ELSEVIER

BOOMSLANG: A program for combinatorial structure generation

David A. Cosgrove and Peter W. Kenny

Zeneca Pharmaceuticals, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, England

An approach to exploiting pharmacophore models is described. Structures are assembled combinatorially from user-defined fragments and flexibly overlaid into the reference frame of the pharmacophore using distance geometry and molecular mechanics. The match with the pharmacophore is quantified by conformational energy and volume of overlap.

Keywords: Molecular design, ligand design, structure generation, pharmacophore

INTRODUCTION

The ultimate objective of computer-aided molecular design (CAMD) is the identification of novel molecular species and materials with specific physicochemical properties. In medicinal chemistry this generally means low molecular weight compounds that bind avidly to their protein or nucleic acid targets. In many CAMD applications construction of molecules in the reference frame of the model using interactive molecular graphics is the most labor-intensive component of the design process.

The process of automatically generating molecules, determining their compatibility with a geometric model and prioritizing compatible structures is termed *de novo ligand design*. Experimentally determined protein structures have provided the starting point for most of the published work[1–4] in this area and with few exceptions[5] flexibility of the binding site is ignored. A pharmacophore[6–8] derived from the common structural features of a set of active compounds can also provide information about the binding site geometry, albeit indirectly and not with the geometric precision of a protein crystal structure. Pharmacophores will become increasingly important starting points for de novo design as technologies mature for high-throughput screening[9,10] of collections of compounds and combinatorial synthesis.[11–13] One strategy for exploiting a pharmacophore for de novo design is to position the appropriate functional groups and then to seek spacers that can link then in low-energy conformations while remaining within the union volume[14,15] of the set of molecules comprising the overlay.

Methodology for using protein structures for the rapid prediction[4,16] of binding affinities is not as well developed as that for structure generation. When a pharmacophore is used as a starting point for de novo ligand design prediction of affinity is not generally possible, although in specific instances comparative molecular field analysis (CoMFA) may be applicable.[17] Without a means of predicting binding affinity, the most easily synthesized compounds become the most attractive targets, although consideration of molecular diversity[18] should form an integral part of any prioritization strategy.

The issue of synthetic accessibility is always important in de novo ligand design but especially so when the information content of the models is low. Synthetic complexity is determined, to some extent, by the size of the molecule although ring fusions, chiral centers, and large numbers of functional groups will all make a target compound less attractive. However, an apparently complex molecule may prove less of a challenge if the appropriate starting materials are available or if procedures of high regiospecificity and stereospecificity can be used, while preparative chemistry that is amenable to automation is especially attractive. Potential synthetic targets should not be considered in isolation; for example, an awkward synthetic intermediate may be worth expending effort on if it can be easily converted to a large number of chemically diverse products.

Ranking potential synthetic targets by accessibility is clearly a complex problem while the acceptable degree of synthetic complexity is dependent on the information content of the geometric model. One way of dealing with the problem is to build synthetic accessibility into the structure generation process. This article describes the program BOOMSLANG, which generates structures combinatorially from a user-defined synthetic strategy and then ranks them by their fit to a geometric model. The methodology is also applicable to model building for quantitative structure–activity relationships (QSARs) as well as the generation of

Color plates for this article are on p. 23.

Address reprint requests to Dr. Cosgrove at Zeneca Pharmaceuticals, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, England.
Received 28 March 1995; accepted 16 May 1995

655 Avenue of the Americas, New York, NY 10010

0263-7855/96/$15.00
SSDI 0263-7855(95)00040-D

databases of spacers for other ligand design programs such as CAVEAT[19] and LUDI.[3]

PROGRAMMING ISSUES

All modules are written for a Silicon Graphics workstation in C with the exception of the SMILES enumerator which is written in FORTRAN 77 and compiled under Irix version 5.2. The graphical user interface and molecule viewing module were created with X Windows X11R4 and Motif version 1.1. All manipulations of SMILES and SMARTS are achieved using routines from the relevant Daylight[20] toolkits, apart from the initial SMILES generation, which is achieved by character string manipulation.

METHODOLOGY

General

The ligand design strategy described in this work imposes synthetic constraints on the structure generation process from the outset. The starting points are a hypothesis that a particular geometric relationship of certain functional groups leads to binding and a synthetic strategy for achieving this relationship. Our approach to the exploitation of pharmacophore models systematically generates molecular connection tables in the manner of a combinatorial library using the elegant SMILES[21] line notation; geometry is considered only after assembly of the connection table is complete, so this is a three-dimensional (3D) database technique.[22,23] The conciseness of SMILES notation and the ease with which atoms can be added to and removed from the molecule by simple insertions into and deletions from the SMILES string make it a particularly suitable data structure for molecule generation. The CHUCKLES[24] method for representing peptide sequences and the DBMAKER[25] programs for generating 3D databases provide excellent illustrations of the power of SMILES notation for assembling connection tables in a controlled manner.

Figure 1 presents an overview of the program. Intermediate results are stored in ASCII files with standard formats facilitating links to other software. For example, the output of the SMILES enumerator can either be filtered geometrically, using a 3D database package such as UNITY,[26] or substructurally using the powerful Daylight[20] toolkit.

SMILES enumeration and processing

Molecules are generated combinatorially as SMILES by inserting character strings representing the variable parts of the structure into specific sites in a character string representing the constant part of the structure. The approach is illustrated in Table 1 while Color Plate 1 shows the graphical interface of BOOMSLANG along with the utility for defining the SMILES enumeration. The number of simultaneous insertions can be controlled, which is useful if the program is used to add substituents to a substructural template for a QSAR study, since optimization within a lead series may not involve simultaneous substitution at all accessible sites in the parent molecule. The SMILES strings are made unique and duplicates are eliminated before assigning a unique structure identifier to each SMILES.

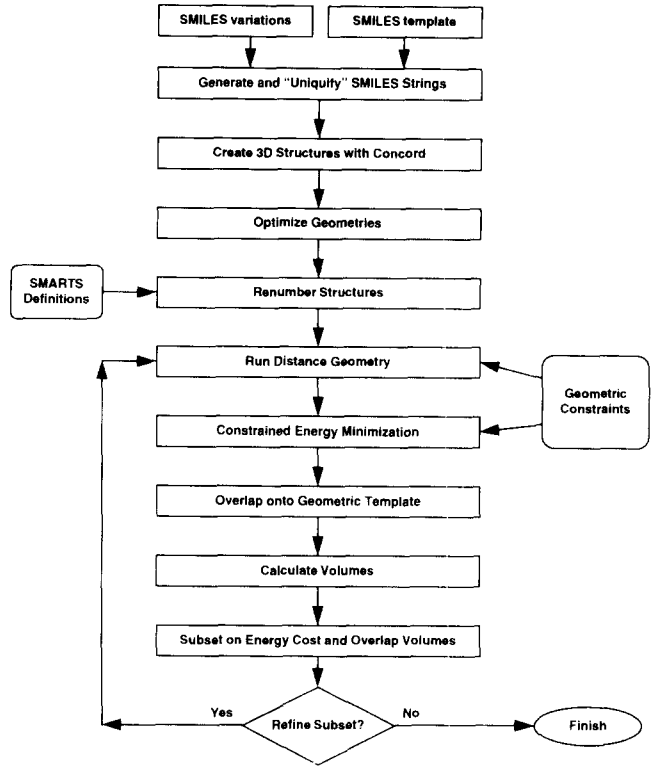


Figure 1. Overview of BOOMSLANG.

Generation of three-dimensional coordinates

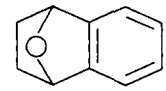
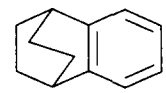
The SMILES are processed with the 3D model-building program CONCORD[28] and the resulting structures are energy minimized with AESOP, a molecular mechanics program, and the resulting energy is stored as an estimate for the global energy minimum of the molecule. The latter program, written by B.B. Masek (Indiana University, Bloomington, IN) is derived from BIGSTRN-3[29] and uses the MM2[30] force field as its default although other functional forms are supported.

Flexible overlay of structures onto geometric template by distance geometry

The molecules are forced onto the geometric template using a combination of distance geometry[31,32] and constrained molecular mechanics energy minimization. The geometric template can be a pharmacophore derived by overlaying a set of known ligands, a model of the protein active site, or the parent structure in a QSAR study. The distance geometry program, DGEOM,[33] allows considerable control over how the structures are forced onto the template and geometric objects such as exclusion zones are easily defined.

In general it will be necessary to specify explicitly the desired geometric relationship between selected substructures in the molecules that are generated. Before performing the distance geometry calculations the molecules are renumbered in terms of some of the substructural elements that they necessarily share, using the powerful SMARTS substructural specification.[20] For example, a series of molecules containing hydroxyl and amide functionality can be

Table 1. Combinatorial generation of structures as SMILES[a]

{*Insert1*}	{*Insert2*}	SMILES	Structure
CC	O	C1CC2CCC1O2	
C=C	O	C1CC2C=CC1O2	
c3ccccc3	O	C1CC2c3ccccc3C1O2	
CC	CC	C1CC2CCC1CC2	
C=C	CC	C1CC2C=CC1CC2	
c3ccccc3	CC	C1CC2c3ccccc3C1CC2	

[a]SMILES consist of a constant part (the SMILES template) and a variable part (the inserts). The SMILES strings in this example are defined by: C1CC2{*Insert1*}C1{*Insert2*}2.

renumbered, using commercially available routines, so that the hydroxyl oxygen is atom 1, the amide oxygen is atom 2, and the amide nitrogen is atom 3. Should more than one hydroxyl oxygen be present then each will be numbered as atom 1 in separate copies of the molecule. The renumbering allows the distance geometry calculation to use the same constraints file for all molecules.

Constrained energy minimization

DGEOM generates a number of conformations up to a limit set by the user and in cases of very poor fit to the template no conformations will be generated at all. The conformations are first refined by constrained energy minimization (AESOP) to meaningful energies and the difference between conformational energy and the estimate for the global energy minimum is computed to give the strain energy. Should a conformational energy be lower than the estimate for the global minimum then the latter quantity is updated appropriately.

Overlay onto geometric template

The molecular mechanics program AESOP does not preserve molecular orientation and the structures must be overlaid onto the template after constrained energy minimization. This is achieved by mapping a set of atom numbers in the molecule onto a set of atom numbers in the template using the method of Diamond.[34]

Overlap and excluded volumes

Following overlay onto the template, overlap volumes and exclusion volumes are computed, using routines from the SKINNY[14] program, for each conformation with respect to a structure that represents the spatial properties of the pharmacophore. This will normally be an overlay of a number active compounds in the same frame of reference as the geometric template. In some cases this reference structure may be identical to the template. In QSAR studies these calculations are helpful for identifying structures that have not overlaid properly, suggesting that the proposed conformation for the analogs is not accessible for these structures or that the overlay strategy is underdefined.

Summary of results and selection of subsets of structures by energy

Results are presented in tabular format listing structure identifier, conformation identifier, strain energy, overlap, and exclusion volumes, and the list can be sorted by these parameters and subsets of the structures can be selected. A viewer is also provided allowing rapid inspection of structures (Color Plate 2). If the number of synthetic targets is sufficiently small the list of structures can be output with strain energies and shape-match data. It also may be appropriate at this stage to subject the molecules to more rigorous conformational analysis to determine whether significantly lower energy conformation exist. Alternatively, the sequence of operations starting from the distance geometry can be repeated on the subset of structures, generating a larger number of conformations for each molecule that will lead to better discrimination between the molecules remaining in the list.

CONCLUDING REMARKS

We have described a simple method for controlled structure generation and an approach to ranking the molecules by their ability to satisfy geometric and steric constraints. Al-

though developed primarily for pharmacophore exploitation, BOOMSLANG finds application as a general purpose model builder in a number of areas.

ACKNOWLEDGMENTS

We thank B.B. Masek (Zeneca, U.S.A.) for making the AESOP and SKINNY codes available and our colleagues J.A. Grant, J.J. Morris, G. Palmer, A.M. Slater, D. Timms, and A.J. Wilkinson for useful discussions. A. Tester programmed the SMILES enumeration module.

REFERENCES

1 Lewis, R.A., and Leach, A.R. Current methods for site-directed structure generation. *J. Comput.-Aided Mol. Design* 1994, **8,** 467–475

2 Rotstein, S.H., and Murcko, M.A. GroupBuild: A fragment-based method for *de novo* drug design. *J. Med. Chem.* 1993; **36,** 1700–1710

3 Böhm, H.-J. The computer program LUDI: A new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol. Design* 1992, **6,** 61–78

4 Bohacek, R.S., and McMartin, C. Highly diverse structures complementary to enzyme binding sites: Results of extensive application of a *de novo* design method incorporating combinatorial growth. *J. Am. Chem. Soc.* 1994, **116,** 5560–5571

5 Caflisch, A., Miranker, A., and Karplus, M. Multiple copy simultaneous search and construction of ligands in binding sites: Application to inhibitors of HIV-1 aspartic protease. *J. Med. Chem.* 1993, **36,** 2142–2167

6 Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoehler, R.A., and Dunn, D.A. The conformational parameter in drug design: The active analogue approach. *ACS Symp. Ser.* 1979, **112,** 205–226

7 Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B., and Marshall, G.R. Constrained search of conformational hyperspace. *J. Comput.-Aided Mol. Design* 1989, **3,** 3–21

8 Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I., and Pavlik, P.A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Design* 1993, **7,** 83–102

9 Leichtfried, F.E. Novel approaches to high throughput screening automation. In: *Proceedings of the First Forum on Data Management Techniques in Biological Screening* (Carter, C., and Freter, F.R., eds.). SRI International Menlo Park, California, 1992, pp. 82–86

10 Burch, R.M. Mass ligand binding and screening for receptor antagonists: Prototype new drugs and blind alleys. *J. Receptor Res.* 1991, **11,** 101–113

11 Gallop, M.W., Barrett, R.W., Dower, W.J., Fodor, S.P.A., and Gordon, E.M. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* 1994, **37,** 1233–1251

12 Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P.A., and Gallop, M.W. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.* 1994, **37,** 1385–1401

13 Jung, G., and Beck-Sickinger, A.G. Multiple peptide synthesis methods and their applications. *Angew. Chem. Int. Ed. Engl.* 1992, **31,** 367–383

14 Masek, B.B., Merchant, A., and Matthew, J.B. Molecular shape comparison of angiotensin II receptor antagonists. *J. Med. Chem.* 1993, **36,** 1230–1238

15 Connolly, M.L. Computation of molecular volume. *J. Am. Chem. Soc.* 1985, **107,** 1118–1124

16 Böhm, H.-J. The development of a simple scoring function to estimate the binding constant for a protein-ligand complex of known structure. *J. Comput.-Aided Mol. Design* 1994, **8,** 243–256

17 Cramer, III, R.D., Patterson, D.E., and Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 1988, **110,** 5959–5967

18 Moos, W.H., Green, G.D., and Pavia, M.R. Recent advances in the generation of molecular diversity. *Annu. Rep. Med. Chem.* 1993, **28,** 315–324

19 Lauri, G., and Bartlett, P.A. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Design* 1994, **8,** 51–66

20 Daylight Chemical Information Systems, Inc., Irvine, California

21 Anderson, E., Veith, G.D., and Weininger, D. SMILES, a chemical language and information system. I. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 1988, **28,** 31–36

22 Martin, Y.C., Bures, M.G., and Willett, P. Searching databases of three-dimensional structures. *Rev. Comput. Chem.* 1990, **1,** 213–263

23 Martin, Y.C. 3D database searching in drug design. *J. Med. Chem.* 1992, **35,** 2145–2154

24 Siani, M.A., Weininger, D., and Blaney, J.M. CHUCKLES: A method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *J. Chem. Inf. Comput. Sci.* 1994, **34,** 588–593

25 Ho, C.M.W., and Marshall, G.R. DBMAKER: A set of programs to generate three-dimensional databases based on user-specified criteria. *J. Comput.-Aided Mol. Design* 1995, **9,** 65–86

26 TRIPOS Associates, 1699 South Hanley Road, St. Louis, Missouri

27 Weininger, D., Weininger, A., and Weininger, J.L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* 1989, **29,** 97–101

28 Rusinko III, A., Skell, J.M., Balducci, R., McGarity, C.M., and Perlman, R.S. *CONCORD: A Program for the Rapid Generation of High Quality Three-Dimensional Structures.* The University of Texas at Austin and TRIPOS Associates, 1699 South Hanley Road, St. Louis, Missouri, 1988

29 Nachbar, R.B., and Mislow, K. BIGSTRN-3: A general-purpose empirical force-field, QCPE Number 514, Quantum Chemistry Program Exchange, Indiana University, Bloomington, Indiana

30 Allinger, N.L., and Yuh, Y.H. MM2: Molecular Mechanics II, QCPE Number 395, Quantum Chemistry Program Exchange, Indiana University, Bloomington, Indiana
31 Crippen, G.M. *Distance Geometry and Conformational Calculations* (Bawden, D., ed.). Research Studies Press (Wiley), New York, 1981
32 Crippen, G.M., and Havel, T.F. *Distance Geometry and Molecular Conformation* (Bawden, D., ed.). Research Studies Press (Wiley), New York, 1988
33 Blaney, J.M., Crippen, G.M., Dearing, A., and Dixon, J.S. DGEOM: Distance geometry, QPCE Number 590, Quantum Chemistry Program Exchange, Indiana University, Bloomington, Indiana
34 Diamond, R. A note on the rotational superposition problem. *Acta Crystallogr.* 1988, **A44,** 211–216