Journal of Computer-Aided Molecular Design, 6 (1992) 513–520
ESCOM

J-CAMD 180

Similarity screening of molecular data sets

A.C. Good[a], E.E. Hodgkin[b] and W.G. Richards[a,]*

[a]*Physical Chemistry Laboratory, Oxford University, South Parks Road, Oxford OX1 3QZ, U.K.*
[b]*British Biotechnology Ltd, Watlington Road, Oxford OX4 5LY, U.K.*

Received 5 June 1992
Accepted 27 July 1992

Key words: Similarity; Electrostatic potential; Shape; 3D-database searches

SUMMARY

Three-dimensional (3D)-database searches are now being widely applied to determine potential new active molecules. Many structural data sets obtained as a result of these searches are still large in size. In this paper we apply molecular similarity calculations as a rapid method to screen two such data sets. In the first investigation, synthetic candidates, produced as a result of a tendamistat β-turn mimic search, were tested for their ability to imitate the β-turn backbone. In the second study, structures extracted through a histamine pharmacophore query search were examined on the basis of their electronic similarity to histamine. Molecular similarity is shown to provide a rapid means of gaining insight into the composition of molecular data sets, with possible implications for future full 3D-database searches.

INTRODUCTION

Molecular similarity calculations have become established as a method for both generating parameters in quantitative structure–activity relationships and finding the optimum position for superimposed structures [1–12].

As originally introduced by Carbo [1–2],

$$R_{AB} = \frac{\int P_A P_B dv}{(\int P_A{}^2 dv)^{1/2} (\int P_B{}^2 dv)^{1/2}}$$

molecular similarity R_{AB} is determined from the structural properties P_A and P_B of the two molecules being compared. The numerator measures property overlap, while the denominator normalizes the similarity result. Quantum mechanically derived electron density was initially used [1–2] as the structural property P. The technique has since been extended to cover electrostatic potentials, electric fields and shape [3–12].

* To whom correspondence should be addressed.

With the most widely used software of this type [7], molecules are surrounded by a rectilinear grid, the structural property is evaluated at each intersection and the integrals are evaluated numerically. Recently, work has been carried out using Gaussian functions to elucidate electrostatic potential and shape similarity [11,12]. These functions permit rapid evaluation of analytical integrals, thereby significantly enhancing the speed of similarity calculations. Here we use this new analytical evaluation technique to extend the application of electrostatic potential similarity calculations to data set screening.

EXPERIMENTAL

Two investigations were undertaken using molecular similarity to evaluate molecular data sets.

Study 1

For the first investigation, the β-turn of tendamistat implicated in inhibition of α-amylase [13] was used as the basis for a 3D-database search. The substructure search query extracted from the β-turn is shown in Fig. 1. This structure has been the subject of a similar 3D-database search by Bartlett et al. [13]. A version of the Cambridge Structural Database [14] (CSD), converted for use in MACCS 3D [15,16] database search software, was used as the search data set. The β-turn was extracted from the PDB [17] file 4AIT using the SYBYL [18] modelling package. The required geometric parameters were measured and the query built within MACCS 3D [15]. The search resulted in a data set containing five structures. Two reasons may account for the small size of the data set obtained compared with that determined by Bartlett et al. [13]. Firstly, the query tolerances set were low. Secondly, the MACCS 3D [15] version used did not have the ability to treat hydrogens, thereby removing a large number of bond vectors that could potentially fit the query. Four of the hits obtained were complex sugar molecules which would be difficult to synthesize. The fifth hit (CSD reference code HXBZLA), however, possessed the required geometric skeleton

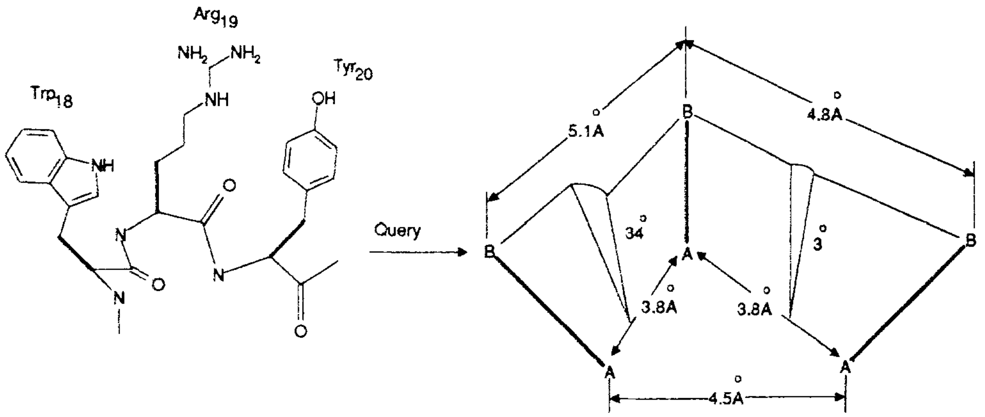
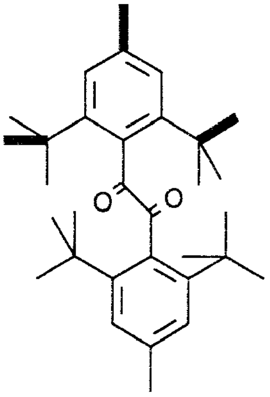


Fig. 1. The 3D substructure query extracted from the α-amylase-inhibiting tendamistat β-turn used in study 1. The query constraints shown were determined from the C_α–C_β side-chain bonds of the β-turn. Query tolerances used were ± 0.2 Å for distances and ± 10 degrees for torsions. A and B atoms could be of any type, but the B atoms could not be part of a ring structure.



HXBZLA

Fig. 2. Structure extracted from MACCS 3D[15] search of the Cambridge Structural Database[14] (CSD) using the β-turn mimic query, together with the internal CSD reference code.

within a relatively simple structure (Fig. 2), and was thus considered suitable as the basis for the design of a β-turn mimic.

Part of this design process was to determine which hetero atom and group substitutions would best allow the molecule to mimic the electrostatic properties of the β-turn backbone. This had to be achieved while maintaining the side chains in the required orientation. To this end, the structure was transferred as a MOL file from MACCS 3D [15] into SYBYL [18]. The molecule was then simplified and substitution points defined as shown in Fig. 3. The substitutions used are shown in Table 1. The substitutions at points in the flexible regions of the molecule (2 and 4) were determined by further MACCS 3D [15] searches of the CSD [14], using the substructure queries shown in Fig. 4. Only hetero atoms and groups found to maintain the required template geometry were used.

Once the template structure had been built and the possible substitutions defined (Fig. 4 and Table 1), the in-house SYBYL PROGRAMMING LANGUAGE [18] (SPL) macro SUBS was run. This macro takes as its input the name of the template structure (i.e., the name of the Fig. 3 structure) and a set of SYBYL [18] log file names. The main group of log files contain the required SYBYL [18] commands to undertake each individual substitution. A fitting log file is also re-

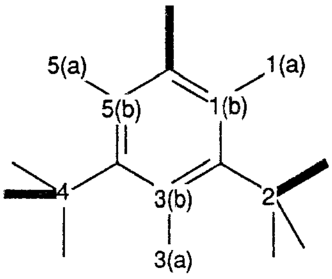


Fig. 3. Structural template extracted from HXBZLA (see Fig. 2) with substitution points shown.

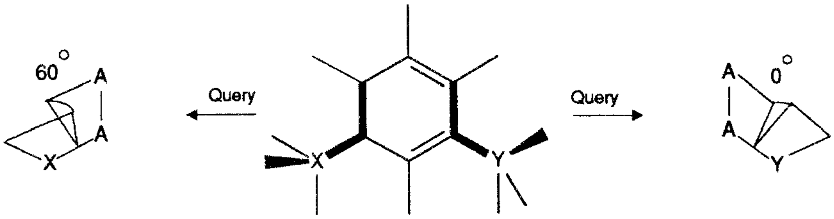


Fig. 4. Substructure queries extracted from the HXBZLA template (see Fig. 3). These queries were used to determine possible substitutions at positions 2 and 4 of the template structure. The torsion query values were obtained from the bonds highlighted in the template structure. The query tolerance used was ± 10 degrees. A atoms had to be aromatic carbons, while the X and Y atoms could be of any type except carbon and hydrogen.

quired to superimpose each resultant structure on top of the lead molecule. The side-chain equivalent atoms of the substituted template were fitted onto the matching side-chain atoms of the β-turn. The lead molecule used was the β-turn, here modified to three alanines. This modification was undertaken so as to remove the effects of the side chains when studying the electrostatic properties of the backbone. C_α and C_β atoms in the β-turn were superimposed onto the equivalent template structure bonds. Given this information, SUBS automatically builds, minimizes and calculates Pullman charges [19] and fits onto the lead each possible substituted molecule. All resultant structures are added to a SYBYL [18] molecular database. A database of 160 structures was created here.

Once SUBS had been run, the in-house SPL program SIMGET was run. SIMGET takes a database of molecules and systematically writes out each structure as a CSSR [14] file. The structure is then compared with the lead molecule, using a modified version of the ASP [7] software. After each similarity calculation, SIMGET extracts the similarity result and places the information, together with a number defining the position of the structure in the database, into a SYBYL [18] table. When the whole database has been processed, a scatter plot of similarity result against structure number is produced. Here a scatter plot with similarity values ranging from 0.004 to 0.331 was created. The structure shown in Fig. 5 showed the highest similarity when compared to

TABLE 1
SUBSTITUTIONS DETERMINED FOR THE HXBZLA-DERIVED STRUCTURAL TEMPLATE (SEE FIG. 3)

Substitution point[a]	Hetero atoms groups used
1 (a)	F
1 (b)	N
2	O, NH_2
3 (a)	F
3 (b)	N
4	O (SO_2)[b]
5 (a)	N
5 (b)	F

[a] Fluorine and nitrogen substitutions were paired since the substitutions are mutually exclusive.
[b] Not used as Pullman charges were not available.

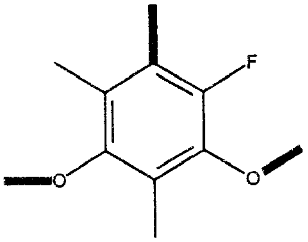


Fig. 5. Substituted template structure showing highest similarity to β-turn backbone. Beta turn side-chain equivalent bonds are shown as thick lines.

the β-turn backbone. In general, substitutions at points 1(a), 2 and 4 (see Fig. 3) increased similarity, while nitrogen substitutions into the benzene ring were found to decrease similarity.

Study 2

For the second study, the histamine pharmacophore [20] shown in Fig. 6 was used as the basis for a 3D substructure search. The Chapman and Hall Dictionary of Drugs [21], converted to run in the CHEM-DBS 3D module of CHEMX [22] modelling software, was used as the search data set.

The substructure was created by building histamine within CHEMX [22], using standard bond lengths and angles. The resultant structure was optimized by using the CHEMX [22] molecular mechanics force field. Atoms not required for the pharmacophore description were then removed. The required geometric constraints were then set (Fig. 6) and the search executed. An initial fast screen search [23] was undertaken to narrow down the range of possible candidate structures. An exact geometric search was then carried out, using the above constraints, to produce the final answer set; 104 hits were obtained in the final data set.

The data set was then processed with the in-house CHEM-LIB [22] program MS. MS takes the results from a CHEM-DBS 3D [22]-database search, extracting each molecule in turn from the resulting CHEM-DBS [22] data set. For each structure extracted, a similarity calculation is made against a given lead compound (in this case histamine), and the result stored as a property field in the data set. CHEM-DBS 3D [22] automatically carries out a rigid fit of each hit between the atoms defining and meeting the geometric search constraints, before adding the structure to the

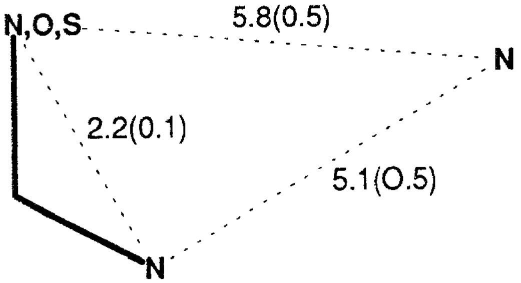
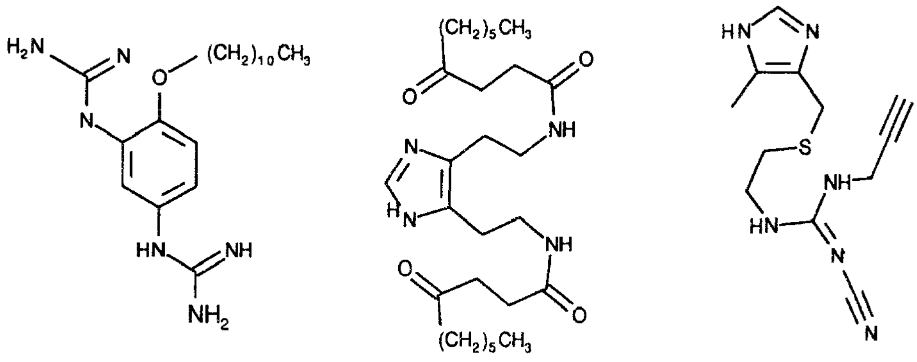


Fig. 6. Histamine pharmacophore substructure query used in study 2. Distance query tolerances shown in brackets.



HOO77KO LOO16KO EOO237KO

Fig. 7. Structures from the results data set from study 2 with a Carbo similarity index value of greater than 0.5 to histamine, together with the internal Chapman and Hall reference codes.

data set. Superposition is thus achieved automatically. Point charges were provided for each structure through Gasteiger [24] charge calculations.

Once the MS run was finished, a property field search was then undertaken on the molecular similarity values stored within the data set. Only hits with a molecular similarity value of 0.5 or greater were accepted. A three structure data set was obtained from this search. The structures are shown in Fig. 7.

DISCUSSION

These studies show that molecular similarity calculations can be used to gain insights into widely differing 3D-search data set problems. In the first study, a molecular design exercise, similarity calculations were used to determine which template modifications would provide the best synthetic β-turn mimic candidate. In the second study, a straightforward database screen, similarity evaluations were used to find the pharmacophore-matching structures that best mimicked the lead molecule used in the pharmacophore construction. Although no experimental tests were made regarding result accuracy, it is of interest that the molecule EOO237KO, detected in study 2, is a known histamine antagonist [21].

There are currently two major problems regarding these data set calculations. The problems relate directly to the conditions required to undertake similarity evaluations. Before a similarity calculation can be made, a method for superposition of lead and database molecule is required. This produces the first difficulty. To achieve superposition, the structural features that make up the database query were used. Thus an initial database search is necessary before similarity evaluations can be made. Some of the restrictions inherent to these searches, for example the reliance on geometric pharmacophore-based queries, therefore carry forward into the results data sets. It would be preferable if the result of the similarity evaluation could be made the major criterion of the 3D search query itself. This could be achieved by an initial superposition of structures, by making use of properties such as dipole moments, mass, biased mass or volume. The initial superpositions could then be optimized on the basis of molecular similarity, although this would produce a significant increase in search time. Database searching on the basis of electronic and steric rather than just geometric similarity would then be possible.

The use of empirical charge calculations introduces our second problem. Previous studies [25] have shown that quantum mechanically derived charges, for example those obtained from molecular electrostatic potential calculations (MEP) [26], generally provide a more reliable basis for similarity evaluations. However, since more accurate charge calculations are time-consuming, quick and less accurate methods, such as those of Gasteiger and Marsili [24] or Berthold and Pullman [19], are used. The way around this would be to determine and store MEP-derived charges when the databases are keyed. This only becomes feasable if a fast method for MEP charge evaluation exists which does not massively increase keying time (the VAMP [27] software of Clark et al. may be useful in this respect).

Once these problems are overcome, the potential for molecular similarity query-based database searches is clear. This has been made possible by the speed of the Gaussian function-based similarity calculations. With regard to time, the SIMGET program used in study 1 took approximately 5 min to run on a Silicon Graphics Iris 4D-20. Although not particularly fast, it should be noted that much of this time is required because of the use of a functional programming language. This is evident when we consider that only 25 s are required to carry out the equivalent 159 similarity calculations from directly within ASP [7]. This is the speed expected if the similarity evaluations are directly integrated into the search software.

As has already been implied, corresponding evaluations are now possible using shape similarity [12]. Shape similarity calculations have fewer inherent difficulties associated with them, since no charge calculations are required for successful elucidation. The use of shape similarity-based database searches would allow the extension of the work undertaken by Desjarlais et al. [28] to systems where the shape of the active site is not known.

CONCLUSIONS

The purpose of this paper was to determine the possible utility of molecular similarity calculations as a tool for searching 3D databases. The two studies undertaken showed that molecular similarity calculations provide the basis for quick and automated data set screening. The method is currently restricted to screening hits from 3D geometric searches. However, the potential exists for the direct use of similarity calculations as part of 3D search queries. This could greatly increase the flexibility of these searches, removing the reliance on purely geometric similarity queries.

ACKNOWLEDGEMENTS

This work is supported by an SERC CASE award from British Biotechnology Ltd, Oxford, U.K.

REFERENCES

1 Carbo, R., Leyda, L. and Arnau, M., Int. J. Quantum Chem., 17 (1980) 1185.

2 Carbo, R. and Domingo, L., Int. J. Quantum Chem., 32 (1987) 517.

3 Hodgkin, E.E. and Richards, W.G., Int. J. Quantum Chem. Quantum Biol. Symp., 14 (1987) 105.

4 Bowen Jenkins, P.E. and Richards, W.G., Int. J. Quantum Chem., 30 (1986) 763.

5 Burt, C. and Richards, W.G., J. Comput.-Aided Mol. Design, 4 (1990) 231.

6 Burt, C., Huxley, P. and Richards, W.G., J. Comp. Chem. 11 (1990) 1139.

7 Automated Similarity Package, Oxford Molecular Ltd, The Magdalen Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, U.K.

8 Richard, A.M., J. Comp. Chem., 12 (1991) 959.

9 Meyer, A.M. and Richards, W.G., J. Comput.-Aided Mol. Design, 5 (1991) 426.

10 Burt, C., Huxley, P. and Richards, W.G., J. Comp. Chem., 11 (1990) 1139.

11 Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Chem. Inf. Comput. Sci., 32 (1992) 188.

12 Good, A.C. and Richards, W.G., J. Chem. Inf. Comput. Sci., in press.

13 Bartlett, P.A., Shea, G.T., Telfer, S.J. and Waterman, S., Caveat: A Program to Facilitate the Structure-derived Design of Biologically Active Molecules, In Roberts, S.M. (Ed.), Molecular Recognition: Chemical and Biochemical Problems, Royal Society of Chemistry, Spec. Public. No. 78, 1989, pp. 183–196.

14 Cambridge Crystallographic Database, University Chemical Laboratory, Lensfield Road, Cambridge CB2 IEW, U.K.

15 Molecular ACCESS System, Molecular Design Ltd, San Leonardo, CA, USA.

16 Baber, J.C. and Hodgkin, E.E., J. Chem. Inf. Comput. Sci., in press.

17 Protein Data Bank, Bernstein, F.C., J. Mol. Biol., 112 (1977) 535.

18 SYBYL, Tripos Associates Inc, 1699 S. Hanley Rd., Suite 303, St Louis, Missouri 63144, USA.

19 Berthold, H. and Pullman, A., J. Chem. Phys., 62 (1965) 942.

20 Ganellin, C.R., Chemistry and Structure–Activity Relationships of Drugs Acting at Histamine Receptors, In Ganellin, C.R. and Parsons, M.E. (Eds.), Pharmacology of Histamine Receptors, Wright-PSG, London, 1982, 35–37.

21 Elks, J. and Ganellin, C.R., Chapman and Hall Dictionary of Drugs. Chemical data, structure and bibliography. Cambridge University Press, 1990.

22 CHEMX, Chemical Design Ltd, Unit 12, 7 West Way, Oxford OX2 OJB, U.K.

23 Murrall, N.W. and Davies, E.K., J. Chem. Inf. Comput. Sci., 30 (1990) 316.

24 Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.

25 Burt, C. and Reynolds, C.A., J. Am. Chem. Soc., submitted.

26 Ferenzcy, G., Reynolds, C.A. and Richards, W.G., J. Comp. Chem., 11 (1990) 159.

27 Clark, T., Erlangen University, Germany, private communication.

28 Desjarlais, R.L., Seibel, G.L., Kuntz, I.D., Furth, P.S., Alvarez, J.C., Ortiz de Montellano, P.S., Decamp, D.L., Babé, L.M. and Craik, C.S., Proc. Natl. Acad. Sci. USA, 87 (1990) 6644.