Journal of Computer-Aided Molecular Design, **12:** 335–349, 1998.
KLUWER/ESCOM

A branch-and-bound method for optimal atom-type assignment in *de novo* ligand design

N.P. Todorov & P.M. Dean
Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K.

Received 8 May 1997; Accepted 5 December 1997

Key words: drug design, free energy, molecular design, protein–ligand interaction, structure generation

Summary

This paper investigates a computational procedure for the determination of the atom types on the vertices of a molecular skeleton to optimize interaction with the receptor site whilst maintaining a synthetically reasonable structure. The connectivity of the skeleton is analysed and appropriate atom types are compiled for each vertex. Receptor ionization and conformational states are generated by varying the positions of hydrogen atoms and electron lone pairs in the carboxyl, rotatable hydroxyl and amino groups. The structure is divided into small non-overlapping substructures. Atom types are assigned exhaustively onto each of the substructures using a depth-first search method; chemical rules are applied to reject unacceptable atom combinations early on. An empirical interaction score is calculated and the representatives of each partial structure are stored in ascending order according to their scores. The branch-and-bound procedure is then used to find the structures with the lowest scores. The method is illustrated using five protein–ligand complexes.

Introduction

The problem of designing new ligands by computer for target proteins of medical interest is a rapidly expanding research field [1–3]. One approach towards the algorithmic solution is to divide the problem into two parts: primary and secondary structure generation [4, 5]. During primary structure generation, molecular scaffolds are generated to fit the shape and some local properties of the receptor site, but the chemical properties of the atoms are not explicitly used. To create a molecular skeleton, it is considered sufficient, for example, to place a geometrically complementary atom at a hydrogen-bonding position, regardless of the atom type and the types of its neighbours [6]. When all geometric requirements are satisfied, during secondary structure generation, the atom types can be re-examined and chemical identities re-assigned to achieve better interaction.

This paper is concerned with secondary structure generation. Three methods for atom assignment onto molecular skeletons have been described in the literature. Chan et al. [7] have proposed an analytical solution to the problem of assigning partial charges at the vertices of the molecular graph. The potential created by the receptor is calculated at a set of points, evenly distributed on the van der Waals surface of the ligand. The method is based on the postulate that electrostatic potentials of the receptor and the ligand must be equal and opposite in sign to each other. In numerous cases this assumption is well justified [8], although it is not always the case [9–11]. The charges placed at the ligand atoms are calculated to minimize the rms difference between these potentials at the set of points.

A procedure for the placement of small fragments with precomputed properties onto the molecular graph has been developed and tested by Barakat and Dean [12–16]. Electrostatic or hydrophobic potential complementarity to the receptor or similarity to known active structures, on the van der Waals surface of the ligand, is optimized using simulated annealing. The method employs a library composed of 3-, 4- and 5-atom fragments commonly occurring in the Cambridge Structural Database [17]. A fair degree of transferability of the CNDO partial atomic charges derived from these fragments has been demonstrated [18,

19]. The algorithm for fragment placement contains a module for graph perception, allowing one to find feasible matches between the fragments and substructures of the skeleton. The skeleton is decorated with atoms by fragment placement and the properties from the fragment definitions are used to calculate the value of the objective function.

A method for atom substitution has also been reported by Gillet et al. [20]. It is used to promote hydrogen-bonding patterns in the interaction between protein and ligand [21].

The algorithm presented here pursues several objectives. First, to provide an empirical estimate of the binding energy [22]; second, to include rules for control of the chemical stability of the proposed structures [23]; third, to be able to generate efficiently the structures with the best binding energies that also comply with the chemical rules provided; and fourth, to take into account different possibilities for ionization states of protein atoms, rotatable hydroxyl and amino groups, water molecules and metal atoms in the site.

Methods

Overview of the algorithm

The atom assignment method uses single-atom templates, with defined atom types and bond orders. A skeleton is first analysed and the appropriate templates are compiled for each atom. Receptor ionization and conformational states are generated by varying the positions of hydrogen atoms and lone pairs in the carboxyl group, rotatable hydroxyl and amino groups. The structure is divided into small non-overlapping partial structures. Atom types are assigned exhaustively onto each of the substructures using a depth-first search method, applying chemical rules to reject unacceptable atom combinations early on. An empirical interaction score [22] is calculated for the representatives of each partial structure against each receptor ionization state. A branch-and-bound search procedure over all representatives and receptor states is used to find the structures with the lowest scores. Each step of the algorithm is described in more detail below. A flowchart of the algorithm is presented in Figure 1.

1. Initialization: read in templates, rules, etc.
2. Read protein.
3. Read ligand.
4. Determine receptor states.
5. Score original ligand.
6. Compile templates for each atom.
7. Find substructures.
8. Systematic atom assignment on substructures.
9. Score substructure representatives.
10. Branch-and-bound search over substructures to find best assignment.
11. Generate all assignments within threshold above best.

Figure 1. Computational steps of the method.

Atom templates

A library of 28 templates was constructed (Figure 2). The template consists of a single atom and the bonds incident upon it; bond orders and dummy electron lone pairs are distinguished. Atom types usually encountered in drug molecules are considered in several geometric and ionization states.

An example definition is shown in Figure 3 where a planar nitrogen atom is described. The first line in the definition contains sequentially the element number, the COSMIC [24] atom type, the atom geometry (4 = tetrahedral, 3 = planar, 2 = linear, 1 = lone pair or hydrogen), formal charge, number of unique bond order permutations, hydrogen-bond donor and hydrogen-bond acceptor participation. Each of the next lines lists one unique bond-order permutation of the incident bonds. In this case all six permutations are unique, but if, for example, the double bond in Figure 3 were to be single, there would be only three unique permutations, depending on the orientation of the lone pair. Formal bonds that correspond to lone pairs are assigned an order of 4. Each permutation is stored as a separate entry when the templates are pre-processed.

Receptor states

Usually, there is ambiguity as to how to position H atoms and lone pairs in terminal OH, NH_2 and COOH groups in the protein. If such groups interact with the ligand, different states will give rise to different complementary atoms. Each skeleton can interact with a particular set of hydrogen-bonding groups in the protein. Consider a hydroxyl group in the protein. If the H atom points towards the ligand then there should be an acceptor group in the ligand adjacent to the hydroxyl. If, however, the H atom points away from the ligand, then the corresponding group should be a donor group. Usually, H atoms are missing in protein structures solved by X-ray crystallography. Even if they are present, the rotational barrier is small and the residues can easily adopt an alternative conformation if that would facilitate ligand binding. Therefore, it is preferable to take into account all possibilities. For a

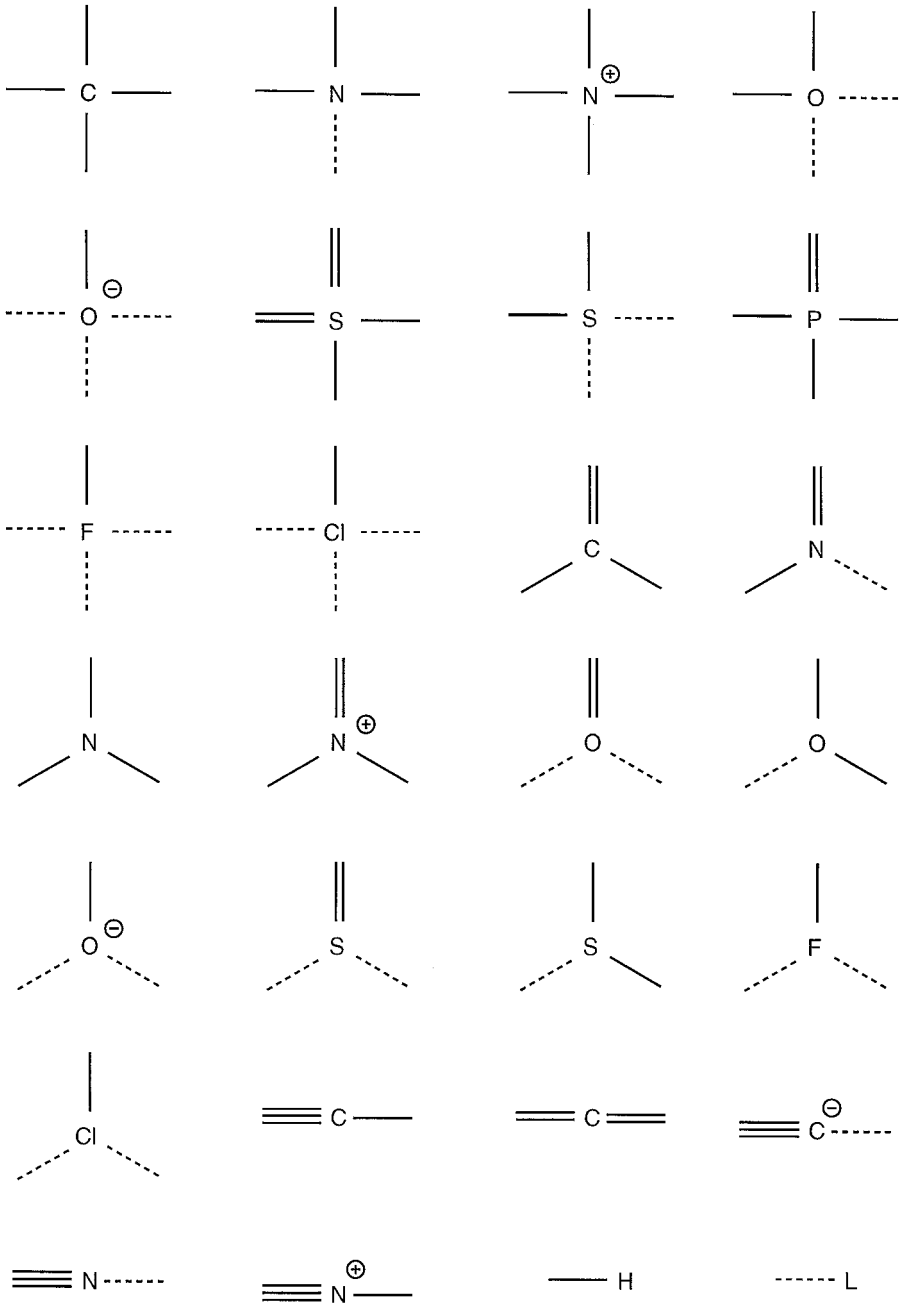

Figure 2. Fragments used in the atom assignment. Virtual bonds that correspond to a lone pair (L) are shown by a dotted line.

single hydroxyl group there are three alternatives to be distinguished by permuting the H atom and the two electron lone pairs. All three have to be considered because it is possible to have several groups in the ligand interacting with the same OH group through the H atom or the lone pairs. The OH can donate one and accept two H atoms. Similar considerations apply to amino and carboxyl groups. For carboxyl groups, ionization states and the position of the H atom depend on the environment, this may change if different ligands are present in the site. The number of possible receptor states is equal to the product of the number of states for each group in the protein with the potential

to form hydrogen bonds with the ligand. These groups may vary from skeleton to skeleton.

All combinations for these groups are tried using a depth-first search method. The search is organized by varying the assignment of H atoms and lone pairs on the terminal nodes of hydrogen-bonding groups of interest in the protein. During the search, unsuitable combinations (e.g. two H atoms attached to both O atoms of a carboxyl group) are identified and rejected; backtracking occurs in such situations. Also, the maximum number of ionized carboxyl groups can be specified.

All receptor states are collected in a list along with the atom types appropriate for the terminal atoms in the complementary ligand group (H or lone pair) and the strength of the bond-estimated as described below. All generated receptor states can be retained. Alternatively, the state could be retained which best fits the atom types specified on the original skeleton, or the original receptor state.

A problem may arise in that metal ions are present in the site and form bonds with both the protein and the ligand. There may be protein and ligand atoms positioned around the metal in such a way that the distances and angles are suitable for hydrogen bonds between them. The donor–acceptor complementarity may be in conflict for such pairs. It is possible, for instance, to have a metal surrounded by anionic oxygens from the protein and the ligand, but it is not possible to maintain the complementarity between the oxygen atoms which are initially assigned to be suitable hydrogen-bond partners. To resolve the ambiguities, in a first pass, atoms bonded to metal cations are identified and labelled and in a second pass the list of potential hydrogen bonds is compiled to exclude those between labelled atoms.

The conformations of rotatable hydrogen-bonding groups in the protein allow different hydrogen-bond states for certain skeleton atoms to be donors or acceptors. One representative for each set of protein conformations which give rise to different hydrogen-bond assignments is generated.

Assignment of templates onto the atoms

In the next step of the algorithm, each atom is examined and a stack of appropriate templates from the library is compiled. The geometric states of the template and the atom must match; tetrahedral-atom templates are assigned to tetrahedral atoms in the skeleton, planar to planar and so on. Hydrogen atoms and electron lone pairs must be present in the skeleton and they can be matched only by hydrogen atoms or lone pairs from the templates.

Optionally, if an atom is in a position to form a hydrogen bond with the receptor, the templates that are unsuitable for this purpose are removed.

Find substructures

The number of possible searches is equal to the product of the number of templates at each node. This number is potentially large, as is the number of acceptable structures, although not all structures are of equal interest; those which interact more strongly with the receptor are preferred. A divide-and-conquer approach is used to generate these assignments with higher priority. The structure is divided into small substructures, and atoms are systematically assigned onto each of them separately. Complete structures are generated as combinations of representatives of each partial structure. The structure with the best score is found using the branch-and-bound search method. This approach is feasible if the ordering of the partial structures is preserved when they are part of a larger structure *i.e.* either there is no interaction between different partial structures, or such an interaction can be neglected.

Substructures are derived by analysis of the molecular graph. A spanning tree is grown using a breadth-first search method [25]. Non-terminal nodes are numbered in the order they are visited and this numbering is used to divide them into substructures of specified size. The default size for the substructure is five non-terminal nodes. Terminal nodes (corresponding to H atoms and lone pairs) are inserted immediately after the non-terminal atom to which they are bonded.

Systematic atom assignment onto substructures

A depth-first search procedure is used to assign atoms onto each substructure taking into account constraints from forbidden atom-string rules, charge accumulation, etc. The representatives generated for each substructure are stored.

A depth-first search [25] is performed over the templates assigned to each atom. In the beginning, the first template of the first atom is assigned, followed by the assignment of the first template of the second atom and so on. When a new template is considered, a check is made to ascertain that the orders of the bonds that are suggested are not in conflict with the orders of the bonds previously assigned; also undesirable combinations of atoms (see next section and Table 1)

are avoided. If a template is rejected, the next of the templates available for the vertex is tested. If none of the suggestions are accepted, the algorithm backtracks to the previous atom and replaces the currently used template with the next in the list and the same procedure is repeated. The order of visiting the atoms helps the early discovery of problems in the structure and pruning the search. When a complete assignment is produced, the structure is scored against each receptor state and stored.

Chemical rules: Forbidden strings of atoms

Certain combinations of atoms are known to be very reactive and thus compounds containing them are unsuitable as synthetic targets. Empirical rules can be derived and expressed as a list of undesirable atomic strings. This idea of keeping a 'bad list' was used in the DENDRAL project for structure elucidation from mass spectroscopic data [23].

The list is consulted each time after a new template is assigned to a vertex. To check whether a particular string of length l is present, each atom, i, in the molecule is considered in turn as a starting point. All different l-atom strings, anchored at i are examined and compared with the target string. If the number of copies of the target string compiled over the whole molecule exceeds a defined maximum number of copies allowed, the rule is violated. The previously added template is withdrawn and the algorithm backtracks. A forbidden combination is defined only once; equivalent strings with shifted frames and reverse orders are automatically checked. Some overlaps between the structures eliminated by one rule or another may be present. The set of rules used in this work is shown in Table 1. The rules are stored in a file which can be modified.

An example definition of a forbidden atom string is shown in Figure 4. This is a rule preventing placing of an oxygen atom next to double bonded carbon atoms in an acyclic part of a molecule.

Row 1 specifies the number of atoms in the string (3), and the maximum number of copies of this string that is allowed in a structure (0). Rows 2–4 contain information about each atom individually. The atom type (C, O, ...), hybridization state (t=tetrahedral, p=planar, l=linear), formal charge and ring participation status (a=acyclic, r=ring) are specified. Wild cards (∗) mean that all states of the corresponding field are considered. Rows 5 and 6 give details about the intermediate bonds. The first number is the order of

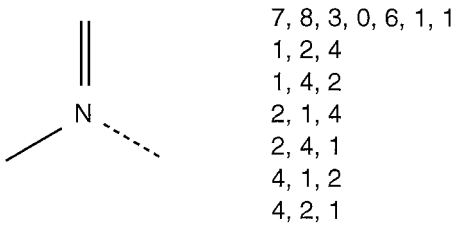

Figure 3. An example definition of a fragment. The virtual bond shown as a dotted line represents a lone pair. The first line in the definition contains sequentially the element number, the COSMIC [24] atom type, the atom geometry (4 = tetrahedral, 3 = planar, 2 = linear, 1 = lone pair or hydrogen), formal charge, number of unique bond order permutations, hydrogen-bond donor and hydrogen-bond acceptor participation. Each of the next lines lists one unique bond-order permutation of the incident bonds. Formal bonds that correspond to lone pairs are assigned an order of 4.

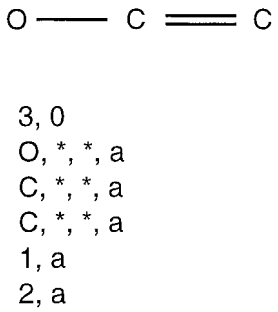

Figure 4. An example definition of a forbidden string. Row 1 specifies the number of atoms in the string (3), and the maximum number of copies of the string allowed in a structure (0). Rows 2–4 contain information about each atom individually. The atom type (C, O, ...), hybridization state (t=tetrahedral, p=planar, l=linear), formal charge and ring participation status (a=acyclic, r=ring) are specified. Wild cards (∗) mean that all states of the corresponding field are considered. Rows 5 and 6 give details about the intermediate bonds. The first number is the order of the bond (1, 2, 3, 4). The second number defines the ring status of the bond (a=acyclic, r=ring).

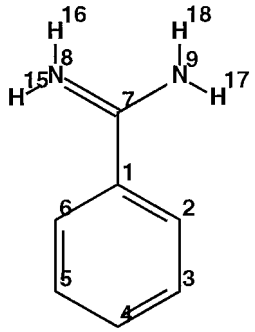

Figure 5. The structure and the numbering scheme of retinol. All ligand drawings were produced with the CACTVS program [35].

Table 1. The chemical rules applied in the atom type assignment procedure. The rules are expressed as alternating strings of atoms and bonds. Each atom is characterized by four values: atom type (C, N, O etc.), geometric state (tetrahedral, planar, etc.), formal charge (−, 0, +) and ring participation (ring or acyclic). Each bond is described by two values: bond order (single, double etc.) and ring participation (ring or acyclic). Copies=maximum number of copies of the string allowed in the molecule. The four numbers that characterize an atom and the two numbers that characterize a bond are listed sequentially below the corresponding atom or bond. Other abbreviations: t = tetrahedral, p = planar, l = linear, a = acyclic, r = ring, ar = aromatic ring, ∗ = any value

String	Copies
(C ,t, ∗,∗) (∗,∗) (O ,∗,−,∗)	0
(O ,∗, ∗,∗) (∗,∗) (O ,∗, ∗,∗)	0
(O ,∗, 0,∗) (1,∗) (S ,∗, ∗,∗)	0
(N ,∗, ∗,a) (∗,a) (N ,∗, ∗,a)	0
(O ,∗, ∗,a) (1,a) (C ,∗, ∗,a) (2,a) (C ,∗, ∗,a)	0
(N ,∗, ∗,a) (1,a) (C ,∗, ∗,a) (2,a) (C ,∗, ∗,a)	0
(S ,∗, ∗,a) (1,a) (C ,p, ∗,a) (2,a) (C ,p, ∗,a)	0
(O ,∗, ∗,∗) (1,a) (C ,∗, ∗,a) (1,a) (O ,∗, ∗,∗)	0
(O ,∗, ∗,∗) (1,a) (C ,∗, ∗,a) (1,a) (N ,∗, ∗,∗)	0
(S ,∗, ∗,∗) (1,∗) (S ,∗, ∗,∗)	0
(O ,∗, ∗,∗) (∗,∗) (S ,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (O ,∗, ∗,∗)	0
(O ,∗, ∗,∗) (∗,∗) (P ,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (O ,∗, ∗,∗)	0
(S ,t, ∗,∗) (∗,∗) (H ,∗, ∗,∗)	0
(P ,∗, ∗,∗) (∗,∗) (H ,∗, ∗,∗)	0
(N ,t, ∗,∗) (∗,∗) (O ,∗, ∗,∗)	0
(N ,p,+,∗) (2,∗) (O ,∗, 0,∗)	0
(N ,p,+,∗) (2,∗) (O ,∗,−,∗)	0
(N ,∗, 0,∗) (∗,∗) (O ,∗, ∗,∗)	0
(N ,∗, ∗,∗) (∗,∗) (N ,∗,+,∗)	0
(N ,∗, ∗,∗) (∗,a) (S ,∗, ∗,∗)	100
(N ,∗, 0,∗) (∗,∗) (P ,∗, ∗,∗)	0
(F ,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (O ,∗, ∗,∗)	0
(F ,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (N ,∗, ∗,∗)	0
(N ,∗, ∗,∗) (∗,∗) (F ,∗, ∗,∗)	0
(O ,∗, ∗,∗) (∗,∗) (F ,∗, ∗,∗)	0
(N ,∗, ∗,∗) (∗,∗) (Cl,∗, ∗,∗)	0
(O ,∗, ∗,∗) (∗,∗) (Cl,∗, ∗,∗)	0
(N ,∗, ∗,∗) (∗,a) (N ,∗, ∗,∗)	0
(N ,∗, ∗,∗) (∗,a) (O ,∗, ∗,∗)	0
(N ,∗, ∗,∗) (∗,∗) (N ,∗, ∗,∗) (∗,∗) (N ,∗, ∗,∗)	0
(Cl,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (O ,∗, ∗,∗)	0
(Cl,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (N ,∗, ∗,∗)	0
(Cl,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (S ,∗, ∗,∗)	0
(Cl,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (P ,∗, ∗,∗)	0
(F ,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (S ,∗, ∗,∗)	0
(F ,∗, ∗,∗) (∗,∗) (C ,∗, ∗,∗) (∗,∗) (P ,∗, ∗,∗)	0
(N ,∗, ∗,r) (∗,r) (N ,∗, ∗,r)	0
(S ,∗, ∗,∗)	1
(P ,∗, ∗,∗)	1
(F ,∗, ∗,∗)	1
(Cl,∗, ∗,∗)	1

Table 1 continued.

String	Copies
(N ,∗, ∗,a) (1,a) (C ,t, ∗,∗) (1,a) (N ,∗, ∗,a)	0
(C ,∗, ∗,a) (2,a) (S ,∗, ∗,a)	0
(S ,t, ∗,∗) (∗,∗) (F ,∗, ∗,∗)	0
(S ,t, ∗,∗) (∗,∗) (Cl,∗, ∗,∗)	0
(N ,∗, ∗,a) (2,a) (S ,∗, ∗,a)	0
(O ,∗, ∗,a) (∗,a) (S ,∗, ∗,a) (∗,a) (F ,∗, ∗,a)	0
(O ,∗, ∗,a) (∗,a) (S ,∗, ∗,a) (∗,a) (Cl,∗, ∗,a)	0
(O ,∗, ∗,a) (∗,a) (P ,∗, ∗,a) (∗,a) (F ,∗, ∗,a)	0
(O ,∗, ∗,a) (∗,a) (P ,∗, ∗,a) (∗,a) (Cl,∗, ∗,a)	0
(F ,∗, ∗,a) (∗,a) (S ,∗, ∗,a) (∗,a) (F ,∗, ∗,a)	0
(F ,∗, ∗,a) (∗,a) (S ,∗, ∗,a) (∗,a) (Cl,∗, ∗,a)	0
(Cl,∗, ∗,a) (∗,a) (S ,∗, ∗,a) (∗,a) (Cl,∗, ∗,a)	0
(F ,∗, ∗,a) (∗,a) (P ,∗, ∗,a) (∗,a) (F ,∗, ∗,a)	0
(F ,∗, ∗,a) (∗,a) (P ,∗, ∗,a) (∗,a) (Cl,∗, ∗,a)	0
(Cl,∗, ∗,a) (∗,a) (P ,∗, ∗,a) (∗,a) (Cl,∗, ∗,a)	0
(S ,∗, ∗,a) (∗,a) (P ,∗, ∗,a)	0
(P,∗,∗,∗)	0
(S,∗,∗,a) (1,a) (N,∗,∗,∗)	0
(C,p,∗,ar) (2,a) (N,∗,∗,∗)	0
(O,∗,∗,a) (1,a) (C,t,∗,a) (1,a) (S,∗,∗,a)	0
(C,p,∗,ar) (2,a) (C,∗,∗,a)	0
(C,t,∗,a) (2,a) (N,p,∗,a) (1,a) (C,p,∗,a)	0
(C,p,∗,r) (1,∗) (8,∗,−,∗)	0
(C,p,∗,a) (2,a) (C,p,∗,a) (1,a) (8,∗,−,∗)	0
(O,∗,∗,a) (2,a) (C,p,∗,∗) (1,∗) (C,p,∗,∗) (2,a) (O,∗,∗,a)	0
(N,p,+,a) (2,a) (C,p,∗,a) (1,a) (O,∗,∗,a)	0

the bond (1, 2, 3, 4). The second number defines the ring status of the bond (a=acyclic, r=ring).

Polar atoms

The ratio of heteroatoms incorporated into the ligand is also controlled. The default value for this ratio, however, is set to one; so no restriction is applied.

Charges

Formal charges, available from the fragment definitions, are also controlled. The positive, negative and total absolute charge (positive minus negative) are accumulated and checked. In addition, when the assignment is completed the total charge (positive plus negative) is also checked. By default no formal charges are allowed.

Scoring the substructures

A fast empirical scoring function has been proposed by Böhm [22] to estimate protein–ligand free energy of interaction. Several adjustable parameters are fitted, using experimentally determined binding constants from a number of complexes. Here we use a similar scheme in which the following contributions are included:

— hydrogen and metal bond: -4.36 kJ mol^{-1}
— lipophilic contact: -0.17 kJ mol^{-1}
— rotatable bond: 0.86 kJ mol^{-1}
— regression constant: -3.61 kJ mol^{-1}

In addition the hydrogen-bond values are multiplied by a correction factor, f, that accounts for distorted geometry:

$$f(\Delta R, \Delta\alpha) = f_1(\Delta R) f_2(\Delta\alpha) \qquad (1)$$

$$f_1(\Delta R) = \begin{cases} 1, & \text{if } \Delta R \leq 0.2 \text{ Å}, \\ 1 - (\Delta R - 0.2)/0.4, & \text{if } 0.2\text{Å} < \Delta R \leq 0.6 \text{ Å}, \\ 0, & \text{if } \Delta R > 0.6 \text{ Å}. \end{cases} \quad (2)$$

$$f_2(\Delta\alpha) = \begin{cases} 1, & \text{if } \Delta\alpha \leq 30°, \\ 1 - (\Delta\alpha - 30)/50, & \text{if } 30° < \Delta\alpha \leq 80°, \\ 0, & \text{if } \Delta\alpha > 80°. \end{cases} \quad (3)$$

where ΔR is the deviation of the hydrogen bond length H-Acceptor from the ideal value of 1.9 Å, $\Delta\alpha$ is the deviation of the hydrogen-bond angle Donor-H-A from 180°.

The contribution of lipophilic contacts is calculated for each C and S atom in the ligand by counting the number of C and S atoms within 5.0 Å in the protein. Thus the scoring function we use is explicitly additive over the ligand atoms. This property is important for the atom assignment method. As rotatable are defined all sp^3–sp^3 and sp^3–sp^2 acyclic bonds, excluding terminal groups.

The coefficients are derived from 30 protein–ligand complexes with crystal structures available in the Brookhaven Protein Databank [26]. Water molecules were removed from the files and not considered in the scoring. Hydrogen atoms and lone pair vectors were added to the protein and retinol molecules using a COSMIC routine [24] and adjusted to optimize to the hydrogen-bond network.

The coefficients were obtained by least-square fitting of the calculated to the experimental binding energies. A value of 10.63 kJ mol^{-1} was obtained for the standard deviation and a linear correlation coefficient of 0.751. The experimental and predicted free energies for the complexes are shown in Table 2. Leave-one-out cross-validation was also performed and a value of 11.45 kJ mol^{-1} was obtained for s_{PRESS} defined as:

$$s_{\text{PRESS}} = \sqrt{\sum_{i=1}^{n} (\Delta G_i^{\text{expt}} - \Delta G_i^{\text{calc}})^2/(n-k-1)},$$

where ΔG_i^{calc} is calculated for each complex in the set from coefficients derived from the other $n-1$ complexes and k is the number of variables.

The scoring function was also tested on five complexes not included in the parameterization and the predicted results are shown in Table 3. The standard deviation for these complexes was 6.60 kJ mol^{-1}. Other expressions of ΔG and different robust statistical estimators [27] are under investigation in order to improve the accuracy of the predictions.

Table 2. PDB complexes used to derive the scoring function parameters. Code is the PDB code of the complex, ΔG_{expt} is the experimentally determined binding energy in kJ mol^{-1}, ΔG_{calc} is the calculated binding energy in kJ mol^{-1}

Code	Protein–Ligand	ΔG_{expt}	ΔG_{calc}
1DWB	Thrombin–Benzamidine	−16.66	−21.91
1FKF	FKFB–FK506	−55.35	−41.04
1MBI	Myoglobin–Imidazole	−10.73	−22.12
1RBP	Retinol BP–Retinol	−38.35	−34.03
1RNE	Renin–CGP38560	−49.64	−47.18
1STP	Streptavidin–Biotin	−76.46	−48.72
1TLP	Thermolysin–Phosphoramidon	−43.08	−33.48
1TMN	Thermolysin–CLT	−41.65	−50.32
1ULB	PNP–Guanine	−30.24	−25.93
2ER6	Endothiapepsin–H256	−41.20	−42.39
2GBP	Galactose BP–Galactose	−43.36	−52.31
2IFB	Fatty acid BP–C_{15}COOH	−30.98	−25.92
2PHH	PHBH–*p*-hydrobenzoic-acid	−26.70	−28.84
2R04	Virus coat protein–Cmpd 4	−35.49	−33.27
2XIS	Xylose isomerase–Xylitol	−33.21	−23.81
2YPI	TIM–phosphoglycolic acid	−27.50	−23.03
3CPA	Carboxypeptidase A–GT	−22.14	−30.50
3DFR	DHFR–Methotrexate	−55.34	−49.46
3PTB	Trypsin–Benzamidine	−27.05	−27.34
4CNA	Concavalin A–Methylmannosid	−11.41	−29.62
4DFR	DHFR–Methotrexate	−55.35	−44.33
4ER4	Endothiapepsin–H142	−38.74	−37.48
4HMG	Hemagglutinin–sialic acid	−14.55	−31.93
4HVP	HIV protease–MVT101	−35.09	−40.18
4PHV	HIV protease–L700,417	−52.21	−67.57
4TLN	Thermolysin–Leu-NHOH	−21.23	−24.37
4TMN	Thermolysin–ZF$^{\text{p}}$LA	−55.12	−54.16
5TMN	Thermolysin–ZG$^{\text{p}}$LL	−45.87	−40.62
6CPA	Carboxypeptidase A–ZAA$^{\text{p}}$(O)F	−65.73	−43.59
9HVP	HIV protease–A74704	−47.64	−56.01

Find best score by branch-and-bound search

The substructures are combined to form complete structures. The number of possible combinations is equal to the product of the numbers of representatives found for each substructure. Some combinations are not viable, due to conflicting bond orders at the

Table 3. PDB complexes used to test the scoring function. ΔG_{expt} is the experimentally determined binding energy in kJ mol^{-1}, ΔG_{calc} is the calculated binding energy in kJ mol^{-1}

Code	Ligand	ΔG_{expt}	ΔG_{calc}
1DWD	Thrombin–NAPAP	−48.61	−41.54
1ETR	Thrombin–MQPA	−42.22	−35.48
1ETT	Thrombin–TAPAP	−35.32	−31.30
1PPC	Trypsin–NAPAP	−36.86	−44.03
1PPH	Trypsin–3-TAPAP	−35.49	−38.83

junctions between the substructures or chemical rule violations; these are identified and rejected.

Both the potential number of combinations and the number of solutions produced may be very large. The lowest scoring structures are of greatest interest and the aim is to find them quickly. The score of each substructure representative is calculated.

The function calculated for the whole structure can be partitioned into contributions of each atom plus a constant term that accounts for the overall loss of translational and rotational freedom. The hydrogen bonds and ionic interactions are associated with particular atoms; all other atoms score zero in these two fields. The rotatable bond term is divided between the atoms that form the bond. The hydrophobic-contact term is also partitioned amongst the atoms.

The score is obtained for each substructure representative against each receptor state. Consider initially the case where there is only one receptor state. The representatives with the lowest and the highest scores are found for each substructure. The score of the complete structure with the best score is bracketed between the sum of the lowest scores and the sum of the highest scores over all substructures. A branch-and-bound procedure [31] is used to find the complete structure with the best score. A tolerance parameter is set and a depth-first search is performed over all combinations of partial structure representatives. Backtracking occurs in three cases: first, if there are mismatched bond orders at the junctions between partial structures; second, if chemical rules are violated in the composite; third, if the sum of the scores of the representatives included so far becomes bigger than the tolerance parameter. To improve the efficiency, the representatives of each partial structure are examined in ascending order according to their scores. If an acceptable complete structure is generated, the tolerance parameter is reset to the sum of the scores of the representatives of the structure above the lowest scores obtained for the corresponding substructures. The search is continued until all possibilities are examined which become fewer and fewer after each tolerance update. After the global minimum is obtained, the tolerance is increased by a specified amount (the default is 0.01 kJ mol^{-1}) and all permitted structures below that threshold are generated.

If more than one receptor state is present then for each receptor state the representatives with the lowest and the highest scores are found for each substructure. The score of the complete structure with the best score is bracketed between the sum of the lowest scores and the sum of the highest scores over all substructures. The receptor is found for which the lowest-score estimate is the lowest amongst all receptor states. The representatives of each substructure are examined in ascending order according to their scores against that receptor state. During the search, the sums of differences between the score of a substructure representative and the lowest score available for that substructure from all receptor states are calculated for each receptor state and the lowest sum amongst all receptor states is compared with the tolerance parameter.

Degeneracy

Degenerate structures can result from two sources. First, symmetry in the scaffold is not considered. The most common situation where symmetry plays a rôle are terminal atoms, which are rotationally invariant and are set to either H atoms or lone pairs. It is undesirable to remove degeneracy before the search since the receptor score will depend on the exact permutation of types. When the interaction energies with the receptor are considered, these may produce different energies. They correspond to different binding modes of the same ligand. Second, alternative patterns on conjugated aromatic systems result in equivalent structures. To take these factors into account, the structures were classified on the basis of unique COSMIC atom types on the nodes. The structures were canonically numbered using the previously described modified Morgan method [6] and assigned into classes.

Implementation

The algorithm has been implemented in a program CHAOS (CHemical Assignment Onto Scaffolds) written in FORTRAN 77. The program was compiled at

optimization level 2 and run on a Silicon Graphics R10000.

Results

Five examples are used to illustrate the method. The coordinates of the complexes were all taken from the Brookhaven Protein Databank [26]. Their codes are: 3PTB, 1RBP, 3DFR, 1TMN and 4PHV. In all cases but 4PHV, all water molecules and cofactors were removed from the complex. For 4PHV, which is a complex of a HIV protease inhibitor, the water molecule that mediates the interaction was retained. Hydrogen atoms and lone-pair vectors were added to the protein and retinol molecules using a COSMIC routine [24] and optimized to the hydrogen-bond network. Results for all complexes are summarized in Table 4.

3PTB complex

3PTB is a trypsin–benzamidine complex [28]. Benzamidine (Figure 5) is a small molecule that displays both hydrogen bonding and hydrophobic interactions and is a very good test example. There are 18 nodes after addition of H atoms. Six receptor models were generated by varying the ionization states of carboxyl groups and rotating hydroxyl groups. The calculated binding energy of benzamidine was -27.34 kJ mol^{-1}; the experimental value is -27.05 kJ mol^{-1} [29].

Templates were assigned at each node. After the initial assignment there were 1.6×10^{12} possible assignments. The structure was automatically divided into four substructures by the program and for each substructure all representatives were generated. At this stage there were 1.1×10^{5} possibilities. The branch-and-bound procedure was used to find the structures with the lowest score.

The best score was found at -28.30 kJ mol^{-1} in 1.1 s. The lowest-score structure was an unprotonated variant of benzamidine (by default no formal charges were allowed). In that structure N9 was both hydrogen bond donor (to Asp189) and acceptor (from Ser190).

To compare the efficiency of the algorithm we also performed an exhaustive search after the initial assignment without using the structure division and branch-and-bound modules. This search produced 2000 structures in 75 s.

Retinol in 1RBP complex

Brookhaven Protein Databank entry 1RBP is a complex between retinol (Figure 6) and retinol binding protein [30]. The interaction is purely hydrophobic. There are 53 nodes after addition of H atoms and lone pairs. One receptor model was generated since there are no interacting hydrogen bonding groups. The calculated binding energy of retinol was -34.03 kJ mol^{-1}; the experimentally determined binding energy is -38.35 kJ mol^{-1}.

Templates were assigned at each node. After the initial assignment there were 2.8×10^{33} possible assignments. The structure was automatically divided into five substructures and for each substructure all representatives were generated. At this stage there were 4.8×10^{9} possibilities. The branch-and-bound procedure was then used to find the structures with the lowest score.

There was one structure with best energy of -35.90 kJ mol^{-1} and it was found in 15.8 s. In this structure O21 was substituted for carbon and the lone pairs attached to O21 for hydrogen atoms. There are no protein hydrogen-bonding groups interacting with the ligand and the energy was optimized by additional hydrophobic interactions through the substitution of O21 for a carbon atom.

3DFR complex

3DFR [31] contains the coordinates of dihydrofolate reductase from *L. casei* co-crystallized with NADPH and methotrexate (MTX). This enzyme catalyses the conversion of dihydrofolate to tetrahydrofolate, which is an important reaction in the biosynthetic pathway of the nucleic acids.

The MTX skeleton (Figure 7) was used in this example. MTX is a relatively large ligand which contains both hydrogen-bonding and hydrophobic moieties as well as a number of rotatable bonds. There are 67 nodes after addition of H atoms and lone pairs. Six receptor models were generated by changing the protonation states of receptor groups. The experimentally determined binding energy of MTX is -55.34 kJ mol^{-1}; the calculated binding energy was -49.46 kJ mol^{-1}.

Templates were assigned at each node. After the initial assignment there were 1.5×10^{45} possible assignments. The skeleton was automatically divided into nine substructures and for each substructure all representatives were generated. At this stage there were 3.9×10^{13} possibilities. The branch-and-bound

Table 4. Test examples. Code is the PDB code, Nodes is the number of nodes, Rec is the number of receptor states, Sub is the number of substructures, Size 1 is the search size after initial template assignment, Size 2 is the search size after substructure assignment, ΔG is the calculated binding energy for the original ligand in kJ mol^{-1}, ΔG_{best} is the calculated binding energy for the best structure in kJ mol^{-1}, Sol is the number of structures with ΔG_{best}, CPU is the CPU time in seconds

Code	Nodes	Rec	Sub	Size 1	Size 2	ΔG	ΔG_{best}	Sol	CPU
3PTB	18	6	4	1.6×10^{12}	1.1×10^{5}	−27.34	−28.30	1	1.1
1RBP	53	1	5	2.8×10^{33}	4.8×10^{9}	−34.03	−35.90	1	15.8
3DFR	67	6	9	1.5×10^{45}	3.9×10^{13}	−49.46	−55.23	90	17.8
1TMN	77	24	10	6.3×10^{50}	5.1×10^{15}	−50.32	−50.56	20	8.9
4PHV	98	36	12	1.0×10^{65}	8.7×10^{19}	−74.95	−76.62	16	11.7

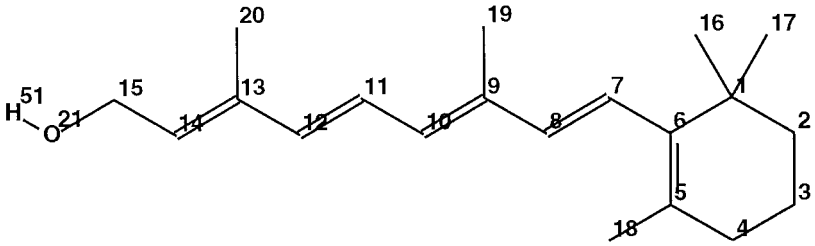

Figure 6. The structure and the numbering scheme of benzamidine.

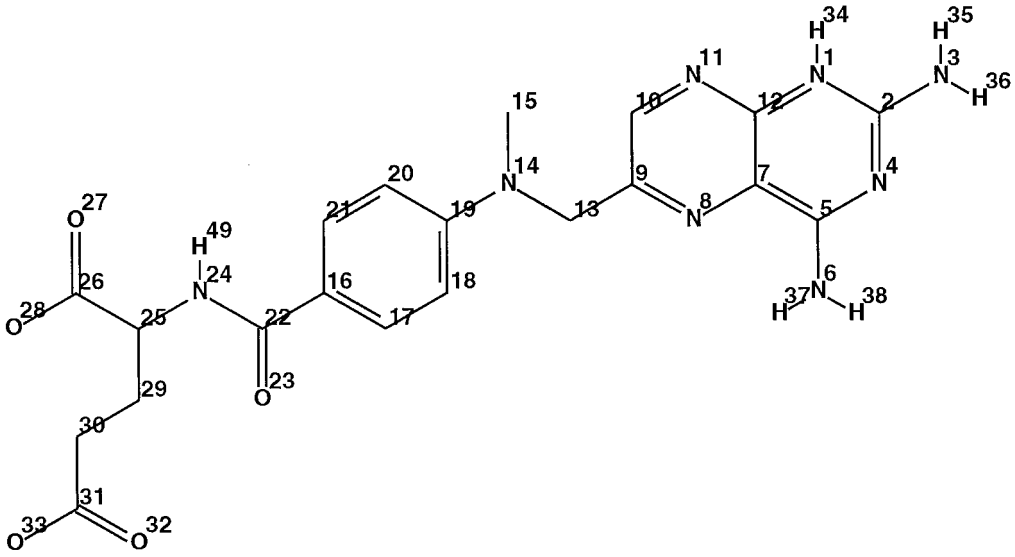

Figure 7. The structure and the numbering scheme of methotrexate.

procedure was then used to find the structures with the lowest score.

The best score was found at −55.23 kJ mol^{-1} in 17.8 s. There were 90 unique structures with the lowest score. From the lowest-score structures, in the one closest to MTX N4, N8, N11 and O23 were set to carbon atoms; they are not involved in hydrogen bonding and a better score was achieved by hydrophobic interaction. N1 was unprotonated (by default no formal charges are allowed) and was a hydrogen-bond acceptor (from Asp26).

Several other features are also interesting to note. The benzene ring was present in the structure; since its constituent atoms do not form hydrogen bonds, the

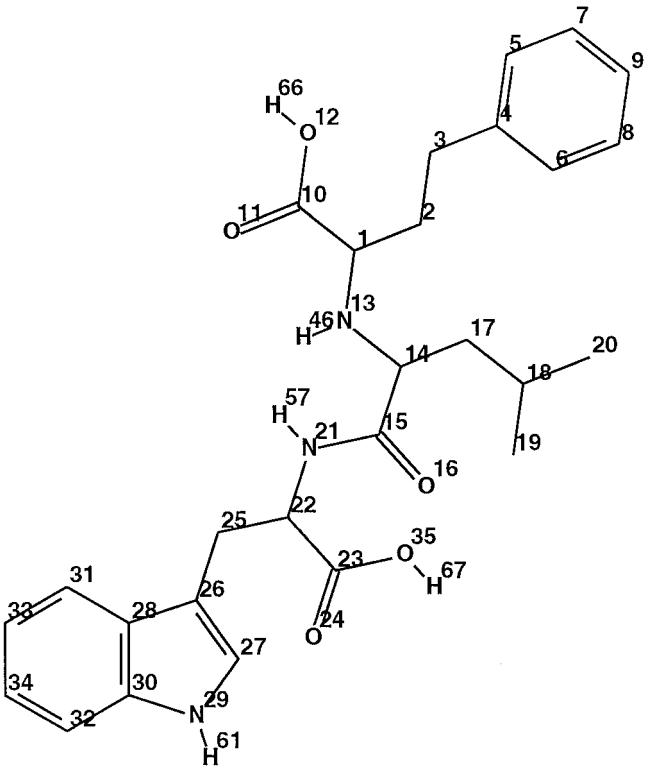

Figure 8. The structure and the numbering scheme of CLT.

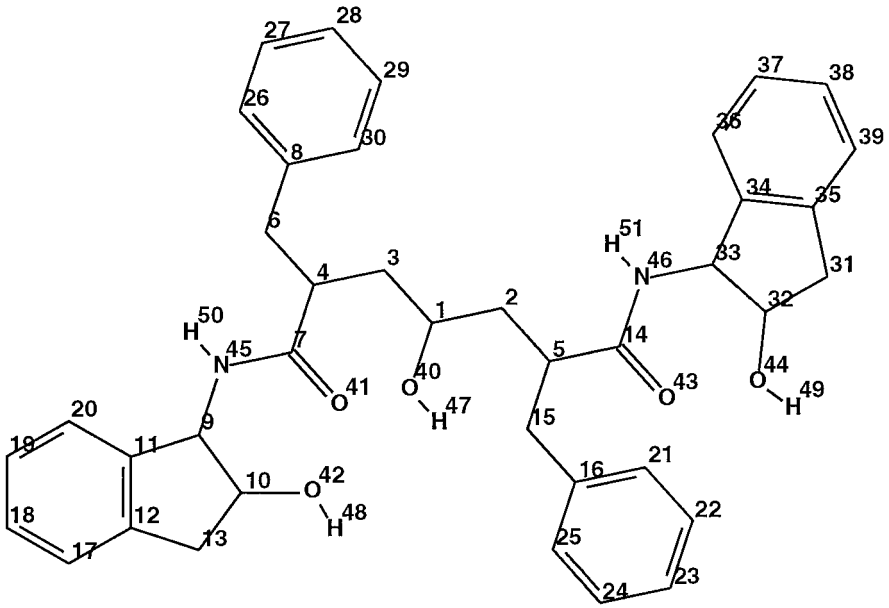

Figure 9. The structure and the numbering scheme of L700,417.

score is minimized through hydrophobic interactions. The presence of N14 and N24 cannot be explained simply. Nitrogen was perhaps forced into this position, when the atoms of the benzene ring were set to carbon atoms, by a combination of matching bond order requirement and chemical rules. In some of the

other lowest-score structures, N24 was mutated to an oxygen atom. Another interesting alteration was the simultaneous replacement of C25 by sp^3 nitrogen, O23 by a planar NH2 group and the migration of the double bond between atoms 22 and 23 to 22 and 24. This arrangement had the same hydrophobic score as the one obtained by single replacement of O23 to CH2 observed in other best score structures. O27, O28, O32 and O33 were sometimes replaced by an acceptor nitrogen so that the hydrogen-bonding pattern was conserved.

1TMN complex

1TMN [32] contains the structure of thermolysin complexed with CLT, N-(1-carboxy-3-phenylpropyl)-L-leucyl-L-tryptophan (Figure 8). The interesting feature of this complex is the presence of a metal ion (Zn) interacting with the ligand. There are 77 nodes after addition of H atoms and lone pairs. 24 receptor models were generated by varying the ionization states of carboxyl groups and rotating hydroxyl groups. The experimental binding energy is -41.65 kJ mol^{-1}; the calculated binding energy was -50.32 kJ mol^{-1}.

Templates were assigned at each node. After the initial assignment there were 6.3×10^{50} possible assignments. The structure was divided into 10 substructures and for each substructure all representatives were generated. At this stage there were 5.1×10^{15} possibilities. The branch-and-bound procedure was then used to find the structures with the lowest score.

The best score was found at -50.56 kJ mol^{-1} in 8.9 s. There were 20 unique structures with the lowest score. The original ligand was one of these structures. The small difference in score is explained by the different position of the H atom attached to N13 (donating to the amide carbonyl of Ala113) in the original ligand.

In some of the structures O12, O24 and O35 were mutated to N atoms or the positions of the double bonds changed from 10–11 to 10–12 and 23–24 to 23–35. These modifications preserve the hydrogen bonding network in the complex. Also, C7 and C34 were mutated to N atoms in some of the structures since they did not contribute to the buried surface area of the complex.

4PHV complex

4PHV [33] is a HIV-1 protease complex with L700,417 (Figure 9). Water molecule 1 was included as part of the receptor model. There are 98 nodes in the ligand scaffold after addition of H atoms and lone pairs. Thirty-six receptor models were generated by varying the ionization states of carboxyl groups and rotating hydroxyl groups. The experimentally determined binding energy is -52.21 kJ mol^{-1}; the calculated binding energy was -67.57 kJ mol^{-1}. In this test, water molecule 1 was considered as part of the receptor and its contribution to the binding energy was not excluded; the calculated energy in this case was -74.95 kJ mol^{-1}.

Templates were assigned at each node. After the initial assignment there were 1.0×10^{65} possible assignments. The structure was divided into 12 substructures and for each substructure all representatives were generated. At this stage there were 8.7×10^{19} possibilities. The branch-and-bound procedure was then used to find the structures with the lowest score.

The best score was found at -76.62 kJ mol^{-1} in 11.7 s. There were 16 unique structures with the best score. In the structure most similar to L700,417, the hydroxyl group at O40 was replaced by an amino group, because it makes an additional hydrogen bond with Asp25 in the protein. In some of the structures, the hydroxyl groups at O42 and O44 were mutated to amino groups. Also carbonyl groups at O41 and O43 were sometimes changed to imino groups. All these modifications preserve the hydrogen-bonding network in the complex.

With the last four test datasets, an exhaustive search without using the structure division and branch-and-bound modules, after 20 h CPU was nowhere near completion.

Discussion and Conclusions

A computational method has been presented for the assignment of atoms onto molecular scaffolds. The algorithm has been extensively tested and results have been presented for five protein–ligand complexes with coordinates available in the PDB. In some cases the original ligands have been recreated as best solutions, in other cases structures with better scores have been obtained with good reasons why the changes have been made.

The method offers several advantages compared with earlier work [7, 12–16, 20, 21]. First, it provides an empirically derived estimate of the free energy change for the receptor–ligand binding which gives an objective way of ranking the solutions. Second, the flexible incorporation of rules for chemical sense,

similar to those used in the DENDRAL project [23] for structure elucidation, makes it possible to direct the generation towards structures which are synthetically reasonable. Third, the method is systematic and guaranteed to find the best solution, unlike stochastic methods. The division of the problem into smaller problems which can be solved and then recombined provides a way of approaching the combinatorial explosion in atom assignment. The method is efficient and a large number of structures can be processed automatically. Fourth, receptor ionization and conformational states are generated and examined so that the most favourable interaction can be utilized. Fifth, water molecules and metal ions can also be handled which makes the algorithm general and applicable to diverse protein complexes.

In future, results will be reported on the application of the algorithm for the optimal atom type assignment on scaffolds generated *de novo* [6]. Such applications are a field test of the usefulness of this and similar methods.

We also plan to provide a number of different DENDRAL type chemical-rule sets derived from analysis of databases of known structures. This will make the rules more objective and will better reflect the current knowledge of chemical reactions available. The rules are easy to edit and can also be modified to suit the preferences of a particular synthetic chemist.

Another direction for future research is the testing of new models for free energy estimation which will lead to improved predictive power. For the purposes of the atom assignment method, the models apart from anything else will be required to satisfy the need for additivity over ligand atoms. This condition is essential for the atom assignment method to work efficiently.

Currently, we are also investigating the possibility of using faster search algorithms which would allow one to process more scaffolds. An implementation of dynamic-programming and application results for atom assignment will be reported shortly. A comparison with the present method will also be presented.

The most direct way to use the method in similarity-directed design [34] is to generate an artificial receptor model around a set of superimposed active molecules. Such a receptor model would consist of suitably placed complementary hydrogen-bonding and hydrophobic groups. The procedure described in this paper could then be applied directly.

In summary, in an earlier paper [6] on *de novo* structure generation techniques we developed a procedure for the creation of diverse molecular skeletons; in the current work this research has been extended to the optimal assignment of atoms onto a chosen skeleton. The procedure has been tested on known protein–ligand complexes. A branch-and-bound procedure has been shown to be an efficient approach for atom assignment. The partitioning of structure generation into a primary step, for creation of scaffolds to fit the site optimally, followed by a secondary step of atom assignment onto the scaffolds, may provide an optimization method for exploring diversity within *de novo* structure generation.

Acknowledgements

The authors wish to thank Rhône-Poulenc Rorer (N.P.T.) and the Wellcome Trust through the PRF scheme (P.M.D.) for personal financial support. Part of this work was carried out in the Cambridge Centre for Molecular Recognition funded by the BBSRC.

References

1. Slater, P.E. and Timms, D., J. Mol. Graph., (1993) 248.
2. Verlinde, C.L.M.J. and Hol, W.G.J., Structure, 2 (1994) 577.
3. Lewis, R.A. and Leach, A.R., J. Comput.-Aided Mol. Design, 8 (1994) 467.
4. Lewis, R.A. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 125.
5. Lewis, R.A. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 141.
6. Todorov, N.P. and Dean, P.M., J. Comput.-Aided Mol. Design, 11 (1997) 175.
7. Chan, S.L., Chau, P.-L. and Goodman, J.M., J. Comput.-Aided Mol. Design, 6 (1992) 461.
8. Nakamura, H., Komatsu, K., Nakagawa, S. and Umeyana, H., J. Mol. Graph., 3 (1985) 2.
9. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8 (1994) 513.
10. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8 (1994) 527.
11. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8 (1994) 545.
12. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 341.
13. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 351.
14. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 359.
15. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 448.
16. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 457.
17. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 385.

18. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 397.
19. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 407.
20. Gillet, V., Johnson, A.P., Mata, P., Sike, S. and Zsoldos, Z., J. Comput.-Aided Mol. Design, 7 (1993) 127.
21. Gillet, V., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z. and Johnson, A.P., J. Comput.-Aided Mol. Design, 7 (1993) 127.
22. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
23. Gray, N.A.B., Computer-Assisted Structure Elucidation, Chapter XI, pp. 399–455, John Wiley and Sons, New York, NY, 1986.
24. Morley, S.D., Abraham, R.J., Haworth, I.S., Jackson, D.E., Saunders, M.R. and Vinter, J.G., J. Comput.-Aided Mol. Design, 5 (1991) 475.
25. Nilsson, N.J., Principles of Artificial Intelligence, Springer-Verlag, Berlin, 1982.
26. 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.
27. Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, Cambridge, 1986.
28. Marquart, M., Walter, J., Deisenhofer, J., Bode, W., Huber, R., Acta Crystallogr. B, 39 (1983) 480.
29. Mares-Guia, M. and Shaw, E., J. Biol. Chem., 240 (1965) 1579.
30. Cowan, S.W., Newcomer, M.E. and Jones, T.A., Proteins, 8 (1990) 44.
31. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257 (1982) 13650.
32. Matthews, B.W., Acc. Chem. Res., 21 (1988) 333.
33. Bone, R., Vacca, J.P., Anderson, P.S. and Holloway, M.K., J. Am. Chem. Soc., 113 (1991) 9382.
34. Dean, P.M., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, pp. 1–23, Blackie Academic and Professional, Cambridge, 1995.
35. Ihlenfeldt, W.D., Takahashi, Y., Abe, H. and Sasaki, S., J. Chem. Inf. Comput. Sci., 34 (1982) 109.