Structure-based selection of building blocks for array synthesis via the World-Wide Web

Andrew R. Leach

Medicines Research Centre, Glaxo Wellcome Research and Development Ltd., Stevenage, Hertfordshire, UK

In this article we are concerned with the selection of chemical entities for array synthesis in a structure-based design project. We have extended our conformational searching algorithm to permit the enumeration of a set of substitutents at a particular position for a fixed template. The conformational space of each of the resulting structures is then explored within the confines of the binding site to identify conformations that do not interact unfavorably with the surrounding protein. The template remains fixed in its original orientation within the binding site. The interaction between each conformation and the binding site can also be quantified using various calculated properties. Each substituent for which one or more acceptable conformations can be found is retained for further analysis. Use of the program is facilitated by a Web-based interface that enables nonexpert molecular modelers to perform searches, view the results in a platform-independent manner (via VRML), and perform simple cluster analysis on the resulting sets of molecules. The approach is illustrated using a series of penicillin-based HIV-1 protease inhibitors. © 1997 by Elsevier Science Inc.

INTRODUCTION

Two of the most significant developments affecting drug design are the ever-increasing number of proteins whose structures have been determined to atomic resolution and the advent of parallel synthesis methods (commonly referred to as combinatorial chemistry or array synthesis). Computational chemistry has a major role to play in assisting the identification of new leads against a macromolecular target and in suggesting modifications to existing lead compounds to enhance their activity. In this article we describe a computational method that can rapidly search the conformational space of a series of related compounds within a protein-binding site to identify which would be sensible selections to include in a chemical library.

A schematic outline of the approach is shown in Figure 1. The algorithm relies on the assumption that a central "core" molecular fragment has been identified and its binding orientation detained. It is further assumed that this core template will remain in essentially the same location within the active site as the substituent at a given position is changed. The template orientation may be derived from a structure of an existing ligand, or may be determined theoretically using a docking algorithm. Each of the chemical entities (or "monomers") for possible substitution is then taken in turn and the connection table corresponding to the appropriate enumerated product is created. For example, a carboxylic acid in the template may be "reacted" with an amine in each monomer to form an amide. It is permissible for more than one functional group to be present in a monomer; in this case all possible products are considered in the enumeration and subsequent conformational analysis.

Having constructed the connection table for each product, the next stage is a conformational search of the ligand within the binding site. This is achieved using a modified version of our rule-based conformational search program COBRA,[1] which can rapidly identify low-energy conformations of drug-like molecules. The core template of the library remains fixed in the conformation originally supplied. Tree-pruning methods are employed to deal with the intermolecular interactions between ligand and protein, as we have used previously.[2] These enable bad contacts between the ligand and binding site to be rapidly identified and can significantly enhance the efficiency of the search. Acceptable structures are then analyzed in a variety of ways; for example, we determine the hydrogen bonds formed between ligand and protein and the amount of solvent-accessible surface buried on binding. Approximately 1000 products can be processed per hour on a single processor.

The result of the calculation is a set of monomers that are compatible with the template orientation and the protein structure. Subsequent processing of these is then possible, for ex-

Color Plates for this article are on page 180.

Received 20 February 1997; revised 5 May 1997; accepted 12 May 1997. Presented as the first MGMS Electronic Conference: http://bellatrix.pcl.ox.ac.uk/mgms/home_fr.html.

655 Avenue of the Americas, New York, NY 10010

1093-3263/97/$17.00
PII S1093-3263(97)00100-9

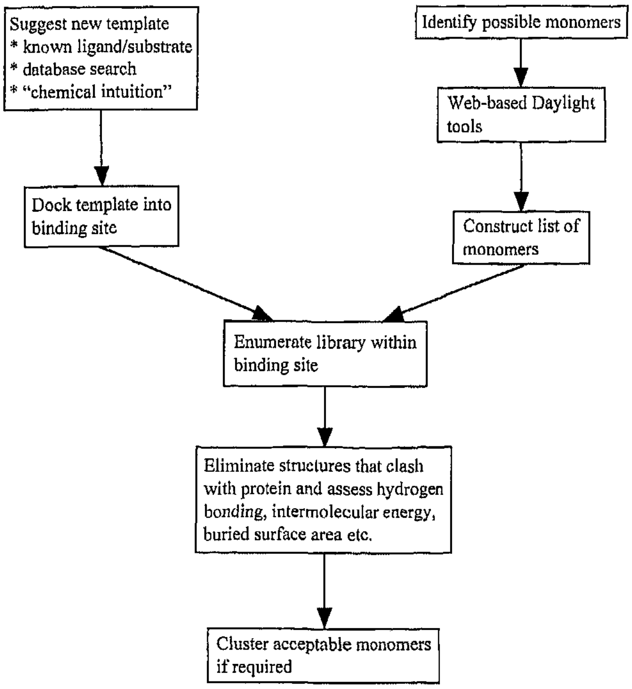
Suggest new template
* known ligand/substrate
* database search
* "chemical intuition"
Identify possible monomers
Web-based Daylight tools
Dock template into binding site
Construct list of monomers
Enumerate library within binding site
Eliminate structures that clash with protein and assess hydrogen bonding, intermolecular energy, buried surface area etc.
Cluster acceptable monomers if required

Figure 1. Schematic outline of the algorithm.

ample, by using clustering techniques or some other form of diversity analysis.

ACCESS VIA A WORLD-WIDE WEB INTERFACE

One of the goals of the project was to deliver relatively sophisticated computational structure-based design tools such as this to our bench chemists in an easy-to-use manner.[3] We achieve this using as our front end a Web browser such as Netscape. This approach has two distinct advantages. First, all our chemists (including nonexpert modelers) are familiar with the use of such browsers. Second, the browser interface enables us to deliver computational chemistry methods to the user via out company intranet independent of hardware platform. Thus the software described in this article can be accessed, and the results viewed, from PC, Macintosh, and Unix platforms. A series of linked Web pages and scripts is presented that leads the user through the process one stage at a time. Thus we have devised Web-based tools that permit the selection of appropriate monomers from a database, tools that perform the conformational search within the binding site, and tools that perform simple cluster analyses on the results. Each of these is described below in turn.

The chemist typically desires to choose building blocks that are readily available either from the internal chemical stores or from a reputable supplier. This information is held within in-house databases that can be accessed by a shell script that is initiated from an HTML form. To facilitate the process prestored lists of compounds containing widely used functional groups (e.g., carboxylic acids, amines, etc.) have been made available. The chemist can refine this initial selection according to the actual chemical reaction that is to be used, by choosing from a menu of functional groups. The chemist is able to specify whether a given functional group must not be present in the monomer (because it would interfere with the reaction), whether it must be present once only, or whether it must be present more than once. A Daylight toolkit program[4] uses this information directly with the substructural targets being defined using the powerful SMARTS language,[5] thus facilitating the definition of a wide variety of functional types. For example, an aliphatic or aromatic primary amine can be defined using the SMARTS [NH2][a,CX4].

Once a set of monomers containing the desired functionality has been identified the library enumeration can be performed. This requires a set of files containing the coordinates of the template, the list of monomers, and the protein, together with specification of a value for the "bump check" that is used to eliminate ligand orientations with unfavorable steric interactions. Variation in this value permits some flexibility in the protein to be implicitly taken into account. A simple HTML form is used to obtain this information. Although COBRA has a wide variety of options that control the generation of conformations an appropriate set of default parameters is used for the Web-driven procedure.

The efficiency of the library enumeration and conformational search is enhanced by dividing the set of monomers equally between the number of available processors on the server, thereby making full use of the available computational resources. Nevertheless, the search is not instantaneous except for very small monomer sets and so facilities are provided that enable the user to monitor the progress of the job. When all processes have finished the user is then able to view a tabular summary of the results. Hyperlinks are also provided to VRML files containing the structures of each of the library products within the protein-binding site (Color Plate 1). The VRML files are generated from the coordinates using the pdb2vrml program.[6] The structures may also be viewed using RASMOL.[7]

The number of monomers that successfully pass through the library enumeration/conformational search procedure may be too many in practical terms, to incorporate into the actual library. Moreover, it is frequently the case that many of the monomers are closely related to each other. We therefore offer the option of performing a simple cluster analysis on the monomer set. As with all other procedures described herein, this is controlled by a Tcl script that executes various in-house software programs. At present the clustering routines are based on a similarity measure determined as the Tanimoto/Jaccard coefficient[8] derived from the Daylight fingerprint. The Ward clustering routine[9] has been found appropriate for the numbers of structures typically under examination. The dendrogram derived from the clustering may be displayed as a VRML object (Color Plate 2). Each node in the dendrogram is hyperlinked to a DEPICT[10] image of the molecule, which is displayed when activated. The clustered structures can also be displayed in various modes such as PostScript or as gif images. It is also possible to select one or more structures from each cluster as it is displayed, with the final set being saved to disk for ordering (Color Plate 3). Again, it is possible to hyperlink where appropriate to VRML or PDB files of the structure of the intermolecular complex.

APPLICATION TO HIV PROTEASE

To illustrate the procedure, we consider here the modification of a series of penicillin-derived C_2 symmetric inhibitors[11] (Figure 2). These inhibitors have already been the subject of a traditional medicinal chemistry optimization program including variation of the amide (R in Figure 2) that binds in the S2 pocket. Our objective here was to determine whether the monomers identified by our automated algorithm could identify reagents similar (if not identical) to those previously investigated by the medicinal chemistry team.

The crystal structure of a related compound (available as pdb entry 1htf) was used to specify the original locations of the core template and the binding site. We used our Web-based monomer selection tools to identify compounds from the Maybridge database that contain a single carboxylic acid but that did not contain any primary or secondary amines, aldehydes, primary halides, or epoxides; groups that might interfere with the amide formation reaction. This resulted in 1939 chemical entities, which were then considered by the conformational search/ monomer selection algorithm described above. Owing to the symmetrical nature of the penicillin inhibitors substitution at just one of the positions was considered.

For each monomer to be considered acceptable at least one low-energy conformation had to be generated. An intermolecular "bump check" of 2.25 Å was used. To ensure that the ligand occupied the S2 pocket all acceptable structures were required to have at least one atom within a sphere centered within this pocket, as in our original method.[2]

Of the 1939 carboxylic acids, 108 were considered acceptable insofar as at least one conformation could be constructed with no unfavorable steric interactions at the binding site. This set of compounds was in good agreement with the potent compounds previously synthesized by the chemists in a traditional medicinal chemistry manner. Some of these are shown in Color Plate 3. This suggests that the list of 108 acids would make a good starting point for enhancement of the initial lead.

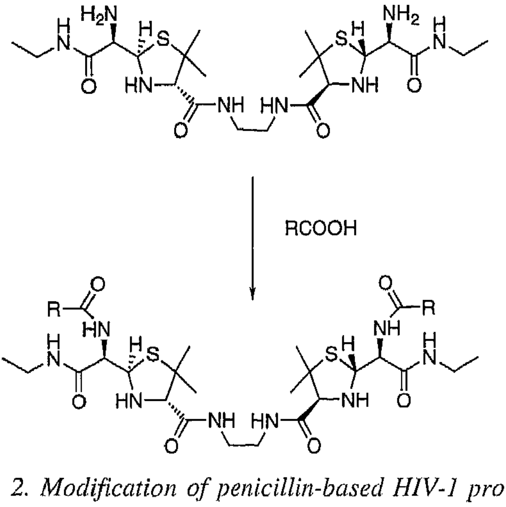

Figure 2. Modification of penicillin-based HIV-1 protease inhibitors by amide bond formation in the S2 binding pocket.

CONCLUSIONS

There are obviously a significant number of approximations with the current approach, such as the use of a rigid template that is fixed within the binding site and the reliance on a simple steric check for intermolecular acceptability. Nevertheless, these approximations are commensurate with our current goal, which is the rapid identification of potential structures for more detailed modeling or subsequent selection technique and the delivery of such techniques to the bench chemist in a platform-independent and user-friendly manner.

ACKNOWLEDGMENTS

My thanks to Neil Taylor and Ryan Smith for assistance with the WWW interface, to Richard Green for providing feedback on its use, and to Mike Hann for comments on the manuscript.

REFERENCES

1 Leach, A.R. and Prout, K. Automated conformational analysis: Directed conformational search using the A* algorithm. *J. Comput. Chem.* 1990, **11,** 1193–1205

2 Leach, A.R. and Kuntz, I.D. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.* 1992, **13,** 730–748

3 Taylor, N.R. and Smith, R. The World Wide Web as a graphical user interface to program macros for molecular graphics, molecular modelling and structure-based drug design. *J. Mol. Graphics* 1997, **14**, 291–296

4 The Daylight toolkit is available from Daylight Chemical Information Systems, Inc. (Irvine, CA)

5 *Daylight Chemical Systems. Daylight Theory Manual.* Daylight Chemical Systems, Inc., Irvine, California, and http://www.daylight.com

6 pdb2vrml. Program written by Horst Vollhardt and http://ws05.pc.chemie.thdarmstadt.de/vrml

7 Sayle, R.A. and Milner-White, E.J. RASMOL—Biomolecular graphics for all. *Trends Biochem. Sci.* 1995, **20,** 374–376

8 See, for example, Everitt, B.S. *Cluster Analysis.* Edward Arnold, London, 1980

9 Ward, J.H. Hierarchical grouping to optimise an objective function. *J. Am. Stat. Assoc.* 1963, **58,** 236–244

10 Weininger, D. SMILES. 3. DEPICT—Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* 1990, **30,** 237–243

11 Humber, D.C., Bamford, M.J., Bethell, R.C., Cammack, N., Cobley, K., Evans, D.N., Gray, N.M., Hann, M.M., Orr, D.C., Saunders, J., Balakrishna, E.V., Shenoy, E.V., Storer, R., Weingarten, G.G., and Wyatt, P.G. A series of penicillin derived C_2-symmetric inhibitors of HIV-1 proteinase: Synthesis, mode of interaction, and structure–activity studies. *J. Med. Chem.* 1993, **36,** 310–312