Journal of Computer-Aided Molecular Design, 9 (1995) 341–350
ESCOM

J-CAMD 297

The atom assignment problem in automated de novo drug design. 1. Transferability of molecular fragment properties

M.T. Barakat and P.M. Dean*

Drug Design Group, Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K.

Received 1 February 1995
Accepted 23 May 1995

Keywords: De novo drug design; Molecular electrostatic potential; Molecular complementarity; Molecular similarity

Summary

This paper is the first of a series which examines the problems of atom assignment in automated de novo drug design. In subsequent papers, a combinatoric optimization method for fragment placement onto 3D molecular graphs is provided. Molecules are built from molecular graphs by placing fragments onto the graph. Here we examine the transferability of atomic residual charge, by fragment placement, with respect to the electrostatic potential. This transferability has been tested on 478 molecular structures extracted from the Cambridge Structural Database. The correlation found between the electrostatic potential computed from composite fragments and that computed for the whole molecule was encouraging, except for extended conjugated systems.

Introduction

Automated computer-aided drug design is a rapidly expanding field of research that holds much promise for lead generation. The research can be divided into two applications. First, site-directed drug design algorithms can be written for a user who has crystallographic coordinates of the site. This design strategy takes all the Cartesian coordinate information from the site, such as the disposition of electrostatic potential, hydrophobicity, hydrogen-bond possibilities and the steric surface, and creates de novo a putative ligand to fit the site with complementary properties. Second, envelope-directed design is a strategy where the input information consists of a supersurface from a set of similar drug molecules, often gleaned from competitors' structures. The algorithm attempts to design a novel structure within the envelope, with the desired properties resembling those obtained from the similarity study. There are obvious differences in input for the two strategies; nevertheless, there is a strong congruence in their approach to structure generation.

De novo generation of molecular structures is plagued by combinatoric steps in many parts of the generation process. A variety of strategies have been elaborated for the many approaches to structure generation, in order to reduce the number of structures that need to be evaluated by the user. The simplest conceptual approach is that used by CAVEAT [1], where the site is specified by a number of site points in 3D space. Although this program does not strictly perform de novo design, it does use a number of concepts to reduce combinatorial problems. In the case of hydrogen-bonding points, vectors can be projected from the sites along the direction of the hydrogen bond to create a set of ligand points. These points are then used as a screen for searching databases of molecular structures to find corresponding hits with appropriate atoms well disposed. These hits can then be used as a basis for the development of a chemical idea into a novel ligand. If no hits are found, then the site can be partitioned and hits sought for the partition. The hits may then be linked together to make a composite structure. Programs like CAVEAT produce a small number of hits of known molecular structures; this is an advantage in assessing ease of synthesis. If the query is poorly styled, the routine may return a large number of hits, for example 1,4 substitutions on a benzene ring. A related alternative, but rather more complex procedure, is to use

*To whom correspondence should be addressed.

hydrogen-bond vectors as input to a constrained structure generator. The vectors are used as seeds to be incorporated into an evolving structure within the site, or envelope; the program LUDI operates through this strategy [2]. An alternative method for creating seed points is to use the GRID program [3] to highlight favourable positions of small chemical groups close to the contact surface of the site. Functional groups can then be linked together using a database of molecular skeletons, for example by the HOOK program [4]. These methods produce a large variety of structures which have to be assessed, usually by a scoring scheme. A different approach, exemplified by the DOCK program [5], is to consider the packing of molecular fragments into the site, or envelope. Small molecular fragments can be approximated by spheres of different radius. A variety of spheres can be moved about in the site to try to optimise their packing and the spheres can be rotated to allow bonding patterns between the fragments. The molecular structures may then be linked together by BUILDER and assessed by a scoring scheme [6].

The methods used in structure generation are either atom- or fragment-based. With atom-based methods, for example with the program LEGEND [7,8], the properties of the evolving structure need to be computed for each atom addition or deletion. If molecular properties are transferable, it should be possible to build large segments of structure without the need for large-scale recalculation. Fragment-based methods, such as LUDI, GROW [9] and GroupBuild [10], are therefore potentially more efficient for automated structure generation. A fundamental question that has to be asked is: how transferable are the fragment properties? Furthermore, if fragment-based methods are to be used, can the combinatoric steps be partitioned into classes of procedures to be optimized separately?

Consider the equation for molecular interaction [11]:

$$E = \sum_i \sum_{j \neq i} \left(-\frac{A_{ij}^{ab}}{r_{ij}^6} + \frac{B_{ij}^{ab}}{r_{ij}^{12}} + \frac{C_{ij}^{ab} q_i q_j}{r_{ij}} \ldots \right) \qquad (1)$$

where A, B and C are fitting constraints between atom i of one molecule and atom j of the other, obtained from quantum chemistry calculations by regression analysis, r_{ij} is the interatomic distance, q_i and q_j are the atomic residual charges, and superscripts a and b refer to the atom type and its class. The first term in the interaction is dispersion attraction, the second represents exchange repulsion, while the third term is Coulombic. From the indices of the r term, it can be seen that the first two terms are short-range and the third term (electrostatic) is long-range. Moreover, the numerator is compounded of atom properties, whilst the denominator is composed only of powers of r_{ij}; the denominator in each term can thus be considered as a distance geometry parameter and determines the geometric fit between the two molecular structures.

The notion of transferable molecular properties, termed the theory of Atoms in Molecules (AIM), has been studied extensively by Bader [12] and was recently summarised for drug design by Popelier [13]. AIM has important consequences for molecular similarity and complementarity. If molecules are partitioned into fragments of charge density, they can be analysed in terms of the gradient vector field and the Laplacian of the charge density. These lead to a topological description of the molecule. It appears that the surface properties of functional groups are readily transferable. With peptides, the properties of the peptide bond are essentially independent of the neighbouring groups [14].

Therefore, it might be possible to consider structure generation as a problem composed of two broad categories of combinatoric processes, i.e., the generation of 3D molecular graphs, which determine the geometry of the structure, and the placement of atoms on the graphs, using molecular fragments, to create a molecular structure. Each process could be optimized separately. Programs for the generation of 3D molecular graphs within the confines of a site have been developed [15,16]. The current series of papers investigates the optimization of atom assignment onto 3D molecular graphs [17–20] and focusses attention predominantly on the electrostatic term. Our objective has been to derive a single method which optimizes the electrostatic potential to be either *complementary to the site* or, in the absence of the site, to be *similar to a reference potential* on the enveloping supersurface of a set of similar molecules.

Electrostatic complementarity between proteins and ligands has recently been studied systematically on 24 high-resolution ligand co-crystal complexes [21–23]. A number of features relevant to drug design problems can be highlighted from this work. First, there is no complementarity between the atomic residual charges of adjacent atoms in contact across the interface between the ligand and the protein; furthermore, there is no significant complementarity between charges across the interface up to 5.89 Å. These findings are valid, independent of whether or not there is strong complementarity in the electrostatic potentials of the ligand and the site. Second, the electrostatic potential complementarity between a ligand and its site is closely similar for the interfacial surface and the whole molecular surface. Third, the use of a distance-dependent dielectric does not appreciably affect the complementarity. Fourth, if the ligands are partitioned into smaller molecular groups, the complementarity in potentials of the groups does not compare well with that of the whole ligand with the site. The whole ligand complementarity is neither a mathematical mean of the group complementarities, nor is there a multiplicative relationship between group complementarity and that of the whole

structure. Fifth, complementarity is a complex relationship which appears to be dominated by: (i) the shape of the ligand and site surfaces; and (ii) the geometric disposition of dominant charges. A simple mathematical model can be constructed to describe the general features of complementarity. However, the model is not sufficiently sophisticated to be used directly in an objective function for complementary atom placement.

One approach to the creation of a complementary electrostatic potential, on the surface of a 3D molecular graph positioned in a site, is to use the image–charge method of the YING program [24,25]. In this technique, the electrostatic potential of the site projected onto the putative ligand surface is taken, using a semi-regular array of points [26], and then an optimal set of Mulliken point charges is determined at the vertices of the molecular graph, such that the complementarity in electrostatic potential is maximized. These are notional charges with no atomic identity. Atoms then have to be placed on the graph so that the atomic residual charges are matched, as well as possible, with the image charges; at the same time the atom placement has to fulfil the valence bond constraints with neighbouring atoms to create a sensible chemical structure.

In this series of papers we develop an alternative procedure which is based on using a comprehensive library of molecular fragments with their charges predetermined [27–29]. Fragments are randomly placed onto the 3D molecular graphs, so that the node requirements of the graph are correctly fulfilled at the placement node. The complementarity between the electrostatic potential of the site projected onto the graph's surface and the potential evolved from the placement process is computed for each placement. Trial fragment placements are made until the complementarity is optimized. In the case where there is no coordinate information about the site, an analogous procedure is used. Instead of an objective function using complementarity, similarity between the evolving surface electrostatic potential and the target potential on the supersurface is used. Paper 1 outlines the approach to fragment placement and examines the transferability of the fragments with respect to electrostatic properties. Paper 2 [17] describes the method of molecular graph perception and fragment placement. Paper 3 [18] enumerates the optimization of fragment placement. Paper 4 [19] describes extensive tests of the method of fragment placement for 3D molecular graphs in known crystallographic sites; complementarity in electrostatic potential is used in the objective function to be optimized. Paper 5 [20] illustrates how the method may be applied to envelope-directed fragment placement using similarity in the objective function.

TABLE 1
THE VARIOUS CONNECTIVITIES ASSOCIATED WITH EACH ATOM

Connectivity	Atoms
1	H, N, O, F, S, Cl
2	C, N, O, S
3	C, N
4	C, N^+, P

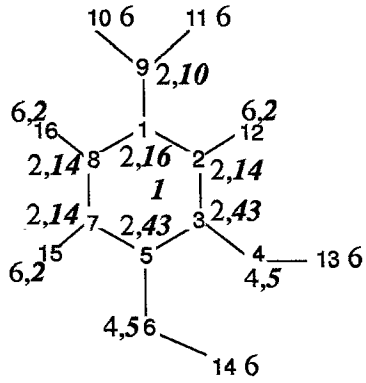


Fig. 1. The combinatorics of atom assignment, using the two approaches illustrated on the molecular graph of 3,4-dihydroxybenzoate. The vertices are numbered in Helvetica small type. (i) The atom-by-atom method; numbers (Roman type) on the graph are the numbers of possible atom placements at the vertices. (ii) The fragment-by-fragment method; numbers on the graph (Italic bold type) are the actual numbers of fragments to be placed at each node, including the necessary rotations; details are given in Table 2.

Theory

This section considers the combinatorial problem associated with atom assignment onto molecular graphs, and outlines two possible approaches for tackling the problem. The decision about which assignment of atoms is the best, for an automated drug design paradigm, depends on how well the predicted properties of the postulated ligand compare with the desired set of properties.

Alternative atom assignment methods

When presented with a molecular graph, there are two ways in which atoms may be assigned to the skeleton, i.e., using either an atom-by-atom or a fragment-by-fragment procedure. For both cases, let the set of possible atoms for use on the skeleton be {H, C, N, O, F, P, S, Cl}. This set of atoms is associated with the connectivity features shown in Table 1.

Atom-by-atom assignment

In the atom-by-atom placement method, each vertex of the molecular graph can be assigned any one of the atoms in the allowed atom set. Consider the simple 16-vertex molecular graph of the 3,4-dihydroxybenzoate molecule. Figure 1 shows the ways in which atoms can be placed. The number of placements, N, is given by:

TABLE 2
NUMBER OF FRAGMENTS ALLOWED AT EACH OF THE 13 NODES IN THE 3,4-DIHYDROXYBENZOATE GRAPH, AND THE TOTAL NUMBER OF FRAGMENT ROTATIONS

Node	Vertices of node	No. of fragments allowed	Total no. of fragment rotations
1	1, 2, 8, 9	5	16
2	2, 12, 1, 3	9	14
3	3, 4, 2, 5	11	43
4	4, 13, 3	5	5
5	5, 6, 3, 7	11	43
6	6, 14, 5	5	5
7	7, 15, 5, 8	9	14
8	8, 16, 1, 7	9	14
9	9, 10, 11, 1	7	10
10	7, 8, 1, 2, 3, 5	1	1
11	15, 7	2	2
12	16, 8	2	2
13	12, 2	2	2

See Fig. 1 for the numbering scheme. The total number of possible arrangements with the fragments is 1.6×10^{11}.

$$N = \prod_{i=1}^{n} m_i \tag{2}$$

where n is the total number of vertices and m_i is the number of atoms allowed at vertex i. It should be noted that most of the 570 million arrangements for the 16-vertex graph illustrated will be rejected because of bonding violations or chemical instability.

In addition to the very large number of possibilities, each arrangement will require calculation of some properties used to optimize the objective function, unless averaged values are taken from precomputed tables.

Fragment-by-fragment assignment

In order to overcome the difficulties with the atom-by-atom approach, a fragment-by-fragment method was considered in which small, frequently occurring molecular fragments are placed onto a molecular graph. The main advantage of such a procedure is the fact that each of the fragments has precalculated atomic properties (e.g., atomic residual charges). The properties can then simply be transferred onto the molecular graph each time a fragment is used. The combinatorics of fragment placement, N, are given by the equation:

$$N = \prod_{i=1}^{j} \sum_{f=1}^{g_i} r_f \tag{3}$$

where the node i can be assigned a fragment f from a set containing g_i members; there are r_f orientations of the fragment f at that node. Figure 1 and Table 2 illustrate the combinatorics of fragment-by-fragment placement. The fragments are taken from the set described earlier [27–29]. Fragment overlap or duplication is allowed, for example, node 10 has six vertices and only one fragment from the library fits it, namely a phenyl ring; whereas node 7, composed of four vertices (three of which are part of node 10), can be fitted by nine fragments from the library. This allows rings to be treated either as entities in themselves, or as composites built up from other fragments. Worked examples of fragment placement are given in more detail in the following paper [17].

Again, inter-fragment connectivity has not been taken into consideration. The increase in the number of possibilities for placement to 1.6×10^{11} is due to the existence of fragments in which the atom types are the same, but some property of the fragment is different (e.g., bond-order or charge); furthermore, fragments can be placed on the graph in distinct orientations. Since this limitation of an increased number of placements is small for large optimization problems compared to the major advantage of precalculated properties, the fragment approach was chosen for the placement of atoms onto a molecular graph.

Optimization of combinatorial problems

Atom assignment onto molecular graphs falls into the group of combinatorial problems where the configurational space is discrete but very large [30]. The aim is to find the configuration or state, s_{opt}, which has the optimum value of the specified objective function, E, given the set S of all possible discrete configurations s. This can be done by optimizing (minimizing) the objective function, E (see Eq. 4):

$$E_{s_{opt}} = \min \{E_s | s \in S\} \tag{4}$$

A method for optimization is given in paper 3 [18].

Methods

Properties for the atom assignment problem

To decide which of the many possible arrangements of atoms on a molecular graph is the most favourable, the properties which the atoms impart to the molecular graph need to be considered. These are: atomic residual charges, hydrogen bonding and atomic hydrophobicity. The latter two features are incorporated into the atom assignment routines as extra options that can be specified by the user.

The van der Waals radii were taken from Bondi's X-ray diffraction data [31]. Surface points were generated by gnomonic projections of a tessellated icosahedron [26]. The total number of points using this method is related to the tessellation frequency, v; this determines the maximum number of points, V, allowed per atom (Eq. 5):

$$V = 2 + 10\mathrm{v}^2 \tag{5}$$

For the work described here, the smallest tessellation frequency used is 1, and the largest is 5, giving a maximum of 12 and 252 points per atom, respectively.

The electrostatic potentials were computed at the van der Waals surface placed on the 3D molecular graph, using Mulliken partial charges taken from the fragment library of charges [29].

In the atom assignment problem, only 'classical' hydrogen bonds were considered, where the atom (X) covalently bonded to the hydrogen is clearly electronegative. It is more difficult, however, to incorporate the more 'esoteric' and weak hydrogen bonds into the design process, where atom X is e.g. carbon (or P, As, Si, Ge), and may even carry a positive charge, or where the acceptor is an aromatic ring. This difficulty stems from the weakly directional and less specific nature of the bond in the latter group.

Hydrophobic potentials were computed by the method proposed by Fauchère, Quarendon and Kaetterer [32]. Φ_j is the hydrophobic potential (in kJ mol^{-1}) at point j:

$$\Phi_j = -2.3RT \sum_{i=1}^{n} f_i e^{-r_{ij}} \tag{6}$$

f_i is the fragmental hydrophobicity value of group i, r_{ij} is the distance between group i and point j, n is the number of groups, R is the gas constant (in kJ mol^{-1} K^{-1}) and T (in K) is the absolute temperature. For the work in this series of papers, the atomic hydrophobic values of Ghose's group [33] were chosen for the potential calculation and placed in the fragment library.

Matching methods for complementarity

When considering interaction properties, a way of assessing the complementarity/similarity of the interaction is needed. Given a set $X=\{x_1, x_2 \ldots x_i \ldots x_n\}$ of n properties from a site, and a set $Y=\{y_1, y_2 \ldots y_i \ldots y_n\}$ of n corresponding properties from a ligand, there are several statistical methods which can be used for the assessment between the two.

Pearson's correlation coefficient, r

Pearson's correlation coefficient is the most widely used measure of association between variables that are ordinal or continuous, rather than nominal [34]. The method correlates the actual values of corresponding pairs, and is given by Eq. 7.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{7}$$

where $\bar{x}$ denotes the mean of the x_i's, and $\bar{y}$ is the mean of the y_i's. For perfect complementarity, $r=-1$, while for perfect similarity $r=1$.

Spearman's rank correlation coefficient, r_s

This method relies on the numerical difference of ranks between the two sets of data. Let set R be the ranks of the set $\{x_1, x_2 \ldots x_i \ldots x_n\}$, and let S be the ranks of the set $\{y_1, y_2 \ldots y_i \ldots y_n\}$. Then Spearman's rank correlation coefficient, r_s, is the linear correlation coefficient of the ranks (Eq. 8).

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i (S_i - \bar{S})^2}} \tag{8}$$

where R_i and S_i are the ranks of x_i and y_i, respectively, and $\bar{R}$ and $\bar{S}$ are the mean ranks in sets R and S, respectively. For perfect complementarity, $r_s=-1$, while for perfect similarity $r_s=1$.

Kendall's rank correlation coefficient, τ

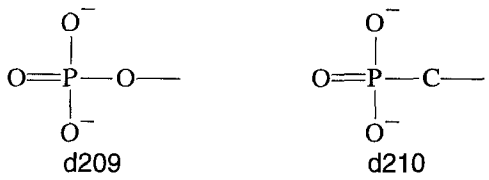
Kendall's τ uses the relative ordering or ranks rather than the numerical difference of ranks. For this reason, there is more information loss than with Spearman's rank, and hence it is more robust. The data points (x_i,y_i) are paired via a subdiagonal matrix to obtain all $n(n-1)/2$ pairings. Then, for each pair, if the relative ordering of the two x values is the same as the relative ordering of the two y's, the pair is *concordant*. If the reverse is true, the pair is termed *discordant*. If there is a tie in the two x values, the pair is called *extra y*; if the tie is in the two y values, the pair is called *extra x*. If the tie is in both x and y values, the pair is ignored. One can then calculate Kendall's τ from Eq. 9:

$$\tau = \frac{\text{concordant} - \text{discordant}}{\sqrt{(\text{concordant} + \text{discordant} + \text{extray})(\text{concordant} + \text{discordant} + \text{extrax})}} \tag{9}$$

The value of τ lies between −1 (complementarity) and 1 (similarity).

The fragment library

The fragment library consists of frequently occurring small molecular fragments from the Cambridge Structural Database (CSD) [35]. It includes electronic charges [29], statistically derived geometry [28], a description of hydrogen-bonding type, and atomic hydrophobic parameters [36]. For the purposes of the work in this series of papers, it was necessary to generate two additional charged aliphatic fragments and some halogenated aromatic ring termini fragments (F-C and Cl-C). The new fragments were named d209, d210, w192 and w193 (see Fig. 2). The advantage of the halogen aromatic ring termini is that



Two new charged fragments;
each carries a new charge of –2.

Two new halogen aromatic ring
termini fragments;
C_{ar} represents an aromatic carbon.

Fig. 2. The additional fragments d209, d210, w192 and w193.

they allow substituted aromatic rings in the placement of a structure; the alternative of using a halogen from part of an aliphatic fragment would result in poor transferability.

CSD search for test molecules

The CSD was searched to obtain structures containing each of the 113 library fragments, as well as the additional fragments d209 and d210 (Fig. 4). The following search criteria were specified:

(1) number of atoms between 20 and 70 (*smols* 2 command);

(2) R-factor between 0.02 and 0.07 (*xprop* 6);

(3) exclude unsuitable groups, such as metal complexes (molecules 12, 60, 61–63, 65–86);

(4) exclude Br, I and Se atoms, since these are not included in the set of atoms making up the fragments in the library (use *atomr* command);

(5) restrict search to diffractometry data structures (*xprop* 1);

(6) remove distorted structures (*xprop* 2), structures with errors (*xprop* 3) and structures whose coordinates are not known (*coord*).

For aliphatic fragments, an atom-centred fragment search (*fprobe* 1) was made, while a search for aromatic fragments was performed which allowed substituents (*rprobe* with level 4 embedding). A slower substructure search (*subss*) was then performed for the aliphatic query fragments, at which stage the maximum number of bonds to non-hydrogen atoms in fragments which contained hydrogens needed to be specified. The reference numbers of the hits were output using the *stash* command.

For a maximum of five random structures per fragment (total of 564), the hybridization states of the constituent atoms were then determined using the program IDATM developed by Meng and Lewis [37], but modified such that the interatomic connectivity was taken as input, rather than calculating this de novo; also, formal charges on atoms were considered in the evaluation of hybridization. The resulting hybridizations were used for the addition of hydrogens to the structures by a COSMIC routine [38]. Since the determination of hybridization states with IDATM was not successful for 100% of the atoms, a quick check of the success of the routine was performed by comparing the total number of hydrogens in the modified structure (sum of the number of hydrogens added to the structure and the number of hydrogens already present) with the expected number of hydrogens (obtained from the chemical formula of the compound). Eventually, 492 structures remained (14 of which were duplicates).

The van der Waals surface of each structure was generated (using a tessellation frequency of 5), onto which the actual electrostatic potential was projected. This could then serve as a test model for investigating how well this electrostatic potential compares with that generated using predicted residual charges obtained from fragment placement onto the structures.

Routine for assigning correct fragments

A routine was written to assign the correct fragments automatically, such that the atom types of a given molecule would be reproduced. The procedure can be summarized as follows:

(1) read in molecular coordinates and atom types;

(2) read in fragment library;

(3) perceive all cyclical and non-cyclical nodes in the molecule (see paper 2 [17]);

(4) assign the correct fragment to each node according

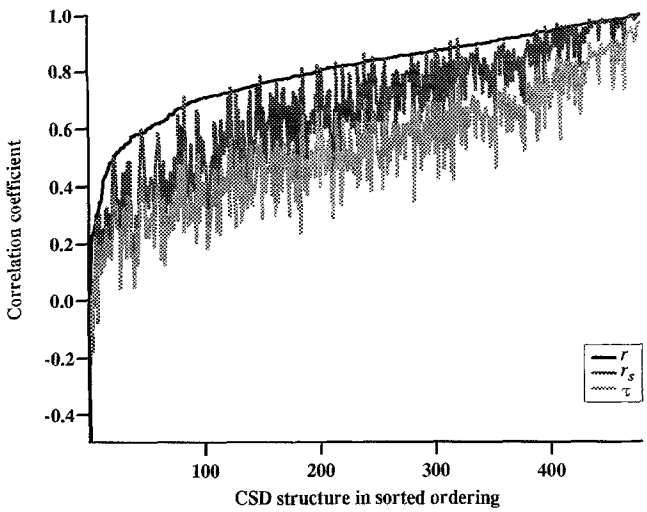


Fig. 3. Test of transferability of fragment charges with respect to the generation of a predictable electrostatic potential on 478 different CSD structures. Assessment methods are Pearson's (r), Spearman's rank (r_s) and Kendall's rank (τ) correlation coefficients. The structures were sorted according to ascending Pearson's correlation coefficient values.

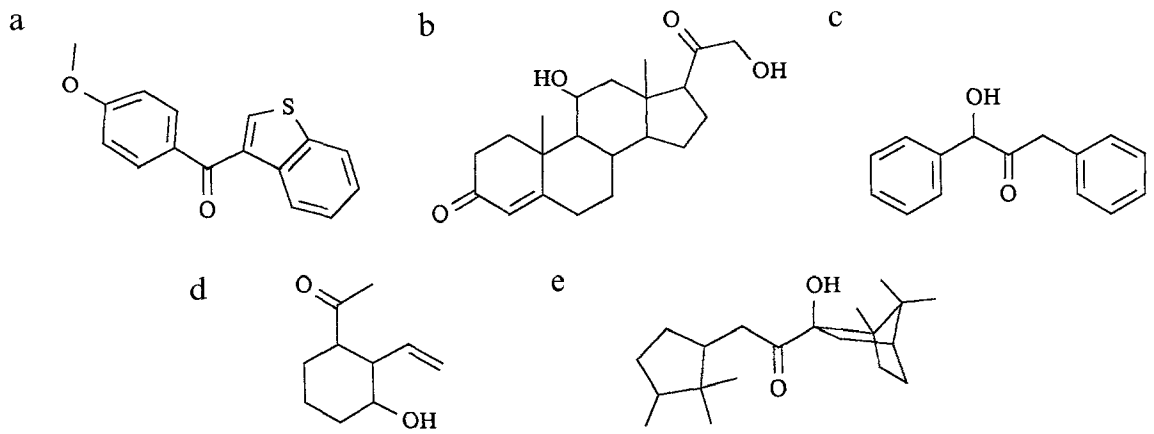


Fig. 4. The five CSD structures randomly selected for the carbonyl fragment (m021).

to the atom types, bond orders and formal charge of the molecule;

(5) transfer the residual charges of each fragment onto the graph of the molecule, taking averages at vertices where overlap of more than one fragment is present.

These fragmental residual charge values were then used to project the predicted electrostatic potential onto the same van der Waals surface (tessellation frequency 5) as the actual electrostatic potential. A direct comparison of potential values on the surface was then made.

Results

If the fragments are to be used as pieces for atom assignment to a molecular graph, the transferability of these properties must be demonstrated. It has been shown that 91% of fragment (predicted) residual charges fell within twice the standard deviation of the mean of the observed values [29]. The transferability of hydrogen-bonding properties is assumed, since they are defined in terms which are independent of whether the involved atoms form part of a fragment or a molecule. And finally, the atomic hydrophobic parameters have been shown to be transferable [33].

Electrostatic transferability

It has already been noted that the electronic property which best indicates complementarity between a site and its ligand is the electrostatic potential rather than adjacent residual charges [21]. It therefore remains to be shown that the residual charges from the fragments can be used in a molecule to generate an electrostatic potential which is comparable to the actual electrostatic potential of that molecule. To investigate this possibility, a large number of CSD structures were obtained and a comparison was made between the actual electrostatic potential and that predicted using the residual charges of the fragments. The following sections summarize the procedure and the results.

Comparison between actual and predicted electrostatic properties

There are several ways in which two sets of data can be statistically compared. To analyse the correlation between the actual and predicted electrostatic potentials of the 492 CSD structures (14 of which are duplicates), Pearson's (r), Spearman's rank (r_s) and Kendall's rank (τ) correlation coefficients were calculated for all observed and predicted sets of potentials. In order to present the results clearly, the structures were sorted into ascending values of Pearson's correlation coefficient (see Fig. 3).

Assuming that coefficients over 0.7 represented good predictability, there appeared to be a marked difference between the parametric method of Pearson's and the two nonparametric (rank) approaches. As an example of the

TABLE 3
NUMBER OF DIFFERENT FRAGMENTS USED IN THE PLACEMENT OF THE FIVE CSD STRUCTURES (FIG. 4) CHOSEN FOR THE CARBONYL FRAGMENT (m021)

	CSD reference code	CSD reference number	No. of different fragments used in placement	Pearson's correlation coefficient, r, for electrostatic potential comparison
Fig. 4a	JEDEUS	71678	7	0.841
Fig. 4b	CORTIC	5449	10	0.901
Fig. 4c	DBEZLM03	65874	5	0.937
Fig. 4d	ACVCHO	25714	8	0.903
Fig. 4e	GETBEM	76580	7	0.807

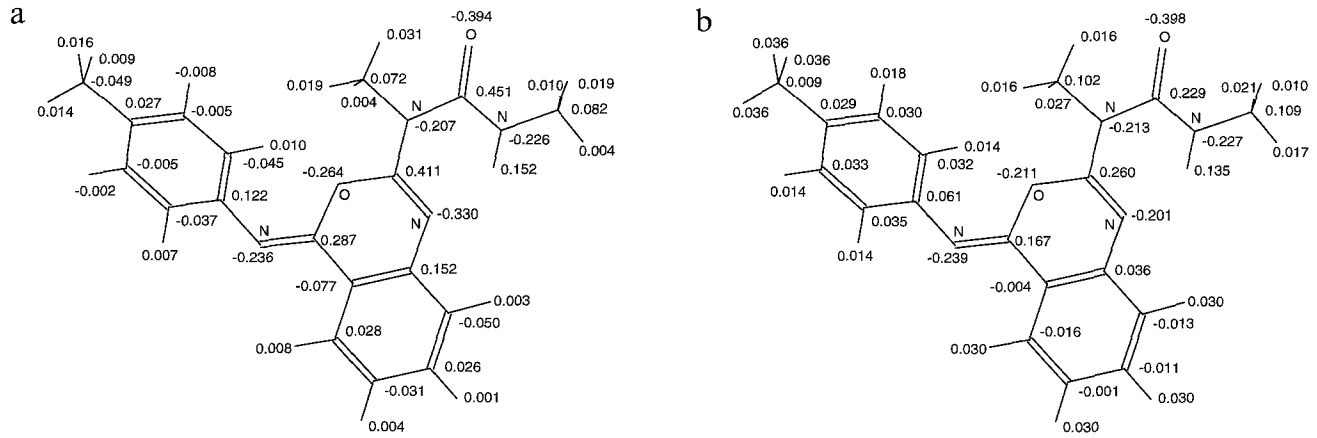


Fig. 5. A CSD structure which gave poor predictability. The reference number is 84624, the reference code name JATWUW. Residual charges are displayed. (a) The structure with actual residual charges derived from CNDO. (b) Predicted residual charges taken from fragments. Pearson's correlation coefficient between the actual and the predicted electrostatic potentials is 0.304, Spearman's rank is −0.077 and Kendall's rank is −0.059.

procedure, Fig. 4 shows the five structures chosen for the carbonyl fragment (m021). The number of different fragments used per structure and the Pearson's correlation coefficient for the comparison between the electrostatic potential obtained from the fragment placement and that calculated de novo from the charges, are listed in Table 3.

It is worth expanding on the CSD structures which fared badly in all three assessment methods. These tended to be structures with extended conjugated systems. As may be expected, the residual charges of small fragments are unlikely to reproduce the residual charges of large complex systems of conjugation. One structure which has a Pearson's correlation coefficient between the actual and the predicted electrostatic potentials of 0.304, a Spearman's rank of −0.077 and a Kendall's rank of −0.059, is illustrated in Fig. 5. The main discrepancy of atomic residual charge occurs at the carbon atoms adjacent to the heteroatoms. In some cases, the actual CNDO charges at these carbons are at least double the value of the predicted fragmental charges. The electrostatic potential range using CNDO charges was −146–126 kJ mol^{-1}, and the range using the fragment charges was −238–174 kJ mol^{-1}.

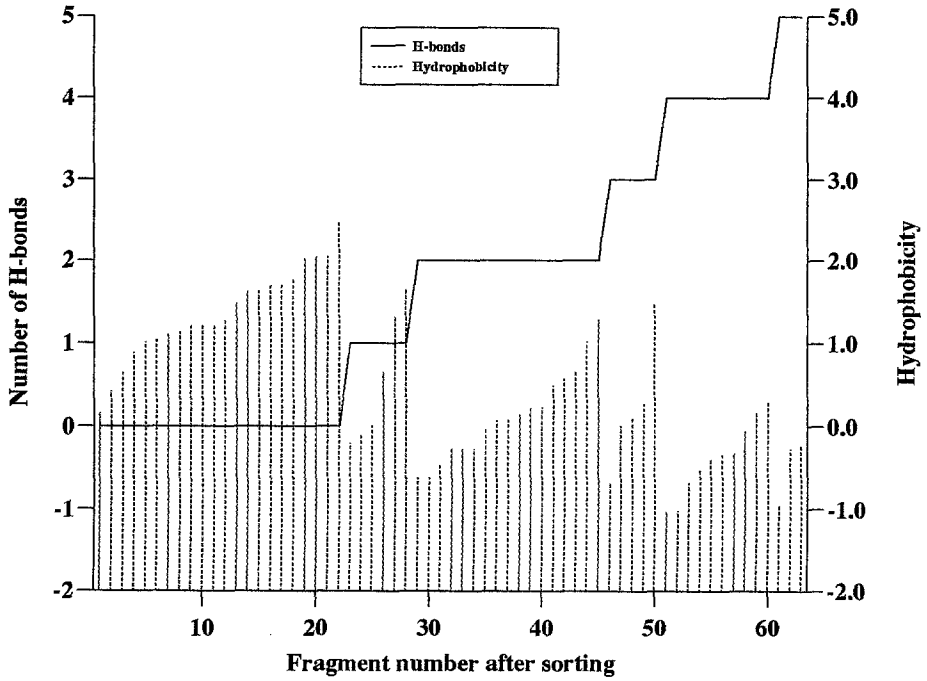


Fig. 6. The relationship between the number of hydrogen bonds in a fragment and its total hydrophobicity value. The fragments used in the analysis were sorted at two levels. The first sort was according to ascending total number of hydrogen bonds; then, for each number of hydrogen bonds, the fragments were sorted again according to ascending hydrophobic values.

Relationship between hydrogen bonding and hydrophobicity

If the hydrogen-bonding ability of the fragments and their hydrophobicity values were anti-correlated, it would simplify the atom placement procedure, since hydrophobicity could be neglected in the objective function, and its presence would be surmised from the absence of hydrogen-bonding properties and vice versa. To test this possibility, the total number of hydrogen-bonding groups in each fragment was compared with the total hydrophobicity value. This could only be done with the fragments whose hydrophobicity values are known (63 fragments). The results are shown in Fig. 6. It is evident that there is no simple anti-correlation between the two properties, and each will have to be considered independently of the other in the fragment placement procedure.

Discussion and Conclusions

Automated de novo drug design is a combinatorial problem with discrete and continuous space components. Combinatorial problems occur at many steps: the choice of subsets of site points from the whole set (discrete space problem) defines the broad classes of shapes to fit the site and has a strong bearing on selectivity; molecular graph generation (discrete space problem) is highly combinatorial and is made more diverse by conformational flexibility (continuous space problem); atom assignment to the molecular graph (discrete space problem) is combinatorial [39]. In principle, it should be possible to develop a routine to optimize the solution of this de novo design problem by a single objective function. However, in order to create an efficient strategy for eventual optimization, it is informative to analyse the problems inherent in each combinatorial step. Our objective in this series of papers has been to explore the optimization of atom placement. Since the problem space is large (10^{30}), a significant sampling of, say, 10^6 points needs rapid calculation for the optimization to be computationally tractable. Hence the need for a strategy to remove as much computation as possible at each step. If small-group properties are transferable between molecules, large reductions in computing resources could be made. This paper has assessed mainly the transferability of atomic residual charges from fragments to create the molecular electrostatic potential.

The theory of Atoms in Molecules (AIM) [12] has gained ground in recent years as a useful way for analysing molecular similarity and complementarity [13]. Our strategy stems from the ideas expressed in AIM, although our approach is much less complete and is defined in a looser manner. Many properties derived from molecular fragments are transferable. The question that has to be addressed is: what fragment size, for a particular arrangement of atoms, leads to transferability? In problems of drug design based on molecular similarity and complementarity, the electrostatic potential frequently plays an important role. This paper has examined the electrostatic potential created by atom assignment to molecular graphs using fragment placement, and compared the potential with that generated from a de novo calculation of the atomic residual charges.

Fragments were investigated by random selection of five examples of molecular structures, taken from the CSD, that contained a particular fragment. In all, 478 structures were examined. The similarity between the electrostatic potential derived from the fragments and that generated by an explicit calculation of the charges was assessed by correlation analysis. The correlations were clearly better when Pearson's correlation coefficient was used to assess the similarity of the values; the ranking methods yielded poorer values. Furthermore, Kendall's rank values were worse than Spearman's ranks. The progressively weaker predictability through the three methods reflects the loss of information incurred in proceeding from a method which correlates actual values, to a method which correlates numerical differences in ranks, and finally to a method which correlates only the relative ordering of the ranks.

In conclusion, from the findings of using fragment residual charges to predict the electrostatic potential of CSD structures, it is clear that, when the method of assessment is Pearson's correlation coefficient, there is good correlation for most structures, in particular those in which there is no extended system of conjugation. Small fragments of three, four or five atoms are not efficient in reproducing the electrostatic potential of a conjugated system. However, where it is possible to use a larger fragment, incorporating conjugation, transferability is improved. These results form the basis of the rationale behind the use of fragments in an atom assignment procedure where the electrostatic potential of the molecule plays an important role. In the future, improvements could be made by reconsidering a larger combinatoric set of aromatic ring fragments.

Other possible transferable properties of interest to a drug designer are hydrogen bonds and atomic hydrophobicity parameters. Both types have been studied extensively. Hydrogen-bonding properties of fragments can be geometrically defined and expressed as probability values within a local coordinate system [40,41]. They are transferable, provided that the geometry of the molecular graph permits hydrogen-bond formation. Atomic hydrophobicity parameters have been studied extensively [42] and are transferable. Surface charge density computed for fragments is not well anti-correlated with hydrophobicity [36]. Similarly, hydrophobicity is not clearly anti-correlated with hydrogen-bonding properties of the fragment. Thus, all properties considered will have to be treated separately in an optimization algorithm for atom assignment to molecular graphs in de novo design.

Transferable fragment properties offer the possibility of constructing an algorithm to optimize atom assignment. The following paper [17] provides an algorithm for the perception of 3D molecular graphs in ways which are useful for fragment placement. Fragments with similar graphs can be classified efficiently for an optimized placement routine, as proposed in paper 3 of the series [18].

Acknowledgements

We wish to thank the Wellcome Trust for a Wellcome Prize Studentship (M.T.B) and for a Principal Research Fellowship (P.M.D). Part of this work was carried out in the Cambridge Centre for Molecular Recognition.

References

1 Lauri, G. and Bartlett, P.A., J. Comput.-Aided Mol. Design, 8 (1994) 51.
2 Böhm, H.J., J. Comput.-Aided Mol. Design, 6 (1992) 593.
3 Goodford, P.J., J. Med. Chem., 28 (1985) 849.
4 Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Protein Struct. Funct. Genet., 19 (1994) 199.
5 DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1988) 722.
6 Lewis, R.A., Poe, D.C., Huang, C., Ferrin, T.E., Langridge, R. and Kuntz, I.D., J. Mol. Graphics, 10 (1992) 66.
7 Nishibata, Y. and Itai, A., Tetrahedron, 47 (1991) 8985.
8 Nishibata, Y. and Itai, A., J. Med. Chem., 36 (1993) 2921.
9 Moon, J.B. and Howe, W.J., Tetrahedron Comput. Methodol., 3 (1990) 681.
10 Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36 (1993) 1700.
11 Clementi, E., Computational Aspects of Large Chemical Systems, Springer, Berlin, 1980.
12 Bader, R.F.W., Atoms in Molecules: a Quantum Theory, Oxford University Press, Oxford, 1990.
13 Popelier, P.L.A., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie Academic and Professional, London, 1995, pp. 217–244.
14 Chang, C. and Bader, R.F.W., J. Phys. Chem., 96 (1992) 1654.
15 Lewis, R.A. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 141.
16 Rotstein, S.H. and Murcko, M.A., J. Comput.-Aided Mol. Design, 7 (1993) 23.
17 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 351.
18 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 359.
19 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) in press.
20 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) in press.
21 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8 (1994) 513.
22 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8 (1994) 527.
23 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8 (1994) 545.
24 Chan, S.L., Chau, P.-L. and Goodman, J.M., J. Comput.-Aided Mol. Design, 6 (1992) 461.
25 Goodman, J.M. and Chau, P.-L., In Doniach, S. (Ed.) Statistical Mechanics, Protein Structure and Protein Substrate Interactions, Plenum Press, New York, NY, 1995, pp. 373–380.
26 Chau, P.-L. and Dean, P.M., J. Mol. Graphics, 5 (1987) 97.
27 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 385.
28 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 397.
29 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 407.
30 Reingold, E.M., Nievergelt, J. and Deo, N., Combinatorial algorithms: Theory and Practice, Prentice Hall, Englewood Cliffs, NJ, 1977.
31 Bondi, A., J. Phys. Chem., 68 (1964) 441.
32 Fauchère, J.-L., Quarendon, P. and Kaetterer, L., J. Mol. Graphics, 6 (1988) 202.
33 Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K., J. Chem. Inf. Comput. Sci., 29 (1989) 163.
34 Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., Numerical Recipes in Fortran, Cambridge University Press, Cambridge, 1992, pp. 630–639.
35 Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rogers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.
36 Chau, P.-L., Ph.D. Thesis, University of Cambridge, Cambridge, 1990.
37 Meng, E.C. and Lewis, R.A., J. Comput. Chem., 12 (1991) 891.
38 Vinter, J.G., Davis, A. and Saunders, M.R., J. Comput.-Aided Mol. Design, 1 (1987) 31.
39 Dean, P.M., Barakat, M.T. and Todorov, N.P., In Dean, P.M., Jolles, G. and Newton, C.G. (Eds.) New Perspectives in Drug Design, Academic Press, London, 1995, pp. 155–180.
40 Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 101.
41 Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 115.
42 Rekker, R.F. and Mannhold, R., Calculation of Drug Lipophilicity: The Hydrophobic Fragmental Constant Approach, VCH, Weinheim, 1992.