De Novo Generation of Molecular Structures Using Optimization To Select Graphs on a Given Lattice

Robert P. Bywater,*,† Thomas A. Poulsen,‡,|| Peter Røgen,§ and Poul G. Hjorth§

Biostructure Group, Novo Nordisk A/S, Novo Nordisk Park, DK-2760 Måløv, Denmark, and Departments of Mechanical Engineering and Mathematics, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

Received October 28, 2003

A recurrent problem in organic chemistry is the generation of new molecular structures that conform to some predetermined set of structural constraints that are imposed in an endeavor to build certain required properties into the newly generated structure. An example of this is the pharmacophore model, used in medicinal chemistry to guide de novo design or selection of suitable structures from compound databases. We propose here a method that efficiently links up a selected number of required atom positions while at the same time directing the emergent molecular skeleton to avoid forbidden positions. The linkage process takes place on a lattice whose unit step length and overall geometry is designed to match typical architectures of organic molecules. We use an optimization method to select from the many different graphs possible. The approach is demonstrated in an example where crystal structures of the same (in this case rigid) ligand complexed with different proteins are available.

INTRODUCTION

Definition of Pharmacophore. A recurrent problem in medicinal chemistry is how to generate libraries of compounds starting from fragmentary but significant information. The most common example of this is the so-called pharmacophore [derived from Greek words meaning "medicine carrier"]—a construct that embodies essential elements required to be present in any ligand designed to elicit a specified physiological response.[1,2] Typically, a pharmacophore model will consist of a list of atom types that have been determined to be functionally significant together with a specification of the geometrical relation between them in the form of inter"atomic" distances. "Functionally significant" in this context means chemical function such as charge, hydrogen bonding propensity (as donor or acceptor), "hydrophobicity" or aromaticity, although the aim is actually to be able to predict or engineer some biological, rather than chemical, function.

In the real world context that this model is intended to represent, ligands that match the pharmacophore are assumed to be capable of being "docked" into a binding pocket in a target protein such that the latter assumes the conformation required for the specified ("active" or "inactive" for example) response. As an example, the ligand methotrexate is shown docked into the binding pocket in the enzyme dehydrofolate reductase[3] along with a calculated[4] pharmacophore in Figure 1a. Ligand binding pockets are complex surfaces containing many protuberances and cavities and maybe re-entrants (see for example Figure 1b). [Titus Lucretius Carus (98−55 BC): *Quorum ita texturae ceciderunt mutua contra, ut cavа convenient plenis haec illius illa huiusque inter se, iunctura haec optima constat* (Translated by RPB as follows: Things whose textures have a mutual correspondence such that cavities fit protuberances and vice versa form the closest union.)] Such a surface may therefore require the inclusion of a (one or only a few) "forbidden zone" for ligand atoms.[5] In this paper we refer to a required "atom" position as a node and a forbidden position as an antinode.

Structures of Organic Compounds. The term "organic compounds" is used here to denote the class of chemical compounds built up on a skeleton of covalently linked, stable, neutral (nonionized), carbon atoms. In the simplest case, that of so-called aliphatic compounds, the quadrivalent nature of carbon atoms naturally lends itself to generating an extended lattice in 3D, the diamond lattice.

This is the simplest kind of network for linking atoms in small molecules (MW < 500) such as typical drug molecules. The adamantane structure, which is the repeating unit in the diamond lattice has been used before,[6] not so much for generating entire ligand structures as we do here but rather for extending existing (sub)structures. But this work[6] was a major inspiration to the present study.

Although we restrict ourselves in this work to aliphatic structures, we shall consider the introduction of atom types other than carbon into the lattice, notably N, O, and S. This is because these elements confer interesting and useful chemical properties to the ligand and because, as long as they display sp_3 geometry, they fit into the standard aliphatic framework (albeit with somewhat different bond lengths). A further complication that we shall not consider in the present work is alternative types/symmetries of lattice network. This would be necessary if we were e.g. to consider aromatic-type geometry. Many if not most pharmacologically

* Corresponding author e-mail: robpbyw@hotmailcom.
† Novo Nordisk A/S.
‡ Department of Mechanical Engineering, Technical University of Denmark.
§ Department of Mathematics, Technical University of Denmark.
|| Present address: NovoZymes A/S, Novo Nordisk Allé, DK-2880 Bagsværd, Denmark.

10.1021/ci0342369 CCC: $27.50
Published on Web 03/23/2004

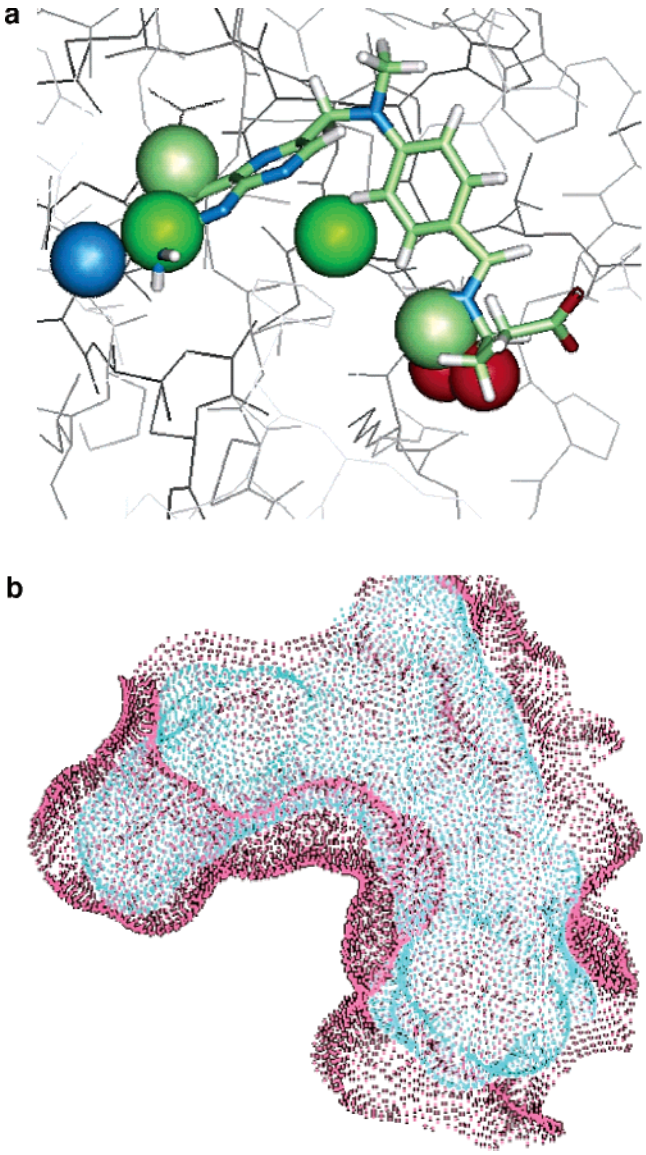

Figure 1. (a) Active site of dihydrofolate reductase with bound methotrexate ligand from crystal structure.[3] A pharmacophore calculated using the PLIM program[4] shows predicted positions of aliphatic carbons (light green spheres), aromatic carbons (dark green), basic nitrogen atoms (blue), and acidic oxygen atoms (red). (b) The same active site with the contact surfaces for enzyme (pink) and ligand (light blue) with all atoms removed. The shape of the binding pocket is complex, and a potential "forbidden zone" is suggested by the region that protrudes into the middle of the binding pocket.

active compounds contain at least one (hetero)aromatic group. Therefore, an obvious requirement for any future development of the method that we describe here would have to cater for this extension.

Using Subgraphs on a Lattice To Generate and Compare Chemical Structures. After constructing a generic diamond lattice we shall define three or four nodes on the lattice from which to "grow" the rest of each "molecule". We shall then find an economic way to navigate in the set of subgraphs that link these starting points. We wish also to find descriptors for each subgraph that encapsulates "shape" information for that subgraph such that each subgraph/molecule can be compared with one another in respect of "shape".

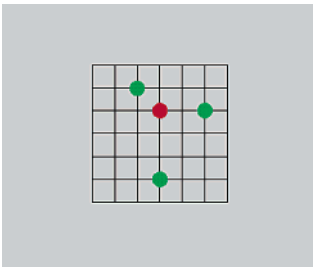

Figure 2. Example of a 2D square lattice showing three nodes (green) and one antinode.

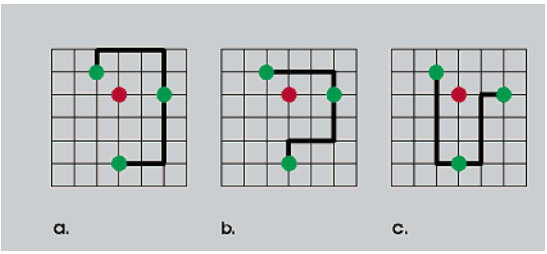

Figure 3. (a−c) Examples of possible paths connecting the nodes and avoiding the antinode.

This problem is about systematic generation of lattice molecules subject to given constraints. We can illustrate the basic idea by looking at a 2D square lattice (Figure 2.). A given set of nodes that must be present possibly along with a set of nodes that must be absent constitutes a pharmacophore, as defined herein. A pharmacophore (here consisting of three nodes and one antinode) has been placed on the lattice.

We construct examples of entire "molecules" by placing along the lattice a graph connecting all the nodes of the pharmacophore and avoiding the antinodes. There are of course several ways to construct a molecule that respects this pharmacophore. Three such ways are illustrated in Figure 3a−c.

How does one determine a reasonable number of such molecules, and ways to distinguish between them, to pick the "optimal" carrier molecule? In the real (pharmaceutical) world, we are interested in organic compounds as defined above. In the simplest case, that of aliphatic compounds, the quadrivalent nature of carbon atoms calls for a specific lattice in 3D, the diamond lattice.

The problem is then more specifically, for a given pharmacophore (not necessarily fitting on the diamond lattice), to

(1) identify "best fit" pharmacophore lattice sites,

(2) find economic ways to find and to navigate through subgraphs/molecules that all incorporate the pharmacophore,

(3) find descriptors for such subgraphs that encapsulates "shape" information such that the subgraphs (or molecules) can be compared with respect to "shape".

For instance, the number of loops present in a given graph is an important measure of the "rigidity" of the associated molecule, and this property should also distinguish among the generated graphs.

METHODS

Generation of the Diamond Lattice. The basic lattice structure for all the experiments reported here is the "base"

diamond lattice in 3D. A data structure representing this lattice was generated in the following way:

First we defined the fundamental *vierbein* of unit vectors **a**, **b**, **c**, and **d** pointing to the corners of the carbon tetrahedron.

$$\mathbf{a} = (0, 0, 1)$$

$$\mathbf{b} = (0, 2\sqrt{2/3}, -1/3)$$

$$\mathbf{c} = (\sqrt{(2/3)}, -\sqrt{2/3}, -1/3)$$

$$\mathbf{d} = (-\sqrt{(2/3)}, -\sqrt{2/3}, -1/3)$$

In addition to this, two linear combinations are frequently employed, namely

$$\mathbf{e} = -\mathbf{d} + \mathbf{a} - \mathbf{c}$$

$$\mathbf{f} = -\mathbf{d} + \mathbf{b} - \mathbf{c}$$

Every lattice site can then be specified by three integer values, *x*, *y*, and *z*. The physical space location of the point labeled (*x*, *y*, *z*) is given by the formula

$$\mathbf{P}(x, y, z) = \mathbf{b}[(x + (1 + z)\%2)/2] + \mathbf{f}[(x + z\%2)/2] - \mathbf{d}[(y + (x + z)\%2)/2] + \mathbf{c}[(y + (x + 1 + z)\%2)/2] + \mathbf{a}[(1 + z)/2) + \mathbf{e}[z/2]$$

Here, %2 denotes **mod** 2 and [ ] denotes integer value.

For a finite (sub)lattice, one can now in a reasonably simple algorithmic way label each point in the lattice with an integer value and label (and number) all the edges and all the loops in the lattice.

A New Method for Generation of Lattice Graphs. With the diamond lattice data structure in place, we can begin to generate graphs. Here, we shall describe a method which to our knowledge has not been used before in molecular design. To manage the large number of possible lattice graphs we borrow the technique from the engineering field of *optimal design*.[7,8] Note that we are using the term "optimal design" in a different way to its usage elsewhere e.g. the statistical D-optimal design algorithms used in the ligand design field[9] and for receptor assays.[10]

Briefly, our method is the following. The computer is given an initial guess for the structure, along with a function which the optimal topology should extremise, subject to a number of constraints. Using numerical iteration techniques for locating such a constrained extremal, the computer goes through a sequence of designs, until an appropriate convergence is achieved. Here the term "design" is used to denote any distribution of "weights" to the nodes in the lattice. This only has a physical interpretation as a molecule in the cases, where all weights are either zero or one, in which case the nodes carrying weight one are interpreted as spacepoints occupied by an atom, and nodes carrying a weight of zero are considered void. The edges between "atoms" can be interpreted as chemical bonds.

In the present case we regard the new molecule containing the pharmacophore as the "design topology" which is to be generated. The design variables are weights ρ_i, where $0 \leq \rho_i \leq 1$, each node of the diamond graph carrying a weight. [For technical reasons, we do not allow the nodes to attain the value zero in the implementation. Empty nodes are derived by applying a cutoff like 0.001).]

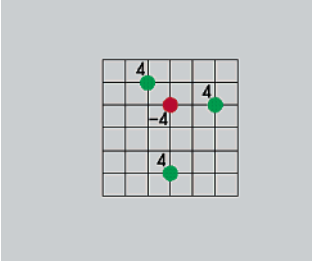

Figure 4. Weighting scheme for nodes and antinodes.

We chose the following function to be maximized:

$$B = \sum \rho_i \rho_j$$

The sum is taken over all pairs of nodes connected by an edge in the lattice (with $i < j$ to avoid duplicate counts). It should be noted that in the limit, where all ρ_i are either 0 or 1, the value of B is exactly the number of edges in the subgraph spanned by the fully occupied nodes.

Further, we impose the constraint

$$N = \sum \rho_i \leq N_o$$

which in the limit of all nodes either fully occupied or void is the number of atoms in the molecule. The number N_o is the prescribed maximum number of atoms desired in the molecule.

The pharmacophore is incorporated into the model simply by specifying from the outset that the nodes known (or specified) to be occupied have the (extraordinary) weight $\rho = 4$, and the antinodes that may not be occupied have the (extraordinary) weight $\rho = -4$ (these values were chosen for the sample model that we used to test the method). The nodes with these extraordinary weights are not changed during the iterations. The other nodes are denoted "active nodes". Figure 4 shows (using again the square lattice in Figure 3) the basic idea.

Starting from some initial guess (e.g. all nodes carrying equal weight), the node values for nodes in the neighborhood of the pharmacophore will decrease or increase until the state is so close to the extremal of the objective function B that iterations produce relatively little change.

Structure of the numerical algorithm:

(1) An initial feasible design is generated, e.g. by assigning a weight of N_o/D to all the active nodes, where D is the number of active nodes.

(2) Evaluate the objective function B.

(3) Compute the "design sensitivity": the derivative of B with respect to all the design variables (see below).

(4) Evaluate the constraint function(s) and compute their sensitivities.

(5) Call an optimization algorithm. In this work we have used Svanberg's Matlab-implementation of his "Method of Moving Asymptotes".[8]

(6) Check for convergence.

(7) If not converged repeat from point 2.

Sensitivity analysis: The design sensitivity of the objective

function B is

$$\frac{\partial B}{\partial \rho_i} = \sum \rho_j$$

where the sum is over all nodes j that are connected to node i.

In practice it turned out that in some cases the method would not converge to a design with only fully occupied and void nodes.

To fix this problem, we replaced the objective function B by B' given by

$$B' = \sum (\rho_i \rho_j)^3$$

The summation is over all neighboring pairs of nodes as before. This trick is well-known in the field of topology optimization.[7]

RESULTS AND DISCUSSION

We implemented the procedure described under **Methods** in the high-level language MATLAB,[11] using as pharmacophore various constellations of node positions that are representative of real-life cases in practical medicinal chemistry. We take the example shown in Figure 1a,b, a pharmacophore determined from the structure of the ligand-binding site for dihydrofolate reductase. Figure 5a−c shows the final output for various pharmacophore configurations (aromatic nodes not included). The large dots represent the (+) part of the pharmacophore, and the diamond represents the (−) part of the pharmacophore. Each node is marked proportional in size to the (average) weight of its adjoining edges.

A further set of examples is shown in Figure 6a,b where the same ligand (dihydrotestosterone) has been co-crystallized with two different proteins, a light chain from the F_{ab} fragment of a monoclonal antibody (PDB i.d. 1d2s—in Figure 6a) and a laminin G-like transporter protein (PDB i.d. 1i9j—Figure 6b). Note that the steroid ligand is compatible with the diamond structure. Nodes from the pharmacophore (shown in blue) are matched by atoms from the ligand but none of the antinodes. The unmatched nodes (yellow) offer opportunities for expanding the ligand structure without making unfavorable contacts with the target protein ("clashes").

These examples demonstrate that the method can link up nodes and avoid antinodes in a convincing way. In particular, the example in Figure 6a,b shows that the correct connectivity in the same ligand can be achieved even when bound to different proteins (and hence having a different set of antinodes). The ligand is the same in each of these two cases, but the target protein and the corresponding antinodes are different in each case. If these are taken into account, the "opportunities for expanding the ligand structure" will enable the medicinal chemist to design analogues of the common ligand that will selectively bind to only one or other of the two target proteins. While this may not be useful in this particular case, the medicinal chemist is frequently confronted with the need to re-engineer a ligand that is common to more than one protein such that it binds selectively to only one of them without cross-reactivity. Examples abound in receptor pharmacology e.g. adrenergic receptors where cross-reaction between subtypes is undesirable. While this could be accomplished by applying the principles expounded here, it becomes an exercise not only in ligand design but also in receptor assay technology (to validate the design), both of which are beyond the intended scope of this paper.

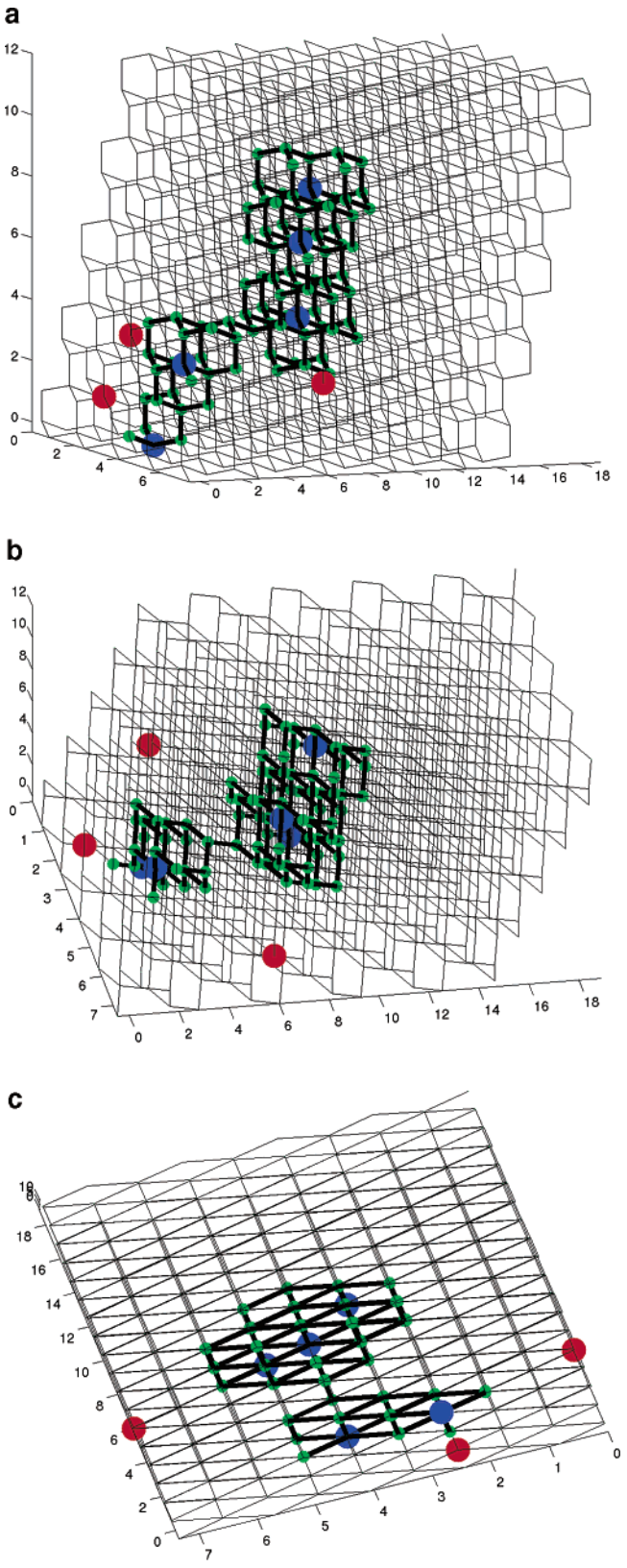

Figure 5. (a−c) Three solutions to "case no. 1".

After running a few further cases we discovered that the procedure without any further constraints could sometimes generate graphs that were not connected. This effect must

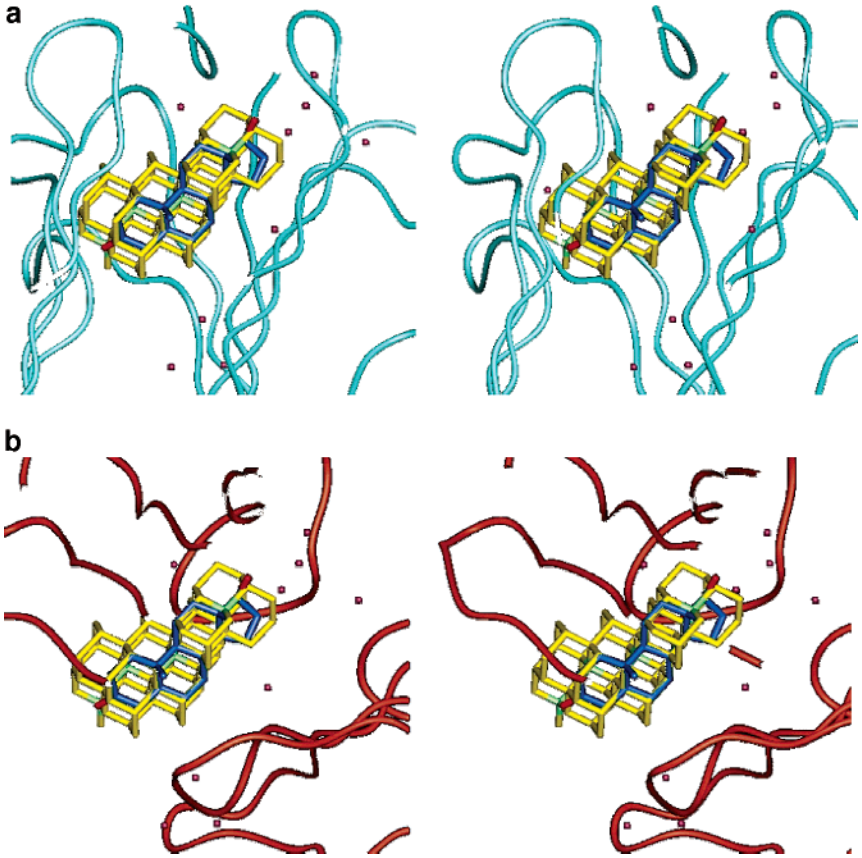

Figure 6. Stereo figures showing dihydrotestosterone mapped to the diamond lattice in the binding sites of (a) monoclonal antibody (PDB i.d. 1d2s) and (b) the laminin G-like transporter protein (PDB i.d. 1i9j).

be avoided, and we must therefore augment the algorithm with a constraint ensuring that the outcome is a connected graph.

Another issue that is important for the method is its "robustness" i.e. how sensitive the outcome is to uncertainty in the specification of the pharmacophore. Future developments such as a scoring method for shape/similarity and flexibility will require running a larger set of test cases.

A most natural descriptor is the number of tight loops in a subgraph. In the case of a weighted graph the number of loops may be defined as

$$L = \sum_{i1 \cdot \Lambda i6 in\ loop} \rho_{i1}\rho_{i2} \cdots \rho_{i6}$$

where the sum is over all loops in the lattice. We note that in the limit of all weights going to zero or one, this function counts the number of loops in the molecule.

As these loops are rigid, the number L is important for the flexibility of the design molecule. Hence adding $(L - L_0)^2$ to the objective function a desired number of loops L_0 may be obtained.

Besides comparing structures through descriptors of "shape", the usual Euclidean metric on the set of weighted graphs

$$d(G_1,G_2) = \sqrt{\sum_{k\ in\ vertices} (q_{k1} - q_{k2})^2}$$

gives a direct metric on the set of candidate graphs as they all are fixed in 3-space relative to the pharmacophore. One or more known (patented or) toxic subgraphs may be avoided by constraining the optimization.

The method presented here is for building a graph of a drug that can block the binding site of a known receptor.

ACKNOWLEDGMENT

This work was initiated during the 41st European Study Group with Industry 13−17 August 2001 at the Odense campus of the University of Southern Denmark. The organizers and sponsors of the Study Group are thanked for their kind support. The Matlab-implementation of the Method of Moving Asymptotes[8] was kindly supplied by Krister Svanberg, KTH, Stockholm. The authors also wish to acknowledge the kind assistance of the Department of Chemistry, University of Southern Denmark for putting at our disposal for the duration of the Study Group a demonstration model of the diamond lattice. This model was an invaluable aid in constructing the lattice labeling algorithm.

REFERENCES AND NOTES

(1) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput.-Aided. Mol. Des.* **1989** 3, 225−251.

(2) Loew, G. H.; Villar, H. O.; Alkorta, I. Strategies for indirect computer-aided drug design. *Pharm. Res.* **1993**, *10*, 475−486.

(3) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Ångstroms resolution. I. General features and binding of methotrexate. *J. Biol. Chem.* **1982**, *257*, 13650−136562.

(4) Harris, M. R.; Kihlén, M.; Bywater, R. P. PLIM−an automatic Protein Ligand Interaction Modeller. *J. Mol. Recognit.* **1993**, *6*, 111−115.

(5) Lewis, R. A.; Dean, P. M. Automated site-directed drug design: the concept of spacer skeletons for primary structure generation. *Proc. R. Soc. London* **1989**, *B236*, 125−140.
(6) Lewis, R. A. Automated site-directed drug design: approaches to the formation of 3D molecular graphs. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 205−210.
(7) Bendsoe, M. P.; Sigmund, O. *Topology Optimization. Theory, Methods, and Applications*, 2nd ed.; Springer 2003; ISBN: 3-540-42992-1.
(8) Svanberg, K. The Method of Moving Asymptotes. *Int. J. Numer. Methods Eng.* **1987**, *24*, 359−373.
(9) Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. Field interaction and geometrical overlap: a new simplex and experimental design based computational procedure for superposing small ligand molecules. *J. Med. Chem.* **2003**, *46*, 1359−1371.
(10) Dunn, G. "Optimal" designs for drug, neurotransmitter and hormone receptor assays. *Stat. Med.* **1988**, *7*, 805−815.
(11) See e.g.: Biran, A.; Breiner, M. *MATLAB 6 for Engineers*, 3rd ed.; Prentice Hall: 2002; ISBN 013-033631-9.

CI0342369