Journal of Computer-Aided Molecular Design, 11 (1997) 193–207
ESCOM

J-CAMD 386

PRO_SELECT: Combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology

Christopher W. Murray*, David E. Clark**, Timothy R. Auton, Michael A. Firth, Jin Li, Richard A. Sykes, Bohdan Waszkowycz, David R. Westhead*** and Stephen C. Young

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

Received 27 September 1996
Accepted 8 November 1996

Keywords: Structure-based drug design; De novo molecular design; Combinatorial chemistry; Synthetic constraint; Thrombin inhibitors

Summary

This paper describes a novel methodology, PRO_SELECT, which combines elements of structure-based drug design and combinatorial chemistry to create a new paradigm for accelerated lead discovery. Starting with a synthetically accessible template positioned in the active site of the target of interest, PRO_SELECT employs database searching to generate lists of potential substituents for each substituent position on the template. These substituents are selected on the basis of their being able to couple to the template using known synthetic routes and their possession of the correct functionality to interact with specified residues in the active site. The lists of potential substituents are then screened computationally against the active site using rapid algorithms. An empirical scoring function, correlated to binding free energy, is used to rank the substituents at each position. The highest scoring substituents at each position can then be examined using a variety of techniques and a final selection is made. Combinatorial enumeration of the final lists generates a library of synthetically accessible molecules, which may then be prioritised for synthesis and assay. The results obtained using PRO_SELECT to design thrombin inhibitors are briefly discussed.

Introduction

In recent years, the process of drug discovery and design has been profoundly affected by the emergence of new methods and technologies. Historically, lead discovery was pursued by means of the random screening of selected compounds against the biological assay of interest. While such an approach is somewhat serendipitous, it provided a starting point for the development of many of the drugs in use today.

Over the last decade, the amount of information concerning the 3D structures of biomolecular targets has increased dramatically with improvements in the experimental techniques of X-ray crystallography and NMR [1]. The availability of such information has encouraged the emergence of structure-based drug design (SBDD), where molecules are designed using the techniques of computer-aided molecular design (CAMD) specifically to fit a binding site of known structure [2–6]. Several lead compounds designed in this manner are now in clinical trials [7–10]. A basic difficulty in most applications of CAMD is that designed molecules are often of uncertain synthetic accessibility. Those that can be synthesised are often challenging, leading to a slow data feedback between experiment and design.

More recently, a different drug discovery paradigm has attracted a great deal of attention. Combinatorial chemistry (CC) has advanced to a stage where large libraries of compounds can be synthesised and tested with moderate cost and effort [11–17]. In particular, very large libraries of peptides and peptoids have been synthesised, leading to the identification of active compounds [18–22]. There are two known difficulties with such libraries: (i) the properties of the peptide-like molecules that make up the libraries; and (ii) the practical difficulties of working with large libraries. Peptide-like molecules are undesirable as

*To whom correspondence should be addressed.
**Present address: Dagenham Research Centre, Rhône-Poulenc Rorer Ltd., Rainham Road South, Dagenham, Essex RM10 7XS, U.K.
***Present address: EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

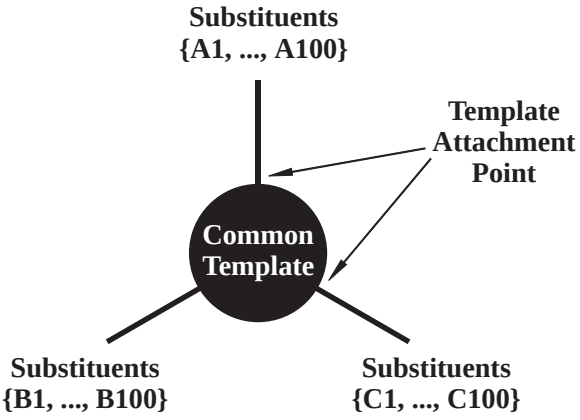


Fig. 1. Illustration of combinatorial elaboration of a template.

drug candidates because they often display poor pharmacokinetic behaviour [23,24]. There is therefore considerable interest in extending combinatorial libraries to a wider range of peptoid systems and especially to small organic molecules. Large libraries are a problem because they often necessitate the synthesis and testing of mixtures of compounds [16]. The use of mixtures is known to lead to false negatives and false positives in some instances [13].

The development of organic array-based (multiple parallel) libraries can potentially circumvent these problems. Small organic molecules are synthesised individually using available chemical reagents and a predefined synthetic route. However, if there are three steps in a synthesis, each involving the introduction of a new chemical chosen from a list of 100 different available reagents, then the number of molecules in the library is 100^3 (see Fig. 1). It would therefore be very difficult to make the entire library without resorting to using mixtures [25]. However, large libraries are necessary if the screening is random because a wide diversity is required to ensure a reasonable hit rate. One approach to avoid the need for a large library is to use diversity analysis or structure–activity information to reduce its size [26–37]. For example, reactions with fluorinated derivatives of phenylalanine would probably give a similar pattern of activity to phenylalanine, and so only one representative need be included in the library. However, it still seems likely that large libraries will be required if active compounds are to be identified in the random screening of a biological target against the library. Another important restriction on large organic libraries is the desire for a robust and generalised synthetic route which in practice is difficult to achieve, whilst maintaining chemical diversity [38,39].

This paper outlines a computational method which uses the structural constraints of the biological target to restrict rationally the size of the combinatorial library. It can be thought of as a method which combines combinatorial chemistry with structure-based drug design. We shall refer to this method as PRO_SELECT (SELECT = Systematic Elaboration of Libraries Enhanced by Computational Techniques). The resulting molecular designs are then synthesised and tested individually without the need for one general and robust synthetic protocol. The automatic generation and assessment of molecular designs is now commonly used in SBDD, where it is referred to as de novo design. A large number of different groups have published de novo design methodologies [40–54] and the subject has been reviewed [55,56]. The extension here is that generated designs are guaranteed to be members of a large virtual combinatorial library, and each member is consistent with a predefined synthetic route derived from available starting reagents. This means that the majority of designs should be relatively easy to synthesise. The method can be described as synthetically constrained de novo design. It differs from peptide de novo design [41, 57] in that more general synthesis routes and chemical reagents are accessible. The method outlined here also differs considerably from the work of Sheridan and Kearsley [31], who have used a genetic algorithm to sort through a virtual combinatorial library according to similarity to known molecules.

The method has been successfully applied to the design of thrombin inhibitors and this will be discussed in detail elsewhere [58]. The main emphasis of this paper will be the detailed description of the technology and methods used in PRO_SELECT.

Overview of PRO_SELECT methodology

The PRO_SELECT procedure consists of the following steps:

(1) Construction of a virtual combinatorial library based around a template chemistry appropriate for the target molecule and amenable to combinatorial synthesis.

(2) Screening of members of the library based on their interaction with a target receptor.

(3) Synthesis and testing of representative elements of the library as single compounds using a variety of synthetic protocols.

If this process can be done efficiently and accurately, and the target molecule structure is known, the approach has the potential to address some of the difficulties associated with applying combinatorial chemistry or structure-based drug design. Firstly, PRO_SELECT offers all the advantages of an array-based combinatorial library (single compounds, wider variety of chemistries) whilst side-stepping the problem of small library arrays. This is because a very large virtual library is considered and screened *computationally*, leaving only a small number of compounds to be synthesised and tested.

Secondly, the need to have one synthetic protocol to cover a wide variety of chemistries is relaxed. The synthesis route can now be tailored to accommodate a larger range of chemistries than could be considered by an automated method. Solution and solid-phase methods can be

used with protection and deprotection steps as required. This means that a larger virtual library can be considered and thus the chance of locating novel active compounds is increased. PRO_SELECT can also consider simple functional group transformations within the starting materials to increase further the diversity of the virtual library.

Thirdly, by restricting the design process to molecules which are accessible by specified synthetic routes, one significantly reduces the problems often associated with rational drug design, i.e. uncertain synthetic feasibility and slow feedback between design and experiment. In fact, application of the PRO_SELECT procedure should yield a set of compounds based around a common template which can be rapidly synthesised and assayed for activity against a given target. Such a set of compounds might form an immediate QSAR (quantitative structure–activity relationship) training set, in contrast to other drug discovery paradigms where further work would be necessary to derive an equivalent QSAR set. The primary advantage over a traditional medicinal chemistry approach, which would obtain a lead by a screening process and synthesise a large number of analogues to provide an SAR set, is one of cost-effectiveness.

PRO_SELECT technology

This section is an introduction to a more detailed discussion of the technology used in PRO_SELECT. Figure 1 serves to define some of the terminology used in this paper. It gives a schematic representation of a virtual library of the type used in PRO_SELECT. Each member of the library consists of a common *template* with different *substituents* attached to it. The substituents are derived from *available chemical reagents* and it is the variation in these substituents at each *template attachment point* that causes the combinatorial explosion in the number of individual molecules in the library. The library has a synthesis route or strategy (sometimes referred to in this paper as *template chemistry*) associated with it, whereby individual members are synthesised from available chemical reagents. The template itself can be an available chemical or can be formed during the chemical reactions (e.g. in ring-forming reactions). The available reagents may also undergo general molecular transformations before they are attached to the template and become substituents.

Figure 2 gives an outline of the PRO_SELECT technology and the approximate order of operation. The technological aspects of the procedure have been divided into four steps. The design specification step determines the constraints which are to be applied and explored during the computational screening of the library. Obviously, these constraints include the actual specification of the library, the 3D structure of the receptor and any specific constraints derived from the receptor to judge the quality of library members. The second step involves selecting chemical reagents and screening the corresponding substituents which are used to form members of the library. The substituent screening is based on the structure of the receptor. The accepted substituents are further assessed and filtered using a variety of computer-aided techniques and chemical considerations. The third step involves enumeration of the library, i.e. production of the full library after all rejected substituents have been deleted. The final computational step of the procedure is to perform simple checks and calculations on the enumerated library and arrive at a ranking of the molecules in the library for synthesis and testing.

Each of the four steps given in Fig. 2 will now be outlined in detail in the following sections. This will be followed by a short discussion of the software architecture.

Design specification

The three aspects of design specification which will be discussed are template selection, template positioning and design criteria, which will be discussed separately despite being interrelated.

Design criteria

Specification of the design criteria involves a careful study of the target macromolecule. Thus, decisions need to be taken at this stage about which X-ray structure(s) of the receptor are to be used (if more than one is avail-

Step 1: Design specifications	- design criteria - template selection - template positioning
Step 2: Substituent selection	- database search for available starting reagents - receptor-based screening of substituents - assessment of accepted substituents
Step 3: Combinatorial enumeration of library	
Step 4: Prioritisation of enumerated molecules for synthesis and testing	

Fig. 2. Overview of PRO_SELECT methodology.

able), and whether some refinement by molecular dynamics/molecular mechanics needs to be carried out in order to generate a more accurate starting point for molecular design. Typically, more than one snapshot of the receptor structure will be used in successive PRO_SELECT experiments. Also, it is necessary at this stage to decide on the key functionalities in the active site with which the substituents are to interact. A 'design model' is then generated for each template attachment point using the Design Model Generation functionality of PRO_LIGAND [53]. A design model consists of a number of *interaction sites* which originate from specified receptor atoms and may be either vectors (denoting favourable positions and directions for hydrogen bond interactions with the active site) or points (denoting positions of favourable lipophilic contact with the active site) [44,59,60]. The vectors and points are labelled to indicate the particular chemistry they represent; thus **D**-**X** and **A**-**Y** vectors represent potential hydrogen bond donor and acceptor positions, respectively. Similarly, **L** and **R** sites represent aliphatic and aromatic lipophilic sites, respectively. The density, positions and orientations of the interaction sites are encoded in a rule-base which can be edited by the user and is based on a statistical examination of experimentally preferred intermolecular contacts [60]. The design model will also contain *link sites* specifying the position where attachment to each template attachment point must occur. The addition of the necessary link sites is made when the appropriately labelled template molecule is read into the program which screens the substituents.

Template selection

The purpose of the molecular template (or scaffold) is to hold in position the substituents which will make hydrogen bonds or lipophilic contacts with the binding site. An advantage of using structural information in the choice of the template chemistry is that a knowledge of the receptor can be used to increase the chances of the library containing active molecules. A number of important issues can be identified in the selection of the template chemistry.

(1) The synthetic chemistry associated with the template should be relatively accessible and capable of delivering a wide diversity of substituents at a number of attachment points. There is no reason why a central template with more than one attachment point has to be chosen. Sometimes it is most appropriate to start the design process from a terminal region, although such templates are not usually capable of delivering as wide a diversity of designs as templates with more attachment points.

(2) Ideally, the template itself should be capable of making a number of favourable contacts with the receptor. This aids in establishing the position of the template and increases the likelihood that the library will contain active molecules.

(3) In some cases it is possible to infer likely templates from known inhibitors or substrates. For example, (i) a known inhibitor of thrombin is PPACK and it contains a central proline moiety which could be used as a template; or (ii) a known substructure with strong binding (e.g. guanidinium in the S1 pocket of thrombin) can be prepositioned and used to search for potential carboxylic acid based templates in an initial run using PRO_SELECT.

(4) Templates can be designed de novo, using structure-based techniques. This could mean using a de novo design method, a receptor-based database screening strategy or even an interactive design method. In our applications, we have also searched reaction databases looking for suitable ring-forming reactions which give rise to β-sheet mimetics.

(5) It is desirable that the template has restricted conformational freedom so that only limited numbers of alternative positions for the template need be considered.

The process of template selection will thus necessarily involve a close collaboration between modellers and synthetic chemists, the former providing expertise about the requirements of the templates in terms of molecular interactions at the binding site and the latter giving guidance concerning the synthetic feasibility of any choices made. The result of the template selection process is a set of scaffolds chosen to achieve the best architecture in the active site and to minimise the synthetic effort required to prepare them. In practice, the decision about which templates to pursue will be a balance between the variety of factors discussed above.

Template positioning

Having chosen the template to be used in the active site under study, the next task is to position the template appropriately within the site. If more than one conformation for the receptor is used, the template will need to be positioned in each conformational snapshot. Additionally, more than one template position can be used for each snapshot. In principle, there may be a very large number of orientations of a given template in the site; however, this number can be reduced if the template is chosen to make a specific interaction with the binding site itself. The latter is certainly a preferred mode of operation. In fact, our experience with thrombin indicates that when the template is chosen to make good contact with the receptor, only a few positions need be considered to obtain similar lists of substituents: the character and diversity space spanned by the resulting list of substituents does not change even though substituents may have different scores and some substituents do not attach to all template positionings and all receptor snapshots. In a way the assumption of a stationary template is a key approximation in PRO_SELECT. The potential position of the template is represented by a very small number of rigid positionings

precalculated in advance. This allows PRO_SELECT to avoid the combinatorial explosion associated with considering all members of the virtual library independently, at the modest extra cost of considering a few snapshots and positionings. This point is discussed in more detail later.

Another important consideration in choosing template positionings is to place the template in such a way as to facilitate the molecular interactions that will be formed by the substituents once they are attached. Only templates capable of giving good 3D substituent diversity when placed in a favourable position in the receptor will be considered during the selection of the template chemistry.

The placement process could be achieved automatically by means of various objective docking protocols based on molecular mechanics energy calculations [61,62] or geometric positioning upon interaction sites [63]. Manual techniques or methods derived from receptor maps [64] are also useful in positioning molecular entities. Crystal structures of related ligand–receptor complexes should be used when practicable. The positioning of seed structures in active sites is the usual starting point in reported successful applications of de novo design (see, for example, Refs. 41, 45, 52, 65 and 66), and the methods applicable in such cases are directly applicable here. It should be realised that docking and molecular dynamics on an isolated template is to be avoided wherever possible. It would be an abnormal design situation in de novo design where no knowledge of favourable substituents or chemistries is available. Normally, one has a good idea of the sort of chemistries that could be attached to the template, and even where such knowledge is sketchy an application of synthetically constrained or unconstrained de novo design would yield putative designs. When positioning the template, it is these original putative designs which contain the template that are used in molecular dynamics and docking studies rather than the isolated template.

The result of the template positioning process is a position, or a number of positions, in 3D coordinate space for each of the receptor snapshots. The chosen orientations are saved for future reference.

Substituent selection

Within PRO_SELECT, the process of substituent selection involves a number of steps:

(1) Searching the Available Chemicals Directory (ACD) [67] (or other sources of available reagents) to find potential substituents for a given template.

(2) Computationally screening these potential substituents using techniques adapted from our in-house de novo design program, PRO_LIGAND [53,57,68–71].

(3) Assessing and deciding on the preferred substituents at each position.

Each of these steps is explained in more detail below. It is important to realise that substituents attached to different attachment points are tested independently of each other at this stage. This makes the process of performing detailed 3D checks on a large virtual library computationally efficient. This and other approximations inherent in the approach are discussed later in the paper.

3D database searches

Given a positioned template, it is possible to infer, for each template attachment point, the nature of the interaction(s) the corresponding substituent is to make with the active site (e.g. hydrogen bond, lipophilic contact, etc.), the nature of the functional group required for a coupling reaction to the template (e.g. acid chloride with a primary amine) and a distance range between the point of attachment to the template and the point of interaction with the active site.

These two (or more) substructural criteria with the associated distance range(s) constitute a viable query for a 3D database search for ISIS/3D (MDL Information Systems, San Leandro, CA, U.S.A.). The query can be made more sophisticated through a consideration of potential molecular transformations, or through the imposition of synthetic constraints on allowed chemistries in specified substructures. By using the ACD, we maximise the chance that all chosen substituents will be commercially available. In general, the search carried out will explore the conformational flexibility of the database molecules [72–74] to ensure that as many as possible of the potential substituents at each position will be retrieved.

For each template attachment point, a file of potential substituents is saved as 2D structures to a file in MDL's SD format [75] and then the Converter program [76] is used to add the necessary hydrogen atoms and generate 3D coordinates for the structures.

Computational screening of substituents

The methods developed for the computational screening of potential substituents are derived in the main part from techniques used in our in-house de novo design package, PRO_LIGAND [53,57,68–71]. As described earlier, each template attachment site has its own design model and the template attachment sites themselves are appended to the design models, according to the labels specified in the template file which is input to the program. By automatically labelling the potential substituents for each template attachment position with appropriate interaction/link sites, it is possible to use rapid algorithms to establish whether they can form good molecular interactions with the active site. For more details, the reader is referred to Refs. 53 and 71.

A pseudo-code description of the computational screening procedure is given in Fig. 3 and the steps are explained in more detail in what follows.

```
Read in template and receptor structure
Do i = 1, Number_of_Template_Attachment_Sites
   Read in design model
   Do j = 1, Number_of_Potential_Substituents_at_Position_i
      Apply specified molecular transformations
      Check initial molecular property screens
      If FAIL then break
      Check subgraph isomorphism match to design model
      If FAIL then break
      Check directed tweak fitting to design model
      If FAIL then break
      Calculate score and internal strain of substituent
      Save substituent
   Enddo
Enddo
```

Fig. 3. Pseudo-code description of computational screening of substituents.

Molecular transformations

The flexibility of PRO_SELECT is enhanced by the program's ability to detect specified functional groups and replace them with another group. This increases the diversity in the virtual library that is screened by PRO_SELECT and so increases the chance of finding novel active compounds. The computational deprotection of protected functional groups is one example of how this feature might be used.

The molecular transformation is controlled by rules containing a SMILES-like notation [77] for the substructures together with a number of integers. Thus, for instance, the rule shown in Fig. 4 indicates that up to three silyl ethers are to be replaced by hydroxyl groups in any molecule. The geometry for the transformed part of the molecule is rebuilt atom by atom using a rule-based procedure and then relaxed with a molecular mechanics minimisation (see below). The molecular transformation procedure is flexible enough to cope with a wide variety of specified transformations such as reduction of nitro or nitriles to amines, addition of glycine (say) to a carboxylic acid or amine substituent, removal of Boc-protecting groups, etc.

A similar approach is used to protonate or deprotonate certain functional groups specified by the user in order that the molecules to be placed in the active site have realistic protonation patterns. Once the molecules have been subjected to all the transformations requested by the user, they are passed on to the initial molecular property screens.

Initial screens

Before subjecting the potential substituents to the more computationally demanding subgraph isomorphism and directed tweak checks, some rapid molecular property screens are used to eliminate unsuitable structures. Thus, the user may set acceptable ranges for a number of properties:

- molecular weight;
- number of atoms;
- log P (calculated using the method of Viswanadhan et al. [78]);
- number of rotatable bonds.

Any substituents which fall outside the acceptable ranges are automatically rejected. This is useful, for example, when the database entry contains more than one component. PRO_SELECT automatically separates the components and treats each one as a separate substituent. The screen based on the number of atoms tends to remove the undesirable component, which is often a counterion. The code can also screen out duplicates.

A further initial screen on substituents is employed for some complex template chemistries. If in a ring-forming reaction one chemical reagent gives rise to two substituents on the template, then the two corresponding template attachment points will have the same list of available chemicals associated with them. Specific checks ensure that only chemical reagents which have provided a substituent to pass all screens for the first template attachment point are considered for the provision of substituents for the second template attachment site.

Subgraph isomorphism matching

The first step in the graph matching process is to label the potential substituent with the appropriate interaction sites. This is accomplished by means of a rule-based procedure where each rule denotes a substructure in the SMILES-like notation mentioned earlier and indicates if

```
("OSi", 1, "OH", 1, 0, 0, 3)

OSi = characteristic substructure of protecting group
1 = atom of protecting group substructure to be unaffected by transformation
OH = unprotected moiety to replace protected version of functional group
1 = atom of unprotected moiety to superimpose upon invariant atom of protecting group
0 = formal charge to be added
0 = atom to which formal charge is to be added
3 = maximum number of invocations of this rule per molecule
```

Fig. 4. Rule for molecular transformation.

and how each of the atoms in that substructure should be labelled. Thus, for example, the first rule in Fig. 5 instructs the program to search the substituent for any matches to the specified substructure (C(= NH)N(H)H) and to label the second and fourth atoms of the match as **X** sites and the third and fifth atoms of the match as **D** sites. A powerful regular expression-based syntax is available within the SMILES-like notation which permits very flexible definitions of the rules; for instance, the second rule in Fig. 5 indicates that any OH or NH group attached to a carbon atom should be labelled as a donor group.

In addition to the interaction sites described earlier, it is also necessary to label each substituent with *link sites*. These denote the vector site in the structure where the potential substituent will join the template. The link sites are assigned in an identical manner to the interaction sites. Thus, for instance, the third rule in Fig. 5 instructs the program to label the C-C bond in a CCO_2H substructure as a link site (link site vectors are denoted as **V**-**W**). The '1' in this instance (in V1 and W1) indicates that the link sites belong to R-group position 1. The operation of this link site labelling when a carboxylic acid on the substituent is combined with a labelled amine on a template is given in Fig. 6a. Notice that the link site does not have to correspond to an attachment point in an actual chemical reaction; in the example given in Fig. 6a, the formation of a peptide bond is the chemical reaction associated with substituent attachment, but by defining the template to already contain the peptide bond, the C-C bond can be used as the computational link site. An alternative and equivalent definition is given in Fig. 6b. The chosen definition is dictated by convenience or computational efficiency, although in templates derived from ring-forming reactions, it is often essential to choose link sites which do not correspond to the bond formed in the chemical reaction.

In general, each rule will give rise to different possible labellings (instantiations) for a substituent (e.g. rule 2 in Fig. 5 will give more than one instantiation for a bisamine). One instantiation is chosen from each rule to form a group labelling scheme for the substituents. All possible groups are formed in this way and are tested for each substituent. If, for any reason, a potential substituent cannot be assigned either interaction sites or link sites, it is automatically rejected. Otherwise, the program proceeds to seek a 3D match between each group of interaction/link sites of the substituent and the interaction/link sites of the design model. This is accomplished using the subgraph isomorphism algorithm of Ullmann [79] which has been used successfully in many chemical structure applications. In order to account for the conformational flexibility of the substituents in this process, distance bounds matrices are calculated using the directed tweak routines which seek to establish the maximum and minimum distances that can be attained between all pairs of atoms through rotation about rotatable bonds [71]. The subgraph isomorphism algorithm then uses these distance ranges in establishing a match in the manner described by Clark et al. [80].

```
("C(=NH)N(H)H", "I", 2, "X", 3, "D", 4, "X", 5, "D")
("C^[ON]$H", "I", 2, "X", 3, "D")
("CC(=O)OH", "L", 1, "W1", 2, "V1")
```

Fig. 5. Rules for substituent labelling. The first string gives a SMILES representation of the substructure to be labelled and the second string indicates whether it is an interaction site or a link site. Subsequent entries in a rule come in pairs, the first is an integer specifying the atom to be labelled (as given by the order in the SMILES string) and the second is the interaction site type to be assigned to the atom.

If no match is found for the substituent, it is rejected and the algorithm returns to consider the next available substituent.

Directed tweak matching

The finding of a match for a substituent in the subgraph isomorphism check is a necessary, but not sufficient, condition for a substituent to be accepted. This is

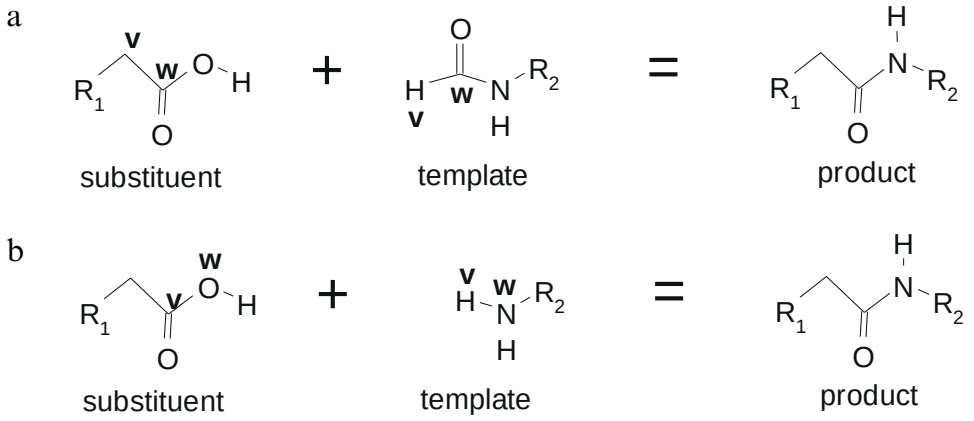
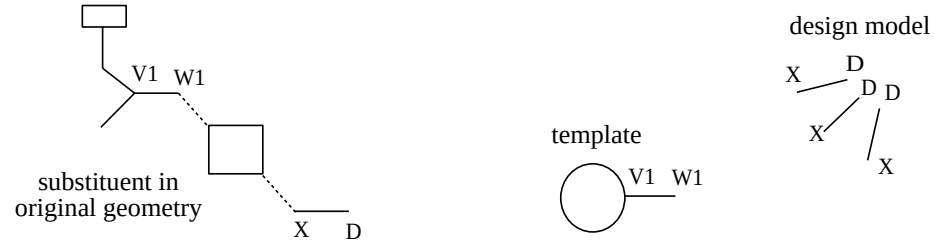
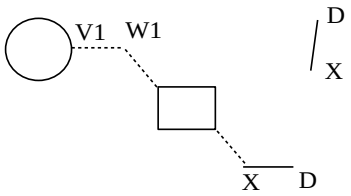


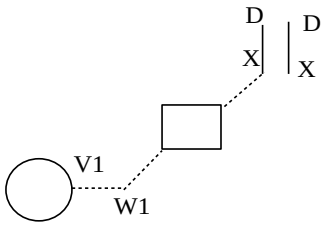
Fig. 6. Alternative definitions for link sites (labelled V and W) on the template and the substituent where the reaction is the formation of a peptide bond.



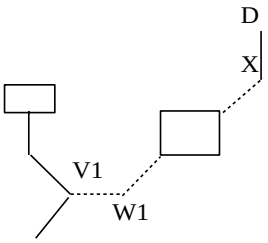
Step 1: Subgraph isomorphism algorithm locates match between sites in substituent and sites in template and design model



Step 2: Substituent is attached to template by overlaying matching V-W sites. Rest of substituent is deleted



Step 3: Rotate about rotatable (dashed) bonds to maximise contact with hit sites



Step 4: Save new geometry of substituent

Fig. 7. Schematic illustration of the flexible fitting of substituents. Rotatable bonds are shown as dashed lines, and non-rotatable bonds as full lines. **D-X** sites indicate hydrogen bond donating functionality on the substituent which is to match similar sites in the design model. **V1-W1** sites indicate the link sites of attachment for the template and the substituent.

because the distance bounds matrix does not include correlation effects, i.e. the effect that one interatomic distance having one value might have on the possible values attainable by the other interatomic distances. Thus, in order to establish whether the substituent is in fact a viable one for the template attachment point in question, a specific matching conformation must be generated using a conformational exploration procedure [80].

The procedure adopted in this work is based on the directed tweak algorithm [73] which was originally developed for 3D database searching applications, where it has been shown to be both efficient and effective [72, 73]. We have recently demonstrated its utility in the field of de novo design [71] and thus it was a natural choice for use in this work. A number of changes were made to the algorithm to improve its efficiency in this application.

The directed tweak algorithm takes the match established by the subgraph isomorphism algorithm and then seeks to verify it by performing a torsional optimisation of the rotatable bonds in the substituent. The process is illustrated schematically in Fig. 7. After a potential match has been located (step 1), the substituent is attached to the template (step 2). The bond where attachment occurs is treated as rotatable. The following cost function is minimised by a steepest descent method:

$$F = \frac{1}{\sqrt{N}} \sum_i a_i d_i^2 \qquad (1)$$

where the summation occurs over all N interaction sites, and d_i is the distance between the *i*th substituent interaction site and the design model interaction site with which it is matched. a_i is a coefficient which depends on the type

of interaction site being matched and is a simple function of the tolerances used in the subgraph isomorphism algorithm. This cost function differs from that used by Hurst [73] and in our previous work [71] in that the distances between pairs of sites are not included, only the absolute distance between the two matched sites. This is possible because the template attachment site provides a fixed point of reference in the design model coordinate space. This means that there are fewer terms in the cost function expression and it is likely that the simpler expression has fewer local minima. There is also no need to check the chirality of the conformations produced. These advantages make the approach considerably faster.

After minimisation, the conformation is accepted if it passes the following criteria. The value of the cost function must be less than a user-defined maximum (typically about 0.5 $Å^2$), and the substituent must not be clashing with the receptor, with the template or with itself. If the conformation fails these checks, the tweak routines are used to find an alternative conformation; the procedure is repeated until an acceptable geometry is located or a user-definable number of attempts has been exceeded [71].

Substituent scoring

The substituent still attached to the template is then optionally minimised using a molecular mechanics energy function. This is done in the presence of the receptor (which is treated as rigid) and a cutoff on the long-range terms of 8 Å is usually applied. An estimate of the strain energy in the receptor-bound conformation is obtained by performing the minimisation (starting from the tweak-generated geometry) in the absence of the receptor, and subtracting from this energy the intramolecular energy of the receptor-bound conformation. During these calculations, the template part of the molecule is held rigid. All molecular mechanics calculations employ the fast and approximate 'Clean' force field developed by Hahn [81]. Partial charges are calculated using the method of Gasteiger and Marsili [82]. The Clean force field bears many similarities to the 'generalised atom' force field incorporated in the Chem-X software [83,84] in that it does not rely on extended force field atom types. Only element type, hybridisation and bond type are used in calculating the energy of a system [81]. A number of minor adjustments were made in our implementation of the force field. The first was that all hydrogen atoms were treated specifically and were assigned an sp^3 atom type. The second was that van der Waals radii for potential hydrogen-bond-forming atom pairs were scaled, typically by 0.8. It should be realised that the purpose of the Clean force field in our application is to provide a rough clean-up of the substituents, which may possess distorted geometries caused by unrealistic torsion angles. The force field must be robust, in the sense that it must be able to cope with any chemistries that are given to it, and this is why a generalised atom force field is the most obvious choice. Additionally, it must meet the approximate accuracy criteria, and, in this context, the accuracy of the intermolecular terms is critical. It was after the analysis of intermolecular geometries obtained using Clean that the scaling of the hydrogen bonding van der Waals radii was introduced. We believe that the force field produces improved and reasonable geometries, at least when some portion of the molecule (in our case, the template) is held fixed in the receptor.

The minimised conformation of the substituent (still attached to the template) is then assigned a score using a scoring function developed by Böhm for use in the de novo design program LUDI [85]. Böhm's scoring function permits an approximate calculation of the binding free energy of the substituent and template in terms of readily calculable quantities such as lipophilic contact surface area, the number and quality of hydrogen bonds formed and the number of rotatable bonds. Following Böhm [85], the form of the equation used is

$$\begin{aligned}\Delta G_{binding} = \Delta G_0 &+ \Delta G_{hb} \sum_{hbonds} f(\Delta R,\Delta\alpha) \\ &+ \Delta G_{ionic} \sum_{ionic} f(\Delta R,\Delta\alpha) \\ &+ \Delta G_{lipo}|A_{lipo}| + \Delta G_{rot} N_{rot}\end{aligned}$$

where

$$f(\Delta R,\Delta\alpha) = f1(\Delta R) f2(\Delta\alpha)$$

and

$$f1(\Delta R) = \begin{cases} 1 & \text{if } \Delta R \le 0.2 \text{ Å} \\ 1 - (\Delta R - 0.2)/0.4 & \text{if } 0.2 \text{ Å} < \Delta R \le 0.6 \text{ Å} \\ 0 & \text{if } \Delta R > 0.6 \text{ Å} \end{cases}$$

and

$$f2(\Delta\alpha) = \begin{cases} 1 & \text{if } \Delta\alpha \le 30° \\ 1 - (\Delta\alpha - 30)/50 & \text{if } 30° < \Delta\alpha \le 80° \\ 0 & \text{if } \Delta\alpha > 80° \end{cases}$$

$f(\Delta R,\Delta\alpha)$ is a function which penalises hydrogen bonds whose geometry deviates from ideality. ΔR is the deviation of the H···O/N hydrogen bond length from 1.9 Å; $\Delta\alpha$ is the deviation of the hydrogen bond angle N/O-H···O/N from its ideal value of 180°. See Ref. 85 for a more detailed discussion. ΔG_0 is a contribution to binding energy which is independent of interactions with the protein. Böhm suggests that this may be rationalised as a

reduction in binding energy due to loss of translational and rotational entropy of the ligand. ΔG_{hb} describes the contribution from an ideal hydrogen bond and G_{ionic} the contribution from an unperturbed ionic interaction. ΔG_{lipo} denotes the contribution from lipophilic interactions which is assumed to be proportional to the lipophilic contact surface between ligand and protein, A_{lipo}. Finally, ΔG_{rot} describes the loss of binding energy due to the freezing of internal degrees of freedom in the ligand. N_{rot} is the number of acyclic sp^3-sp^3 and sp^3-sp^2 bonds excluding the rotations of terminal methyl and amine groups.

The values used for the various coefficients are those adopted by Böhm for the LUDI program [85]: $\Delta G_0 = 5.4$, $\Delta G_{hb} = -4.7$, $G_{ionic} = -8.3$, $\Delta G_{lipo} = -0.17$ and $\Delta G_{rot} = 1.4$. The coefficients were obtained by fitting the equation to the activities for ligand–receptor binding where crystallographic structures for the complexes were available (although a few geometries were obtained by docking the ligands into the receptor). The accuracy of the function is not expected to be better than 1.5 orders of magnitude in the binding affinity.

Using this scoring function, it is possible to rank the accepted substituents according to the strength of interaction they are likely to make with the receptor by subtracting the precomputed score for the template from the total score for the template–substituent combination.

Since the first-found conformation is not necessarily the highest scoring one available to the substituent, a user-specified number of acceptable conformations (typically 10 or more) will be sought and scored. After these conformations have been examined, the substituent geometry with the highest score is saved for future reference.

Substituent selection

Once potential substituents have been located for each template attachment point, the program can automatically enumerate the possibilities to produce the full library for all combinations of those substituents and the template. However, it is usually advisable to consider the substituent lists further so as to reduce the size of the enumerated library. This section outlines some of the methods that we have developed and used for further reducing the size of the substituent lists.

The output from PRO_SELECT for each template attachment point is a directory of substituent files, each containing scoring and database information. A directory of substituents, the receptor structure and the template molecule are read into a graphical visualisation package. The package has been designed to allow the user to scroll quickly through the substituent list in any order whilst displaying the file name and any molecular properties that are present in the substituent files (e.g. the strain energy, the Böhm score or the components of the score). The properties are displayed in a spreadsheet running alongside the molecular visualisation. Substituents can be visualised in isolation, with the template or with the receptor structure.

A set of substituent structures is treated as a *list* on which operations can be performed by the user. For instance, one would probably want to store all structures with Böhm scores less than a given value in a new list of 'Good Scores'; one might also want to exclude all structures with high strain energies, and possibly remove bad structures judged by more subjective criteria (e.g. bad chemistries or geometries). The user has full control over which list of structures is displayed by the package. At any time, the user can write a list to a new or old directory or remove a list from an old directory.

Coupled to the list functionality is a clustering facility which allows one to cluster a specified list on the basis of 2D chemical functionality. The clustering is based on similar functionality available in PRO_LIGAND [70] which measures similarity by Tanimoto coefficients derived from bit-string representations of the chemical structures [86,87]. The bit strings are specified by 172 atom-centred fragments generated from an analysis of 5000 structures in the Cambridge Structural Database [88]. Several different clustering algorithms are available, and we have tended to use a hierarchical clustering method such as Complete Linkage or Ward's. (In our applications, the number of structures clustered has been about 100 or less, so CPU time is not an issue.) A number of tools are available to help decide on the appropriate number of clusters for the specified lists. The output from the clustering is a new set of lists, each containing an individual cluster. These can be browsed and operated on as described above. Whilst the clustering is not always perfectly in line with chemical intuition, it is an extremely useful way of navigating through and keeping track of a fairly large number of substituents.

The final facility provided by the molecular browser is to rescore a list of substituents using the empirical Böhm score. Rescoring in this way is practical because tens of structures can be scored per second, and is useful because information gained during the scoring can be used to provide a graphical representation of the score. Hydrogen bonds or ionic interactions are located, marked and annotated with the contribution they make to the predicted binding affinity. This saves a lot of time in deciding which hydrogen bonds are formed and how good they are. It also points out hydrogen bonds which may be contributing to the score in an unrealistic way. Bonds that are considered rotatable are also marked so that the user can see which bonds are (or are not) contributing to the score. Finally, the grid used to establish the lipophilic contribution to the score is displayed graphically. Grid points can be displayed directly or can be interpolated to allow their visualisation as surfaces. Relevant grid points/surfaces fall into several categories:

lipophilic ligand atom in contact with lipophilic receptor atom;
lipophilic ligand atom in contact with polar receptor atom (or vice versa);
polar ligand atom in contact with polar receptor atom;
lipophilic ligand atom in contact with nothing (i.e. solvent);
polar ligand atom in contact with nothing (i.e. solvent);
volume of ligand.

The user can colour each of these grid point types, though, in practice, we have tended to use colours for the first three types only. The visualisation is useful because it displays aspects of ligand–receptor contact which are often difficult to assess quickly from looking at the complex alone.

After application of these tools, a smaller set of substituents is decided on for each of the template attachment points. The aspects which are considered in producing this list are:

2D diversity: Using the clustering tools and chemical knowledge, a diverse set of substituents must be chosen. For example, if there are 10 fluorinated derivatives of phenylalanine only one need be chosen. The exploration of different chemistries is important because the scoring functions can only be expected to deliver approximate accuracy in the prediction of the binding affinity.

3D contacts: It is important to look at the contacts a substituent is predicted to make with the receptor and to form a judgement as to whether these seem reasonable or not. In particular, substituents which have a large amount of polar–nonpolar contact are suspect. There must also be an awareness of 3D diversity and there should be an attempt to target molecules which explore different forms of receptor contact to make up for deficiencies in the scoring criteria.

Synthetic considerations: There must be a consideration of synthetic feasibility. Although the strategy of making single compounds by the most appropriate protocol means that a larger diversity of substituents are synthetically accessible, there will still be some substituents which contain functionalities that are difficult to incorporate in any synthetic protocol. Additionally, where one compound is to be chosen from several similar possibilities, choices could be made on the basis of ease of availability or price of the compounds.

Scores: The scores of the substituents (e.g. Böhm scores, force-field energies, etc.) can be used to choose preferred substituents from among lists of similar compounds.

Combinatorial enumeration

The process of combinatorial enumeration simply involves forming a list of all the remaining substituents at each R-group position and then creating all possible combinations of them. Thus, given a template with three R-group positions and three substituents for each, the combinatorial enumeration procedure will produce 27 different molecules. It is important to realise that the number of substituents at each position will have been reduced so as to allow the enumeration to be feasible. The enumeration typically involves the formation of hundreds of structures. The geometries produced by the enumeration are based on the highest scoring geometries of the corresponding substituents. At this stage it is possible that there will be clashes between substituents associated with different template attachment points. Such clashes are unlikely, but will be detected when the enumerated structures are analysed further and ranked. The resulting molecules are stored for further analysis or transfer to a 3D database.

Ranking of enumerated molecules

The resulting molecules can be, but are not usually, minimised with the Clean force field and are then rescored in the same manner as the substituents. The estimated log P and molecular weight are also routinely calculated for the complete molecules.

In our applications, the complete molecules have also been subjected to evaluation using the CFF95 force field [90] in Discover [89]. Simplified cut-down models of the receptor are used and minimisation and/or molecular dynamics are used to assess the quality of the designs. If the designs are reasonably stable during minimisation and dynamics, and possess high-scoring snapshots, then they are considered suitable for synthesis. The approach is discussed in detail in another paper [58].

The final decision about which molecules to synthesise is made by considering all the data collected for the substituents and enumerated molecules. The full library could be synthesised, or selected molecules can be chosen from the full library. The possibility of using experimental design to choose the best candidates has been explored. The method used was D-optimal design which attempts to maximise the coverage of a specified property space in a subset of molecules chosen from a larger library. In our explorations, the spread in the following properties was approximately maximised:

(1) the substituents from which each library member was derived;

(2) the estimated value of log P for each library member; and

(3) the hydrogen bond, the rotatable bond and lipophilic contributions to the Böhm score for each library member.

Several constraints can be imposed on the design such as inclusion or exclusion of compounds which are outside a specified range of a molecular property. The general conclusion of this application of experimental design was that, although it was useful, practical considerations, such

as ease of synthesis of particular classes of compounds from the full library, were usually more important. For this reason, experimental design has not generally been adopted during applications of PRO_SELECT.

Software architecture

This section considers the architecture of the PRO_SELECT software. Most of the computationally intensive routines are written in Fortran, the data structure and data handling code is written in C, and the drivers and user interface parts are written in Global. Global is a proprietary interpreted language designed for application to computer-aided molecular design. The main use of Global is that, together with the chemical utilities and their associated data structure routines, it provides a flexible environment for the operation of PRO_SELECT. A language which allows high-order chemical design features and user input to be expressed succinctly and naturally makes the methods easy to program, amend and debug. Global also makes mundane tasks such as IO and memory management straightforward, and frees the programmer to concentrate on the chemical design aspects of a programming task. Because there is no compilation for the interpreted language, it is easy to adapt the drivers and run them interactively or in batch mode. The users can either treat the GLOBAL files as input decks in the traditional sense or, if they have more confidence, can make fairly significant changes to the order of operation of the drivers, introducing different screens for the substituents as they see fit. Higher level languages have shown their worth before in CAMD applications as illustrated by Tripos's SPL language or the various languages offered to MSI users. The PRO_SELECT software discussed in this paper was constructed by four full-time programmers in 3 months and this illustrates the advantages of the software architecture.

Application to thrombin inhibitor design

Thrombin is a key serine protease within the blood coagulation cascade and inhibitors are useful in anticoagulant prophylaxis or therapy [91–93]. We have applied PRO_SELECT to the design of reversible inhibitors of thrombin and full details of the application are to be published elsewhere [58]. Here the intention is to discuss the results generally so as to set the description of the technology in context.

The template was L-proline, which is the central portion of a known covalent inhibitor of thrombin, PPACK. PPACK is a tripeptide analogue of D-Phe-Pro-Arg and its crystal structure bound to thrombin is available [94]. A template positioning was taken directly from the crystal structure and an alternative positioning was generated by the modelling of a non-covalently binding analogue of PPACK. The crystal structures were used to construct 3D database searches. At the carboxy terminus of proline (the arginine end), amines were required which also possessed hydrogen bond donor functionality at a specified distance from the attachment site. At the N-terminus of proline (the phenylalanine end), carboxylic acids were required which also possessed a donor and a hydrophobe at specified distances from the attachment site. The size of the virtual library resulting from the combination of the substituents extracted from the 3D database search is 400 000. Operation of PRO_SELECT and subsequent scoring and diversity analysis brought the library down to a manageable size. Eight substituents at the phenylalanine end and nine at the arginine end were eventually chosen for synthesis. Other synthesis candidates were produced by PRO_SELECT runs on sulphonic acids at the phenylalanine end (producing sulphonamides), and by PRO_SELECT runs on nitro and nitrile compounds at the arginine end. In the latter case, the nitro and nitrile compounds were treated as protecting groups for amines.

Over 30 molecules were selected from the resulting libraries and were synthesised. About half of the synthesised molecules had micromolar activity and the most active compound was pBr-D-Phe-Pro-Agmatine, with a K_i of 40 nM. Agmatine is an analogue of arginine, and is therefore likely to have poor pharmacokinetic properties and side-effect profile. However, several active compounds contain different functional groups and, in particular, an aniline has promising activity for thrombin and excellent selectivity compared to trypsin. At the phenylalanine end, several interesting substituents were located and, in particular, several analogues of D-Phe containing additional hydrogen bonding functionality were active.

Discussion

The integration of ideas from combinatorial chemistry and structure-based drug design is currently one of the most active areas in computer-aided molecular design. To our knowledge, this paper represents the first published strategy for a synergistic coupling of these two complementary drug discovery paradigms. PRO_SELECT builds upon technology and expertise we have acquired in the process of developing and using our in-house de novo design package, PRO_LIGAND, and has been heavily influenced by close collaboration with our synthetic chemists. The result is an approach which is rapid, utilises all available structural information and produces novel molecules which are amenable to organic synthesis.

There are several particular strengths inherent in the PRO_SELECT approach. Firstly, following PRO_LIGAND, we have used discrete interaction sites to represent favourable binding regions of the active site. By labelling the potential substituents with appropriate interaction sites, we are able to use rapid subgraph isomorph-

ism and directed tweak algorithms to search for substituent conformations which are low in strain energy and score well on the basis of an empirical scoring function. These techniques enable us to screen molecules faster than if a purely molecular mechanics-based approach was employed. Secondly, the 'template elaboration' approach is conceptually simple and permits a wide variety of molecular scaffolds to be employed, depending on the constraints of the active site under question. Thirdly, the use of a rapid, approximate empirical scoring function enables us to obtain a quick estimate of the (contribution to) binding affinity of a particular substituent or molecule. In addition, the rule-based molecular transformations allow us to use protected molecules directly from the ACD, significantly increasing the number of structures available for assessment by PRO_SELECT.

The underlying assumptions of the method are that the template and the receptor do not move and that substituents can be assessed independently of each other. These assumptions are important if the computational expense is to be minimised. A list of 1000 reasonable substituents can be screened and scored in an overnight run on an HP-735 workstation when molecular mechanics minimisations are employed. The run times are variable according to the quality of the substituent lists (lists containing fewer appropriate substituents are screened more quickly) and the type of job being performed (how many tweak conformations to locate, etc.). The assumption of template rigidity is important, but could perhaps be avoided. One strategy would be to assume that it could move and perform a local docking of the substituent template complex into the active-site region after initial graph-based screenings had been passed. This would require a more complex enumeration strategy probably involving further docking runs. However, we believe that our approach of multiple template and receptor positions is probably as effective and is certainly quicker. Whilst some individual substituents are possibly missed, we believe that whole classes will not be, provided the binding modes considered for the library are reasonable. This has been our experience with thrombin and other serine proteases where the use of different template positions does not change the character of the results since, although substituents get slightly different scores and some substituents are not selected in all template positions, the targetted diversity space associated with the list of substituents remains the same. The assumption of a rigid receptor will probably cause some solutions to be missed, although again our use of multiple receptor snapshots will mitigate this. It should be possible to amend the strategy in a variety of ways to afford improved conformational sampling of the receptor side chains and this is a focus of our current research. The assumption that substituents can be rejected independently of each other should be a good one, provided the binding mode does not change significantly and unexpectedly on the addition of different substituents. This latter problem relates to deficiencies inherent in structure-based drug design; actives will be missed if they bind in unanticipated ways. However, the alternative of synthesising and testing all the molecules in the virtual library, just in case they bind in an unexpected way, is not practical. The possibility of cooperative effects between substituents, such as clashes or intramolecular interactions, is dealt with when the library of synthesis candidates is enumerated and ranked.

The main weakness of the method is perhaps the scoring function used to assess the binding affinity of the designed molecules. We believe that the Böhm scoring function is probably the best quick method for general application to ligand–receptor binding which has been published so far; however, it does have a number of limitations. These are discussed in detail in a recent review article on binding affinity prediction by Ajay and Murcko [95]. A new method which adopts a similar approach has been developed by Marshall and co-workers [96] and this may offer some improvements. However, it is to be expected that methods based on known receptor–ligand complexes may have poor predictive power when applied to ligand–receptor geometries which are unrealistic, i.e. when applied to ligands which are not active. It is because of reservations about the accuracy of the scoring function that it is important to take diversity into account in the choice of substituents and synthesis candidates. Improvements to the scoring function should improve the PRO_SELECT process.

The application of PRO_SELECT to the design of thrombin inhibitors indicates the ability of the method to generate novel active molecules which are both synthetically accessible and have a high likelihood of showing activity. This in itself is an indication of the success of the method and its approximations. Of particular importance is the ability of PRO_SELECT to use structural information to reduce the potentially large substituent lists at each substituent position on the template. It has been shown in the thrombin case [58] how initial substituent lists based on the template chemistry alone contained 4000 amines and 9000 carboxylic acids at the arginine and phenylalanine ends, respectively. The virtual library associated with these is immense, and it is difficult to see how a non-receptor-based diversity analysis could hope to span the associated space with a manageable number of synthesis candidates. Even after a fairly sophisticated conformationally flexible search, there were still 400 and 900 substituents respectively yielding a library of about 400 000 compounds. The PRO_SELECT procedure then reduced the lists to eight and nine, respectively [58]. These molecules themselves were not amenable to one synthetic protocol and therefore would have been difficult to include in a single application of a combinatorial method. In this manner, we believe that PRO_SELECT can rapid-

ly direct synthesis and testing into areas likely to bear fruit.

Conclusions

We have presented a novel methodology embodying the concepts underlying combinatorial chemistry and structure-based drug design. The approach, PRO_SELECT, is rapid in operation and is capable of generating novel molecules which are amenable to synthesis. It therefore addresses a major problem in some applications of structure-based drug design, i.e. the problem of slow data feedback between synthesis and experiment. PRO_SELECT operates by computationally screening a virtual combinatorial library against the receptor of interest. The process is computationally efficient for large libraries because the 3D receptor-based screens are performed mainly on individual substituents. PRO_SELECT also contains a strategy for performing molecular transformations on the available starting materials which mirror simple synthetic reactions. This has the effect of increasing the diversity of the virtual library screened by PRO_SELECT.

There are also advantages over combinatorial chemistry for applications to biological targets where the 3D structure is known. The first is that diversity is targetted to specific binding modes for the macromolecule of interest and so far fewer molecules need to be synthesised to span the diversity space of the full library for the targetted binding modes. Additionally so few molecules need to be synthesised that different synthetic protocols for different molecules can easily be considered; there is no need for one or two specific synthetic routes to all members of the library. This means that a greater diversity of molecules can be considered for inclusion in the library, and so paradoxically, although fewer molecules need to be made, this set of molecules should possess greater diversity.

The method has been successfully applied to the design of thrombin inhibitors and this illustrates the potential of the approach.

References

1 Whittle, P.J. and Blundell, T.L., Annu. Rev. Biophys. Biomol. Struct., 23 (1994) 349.
2 Greer, J., Erickson, J.W., Baldwin, J.J. and Varney, M.D., J. Med. Chem., 37 (1994) 1035.
3 Verlinde, C.L.M.J. and Hol, W.G.J., Structure, 2 (1994) 577.
4 Guida, W.C., Curr. Opin. Struct. Biol., 4 (1994) 777.
5 Colman, P.M., Curr. Opin. Struct. Biol., 4 (1994) 868.
6 Bohacek, R.S., McMartin, C. and Guida, W.C., Med. Res. Rev., 16 (1996) 3.
7 Montgomery, J.A., Niwas, S., Rose, J.D., Secrist III, J.A., Babu, S., Bugg, C.E., Erion, M.D., Guida, W.C. and Ealick, S.E., J. Med. Chem., 36 (1993) 55.
8 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. and Bartlett, C.A., J. Med. Chem., 36 (1993) 733.
9 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. and Penn, C.R., Nature, 263 (1993) 418.
10 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. and Erickson-Viitanen, S.E., Science, 263 (1994) 380.
11 Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gordon, E.M., J. Med. Chem., 37 (1994) 1233.
12 Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gallop, M.A., J. Med. Chem., 37 (1994) 1385.
13 Terrett, N.K., Gardner, M., Gordon, D.W., Kobylecki, R.J. and Steele, J., Tetrahedron, 51 (1995) 8135.
14 Thompson, L.A. and Ellman, J.A., Chem. Rev., 91 (1996) 555.
15 Gordon, E.M., Gallop, M.A. and Patel, D.V., Acc. Chem. Res., 29 (1996) 144.
16 Janda, K.D., Proc. Natl. Acad. Sci. USA, 91 (1994) 10779.
17 Ohlmeyer, M.H., Swanson, R.N., Dillard, L.W., Reader, J.C., Asouline, G., Kobayashi, R., Wigler, M. and Still, W.C., Proc. Natl. Acad. Sci. USA, 90 (1993) 10922.
18 Koppel, G., Dodds, C., Houchins, B., Hunden, D., Johnson, D., Owens, R., Chaney, M., Usdin, T., Hoffman, B. and Brownstein, M., Chem. Biol., 2 (1995) 483.
19 Zuckermann, R.N., Martin, E.J., Spellmeyer, D.C., Stauber, G.B., Shoemaker, K.R., Kerr, J.M., Figliozzi, G.M., Goff, D.A., Siani, M.A., Simon, R.J., Banville, S.C., Brown, E.G., Wag, L., Richter, L.S. and Moos, W.H., J. Med. Chem., 37 (1994) 2678.
20 Wang, G.T., Li, S., Wideburg, N., Krafft, G.A. and Kempf, D.J., J. Med. Chem., 38 (1995) 2995.
21 Pirrung, M.C., Chau, J.-L. and Chen, J., Chem. Biol., 2 (1995) 621.
22 Murphy, M.M., Schullek, J.R., Gordon, E.M. and Gallop, M.A., J. Am. Chem. Soc., 117 (1995) 7029.
23 Schnebli, H.P. and Braun, N.J., In Barrett, A.J. and Salvensen, G. (Eds.) Proteinase Inhibitors, Elsevier, New York, NY, U.S.A., 1986, pp. 613–617.
24 Bieth, J.G., Biochem. Med., 32 (1984) 387.
25 Carell, T., Wintner, E.A., Sutherland, A.J., Rebek, J., Danayerskiy, Y.M. and Vouros, P., Chem. Biol., 2 (1995) 171.
26 Madden, D., Krchnak, V. and Lebl, M., Perspect. Drug Discov. Design, 2 (1994) 269.
27 Legion, Unity and Selector, Tripos Associates Inc., St. Louis, MO, U.S.A., 1996.
28 Project Library, MDL Information Systems Inc., San Leandro, CA, U.S.A., 1996.
29 ChemDiverse, Chemical Design Ltd., Chipping Norton, Oxfordshire, U.K., 1996.
30 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., J. Med. Chem., 38 (1995) 1431.
31 Sheridan, R.P. and Kearsley, S.K., J. Chem. Inf. Comput. Sci., 35 (1995) 310.
32 Ashton, M.J., Jaye, M.C. and Mason, J.S., Drug Discov. Today, 1 (1996) 71.
33 Holland, J.D., Ranade, S.S. and Willett, P., Quant. Struct.–Act. Relatsh., 14 (1995) 501.

34 Sadowski, J., Wagener, M. and Gasteiger, J., Angew. Chem. Int. Ed. Engl., 34 (1996) 2674.
35 Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
36 Weber, L., Wallbaum, S., Broger, C. and Gubernator, K., Angew. Chem. Int. Ed. Engl., 34 (1995) 2280.
37 Singh, J., Ator, M.A., Jaeger, E.P., Allen, M.P., Whipple, D.A., Soloweij, J.E., Chowdhary, S. and Treasurywala, A.M., J. Am. Chem. Soc., 118 (1996) 1669.
38 Bunin, B.A. and Ellman, J.A., J. Am. Chem. Soc., 114 (1992) 10997.
39 DeWitt, S.H., Kiely, J.S., Stankovic, C.J., Schroeder, M.C., Cody, D.M.R. and Pavia, M.R., Proc. Natl. Acad. Sci. USA, 90 (1993) 6909.
40 Lewis, R.A. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 141.
41 Moon, J.B. and Howe, W.J., Proteins Struct. Funct. Genet., 11 (1991) 314.
42 Nishibata, Y. and Itai, A., Tetrahedron, 47 (1991) 8985.
43 Lewis, R.A., Roe, D.C., Huang, C., Ferrin, T.E., Langridge, R. and Kuntz, I.D., J. Mol. Graph., 10 (1992) 66.
44 Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 61.
45 Caflisch, A., Miranker, A. and Karplus, M., J. Med. Chem., 36 (1993) 2142.
46 Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36 (1993) 1700.
47 Gillet, V.J., Johnson, A.P., Mata, P., Sike, S. and Williams, P., J. Comput.-Aided Mol. Design, 7 (1993) 127.
48 Tschinke, V. and Cohen, N.C., J. Med. Chem., 36 (1993) 3863.
49 Ho, C.W.M. and Marshall, G.R., J. Comput.-Aided Mol. Design, 7 (1993) 623.
50 Leach, A.R. and Kilvington, S.R., J. Comput.-Aided Mol. Design, 8 (1994) 283.
51 Bohacek, R.S. and McMartin, C., J. Am. Chem. Soc., 116 (1994) 5560.
52 Gehlhaar, D.K., Moerder, K.E., Zichi, D., Sherman, C.J., Ogden, R.C. and Freer, S.T., J. Med. Chem., 38 (1995) 466.
53 Clark, D.E., Frenkel, D., Levy, S.A., Li, J., Murray, C.W., Robson, B., Waszkowycz, B. and Westhead, D.R., J. Comput.-Aided Mol. Design, 9 (1995) 13.
54 Glen, R.C. and Payne, A.W.R., J. Comput.-Aided Mol. Design, 9 (1995) 181.
55 Lewis, R.A. and Leach, A.R., J. Comput.-Aided Mol. Design, 8 (1994) 467.
56 Müller, K. (Ed.) De Novo Design, Perspectives in Drug Discovery and Design, Vol. 3, ESCOM, Leiden, The Netherlands, 1995.
57 Frenkel, D., Clark, D.E., Li, J., Murray, C.W., Robson, B., Waszkowycz, B. and Westhead, D.R., J. Comput.-Aided Mol. Design, 9 (1995) 213.
58 Young, S.C., Waszkowycz, B.W., Clark, D.E., Li, J., Liebeschuetz, J.W., Lowe, R., Mahler, J., Martin, H., Murray, C.W., Rimmer, A.D. and Westhead, D.R., J. Med. Chem., to be submitted.
59 Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 593.
60 Klebe, G., J. Mol. Biol., 237 (1994) 212.
61 Blaney, J.M. and Dixon, J.S., Perspect. Drug Discov. Design, 1 (1993) 301.
62 Kuntz, I.D., Meng, E.C. and Shoichet, B.K., Acc. Chem. Res., 27 (1994) 117.
63 Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 623.
64 Goodford, P.J., J. Med. Chem., 28 (1985) 849.
65 Pisabarro, M.T., Ortiz, A.R., Paolmer, A., Cabré, F., Garcia, L., Wade, R.C., Gago, F., Mauleón, D. and Carganico, G., J. Med. Chem., 37 (1994) 337.
66 Babine, R.E., Bleckman, T.M., Kissinger, C.R., Showalter, R., Pelletier, L.A., Lewis, C., Tucker, K., Moomaw, E., Parge, H.E. and Villafranca, J.E., Biomed. Chem. Lett., 5 (1995) 1719.
67 Available Chemicals Directory, MDL Information Systems Inc., San Leandro, CA, U.S.A., 1995.
68 Waszkowycz, B., Clark, D.E., Frenkel, D., Li, J., Murray, C.W., Robson, B. and Westhead, D.R., J. Med. Chem., 37 (1994) 3994.
69 Westhead, D.R., Clark, D.E., Frenkel, D., Li, J., Murray, C.W., Robson, B. and Waszkowycz, B., J. Comput.-Aided Mol. Design, 9 (1995) 139.
70 Clark, D.E. and Murray, C.W., J. Chem. Inf. Comput. Sci., 35 (1995) 914.
71 Murray, C.W., Clark, D.E. and Byrne, D.G., J. Comput.-Aided Mol. Design, 9 (1995) 381.
72 Clark, D.E., Jones, G., Willett, P., Kenny, P.W. and Glen, R.C., J. Chem. Inf. Comput. Sci., 34 (1994) 197.
73 Hurst, T., J. Chem. Inf. Comput. Sci., 34 (1994) 190.
74 Moock, T.E., Henry, D.R., Ozkabak, A.G. and Alamgir, M., J. Chem. Inf. Comput. Sci., 34 (1994) 184.
75 Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D., Leland, B.A. and Laufer, J., J. Chem. Inf. Comput. Sci., 32 (1992) 244.
76 Converter, v. 2.3, Molecular Simulations Inc., San Diego, CA, U.S.A., 1995.
77 Weininger, D., J. Chem. Inf. Comput. Sci., 28 (1988) 31.
78 Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K., J. Chem. Inf. Comput. Sci., 29 (1989) 163.
79 Ullmann, J.R., J. Assoc. Comput. Machin., 23 (1976) 31.
80 Clark, D.E., Willett, P. and Kenny, P.W., J. Mol. Graph., 10 (1992) 194.
81 Hahn, M., J. Med. Chem., 38 (1995) 2080.
82 Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.
83 Chem-X, Chemical Design Ltd., Chipping Norton, Oxfordshire, U.K., 1995.
84 Davies, E.K. and Murrall, N.W., Comput. Chem., 13 (1989) 149.
85 Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
86 Willett, P., Winterman, V. and Bawden, D., J. Chem. Inf. Comput. Sci., 26 (1986) 109.
87 Barnard, J.M. and Downs, G.M., J. Chem. Inf. Comput. Sci., 32 (1992) 644.
88 Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C., Mitchell, E.M., Mitchell, G.F., Smith, J.M. and Watson, D.G., J. Chem. Inf. Comput. Sci., 31 (1991) 187.
89 Discover, v. 2.9.5, Molecular Simulations Inc., San Diego, CA, U.S.A., 1995.
90 CFF95 force field, implemented in Discover 2.9.5, Molecular Simulations Inc., San Diego, CA, U.S.A., 1995.
91 Davie, E.W., Fujikawa, K. and Kisiel, W., Biochemistry, 30 (1991) 10363.
92 Anderson, H.V. and Willerson, J.T., New Engl. J. Med., 329 (1992) 703.
93 Beck, W.S., Hematology, MIT Press, Cambridge, MA, U.S.A., 1991.
94 Bode, W., Turk, D. and Karshikov, A., Protein Sci., 1 (1992) 426.
95 Ajay and Murcko, M.A., J. Med. Chem., 38 (1995) 4953.
96 Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., J. Am. Chem. Soc., 118 (1996) 3959.