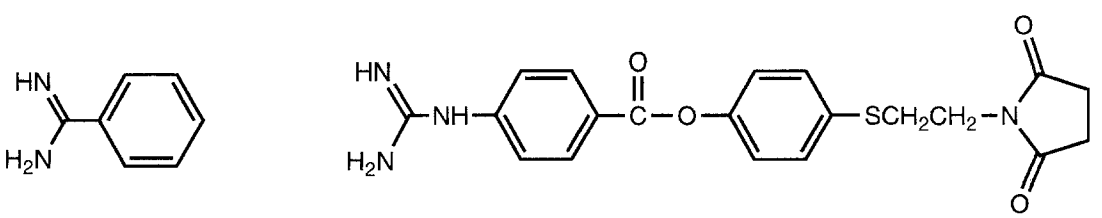


Fig. 4. Synthetic inhibitors of trypsin. Left: benzamidine; right: PATM.

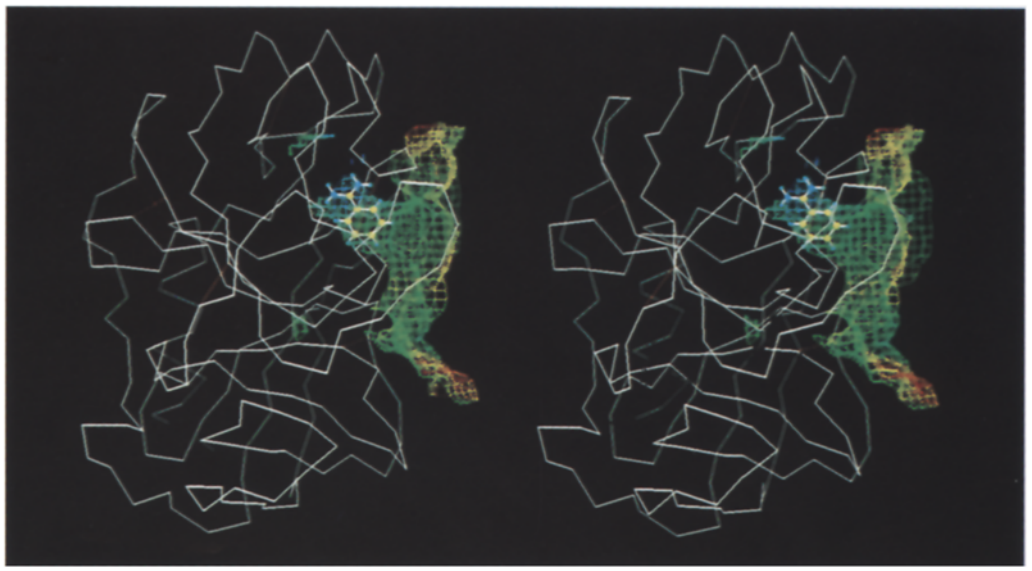


Fig. 5. Stereoview of the atom-acceptable region, displayed as a bird-cage model at the ligand-binding site of bovine trypsin. The colors of the cage indicate the values of the electrostatic potential in the grid-point data from blue (G_{elc} < −28 kcal/mol) to red (G_{vdW} ≥ 28 kcal/mol). The enzyme molecule is shown as a 'pseudo-skeleton' model. The inhibitor benzamidine, bound in the crystal, is shown as a ball-and-stick model.

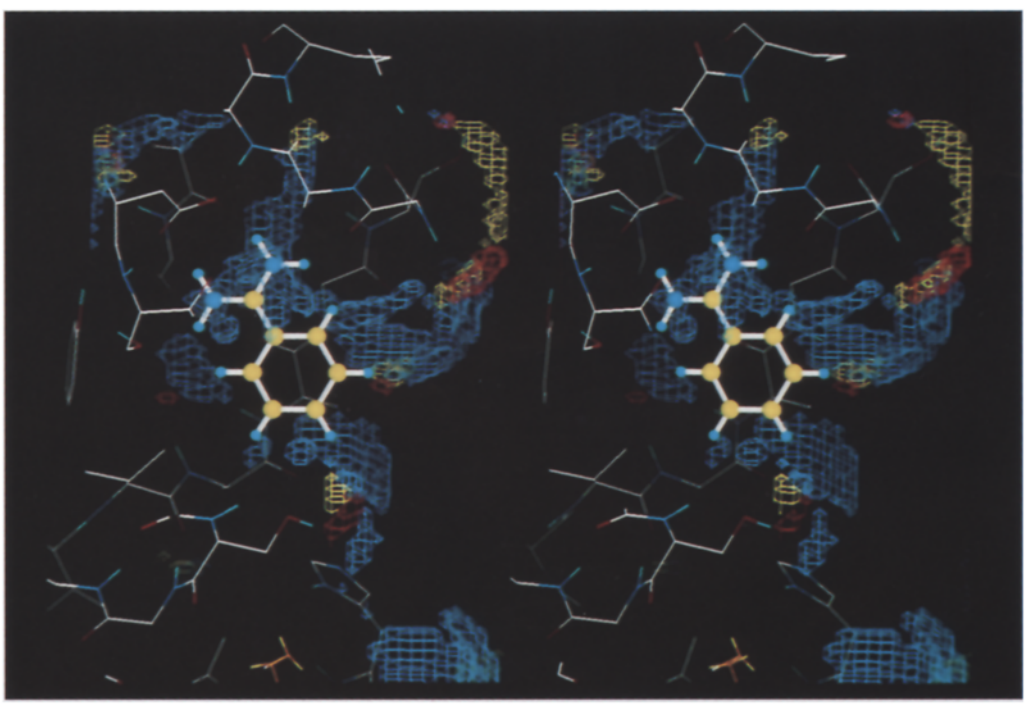


Fig. 6. Stereoview of the hydrogen-bonding region, displayed at the ligand-binding site of trypsin. The colors of the cages indicate properties of the protein functional group affecting the region (blue: hydrogen acceptor; red: hydrogen donor; yellow: ambivalent hydrogen-bonding group such as hydroxyl). This is a close view of the ligand-binding site with the same orientation as in Fig. 5. The inhibitor benzamidine is shown with carbon atoms in yellow.

method, because its structure and functions have been extensively studied by experimental [15–18] and theoretical [19] methods. As a test application of the GREEN package, we have performed a docking study of a small-molecule inhibitor, PATM (Fig. 4), to bovine trypsin. PATM is a synthetic trypsin inhibitor that is used for the treatment of acute pancreatitis. The molecule has a guanidyl group which is supposed to be important for specific binding to the P1 binding site of trypsin. Kinetic measurements have shown that the ester bond of PATM is slowly hydrolyzed upon binding to trypsin [20].

Atomic coordinates of bovine trypsin were taken from the crystal structure complexed with an inhibitor, benzamidine (PDB 3PTB) [21]. Hydrogen atoms attached to heteroatoms were added by the program PDBFIL; aliphatic carbon atoms were treated as united atoms. The total number of atoms in the protein model was 2039 (1629 non-hydrogen atoms, 410 hydrogen atoms). Grid-point data were calculated on 73 736 grid points within a box of dimensions $14.4 \times 16.8 \times 23.2$ Å, which sufficiently includes the active site and substrate-binding site. The grid spacing was set to 0.4 Å. The calculation of the grid-point data required about 650 s on an IRIS 4D-35G workstation (CPU 36 MHz R3000).

The atom-acceptable region of the ligand-binding site is shown in Fig. 5. It can be clearly seen that the inhibitor benzamidine, bound in the crystal, is accommodated in a well-formed pocket with negative electrostatic potential. The pocket corresponds to the binding site of P1 residues (lysine or arginine) of peptide substrates, and the catalytic triad residues (Ser^{195}, His^{57} and Asp^{102}) of the enzyme are located near the opening of the pocket. Figure 6 shows a closer view of the binding site, with the representation of the hydrogen-bonding regions. Hydrogen-donating amidine nitrogen atoms of the inhibitor are located within the hydrogen-acceptor region (blue cages), indicating the possible formation of hydrogen bonds between protein and inhibitor.

Atomic coordinates of PATM were taken from the crystal structure (T. Kawai, personal communication). Atomic charges were calculated by the MNDO method in the MOPAC package [14]. In docking PATM, it was assumed that the guanidinobenzoate moiety of the inhibitor is accommodated in the P1-site pocket, as it is a hydrogen-bond donor moiety similar to the side chain of arginine. The docking operation was started by fitting the guanidinobenzoate moiety into the pocket. A snapshot from an early stage of the docking operation is shown in Fig. 7. By employing the 'RESTRICT ATOM' command of the 'ENERGY' menu, only the guanidinobenzoate moiety was used for calculating intermolecular energy; the other part of the inhibitor, which is protruding from the pocket, was ignored at this stage. The energy graph of the trajectory data is displayed on the upper part of the screen and several low-energy structures were selected for further docking.

Figure 8 shows the advanced stage of the docking operation. The energy contribution of each ligand atom to the total interaction is monitored by the 'ATOM ENERGY' command. Here, a part of the succinimide moiety clashes with the atom-acceptable region and high energy values are shown at those atoms. Both the 'ATOM ENERGY' monitor and the energy graph are updated in real time. By employing this function, we obtained proper binding modes, in which all parts of the ligand molecule have a favorable interaction with the protein molecule.

Since the flexible moiety of PATM is accommodated in a wide pocket, as shown in Fig. 8, there are several possible binding modes with different conformations of the moiety. We tested various conformations and selected several models on the basis of the intermolecular and intramolecular energies. The models were further refined by molecular mechanics calculation with the

AMBER package, allowing movements of atoms of the inhibitor and its neighboring residues. The lowest energy model after energy refinement is shown in Fig. 9. In the model, the carbonyl oxygen of the ester group of the inhibitor has a hydrogen-bonding interaction with the 'oxyanion hole' which is formed by the amide N-H groups of the Gly^{193} and Ser^{195} residues. The Oγ atom of Ser^{195} closely approaches the carbonyl carbon of the inhibitor. This model is consistent with the proposed mechanism of trypsin catalysis, in which Oγ of Ser^{195} attacks the carbonyl carbon of the substrate as a nucleophile [19]. Thus, with this model we can well explain the mechanism of the hydrolysis of this inhibitor by trypsin.

DISCUSSION

We have developed the program package GREEN for performing rational and efficient docking studies. It uses 3D grid-point data to express the physical and chemical environment of the ligand-binding site of the receptor. The package is useful for obtaining a stable model of the complex between a receptor and a drug molecule and allows us to understand relationships between drug structures and their activities. The package is also useful for modifying the structure of the docked drug molecule in a logical way, based on the environment of the binding site.

With rather simple software, it is possible to perform a docking operation in which a ligand molecule is manipulated by rotation, translation and bond rotations, while checking for clashes between the ligand and the receptor. Most commercially available software packages for computer-aided drug design have such functions. Recently, several programs have become available that can calculate protein–ligand interaction energies by using a grid-point approximation method similar to that used in our method. However, even with real-time monitoring of the interaction energy, problems of reproducibility and coverage of the possibilities remain in the interactive method. We have developed the ENERGY GRAPH function as an approach to deal with these problems. From the trajectory of the movement of the ligand and the corresponding energy values stored during the docking operation, one can locate and reproduce several low-energy models, which can be further refined with the built-in minimizer.

Using the interactive docking method, it is very difficult to search completely all possible binding modes and conformations of a ligand molecule, especially when the ligand molecule is flexible, with many rotatable bonds. In order to cover all possibilities, we need an automated docking method that can take into account all possible binding modes of the ligand molecule, including the conformational freedom. For such purposes, we have developed a method of automatic docking [22,23]. The method uses the grid-point data of the GREEN package to describe the binding-site environment and to evaluate the protein–ligand interaction energy. Methods of automatic docking which use different algorithms have also been reported by other groups [24–26].

Even with the automatic docking method, we cannot always unambiguously determine the most promising docking model. The automatic docking method is suitable for selecting several promising models from the vast number of possible models, based on some score such as the interaction energy. However, in order to select the most promising model, we need to examine carefully each candidate model in the light of other data, such as binding strengths of related compounds. For such purposes, we still need an interactive docking method next to the automatic method.

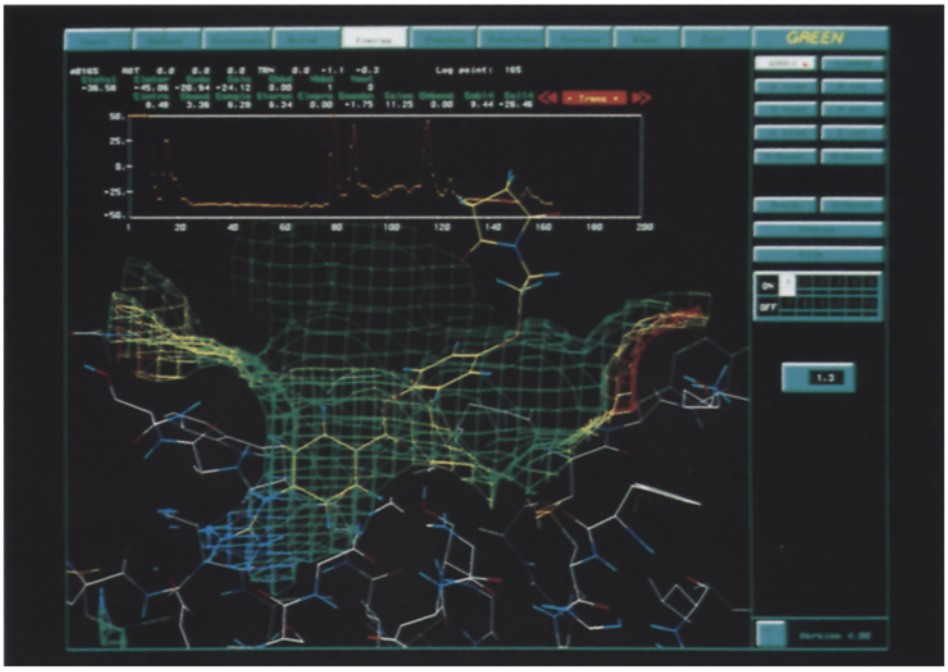


Fig. 7. A snapshot from an early stage of the docking of the inhibitor PATM (carbon atoms shown in yellow) to trypsin. The atom-acceptable region is displayed in the same way as in Fig. 5. The energy graph (red lines with yellow points) is displayed on the upper part of the screen.

Fig. 8. A snapshot from the advanced stage of the docking operation. The energy contribution of each drug atom to the total interaction is shown as a label near the atom.

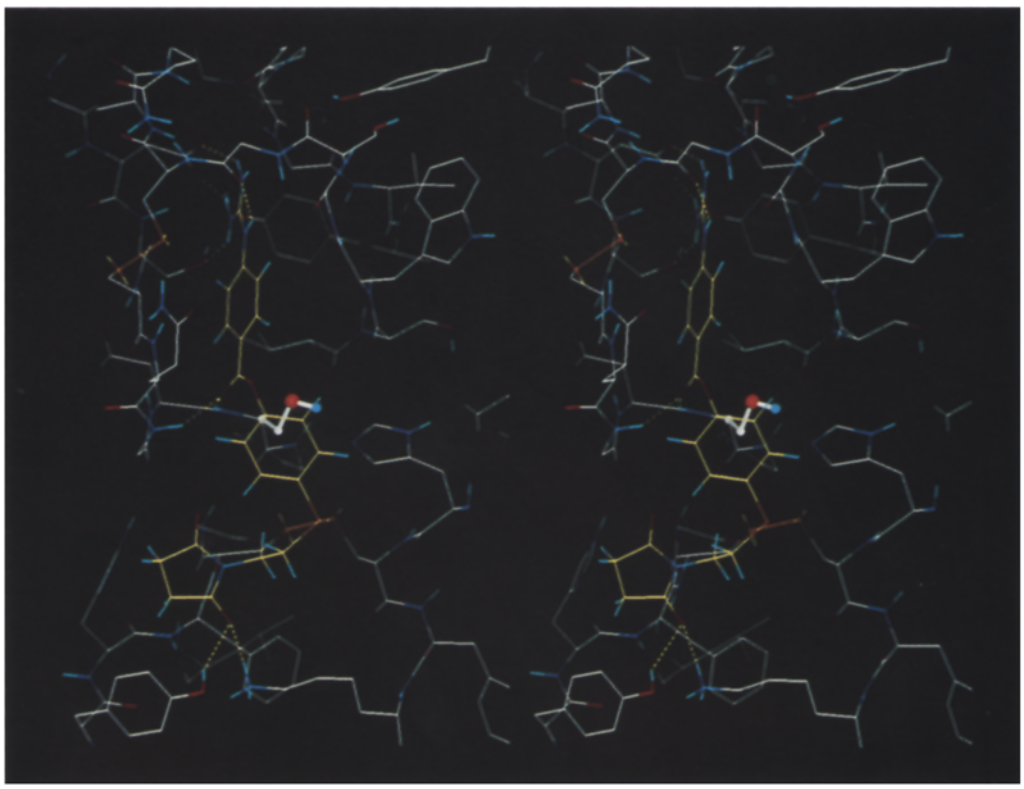


Fig. 9. Stereoview of the lowest energy model of the trypsin–PATM complex. The side chain of the catalytic Ser^{195} residue is shown in a ball-and-stick representation. Hydrogen bonds between trypsin and the inhibitor are shown as yellow dotted lines.

Since the atom-acceptable region is defined for the center of the probe atom, the goodness of fit between the ligand and protein can easily be seen with simple wire-frame pictures of the molecules. At the same time, the electrostatic character of the binding site is easily seen by the color-coding of the atom-acceptable region according to the electrostatic potential. These features are advantageous for docking and modeling operations compared to other expressions of cavity shape, such as the van der Waals surface and solvent-accessible surface [27], which have often been used in docking studies. The expression of the hydrogen-bonding region is also a unique feature of the GREEN package. It helps one to design appropriate positions for introducing functional groups that might form stable hydrogen bonds to the protein molecule.

The grid-point approximation method reduces the computation time of protein–ligand interaction energy, with a trade-off in precision of the calculated energy. Due to recent advances in computer power, such an approximation method seems to have become less important for the treatment of small-molecule ligands, for which the interaction energy can be calculated almost in real time even with the conventional method. Nevertheless, the method is still useful because of its efficiency and it is often used for calculating intermolecular energy in exhaustive search problems such as automatic docking [22,23,28] and de novo ligand design [29]. The grid-point

approximation would be useful also for the treatment of protein–protein interactions, where an enormous number of pairwise interactions would have to be calculated if we used the conventional method.

The precision of the energy values calculated by the grid-point approximation method is affected not only by the grid spacing, but also by the shape and other properties of the ligand-binding site. The approximated energy is adequate for locating local minimum binding modes of the ligand, but it is not sufficiently precise to be used for quantitative evaluation of the binding modes. Therefore, we have also included the 'PRECISE' option of intermolecular energy calculation by the conventional pairwise method in the GREEN package. The intramolecular energy of the ligand molecule is always calculated by the precise method, since this calculation takes much less time than that of the intermolecular energy. The user can toggle between the grid-point and precise methods in all the energy-related functions, such as manual fitting and energy minimization. A typical use of the two energy evaluation methods would be to perform the initial docking for obtaining several candidate models by using the grid-point method, which is much faster than the precise method, and then to refine and select these models by using the precise method, which is slower but more reliable. In the docking of PATM, simplex minimization of the binding mode shown in Fig. 9 resulted in almost the same conformation (maximum deviation of atom positions 0.2 Å), both with the grid method and the precise method. This shows that the grid method is reliable for the purpose of searching local minimum binding modes. The error in the intermolecular energy by the grid approximation (compared to the 'precise' energy) was about 10 kcal/mol, which is relatively large for the purpose of comparing different binding modes quantitatively.

Docking studies are useful not only for explaining the structure–activity relationships of existing drug molecules, but also for modifying the drug structure in order to improve the binding affinity to the receptor. The interactive docking method is suitable for bringing the ideas and intuition of the user (especially chemists) into the design process. With the GREEN package, one can modify a ligand structure by using the incorporated modeling functions and evaluate the modified structure immediately by using the energy calculation functions. With the manual fitting and energy minimization functions, the modified molecule can be refitted to the binding site and the intermolecular and intramolecular energy can be obtained. By the energy contribution analysis, one can evaluate the energetic contribution of the modified moiety to the stability of the complex. The various representations of the binding-site environment also facilitate the design process. With the representation of the atom-acceptable region, one can easily see the goodness of fit of the designed structure to the binding site. From the representation of the hydrogen-bonding region, one can determine suitable positions for introducing functional groups that might form stable hydrogen bonds to the receptor.

In the initial stage of a docking study, where we have to consider many possible binding modes of the ligand, we must inevitably treat the receptor structure as a rigid body. Nevertheless, if the conformational change in the receptor induced by the ligand binding (induced fit) is not very large, binding modes obtained by keeping the receptor rigid can be good initial models for further simulations. In such cases, a correct model of the protein–ligand complex might be obtained by optimizing these models with inclusion of the flexibility of the protein moiety. However, we cannot apply this strategy to highly flexible proteins such as allosteric enzymes. Furthermore, there are many cases where we cannot use the crystal structure of a protein as it is, because of

the presence of a significant induced fit or a structural distortion due to the crystal packing. Before initiating a docking study, it is important to examine carefully the available experimental information about the structure of the target protein. Results of crystallographic analyses of the structures of complexes with different ligand molecules are especially useful for assessing the flexibility of the target protein.

A difficult but important problem about docking is a proper evaluation of the free energy of binding, including the entropic contribution. In the GREEN package, we calculate only the enthalpic contribution of the binding energy by using a simple force field. Although such a treatment allows qualitative comparison of binding modes of the same or similar ligand molecules, it cannot be used to compare molecules with quite different structures. In order to evaluate the free energy of binding more accurately, we might need more elaborate approaches such as the free-energy perturbation and dielectric continuum method [30]. Since such calculations require a proper starting model of the ligand binding, a binding mode determined by the GREEN package will be a good starting point for performing calculations.

Our method of expressing the binding-site environment by using 3D grid-point data is applicable for many other purposes. We also use the grid-point data of the GREEN package for automatic docking [22,23] and for a method of de novo ligand design that constructs new ligand structures automatically based on the receptor structure [31,32]. These methods will be integrated into the GREEN package in the near future.

Readers interested in using the program are requested to contact the author.

CONCLUSIONS

The program package GREEN facilitates rational docking studies by incorporating various functions, such as real-time evaluation of the protein–ligand interaction energy and representations of the binding-site environment. The program is available for designing new ligands, based on the 3D structure of the target protein.

REFERENCES

1 Pattabiraman, N., Levitt, M., Ferrin, T.E. and Langridge, R., J. Comput. Chem., 6 (1985) 432.
2 Goodford, P.J., J. Med. Chem., 28 (1985) 849.
3 Tomioka, N. and Itai, A., In Abstracts of Papers, 6th Symposium on Medicinal Chemistry, Pharmaceutical Society of Japan, Tokyo, 1984, p. 105.
4 Tomioka, N., Itai, A. and Iitaka, Y., In Iitaka, Y. and Itai, A. (Eds.) Proceedings of the Symposium on Three-Dimensional Structures and Drug Design, University of Tokyo Press, Tokyo, 1987, pp. 186–194.
5 Tomioka, N., Itai, A. and Iitaka, Y., J. Comput.-Aided Mol. Design, 1 (1987) 197.
6 Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.
7 Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7 (1986) 230.
8 Taylor, R. and Kennard, O., Acc. Chem. Res., 17 (1984) 320.
9 Murray-Rust, P. and Glusker, J.P., J. Am. Chem. Soc., 106 (1984) 1018.
10 Baker, E.N. and Hubbard, R.E., Prog. Biophys. Mol. Biol., 44 (1984) 97.
11 Vedani, A. and Dunitz, J.D., J. Am. Chem. Soc., 107 (1985) 7653.
12 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.

13 Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.
14 Stewart, J.J.P., J. Comput.-Aided Mol. Design, 4 (1990) 1.
15 Stroud, R.M., Kay, L.M. and Dickerson, R.E., J. Mol. Biol., 83 (1974) 185.
16 Fehlhammer, H. and Bode, W., J. Mol. Biol., 98 (1975) 683.
17 Tanizawa, K., Kasaba, Y. and Kanaoka, Y., J. Am. Chem. Soc., 99 (1977) 4485.
18 Sprang, S., Standing, T., Fletterick, R.J., Stroud, R.M., Finner-Moore, J., Xuong, N.-H., Hamlin, R., Rutter, W.J. and Craik, C.S., Science, 237 (1987) 905.
19 Weiner, S.J., Seibel, G.L. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 83 (1986) 649.
20 Nochi, S., Shimomura, N., Hattori, T., Sato, T., Miyake, Y. and Tanizawa, K., Chem. Pharm. Bull., 37 (1989) 2855.
21 Marquart, M., Walter, J., Deisenhofer, J., Bode, W. and Huber, R., Acta Crystallogr., B39 (1983) 480.
22 Yamada, M. and Itai, A., Chem. Pharm. Bull., 41 (1993) 1200.
23 Yamada, M. and Itai, A., Chem. Pharm. Bull., 41 (1993) 1203.
24 Goodsell, D.S. and Olson, A.J., Protein Struct. Funct. Genet., 8 (1990) 195.
25 Jiang, F. and Kim, S.-H., J. Mol. Biol., 219 (1991) 79.
26 Leach, A.R. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 730.
27 Connolly, M.L., J. Appl. Crystallogr., 16 (1983) 548.
28 Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.
29 Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36 (1993) 1700.
30 Van Gunsteren, W.F. and Berendsen, H.J.C., Angew. Chem., Int. Ed. Engl., 29 (1990) 992.
31 Nishibata, Y. and Itai, A., Tetrahedron, 47 (1991) 8985.
32 Nishibata, Y. and Itai, A., J. Med. Chem., 36 (1993) 2921.