PRO_LIGAND: An Approach to *de Novo* Molecular Design. 5. Tools for the Analysis of Generated Structures

David E. Clark* and Christopher W. Murray

Proteus Molecular Design Ltd., Proteus House, Lyme Green Business Park, Macclesfield, Cheshire, SK11 0JL, United Kingdom

Received April 6, 1995®

A pressing problem facing users of *de novo* design programs and 3-D database searching software is how to cope with the large numbers of structures which can be produced as answers to a given design question. In this paper, an integrated suite of tools for molecular structure analysis is described which can assist in the evaluation of large answer sets. The suite includes tools for clustering and ranking of structures by various criteria and for 2-D substructure searching. The use of the tools is illustrated with reference to the evaluation of two large answer sets generated by our in-house *de novo* design system, PRO_LIGAND.

INTRODUCTION

Since the beginning of the decade, there has been a sharp rise in the number of protein structures solved using X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.[1] This increase has enabled the wider application of structure-based techniques in drug discovery research.[2-5] Several reports of lead compounds generated using structure-based drug design (SBDD) methodologies have been published, e.g.,[6-9] and some of these are, at the time of writing, in clinical trials.

An emerging class of computational tools for SBDD are programs for *de novo* design, i.e., the design of novel molecules to satisfy a set of steric and/or chemical constraints imposed, for example, by a receptor active site, a CoMFA model or a pharmacophore. Many such *de novo* design programs have been developed,[10-35] and the field has been recently reviewed.[36] Our in-house facility for *de novo* design, PRO_LIGAND, has also been described in a number of recent papers.[37-40]

Using PRO_LIGAND, we find it is possible to generate hundreds of structures per hour as potential solutions to a design problem. However, this in itself presents a further problem: how to deal with such large answer sets? This difficulty is not unique to *de novo* design programs; it is also encountered in 3-D database searching, particularly with the advent of systems capable of conformationally flexible searching (CFS) on large databases.[41-45] This fact has been noted by Pearlman[46] who suggests, *inter alia*, that 3-D searching software packages (and thus, by implication, *de novo* design programs) will in the future be distinguished not by their ability for generating answer sets, but rather by the tools they provide to aid the user in dealing with the voluminous output. Two such tools developed for the 3-D database context are FAMILY[47] which employs clique detection to cluster search output into structural families according to their similarity in specified interatomic distances and MODSMI[48] which can identify unique "core" molecules from the output of a 3D database search. Commercial software packages also offer tools for prioritizing 3D search output. For instance, MDL's ISIS/3D[49] allows the user to rank hits from a 3D search by the value of the RMS deviation of the fitted conformation from the query constraints, the number of rotatable bonds used in the torsional fitting process or a van der Waals energy difference between the fitted and the stored conformation.[44] Tripos' Molecular Spreadsheet functionality also permits the ranking and clustering of molecular structures from a database hit list[50] or the LeapFrog *de novo* design program.[51]

In general, in the case of *de novo* design programs, the problem of analyzing output is even more acute than with 3-D database searching. In the latter, the molecules retrieved are, in general, known compounds which may even have been previously synthesized in-house. However, *de novo* design programs are capable of suggesting entirely new chemical entities which require further considerations in the evaluation process such as chemical stability and synthetic accessibility. Early *de novo* design programs described in the literature seemed to offer little to aid the user in this way apart from various scoring functions. More recently, the problem has begun to be addressed by various groups: Eisen *et al.*[31] who have implemented a spreadsheet-like TABLE function in their HOOK program for sorting generated structures by various energetic and structural criteria, Bohacek and McMartin[32] have described an apparently effective strategy for evaluating structures generated by their GROWMOL program using a series of "filters" followed by clustering. Finally, the SPROUT program of Johnson *et al.* allows clustering of generated skeletons into groups for ease of viewing.[24] The same workers have also sought to analyze structures in terms of synthetic accessibility by means of their expert system-based CAESA program.[52]

In this paper, we describe the further development of PRO_LIGAND, specifically, an integrated suite of tools (the Analysis module) to facilitate the evaluation of the structures generated by the program. First, however, we give a brief description of PRO_LIGAND to set the work in context. Full details may be found in refs 37-39.

OVERVIEW OF PRO_LIGAND

PRO_LIGAND consists of five modules which operate in sequence to generate and analyze molecular structures. The first module to operate in the design process is *Design-base Generation*. As the name suggests, this module produces a *design base* from one or more input molecular

* To whom all correspondence should be addressed.
® Abstract published in *Advance ACS Abstracts*, August 15, 1995.

0095-2338/95/1635-0914$09.00/0

structures. The design base represents the key structural features which will guide the design process and typically requires the extraction of the active site from a receptor or the generation of a pharmacophore from a set of active analogues or a molecular field analysis (MFA) grid.

Next, a *design model* is constructed by the *Design-model Generation* module. The design model is a 3D template that describes the idealized steric and hydrogen-bonding features of the chemical structures to be designed. These features are represented by *interaction sites*.[16,17,53] Hydrogen bond acceptors and donors are represented by **A**−**Y** and **D**−**X** vectors, respectively, while lipophilic regions are characterized by **L** or **R** points according to whether they are aliphatic or aromatic in nature. These sites are generated to be either complementary or similar to the design base atoms, depending on whether the object is to design a molecule to fit into a known receptor or to mimic a set of active analogues. The type and location of these sites are generated *via* a user-definable rule base.

The *Structure Generation* module produces molecular structures consistent with the design model by assembling small 3-D molecular fragments from preconstructed libraries. These library fragments are labeled to indicate the types of interaction site they may match, and a rapid graph-theoretical algorithm is used to seek fits of the fragments onto the design model. The fitting procedure also corrects or eliminates any bad inter- or intramolecular van der Waals' clashes. A great variety of modes of fragment assembly are available to the user, including a continuous growth procedure and procedures for inter- and intrafragment bridging. In addition, the user also has full control over the structuring and ranking of the fragment libraries. Each generated solution is scored on the basis of the number of design model features it has succeeded in fulfilling and on certain structural characteristics, such as the number of rings or asymmetric carbon atoms.

Once structure generation is complete, the user may submit (a subset of) the built structures to the *Structure Refinement* module, in which a genetic algorithm approach is employed to breed further high-scoring structures from those produced by structure generation.

Finally, the structures produced by either the Structure Generation or the Structure Refinement module can be evaluated by means of the Analysis module which is the subject of this paper.

OVERVIEW OF THE ANALYSIS MODULE

The tools contained within the Analysis module of PRO_LIGAND fall into five classes:

- tools for clustering and ranking structures according to 2-D structural similarity
- tools for clustering and ranking structures according to the region of the design model they occupy or the design base (receptor active site) atoms with which they interact
- tools for grouping and ranking structures according to molecular property values
- tools for grouping structures according to the presence or absence of user-specified substructures
- utility tools

The following sections will describe each of these classes more fully.

CLUSTERING AND RANKING BY 2-D STRUCTURAL SIMILARITY

Given a set of molecular structures, two methods by which an overview of their structural types may be obtained are *clustering* and *dissimilarity ranking*. Once a structure of interest has been located, perhaps by one of these methods, *similarity ranking* can be used to order the remaining structures according to their similarity to that structure. Each of these methods will be described in more detail below. Since, however, all of them depend on some means of quantifying the similarity of any pair of structures, the *similarity measures* employed in this work will be discussed first.

Similarity Measures. To enable the clustering of chemical structures by structural similarity, two *fragment-based* similarity measures have been implemented. Both measures are types of 2-D descriptors which have been previously applied in the context of chemical structure database searching.[54,55]

The first descriptor set consists of 172 *atom-centered* fragments which were generated from an analysis of 5000 structures contained in the Cambridge Structural Database (CSD) (version 5.07).[56] The descriptors consist of a character string containing the element type, the number of attached hydrogen atoms, and then the attached heavy atoms in reverse alphabetic order. Thus, for instance, the string "C 1 O C C" indicates a carbon atom attached to one hydrogen atom, one oxygen atom, and two other carbon atoms.

The second descriptor set is based on the work of Carhart *et al.*[57] and consists of a set of 1643 *atom-pair* descriptors. An initial set of 1106 descriptors were generated by an analysis of 2600 CSD structures; there were added to manually as gaps in the descriptor set were revealed during use. Each descriptor string comprises two atoms described in terms of their element type and the number of attached hydrogen atoms and also the shortest path (in terms of the number of bonds) between them. For example, "C 1 5N 2" denotes a carbon atom with one attached hydrogen separated by five bonds from a nitrogen atom with two attached hydrogens.

At this point, it should be noted that while the CSD is a useful research tool, many of the structures it contains are atypical of those encountered in pharmaceutical research. Thus, other databases, such as MDL's MDDR,[58] might give rise to descriptor sets that would better describe the types of structures generated by general *de novo* design programs. However, the CSD was the only available database at the time this work was carried out.

To represent a chemical structure in terms of either of these sets of descriptors, a bit string of length *NDES* is set up, where *NDES* is the number of descriptors in the set. The structure's heavy atoms are then analyzed to see which of the descriptors the structure contains, and the bits corresponding to those descriptors are then set to 1. If the program cannot find a particular descriptor contained in a structure in its descriptor set, it will issue a warning message indicating the missing descriptor. The descriptor(s) in question can then be manually added to the relevant files of descriptors and the number of descriptors incremented for future runs.

By comparing the bit strings calculated for two structures, it is possible to determine rapidly a quantitative measure of

Figure 1. Two simple structures to illustrate the calculation of fragment-based molecular similarity.

their structural similarity. The similarity measure used in the Analysis module is the *Tanimoto coefficient* which has been shown to work well for chemical structure applications.[54,55] The Tanimoto coefficient, *TC*, for two structures A and B is calculated from their bitstrings as follows

$$TC = \frac{COMMON}{NBITS\_A + NBITS\_B - COMMON} \quad (1)$$

where *COMMON* is the number of bits set to 1 which the structures have in common, *NBITS_A* is the total number of bits set to 1 in structure A, and *NBITS_B* is the total number of bits set to 1 in structure B. In our application we have used the efficient algorithm described by Willett[59] for the calculation of molecular similarities from such bitstring representations.

The maximum value of *TC* is 1.0 which occurs when *NBITS_A* equals *NBITS_B* equals *COMMON*, i.e., the structures are identical in terms of the descriptors applied to them. Correspondingly, the lowest value of *TC* is 0.0, when *COMMON* is zero, i.e., the structures have no descriptors in common.

As an example of this, consider the two simple structures in Figure 1. **A** contains four different atom-centered descriptors, *viz.*, "C 3 C", "C 2 C C", "C 1 C C", and "C 0 C C C". **B**, on the other hand, contains only three descriptors: "C 3 C", "C 1 C C", and "C 0 C C C". Thus, referring to the equation above, *NBITS_A* is 4, *NBITS_B* is 3, and *COMMON* is 3. The resulting similarity value is therefore 0.75 which would seem to be reasonable.

Using similarity measures calculated in this manner, enables the use of clustering and ranking tools such as those described below.

Clustering. The purpose of *clustering* is to group together the structures which have been generated by PRO_LIGAND so that each group, or *cluster*, holds structures which are *similar* to one another according to the similarity measure in use and different from the molecules in the other clusters by that same definition. Each cluster can then be represented by a single *representative structure*, and, by looking at these representatives, the user can gain a rapid overview of the structural classes produced by a PRO_LIGAND job. An example of this kind of application of clustering in a database context has recently been reported by Johnson and Maliski.[60]

There are many available algorithms for clustering sets of data, but that due to Jarvis and Patrick[61] has been shown to work best for the clustering of chemical structures by fragment-based similarity measures.[54,55] Consequently, this has been implemented in the Analysis module. For full details of the algorithm's operation, the reader is referred to refs 54 and 62 but, in brief, it works as follows:[63]

1. Calculate the *nearest-neighbour table, NNTABLE.* This is an array of dimension *KNN* × *NSTRUC* which identifies the *KNN* nearest neighbors (i.e., most similar structures) of each of the *NSTRUC* structures. The intermolecular similarities are calculated using one the fragment-based measures described above.

2. Two structures, A and B, are placed in the same cluster if A is a nearest neighbor of B, B is a nearest neighbor of A, and A and B share at least *KMIN* nearest neighbors in common. (A further criterion may also be imposed at this point, i.e., that A and B are more similar than a threshold similarity value, *TSIM*.)

The parameters *KNN*, *KMIN*, and *TSIM* are supplied by the user and may be used to control (to some extent) the number and size of the clusters produced by the algorithm.[55]

Dissimilarity Ranking. The purpose of *dissimilarity* ranking is similar to that of clustering, i.e., it enables a rapid overview of the structural types generated by PRO_LIGAND. Such an overview is gained by ranking the set of structures so that the top *N* from the list is the most diverse *N* possible.[64] This approach has the advantage over clustering that one can specify exactly how many structures one wants to look at, whereas the Jarvis–Patrick algorithm does not permit the prior specification of the number of clusters to be produced.

The way that dissimilarity ranking works is as follows:[65]

1. Select a starting structure—this may be user-defined or selected at random by the program. (In practice, the choice of the starting structure does not have a great effect on the ranking obtained after the first few structures.[65])

2. Place this structure in the first available position in the ranking.

3. Choose the next structure from those structures remaining so that it is the most dissimilar to the structures already selected.

4. Go to 2 until the required number of structures has been chosen.

The key dissimilarity calculation in step 3 is accomplished using the similarity measures mentioned earlier. For any given pair of structures, the dissimilarity value is given simply by 1.0−*TC*, where *TC* is the relevant Tanimoto coefficient.

Similarity Ranking. If, by browsing through the generated answer set using clustering or dissimilarity ranking, the user happens upon a structure of potential interest, it would then be desirable to pick out other structures which are similar to it. This may be accomplished by *similarity ranking*.

Similarity ranking is simply effected by calculating the similarity of every molecule in the answer set with the specified molecule of interest and then ranking the answer set in order of decreasing similarity. The user may then specify the top *N* structures depending on how many he/she wishes to examine.

TOOLS FOR "IN SITU" CLUSTERING AND RANKING

The facilities mentioned in the previous section may be thought of as analyzing the structures *ex situ*, i.e., without any reference to the design model upon which they were built, but simply according to intrinsic structural characteristics. However, it is also of great interest to examine the molecules *in situ*, i.e., in the context of the design model whose constraints they are intended to satisfy. What is required here is some similarity measure which reflects how alike two molecules are in terms of their position on the design model. How such a measure is calculated depends

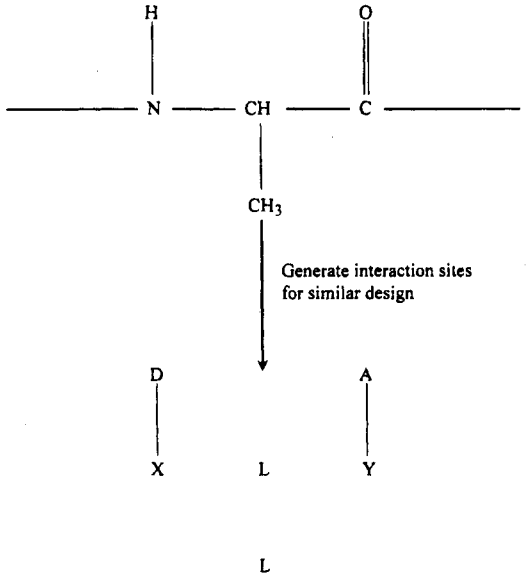

Figure 2. Construction of a design model for similar design. Reprinted (with adaptations) with permission: Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. *J. Comput.-Aided Mol. Design* **1995**, *9*, 13. Copyright 1995 ESCOM Science Publishers B.V.

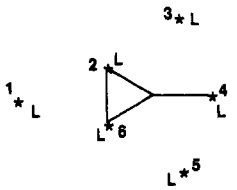

Figure 3. A schematic design model for similar design showing sites hit by a designed structure. All the design model sites are lipophilic, denoted by **L**.

on whether the design model is constructed for *similar design* to the original target, e.g., designing mimics to a known inhibitor, or for *complementary design*, e.g., building putative ligands within an active site. We shall examine these two cases in turn.

The Case of Similar Design. In the case of similar design, each atom in the target structure gives rise to a single interaction site in the design model of an appropriate generic type at the same point in space (see Figure 2). The structures generated by PRO_LIGAND will "hit" a certain number of these interaction sites as they are placed by the graph-theoretical construction algorithm.

If there are *NDM* design model interaction sites, it is then possible to set up a bitstring description of each designed molecule in terms of the sites hit by it. Thus, the setting of bit *I* to 1, indicates that the molecule has hit interaction site *I*. Figure 3 illustrates this schematically using a simple design model consisting of six lipophilic sites. The molecule shown has hit design model sites 2, 4, and 6, and so, in the 6-bit string, bits 2, 4, and 6 would be set to 1. Given such a bitstring representation of all the molecules in the answer set, it is easy to use the Tanimoto coefficient to calculate a similarity value for any pair and thus to cluster the molecules or rank them by (dis)similarity as described in the previous section.

In this way, it is possible to examine the generated structures to see how they span the design model and to examine how the same region of the design model may be satisfied by different functionalities.

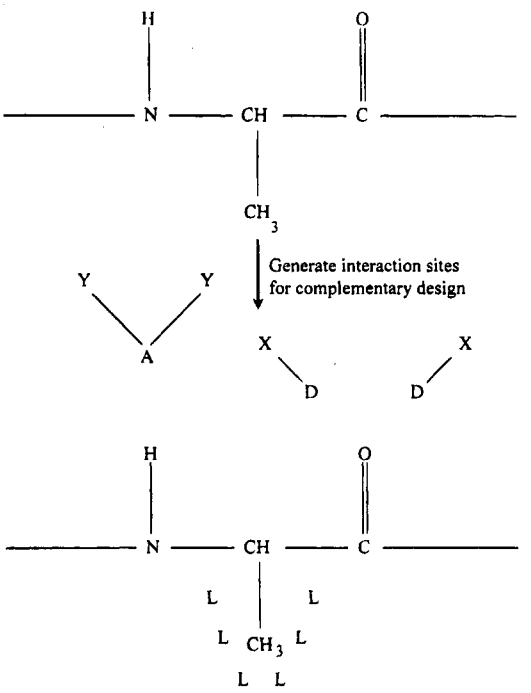

Figure 4. Construction of a design model for complementary design. Reprinted (with adaptations) with permission: Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. *J. Comput.-Aided Mol. Design* **1995**, *9*, 13. Copyright 1995 ESCOM Science Publishers B.V.

The Case of Complementary Design. A design model for complementary design generally consists of groups of interaction sites each of which arises from a specific feature in the design base of interest. Thus, a carbonyl group in the design base will give rise to a set of **D−X** sites on the design model representing the permitted distances and angles of interaction with it by a designed ligand (see Figure 4). In this instance, it is more interesting to employ a similarity measure based not on the design model sites themselves but on their "parent atoms". By parent atoms it is meant the atom in the design base which gives rise to the design model sites in question. Thus, in this example, the **D−X** sites in the design model have the carbonyl oxygen atom as their parent atom.

The parent atom information for each design model interaction site is carried in the design model file, and so it is simple to work out the number of parent atoms involved in the production of the design model and then set up a bitstring for each designed molecule indicating the design base atoms with which it interacts. A simple example of this is given in Figure 5. In this figure, there are a number of design model sites indicated by vectors (hydrogen bond acceptor/donor sites) and stars (lipophilic sites) resulting from three parent atoms in the design base. These sites represent acceptable positions for functional groups to be placed in order to form hydrogen bonds or lipophilic contacts with the design base (active site) atoms. The designed molecule can be seen to have hit design model sites from the parent atoms 1 and 3. Thus, in the 3-bit string, bits 1 and 3 would be set to 1. Once again, such a bitstring representation permits clustering and ranking operations to be carried out in a similar manner to the similar design case.

An added feature in the complementary design case is what we have termed "targeted" ranking. In this option, the user specifies the design base atoms (i.e., the parent atoms) which

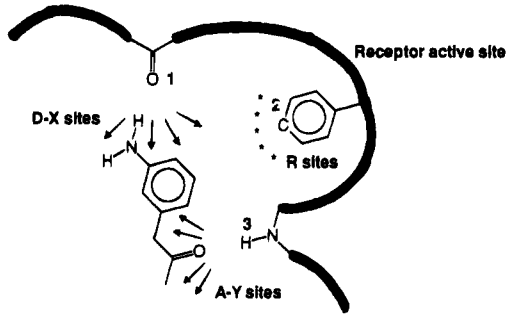

Figure 5. A schematic design model for complementary design showing sites hit by a designed structure and their design base parent atoms.

are of particular interest. A bit string is then set up with these bits set to 1. This, in effect, represents a pseudo-molecule interacting with these atoms. This bitstring may then be compared to those for all the molecules in the answer set and the set can be ranked by similarity to this pseudo-molecule. Any molecule with a similarity value greater than zero will be interacting with at least one of the specified design base atoms. Using this facility, the user can find all the molecules in the answer set which interact with a particular set of features in the receptor active site.

GROUPING AND RANKING BY MOLECULAR PROPERTY VALUES

A third set of analysis options involves the grouping or ranking of the answer set according to a range of molecular property values. The list of currently available properties is as follows:

- the number of atoms in a structure
- the molecular weight of a structure
- the number of rings in a structure as given by the *nullity* of the molecular graph
- the number of asymmetric carbons in a structure as an approximate indication of synthetic accessibility
- the molecular flexibility index (GS) of Fisanick *et al.*[66]
- various graph-theoretical indices such as the Wiener number,[67,68] the Randić index, χ (a measure of molecular branching),[69] and Kier's $^2\kappa$ shape index for a structure[70] which may be viewed as a rough approximation to the ratio of the long dimension of the molecule to the shorter lateral radial dimension
- PRO_LIGAND score,[37] S, given by

$$S = \sum_{1}^{N_A} W_A + \sum_{1}^{N_D} W_D + \sum_{1}^{N_{al}} W_{al} + \sum_{1}^{N_{ar}} W_{ar} + \sum_{1}^{N_{rot}} W_{rot} + \sum_{1}^{N_{ring}} W_{ring} + \sum_{1}^{N_{asym}} W_{asym}$$

where S is the structure's score, N_A and N_D are the numbers of hydrogen-bond acceptor and donor sites hit by the structure, N_{al} and N_{ar} are the numbers of lipophilic aliphatic and lipophilic aromatic interaction sites hit by the structure and N_{rot}, N_{ring} and N_{asym} are the number of rotatable bonds, the number of rings, and the number of asymmetric carbon atoms in the structure respectively. W_A is the contribution to the score for each hit hydrogen-bond acceptor site hit and the other weights refer to their respective features.

All the weights above may be specified by the user so that those structures which best meet the user's requirements, both in terms of the design model constraints and intrinsic structural features, will be assigned the highest scores. This feature permits rescoring of structures using score function weights different to those employed during Structure Generation or Structure Refinement.

- a calculation of the log P of a structure using the parameters of Viswanadhan *et al.*[71]
- the molecular volume, van der Waals' molecular surface area and solvent-accessible surface area of a structure[72]

The set of structures in the answer set may be ranked in ascending or descending order according to the value of any one of the above properties. Alternatively, the user may specify a range of values and a number of groups into which the set should be partitioned and the structures will be grouped appropriately. Successive applications of this type of strategy with various property types can quickly reduce the list of structures considerably.

SUBSTRUCTURE SEARCHING

As was suggested earlier, *de novo* design programs can generate a great variety of structures and it is inevitable that, unless sophisticated rules are employed, some of the generated solutions will contain substructures that are associated with toxicity or chemical instability.

PRO_LIGAND offers the user two options for dealing with this problem. Substructure searching for a specified set of substructures can be carried out during Structure Generation and any evolving molecules which are found to contain one or more of the substructures can be aborted. In this manner, all the molecules in the answer set can be guaranteed to be free from the specified substructures. Alternatively, substructure searching can be used in the Analysis module as a means of grouping the generated structures according to the presence or absence of a given substructure. This option allows the user to search through the answer set and remove any structures which contain substructures deemed undesirable, perhaps on grounds of stability or toxicity. Both of these options may be useful, since it is not always possible to say that a given substructure is undesirable in all circumstances. Substructure searching during Structure Generation can be used to avoid the inclusion of substructures which are *definitely* undesired in the context in question or because they are always unstable or toxic. The substructure searching in the Analysis module may be used to search for the substructures which are more borderline cases or indeed to look for substructures which are particularly desirable.

The substructures are specified by the user in a file using a simple SMILES-like notation[73] (see Figure 6), and the substructure searching is effected using a two-level search. Firstly, a rapid formula check examines each molecule to see if it contains the requisite number and types of atoms to match the substructure in question. Those structures that are potential matches are then passed on to the more computationally demanding subgraph isomorphism algorithm of Ullmann[74] which establishes the presence or absence of an exact match. At present, only 2-D substructure searching is supported, but the extension to 3-D searching is straightforward to implement if desired.

If the user specifies *NSUBS* substructures, the molecules will be grouped into (*NSUBS* + 1) groups; one group for

```
COMMENT Gem di-substituted species

SUBSTRUCTURE C(OH)N(H)H
SUBSTRUCTURE C(OH)OH
SUBSTRUCTURE C(N(H)H)N(H)H
SUBSTRUCTURE C(C(=O)OH)C(=O)OH
SUBSTRUCTURE C(C(=O)OH)C#N

COMMENT Aldehyde

SUBSTRUCTURE C(=O)H

COMMENT Alpha keto-acid and ester

SUBSTRUCTURE C(=O)C(=O)OH
SUBSTRUCTURE C(=O)C(=O)OC

COMMENT Natural electrophiles
COMMENT See Ashby, J. and Tennant, R.W., Mutation Research, 257, 229-306,
1991

SUBSTRUCTURE C(H)(H)Cl
SUBSTRUCTURE C(H)(H)Br
SUBSTRUCTURE C(H)(H)I
SUBSTRUCTURE NC(H)(H)C(H)(H)Cl
SUBSTRUCTURE OC(H)(H)C(H)(H)Cl
SUBSTRUCTURE C1OC1
SUBSTRUCTURE C1(=O)OCC1
SUBSTRUCTURE N=C=S
SUBSTRUCTURE N=C=N
SUBSTRUCTURE C(=O)C(H)(H)Cl
```

Figure 6. An example substructures file.

each of the substructures and a group containing those molecules which contain *none* of the specified substructures.

UTILITY TOOLS

Finally, the Analysis module contains some utility tools. One of the most important of these is the ability to rename molecules according to group, ranking, or cluster. This facility, coupled with a flexible file reader for input which reads all files matching a specified *regular expression*, allows the answer set to be manipulated and pruned in successive steps. For instance, the set may first be ranked according to the number of atoms and the top 100 renamed. This renamed set may then be read in on the next run, and some other operation performed on the subset of structures. This process may be repeated iteratively until a manageable number of structures is attained.

It is also possible to write out the groups/rankings/clusters as command files which can be read in by our in-house molecular graphics package to facilitate viewing of the results of any particular option.

RESULTS

The use of some of the tools described above will now be illustrated with reference to two examples.

Similar Design to Distamycin. Distamycin is a naturally occurring antibiotic which binds to the minor groove of DNA. In a recent paper,[39] we have described the use of PRO_LIGAND's structure generation and structure refinement modules to generate novel designs to replace the *N*-methylpyrrolecarboxamide ring system of distamycin (Figure 7). Here, we use the same example system but to illustrate the use of the Analysis module.

The design model for the example is shown in Figure 8. PRO_LIGAND's structure generation module was used to generate 974 structures to fit this design model. This procedure took 5976 CPU s (HP 735 workstation). Of these 974 structures, only 248 were unique, a consequence of the constrained nature of the design model and the ranking of the fragment libraries employed. All 974 structures (i.e., the duplicates were included) were then submitted to the

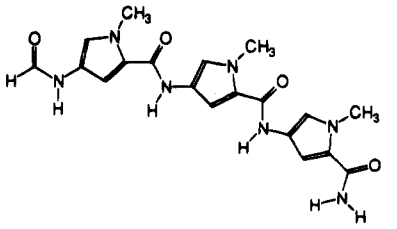

Figure 7. The *N*-methylpyrrolecarboxamide ring system of distamycin. Reprinted with permission: Westhead, D. R.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B. *J. Comput.-Aided Mol. Design* **1995**, *9*, 139. Copyright 1995 ESCOM Science Publishers B.V.

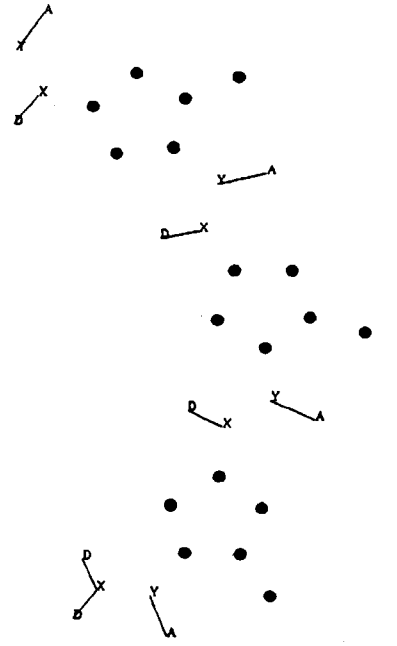

Figure 8. PRO LIGAND design model for similar design to the *N*-methylpyrrolecarboxamide ring system. Black circles indicate lipophilic interaction sites. Hydrogen bond donor and acceptor sites are indicated by **D**−**X** and **A**−**Y** vectors, respectively.

Table 1. Fitness Statistics for Structure Refinement Run

	initial population	final population
min. fitness	0.90	8.30
av fitness	2.82	8.65
max. fitness	6.75	9.50

structure refinement module which, after initial filtering, formed a population of 957 structures with which to work. The steady-state mode of the genetic algorithm as described in ref 39 was employed and run for 500 000 genetic operations. Two new program options were also invoked: an atom-type mutation procedure which mutates element types in the designed structures according to a user-defined rulebase and a Boltzmann-type acceptance criteria which permits the acceptance of child structures which are less fit than the present least fit structure. This latter feature has been found to help prevent premature convergence, particularly in the steady-state mode. The 500 000 genetic operations were completed in 7687 CPUs (HP 735), and the population statistics were improved as shown in Table 1. The final population of 957 structures was also free from duplicate compounds which is an indication of the GA's ability to create diversity or eliminate redundancy.

The Analysis module was then used to select 10 structures from this final population of 957. The steps involved in this

```
COMMENT Undesirable substructures for distamycin mimics

SUBSTRUCTURE N(H)C(H)(H)OC
SUBSTRUCTURE O(H)CN(H)
SUBSTRUCTURE OCO(H)
SUBSTRUCTURE NC(=O)O
SUBSTRUCTURE COC
SUBSTRUCTURE C(=O)C=N
```

Figure 9. Undesirable substructures. Structures containing these are eliminated by substructure search.

process are detailed below. All CPU times refer to an R3000 SGI Indigo machine and include the time required to read, rename, and write structure files.

1. Using the PRO_LIGAND score as a property type, a group was formed of all those structures which hit the four hydrogen-bond donating (**D−X**) interaction sites on the design model. This was accomplished by setting the scoring function weight for hitting a **D−X** site to 10.0 and all other score function weights to zero. Any structure with a score of 40.0 or more was designated a member of the group. This group was formed in 141 CPUs and consisted of 434 members.

2. The group formed above was then further subdivided using the molecular flexibility index. The *N*-methylpyrrole-carboxamide system hss a flexibility index of 4.96, and so all structures with a flexibility greater than 5.00 were discarded. This step took 81 CPUs and left a group of 108 structures for further analysis.

3. Structures containing any of a set of six undesirable substructures (Figure 9) were then eliminated using the substructure searching option. This is particularly useful in this instance as the atom-type mutation procedures can generate unstable moieties (e.g., hemiacetals). Forty-nine of the 108 structures were free from any of the specified substructures. The search took 26 CPU s.

4. Finally, a set of 10 structures were selected using the dissimilarity ranking option. The atom-centered descriptor set was used, and the procedure took 7 CPU s.

The set of 10 structures (**1−10**) selected is shown in Figure 10. As can be seen, they exhibit interesting diversity within the constraints specified by the user concerning molecular flexibility and functionality. The total CPU time required to select this set from the 957 members of the final population was only 255 s, which is obviously considerably more efficient than any manual procedure.

Complementary Design to HIV-1 Protease. In the search for therapeutics and vaccines to combat acquired immunodeficiency syndrome (AIDS), much attention has focussed upon its causative agent—human immunodeficiency virus (HIV). In the life-cycle of HIV, the processing of *gag* and *gag-pol* polyproteins by the enzyme HIV protease has been shown to be essential for viral replication. Thus, it is generally believed that if the activity of the protease can be inhibited, the spread of viral infection can be attenuated.[75,76] The protease has thus become a popular target for rational drug design efforts, and a number of novel inhibitors have been designed using structure-based approaches.[77−80] The application of the Analysis module to a set of ligands designed as potential HIV-1 protease inhibitors thus represents a test case of some timeliness and realism.

The crystal structure used in this example was that of HIV-1 protease complexed with the inhibitor acetyl pepstatin (PDB entry 5HVP).[82] As described in our earlier paper,[37] the active site was defined using the positions of selected

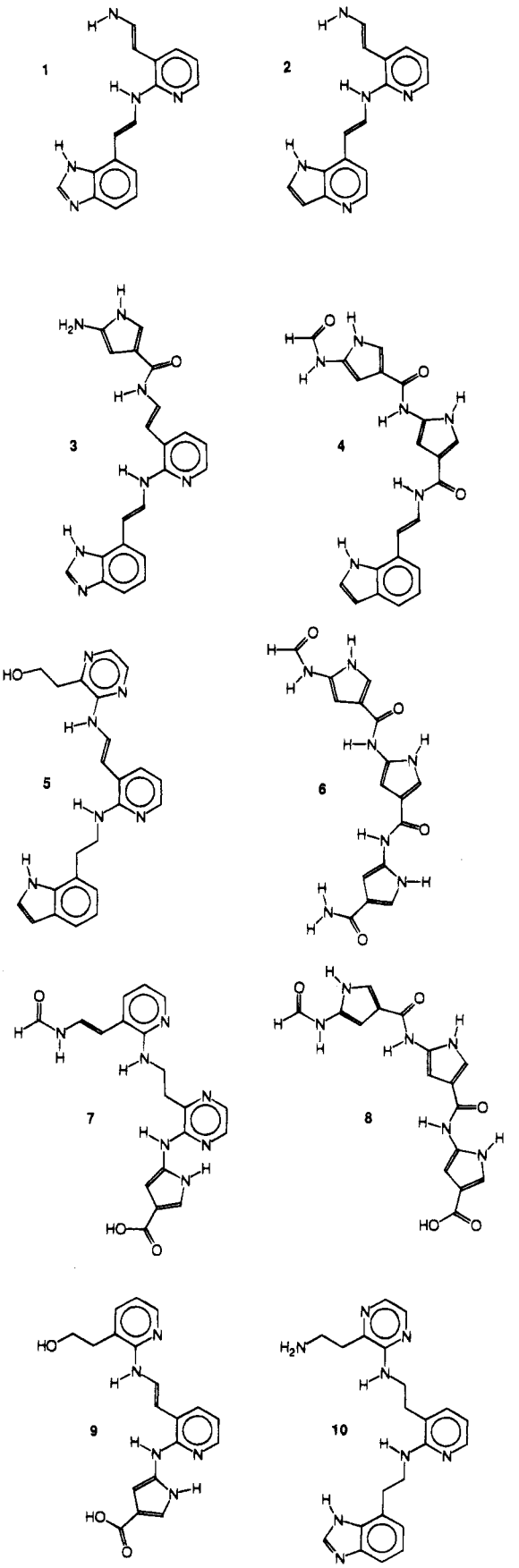

Figure 10. Ten structures designed as mimics of the *N*-methylpyrrolecarboxamide ring system.

inhibitor atoms as centers for spheres, in this case of 7.5 Å radius. With the addition of a water molecule necessary for mediation in the contact between Val3 and Sta4 in the

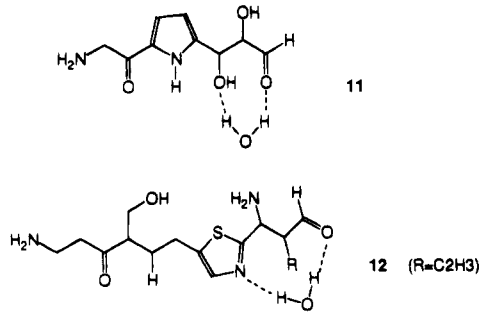

Figure 11. Seed structures showing the trapping of the active site water molecule.

inhibitor and Ile50 and Ile250 in the protease,[82] a total of 491 atoms were present in the final design base. After the operation of the Design-model Generation module, a design model consisting of 855 interaction sites was produced.

A number of Structure Generation runs were carried out to build structures based upon this design model. From a total of 1669 structures, the Analysis module was used to select 10 for detailed viewing and further consideration. The steps used to separate out these structures are explained in what follows. As before, all CPU times refer to an R3000 SGI Indigo machine and include the time required to read, rename, and write structure files.

1. As a first step, the PRO_LIGAND score function was used to form a group of structures which make six or more hydrogen bonds with the active site of HIV-1 protease. This procedure reduced the number of structures from 1669 to 319 in 244 CPU s.

2. The group of 319 were then examined to find those structures which "pincer" the active site water molecule. This was accomplished by specifying the water molecule atoms as targets in the targeted ranking mode. A similarity score of 2.0 for any structure indicates that it is forming two hydrogen bonds to the water molecule. An initial ranking showed that 86 structures satisfied this criterion, and these were written out as a separate group in a subsequent run. The total CPU time for these actions was 112 s.

3. An initial browse through a few structures at this stage revealed the presence of two undesirable substructures (acyclic C=N and O−C−N). Thus substructure search was carried out to eliminate any structures containing these substructures. This reduced the total from 86 to 43 in 20 s.

4. Finally, since the large, hydrophobic nature of the HIV-1 protease active site tends to produce less desirable long flexible aliphatic groups, the 10 most rigid structures from this group of 43 were selected by ranking the set in inverse order of molecular flexibility. This operation required 13 s.

While none of the final 10 structures could be considered as final candidate designs, at least two contain moieties that would make useful "seeds" for further structure generation experiments. These are shown as **11** and **12** in Figure 11. Of particular interest, is the recurrence of the thiazole ring which featured in a potential inhibitor (Figure 12) whose construction was detailed in an earlier paper.[37] What is of importance here, however, is the speed with which the Analysis module allows the user to focus upon useful (partial) structures thus expediting the iterative design process. The total CPU time required for this analysis was 389 s on a modest workstation, and the actual elapsed time was about half-an-hour.

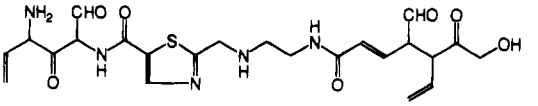

Figure 12. A potential inhibitor for HIV-1 protease designed by PRO LIGAND. Reprinted with permission: Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. *J. Comput.-Aided Mol. Design* **1995**, *9*, 13. Copyright 1995 ESCOM Science Publishers B.V.

DISCUSSION

While other workers in the field of *de novo* design have mentioned tools for the analysis of structures, we believe this paper is the first to present in detail a set of integrated software functions which aid the user in sorting through a large answer set. The Analysis module of PRO_LIGAND is able to accept structures either directly from the Structure Generation module or after genetic algorithm-based Structure Refinement. Ranking or clustering operations may be carried out based on the molecular properties of the generated structures alone or with reference to the environment in which they were created. In addition, both similar and complementary design situations are catered for. Finally, the substructure searching tools permit the partitioning of structure sets according to the presence or absence of user-specified 2D substructures.

The examples given show how rapidly a set of solutions can be narrowed down using a series of filters based on various molecular properties. In both cases, a large set of structures was narrowed down to just 10 in a matter of a few CPU minutes. The final set thus produced may then be taken forward for more detailed and time-consuming analysis using molecular mechanics/dynamics or used as seed structures for subsequent *de novo* design experiments. The use of a synthetic accessibility estimation program, such as CAESA,[52] would also be valuable at this stage.

The speed of the Analysis module and the variety of tools means that a given set of structures can be analyzed a number of times in the search for a satisfactory subset of solutions. Usually, a first pass at analysis will serve simply to give a feel for the type of structures generated, and this information can be used to direct subsequent analyses. For instance, an initial analysis might show the presence of undesired substructures which could be weeded out in a later run. Even if several iterations of analysis are required, the procedure is still quick enough to make it far more efficient than an entirely manual assessment.

Future work in developing the Analysis module will concentrate on the incorporation of better empirical scoring functions such as those developed by Böhm[18] and Bohacek and McMartin.[32] The development of a more sophisticated, graphical user interface would also enhance and facilitate the analysis process.

CONCLUSIONS

Many branches of drug discovery research are now turning from the problem of idea *generation* to that of idea *evaluation*. Technologies such as 3D database searching, *de novo* design programs, and combinatorial libraries are all capable of rapidly producing many thousands of molecules as potential solutions to a given design problem. The problem now to be addressed is how to prioritize the proffered solutions in an efficient manner. In this paper,

we have described the Analysis module of our in-house *de novo* design package, PRO_LIGAND, and have demonstrated how it may be used to facilitate the evaluation of sizeable sets of solutions. The fast analysis of such output should be of significant value in the process of rational drug design.

ACKNOWLEDGMENT

The authors gratefully acknowledge the provision by Dr. L. R. Dodd of the code for the calculation of molecular volume and surface area and programming support by Dr. M. A. Firth. We are also grateful to ESCOM Science Publishers B. V. for permission to reuse figures from earlier papers.

REFERENCES AND NOTES

(1) Whittle, P. J.; Blundell, T. L. Protein Structure-Based Drug Design. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 349−375.
(2) Greer, J.; Erickson, J. W.; Baldwin, J. J.; Varney, M. D. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design. *J. Med. Chem.* **1994**, *37*, 1035−1055.
(3) Verlinde, C. L. M. J.; Hol, W. G. J. Structure-Based Drug Design: Progress, Results and Challenges. *Structure* **1994**, *2*, 577−587.
(4) Guida, W. C. Software for Structure-Based Drug Design. *Curr. Opin. Struct. Biol.* **1994**, *4*, 777−781.
(5) Colman, P. M. Structure-Based Drug Design. *Curr. Opin. Struct. Biol.* **1994**, *4*, 868−874.
(6) Montgomery, J. A.; Niwas, S.; Rose, J. D.; Secrist III, J. A.; Babu, S.; Bugg, C. E.; Erion, M. D.; Guida, W. C.; Ealick, S. E. Structure-Based Design of Inhibitors of Purine Nucleoside Phosphorylase. 1. 9-(Arylmethyl) Derivatives of 9-Deazaguanine. *J. Med. Chem.* **1993**, *36*, 55−69.
(7) Webber, S. E.; Bleckman, E. M.; Attard, J.; Deal, J. G.; Kathardekar, V.; Welsh, K. M.; Webber, S.; Janson, C. A.; Matthews, D. A.; Smith, W. M.; Freer, S. T.; Jordan, S. R.; Bacquet, R. J.; Howlan, E. F.; Booth, C. L. J.; Ward, R. W.; Hermann, S. M.; White, J.; Morse, C. A.; Hilliard, J. A.; Bartlett, C. A. Design of Thymidylate Synthase Inhibitors Using Protein Crystal Structures: The Synthesis and Biological Evaluation of a Novel Class of 5-Substituted Quinazolinones. *J. Med. Chem.* **1993**, *36*, 733−746.
(8) von Itzstein, M.; Wu, W.-Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Phan, T. V.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R. Rational Design of Potent Sialidase-Based Inhibitors of Influenza Virus Replication. *Nature* **1993**, *263*, 418−423.
(9) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.-H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. E. Rational Design of Potent, Bioavailable, Nonpeptide Cyclic Ureas as HIV Protease Inhibitors. *Science* **1994**, *263*, 380−384.
(10) Moon, J. B.; Howe, W. J. Computer Design of Bioactive Molecules: A Method for Receptor-Based De Novo Ligand Design. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 314−328.
(11) Moon, J. B.; Howe, W. J. Recent Advances in De Novo Molecular Design. In *Trends in QSAR and Molecular Modelling 92*; Wermuth, C. G., Ed.; ESCOM: Leiden, 1993; pp 11−19.
(12) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 29−34.
(13) Caflisch, A.; Miranker, A.; Karplus, M. Multiple Copy Simultaneous Search and Construction of Ligands in Binding Sites: Application to Inhibitors of HIV-1 Aspartic Protease. *J. Med. Chem.* **1993**, *36*, 2142−2167.
(14) Nishibata Y.; Itai, A. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron* **1991**, *47*, 8985−8990.
(15) Nishibata, Y.; Itai, A. Confirmation of Usefulness of a Structure Construction Program Based on Three-Dimensional Receptor Structure for Rational Lead Generation. *J. Med. Chem.* **1993**, *36*, 2921−2928.
(16) Böhm, H.-J. The Computer Program LUDI: A New Method for the De Novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.
(17) Böhm, H.-J. LUDI: Rule-Based Automatic Design of New Substituents for Enzyme Inhibitor Leads. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593−606.
(18) Böhm, H.-J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(19) Böhm, H.-J. On the Use of LUDI to Search the Fine Chemicals Directory for Ligands of Proteins of Known Three-Dimensional Structure *J. Comput.-Aided Mol. Des.* **1994**, *8*, 623−632.
(20) Lewis, R. A.; Roe, D. C.; Huang, C.; Ferrin, T. E.; Langridge, R.; Kuntz, I. D. Automated Site-Directed Drug Design Using Molecular Lattices. *J. Mol. Graphics* **1992**, *10*, 66−78.
(21) Rotstein, S. H.; Murcko, M. A. GenStar: A Method for De Novo Drug Design. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 23−43.
(22) Rotstein, S. H.; Murcko, M. A. GroupBuild: A Fragment-Based Method for De Novo Drug Design. *J. Med. Chem.* **1993**, *36*, 1700−1710.
(23) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: A Program for Structure Generation. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127−153.
(24) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent Developments in the De Novo Design of Molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207−217.
(25) Mata, P.; Gillet, V. J.; Johnson, A. P.; Lampreia, J.; Myatt, G. J.; Sike, S.; Stebbings, A. L. SPROUT: 3D Structure Generation using Templates. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 479−493.
(26) Pearlman, D. A.; Murcko, M. A. CONCEPTS: New Dynamic Algorithm for De Novo Drug Design. *J. Comput. Chem.* **1993**, *14*, 1184−1193.
(27) Tschinke, V.; Cohen, N. C. The NEWLEAD Program: A New Method for the Design of Candidate Structures from Pharmacophoric Hypotheses. *J. Med. Chem.* **1993**, *36*, 3863−3870.
(28) Ho, C. W. M.; Marshall, G. R. SPLICE: A Program to Assemble Novel Partial Query Solutions from Three-Dimensional Database Searches into Novel Ligands. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 623−647.
(29) Leach, A. R.; Lewis, R. A. A Ring-Bracing Approach to Computer-Assisted Ligand Design. *J. Comput. Chem.* **1994**, *15*, 233−240.
(30) Leach, A. R.; Kilvington, S. R. Automated Molecular Design: A New Fragment-Joining Algorithm. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 283−298.
(31) Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E. HOOK: A Program for Finding Novel Molecular Architectures that Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site. *Proteins: Struct., Funct. Genet.* **1994**, *19*, 199−221.
(32) Bohacek, R. S.; McMartin, C. Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a De Novo Design Method Incorporating Combinatorial Growth. *J. Am. Chem. Soc.* **1994**, *116*, 5560−5571.
(33) Cohen, A. A.; Shatzmiller, S. E. Implementation of Artificial Intelligence for Automatic Drug Design. I. Stepwise Computation of the Interactive Drug-Design Sequence. *J. Comput. Chem.* **1994**, *15*, 1393−1402.
(34) Gehlhaar, D. K.; Moerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. De Novo Design of Enzyme Inhibitors by Monte Carlo Ligand Generation. *J. Med. Chem.* **1995**, *38*, 466−472.
(35) Glen, R. C.; Payne, A. W. R. A Genetic Algorithm for the Automated Generation of Molecules within Constraints. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 181−202.
(36) Lewis, R. A.; Leach, A. R. Current Methods for Site-Directed Structure Generation. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 467−476.
(37) Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. PRO_LIGAND: An Approach to De Novo Molecular Design. 1. Application to the Design of Organic Molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 13−32.
(38) Waszkowycz, B.; Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Westhead, D. R. PRO_LIGAND: An approach to De Novo Molecular Design. 2. Design of Novel Molecules from Molecular Field Analysis (MFA) Models and Pharmacophores. *J. Med. Chem.*, **1994**, *37*, 3994−4002.
(39) Westhead, D. R.; Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B. PRO_LIGAND: An approach to De Novo Molecular Design. 3. A Genetic Algorithm for Structure Refinement. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 139−148.
(40) Frenkel, D.; Clark, D. E.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. PRO_LIGAND: An approach to De Novo Molecular Design. 4. Application to the Design of Peptides. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 213−225.
(41) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.
(42) Clark, D. E.; Jones, G.; Willett, P.; Kenny, P. W.; Glen, R. C. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197−206.

(43) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.
(44) Moock, T. E.; Henry, D. R.; Ozkabak, A. G.; Alamgir, M. Conformational Searching in ISIS/3D Databases. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 184−189.
(45) Bures, M. G.; Martin, Y. C.; Willett, P. Searching Techniques for Databases of Three-Dimensional Chemical Structures. *Topics in Stereochemistry* **1994**, *21*, 467−511.
(46) Pearlman, R. S. 3D Molecular Structures: Generation and Use in 3D Searching. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 41−79.
(47) Bures, M. G.; Danaher, E.; DeLazzar, J.; Martin, Y. C. New Molecular Modeling Tools Using Three-Dimensional Substructures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 218−223.
(48) Martin, Y. C.; van Drie, J. H. Identifying Unique Core Molecules from the Output of a 3-D Database Search. In *Chemical Structures 2: The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1993; pp 315−326.
(49) *ISIS/3D*. MDL Information Systems, Inc., 2132 Farallon Drive, San Leandro, CA 94577, U.S.A.
(50) *SYBYL*. TRIPOS Associates, Inc., 1699 South Hanley, Suite 303, St. Louis, MO 63144, U.S.A.
(51) *LeapFrog*. TRIPOS Associates, Inc., 1699 South Hanley, Suite 303, St. Louis, MO 63144, U.S.A.
(52) Johnson, A. P. New Developments in the SPROUT Program for De Novo Design and the CAESA System for Estimation of Synthetic Accessibility. Presented at the 13th Annual Conference of the Molecular Graphics Society, Evanston, IL, U.S.A., July 1994.
(53) Klebe, G. The Use of Composite Crystal-field Environments in Molecular Recognition and the *de Novo* Design of Protein Ligands. *J. Mol. Biol.* **1994**, *237*, 212−235.
(54) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.
(55) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.
(56) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.
(57) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(58) *MACCS Drug Data Report*. MDL Information Systems, Inc., 2132 Farallon Drive, San Leandro, CA 94577, U.S.A.
(59) Willett, P. *Similarity and Clustering in Chemical Information*; Research Studies Press: Letchworth, 1987; p 218.
(60) Johnson, P.; Maliski, E. Processing the Ore of Database Mining: Evaluate and Edit Large Hitlists through Cluster Browsing. *Chem. Des. Auto. News* **1994**, *9*, *(Nos. 11/12)*, 1−27.
(61) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbours. *IEEE Trans. Computing* **1973**, *C-22*, 1025−1034.
(62) Willett, P. *Similarity and Clustering in Chemical Information*; Research Studies Press: Letchworth, 1987.
(63) Willett, P. Algorithms for the Calculation of Similarity in Chemical Structure Databases. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley-Interscience: New York, 1990; pp 43−64.
(64) Bawden, D. Applications of Two-Dimensional Chemical Similarity Measures to Database Analysis and Querying. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley-Interscience: New York, 1990; pp 65−76.
(65) Bawden, D. Molecular Dissimilarity in Chemical Information Systems. In *Chemical Structures 2: The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Heidelberg, 1993; pp 383−388.
(66) Fisanick, W.; Cross, K. P.; Rusinko III, A. Characteristics of Computer Generated 3D and Related Molecular Property Data for CAS Registry Substances. *Tetrahedron Comput. Methodol.* **1990**, *3*, 635−652.
(67) Balaban, A. T. Applications of Graph Theory in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334−343.
(68) Trinajstić, N.; Nikolić, S.; Knop, J. V.; Müller, W. R.; Szymanski, K. *Computational Chemical Graph Theory: Characterization, Enumeration and Generation of Chemical Structures by Computer Methods*; Ellis Horwood: New York, 1991; pp 252−267.
(69) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(70) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.
(71) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(72) Dodd, L. R.; Theodorou, D. N. Analytical treatment of the volume and surface area of molecules formed by an arbitrary collection of unequal spheres intersected by planes. *Mol. Phys.* **1991**, *72*, 1313−1345.
(73) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(74) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. ACM* **1976**, *23*, 31−42.
(75) Kohl, N. E.; Emini, E. A.; Schlief, W. A.; David, L. J.; Heimbach, J. C.; Dixon, R. A. F.; Scolnick, E. M.; Sigal, I. S. Active Human Immunodeficiency Virus Protease is Required for Viral Infection. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4686−4690.
(76) McQuade, T. J.; Tomaselli, A. G.; Liu, L.; Karacostas, V.; Moss, B.; Sawyer, T. K.; Heinrikson, R. L.; Tarpley, W. G. A Synthetic HIV-1 Protease Inhibitor with Antiviral Activity Arrests HIV-like Particle Maturation. *Science* **1990**, *247*, 454−456.
(77) Appelt, K. Crystal Structures of HIV-1 Protease-Inhibitor Complexes. *Perspectives in Drug Discovery and Design* **1993**, *1*, 23−48.
(78) Fitzgerald, P. M. D. HIV Protease-Ligand Complexes. *Curr. Opin. Struct. Biol.* **1993**, *3*, 868−874.
(79) Redshaw, S. Inhibition of HIV Proteinase. *Exp. Opin. Invest. Drugs* **1994**, *3*, 273−286.
(80) Erickson, J.; Kempf, D. Structure-based Design of Symmetric Inhibitors of HIV Protease. *Arch. Virol.* **1994**, *9*, 12−29.
(81) West, M. L.; Fairlie, D. P. Targeting HIV-1 Protease: A Test of Drug-Design Methodologies. *TIBTECH* **1995**, *13*, 67−75.
(82) Fitzgerald, P. M. D.; McKeever, B. M.; VanMiddlesworth, J. F.; Springer, J. P.; Heimbach, J. C.; Leu, C.-T.; Herber, W. K.; Dixon, R. A. F.; Darke, P. L. Crystallographic Analysis of a Complex between Human Immunodeficiency Virus Type 1 Protease and Acetyl-Pepstatin at 2.0 Å Resolution. *J. Biol. Chem.* **1990**, *265*, 14209−14219.

CI950203K