Superposition of Three-Dimensional Chemical Structures Allowing for Conformational Flexibility by a Hybrid Method

Sandra Handschuh,† Markus Wagener,‡ and Johann Gasteiger*

Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received December 2, 1997

The superposition of three-dimensional structures is the first task in the evaluation of the largest common three-dimensional substructure of a set of molecules. This is an important step in the identification of a pharmacophoric pattern for molecules that bind to the same receptor. The superposition method described here combines a genetic algorithm with a numerical optimization method. A major goal is to adequately address the conformational flexibility of ligand molecules. The genetic algorithm optimizes in a nondeterministic process the size and the geometric fit of the substructures. The geometric fit is further improved by changing torsional angles combining the genetic algorithm and the directed tweak method. This directed tweak method is based on a numerical quasi-Newton optimization method. Only one starting conformation per molecule is necessary. Molecules having several rotatable bonds and quite different initial conformations are modified to find large structural similarities. A set of angiotensin II antagonists is investigated to illustrate the performance of the method.

1. INTRODUCTION

The investigation of a series of ligands binding to the same receptor is usually performed by defining the similarities between the ligands through a common pharmacophoric pattern. With the development of effective 2D to 3D structure generators,[1,2] the pharmacophore determination problem can be replaced by a three-dimensional substructure search. Initially, the methods used for searching for the 3D common substructure were performed only on a single, rigid conformation, not taking into account conformational flexibility of the substrates.[3−6]

Many of these substructure search programs are based on interatomic distance information. The first detailed study of distance-based methods for 3D similarity searching was published by Pepperrell and Willett.[7] More recently, Sheridan et al.[8] reported on distance based methods using several conformations per structure. Angle-based and fragment-based methods[9,10] like those of Fisanick et al.[11] were also used to calculate 3D similarities.

Wagener et al. developed an approach for MCSS (*m*aximum *c*ommon *s*ubstructure) search that is based on atom mappings. Atoms of the same atomic number or, alternatively, those having values of an optional physicochemical property within a given range are matched onto each other. The approach was initially developed to be applied to the topology of a molecule as given by a connection table.[12] However, it was shown that this method can be extended to the 3D structure of a molecule.[12]

However, the majority of substances having a certain biological activity possess at least one, usually several rotatable bonds and can exist in a large number of different low energy conformations. Thus, the problem of flexible 3D substructure searching has to be addressed.

3D substructure searches in large databases containing only a single conformation for each molecule can extract not all of the structures which would fit the query geometry when other reasonable conformations would be considered. To increase the number of hits a set of multiple low energy conformations for each structure is stored in databases.[3,4] Databases containing several low energy conformations to span the conformational space of a molecule rapidly demand sizable amounts of memory space. Nevertheless, it must be realized that even these databases do not completely cover the conformational space. Many biologically active compounds can attain hundreds of thousands of conformations within 5−10 kcal/mol of the conformation with the lowest energy. Thus, searches for the 3D-MCSS or for a pharmacophore pattern in databases that have a limited set of conformations for each molecule are bound to miss interesting hits for new lead structures.

Therefore, it is advantageous to start the 3D substructure search with one conformation for each structure and to investigate conformational flexibility during the optimization process. Hence a "query-directed" conformational search technique was implemented.[13] Several methods for flexible 3D searching have been published including distance geometry,[14] systematic search,[15] genetic search,[16−22] and directed tweak.[13] The results of comparisons between these methods developed up to 1993 were reported by Clark et al.[23] Their studies came to the conclusion that genetic algorithms and the directed tweak technique are most appropriate for flexible 3D substructure searching. In agreement with their results we have implemented an algorithm combining evolutionary theories with a numerical optimizer. This hybrid technique of a genetic algorithm and

* To whom correspondence should be addressed. E-mail address: Gasteiger@ccc.chemie.uni-erlangen.de.
† E-mail address: Handschuh@torvs.ccc.uni-erlangen.de.
‡ Present address: N.V. Organon, Dept. of Computational Medicinal Chemistry, 5340 BH Oss, NL.

S0095-2338(97)00438-1 CCC: $15.00
Published on Web 02/28/1998

a directed tweak method based on numerical optimization is a flexible search system that accounts for conformational flexibility by rotation around single bonds during the optimization process. This superposition method reveals similarities even between those ligands which show no sizable common 3D-fragments at first sight.

In this approach, the superposition of 3D structures is monitored by matching atoms, and an optimum assignment of the atoms is determined by the algorithm. This algorithm not only optimizes the mutual assignment of the atoms but also automatically identifies rotatable bonds. The conformations of the superimposed molecules are then adapted to each other during the flexible overlay. The assignment of the atoms and the geometric fit of the overlay is first optimized by the genetic algorithm, and, afterwards, the geometric fit of each solution is improved by the directed tweak method.

2. METHODS

2.1. Overview of Genetic Algorithms. Genetic algorithms (GAs) represent robust optimization methods that are based on the mechanics of natural selection and genetics.[24−28] They can efficiently solve problems involving large search spaces and, thus, can even be applied to problems beyond the reach of classical exhaustive search methods.[26−28] A GA imitates nature's methods for adapting to a changing environment, and optimization does therefore not start from a single point but from a population of starting points generated randomly. These starting points correspond to the chromosomes or individuals of a population representing potential solutions to the search problem. In these individuals the parameters of the function to be optimized have to be encoded, e.g., in the form of a bit string. The genetic operators *selection*, *mutation*, and *crossover* are iteratively applied to the population. In the method presented here, two additional operators that are tailored to the specific problem are implemented, called *creep* and *crunch* (see 2.6.3). The course of the program (Figure 1) starts with selection. In general, individuals are selected with a probability proportional to their fitness, i.e., how closely they approach the solution to the problem (Figure 1). In the approach presented here, another selection type is used, called *restricted tournament selection.*[29] This selection type preserves variety in genetic information during the GA run. After selection, the genetic operators are applied to the chromosomes, and a new population forms the offspring generation. Consequently, one complete GA run begins with initialization of the individuals and ends by generating one set of optimized solutions after each run through all generations.

Genetic operators are not based on a deterministic procedure. Therefore, optimization by a GA does not necessarily arrive at the optimum solution. To alleviate this problem, an additional method, the *directed-tweak* procedure was implemented. It helps to assess the geometric fitness of the offspring population by minimizing differences in the conformations (Figure 1).

2.2. An Individual of a Population—The Data Structure. A major task in adapting a genetic algorithm to a specific problem is the encoding of the individuals of the population, i.e., the representation of the chromosomes. The search for a common substructure matches atoms from two molecules to each other. Therefore, such a mapping was

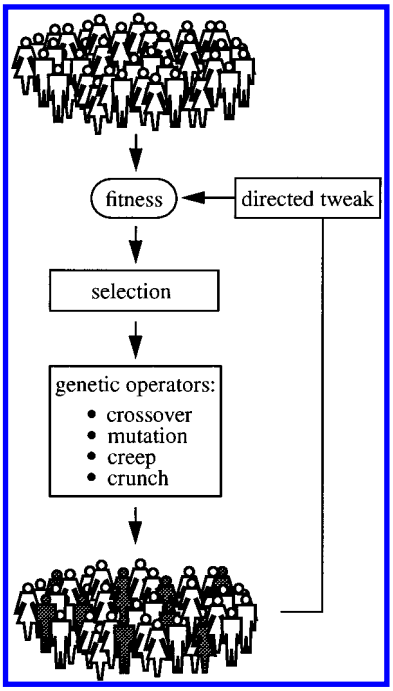

Figure 1. The combination of a genetic algorithm with the directed tweak technique. The figures shown in gray indicate those altered by the application of the genetic operators.

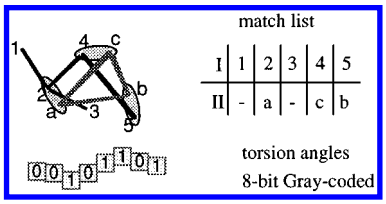

Figure 2. An individual or chromosome of a population that consists of the match list (top) and the representation of the torsion angles (bottom). The match list is as long as the number of atoms in the largest molecule. In the case shown it indicates that atom 2 of molecule I is matched to atom *a* of molecule II, 4 to *c*, and 5 to *b*. The remaining atoms do not have any matching partner. Torsion angles (0°−360°) are represented by 8 bits; the Gray coding has the result that a change in one bit drives the angle by $360°/256(2^8) = 1.4°$.

chosen as an individual of the GA. GA techniques described in the literature use integer schemes to encode individuals.[28,30] We have chosen an approach that represents an individual by a data structure consisting of two parts. In the first part, the mapping of substructures is coded by an integer string and is represented as a fixed-length linked list of matching atom pairs. Since GAs work by combining partial solutions, e.g., parts of substructures that are relevant to the problem, an explicit representation of the matching atom pairs is better suited to the problem than a binary one. Thus, one individual consists of a match list (Figure 2 top right), which contains the atom pairs of the two molecules mapped onto each other.

The second part of the data structure consists of two lists of torsional angles representing the conformations of the two three-dimensional molecules. Eight-bit binary coded integers were found to be quite suitable for encoding these torsion angles. Two main problems arise: the distribution of the torsion angles must be large enough to find useful minima during the superposition process. Second, too wide a distribution leads to quite large computation times and to convergence problems. A Gray-coded representation[20,24] was

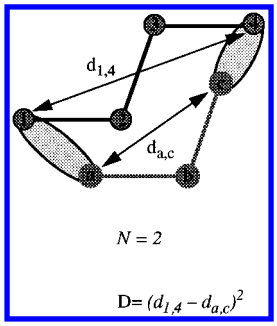

Figure 3. The calculation of the distance parameter *D* for a superposition and the size of the substructure *N*. The atoms marked in gray are assigned to each other.

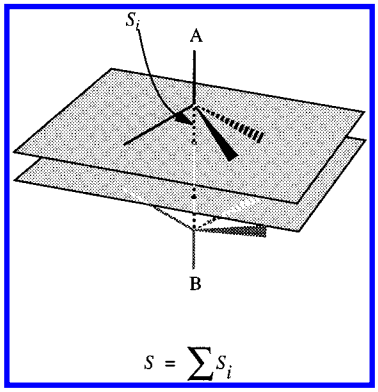

Figure 4. The calculation of the stereochemical parameter *S*. Atoms A and B are part of a match pair and only differ in their stereochemistry.

chosen instead of a normal two-complement one (Figure 2 bottom left) because it offers the advantage that adjacent integers differ by only a single bit.

At the start of the optimization process a population of individuals having this data structure is initialized with random matches of atoms and torsion angles. Each randomly built chromosome then represents an alignment and contains information on torsional angles.

2.3. Optimization Criteria. The search for the MCSS of a set of molecules takes into account two optimization criteria: the size of the substructure, as given by the number *N* of atoms of the substructure (Figure 3), and the geometric fit of the matching atoms. The geometric fit is represented by a distance parameter, *D* (Figure 3), and a stereochemical parameter, *S* (Figure 4). The distance parameter *D* consists of the sum of the squared differences of corresponding atom distances in the two substructures in molecule I and II (eq 1).

$$D = \sum_{1}^{(N-1)!} (d_I(i,j) - d_{II}(i,j))^2 \qquad (1)$$

with $d_I(i, j)$, $d_{II}(i, j)$ = atom distances in molecule I and molecule II and N = number of match pairs (size of the substructure).

D is related to the *r*oot *m*ean *s*quare error (rms) (eq 2) of the distances of corresponding atoms in an optimized superposition.

$$\mathrm{rms} = \sqrt{\frac{1}{N}\sum_{i}^{N}\sum_{j=1}^{3}(a_{ij} - b_{ij})^2} \qquad (2)$$

with N = number of match pairs to be compared and a_{ij}, b_{ij} = atom coordinates of substructures I and II in a three-dimensional Cartesian coordinate system.

The rms value, however, is subject to large changes even when the mapping changes only slightly. Therefore the distance value *D* is better adapted to the specific use during a GA optimization than the rms value. The rms value of the obtained superposition is calculated only at the end of each GA run in order to present the results.

Enantiomers are identical in distance space but differ in stereochemistry. A stereochemical parameter *S* (Figure 4) was defined to distinguish enantiomers. *S* consists of the sum of the single stereochemical descriptors S_i (Figure 4) of each match pair ($S = \sum S_i$). The stereochemical descriptor S_i of a match pair is determined by, first, defining two planes spanned by the atoms of the three first match pairs in the match list. One plane is spanned by the three atoms of the first molecule, and the other plane is spanned by the assigned atoms of the second molecule. The distances and the orientation of the two atoms A and B in molecule I and II of each match pair (Figure 4) to the corresponding plane are calculated. If these central atoms A and B are on opposite sides of the planes, they differ in stereochemistry. The parameter S_i is taken as the larger distance of the two distances between atom A and atom B to the corresponding plane. *S* is different from zero only if the considered structure has at least four atoms.

2.4. Pareto Optimality. Maximum common substructure (MCSS) search is a multicriteria optimization problem, where the notion of optimality is difficult to define. Two main contradictory parameters contribute to the fitness of a superposition and have to be optimized: the size of the substructure and its geometric fit. The obtained substructure size has to be as high as possible. An optimum must be found which takes into account both criteria. The optimization must respect the integrity of each of the separate parameters. Vilfredo Pareto developed a concept for solving multicriteria optimization.[24,25] Pareto optimization means that an optimized state is reached if none of the parameters can be improved further without making another one worse.

If one solution corresponds to a vector which is assumed to be better than another one by being partially less, a mathematical definition of Pareto optimality is that a vector **u** is partially less than **v**, symbolically **u** < p**v**, if the following conditions are kept (eq 3):

$$(\boldsymbol{u} < p\boldsymbol{v}) \Leftrightarrow (\forall i)(u_i \le v_i) \wedge (\exists i)(u_i < v_i) \qquad (3)$$

with x_i, y_i = components of the vectors **u** and **v** ∀, ∃ = allquantor (all of them...), existentialquantor, (at least one of them...).

Under these conditions it is possible to say that vector **u** dominates vector **v**. If vector **u** is neither dominated nor nondominated, the two vectors are equal solutions.

Pareto optimality applied to the MCSS search problem in a three-dimensional space results in simultaneously maximizing the size of the substructure and optimizing the geometric fit. This does not result in obtaining only one probably perfect substructure, but for each possible size of the common substructure an optimal geometric fit is produced.

The application of Pareto optimization to the superposition of vinylcyclobutane and propylcyclobutane is shown in

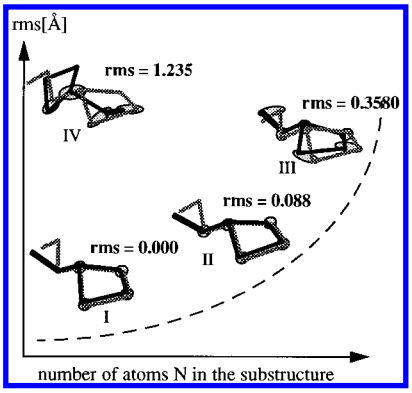

Figure 5. Results of the Pareto optimization of the superposition of vinylcyclobutane and propylcyclobutane. The plot shows the *rms* value of the superpositions versus the size *N* of the substructures. The dashed line marks the set of the Pareto solutions which cannot be improved further.

Figure 5. The atoms marked in gray are those of the substructure. The result of a Pareto optimization is a set of common substructures for which the geometric deviation cannot be further minimized. In Figure 5 four different superpositions are shown, three of them corresponding to Pareto optimality.

Superposition IV shows large geometric distances, whereas substructure I, which has the same number of matching atoms, shows a much better geometric fit. Obviously, superposition I dominates superposition IV (Figure 5). In this sense, superposition I represents a Pareto optimal solution, because no other substructure can be found which has a better geometric correspondence with this number of matching atoms. Superpositions II and III are also members of the Pareto set, and no other superpositions having the same sizes and better geometric fits can be found. Taken together, superpositions I, II, and III represent the set of equivalent Pareto solutions, and none dominates the other.

4.5. Restricted Tournament Selection. Selection drives the optimization and causes evolutionary pressure: The selection operator moves individuals from one generation into the next one based on their relative fitness. This corresponds to Darwin's evolution theory of "survival of the fittest". Most of the GAs described in the literature make use of the procedure of *roulette wheel selection*.[20,26−30] Each individual is assigned to a sector of a roulette wheel with the sector being proportional to the fitness of the individual: the better the fitness the larger the sector. Hence, the size of the sector corresponds to the probability of an individual being selected as a parent of the next generation. To prevent convergence to a suboptimal solution the population must consist of diverse and relevant members, and the rapid decrease of genetic variety is to be prevented. Some genetic algorithms use the concept of niches to prevent convergence to a suboptimal solution.[12,24,30] A niche is a local optimum, which is occupied by several similar individuals. Crowding is a method to arrive at niches:[31,32] A new individual generated by the genetic operators does not replace its direct ancestor but replaces that individual from a randomly chosen subpopulation that is most similar to the new one.

We have decided to choose another selection type to prevent premature loss of genetic information that might occur in a roulette wheel selection procedure. This alternative is called *restricted tournament selection* (RTS) (Figure 6).[29] Restricted tournament selection is found to be useful for solving multimodal problems and is a modification of a binary tournament selection. In a binary tournament selection, tournaments for a place in the new population are held between pairs of individuals chosen at random from the entire population. In this sense "restricted" means that tournaments are not made between any individuals chosen at random from the entire population but only between similar individuals.

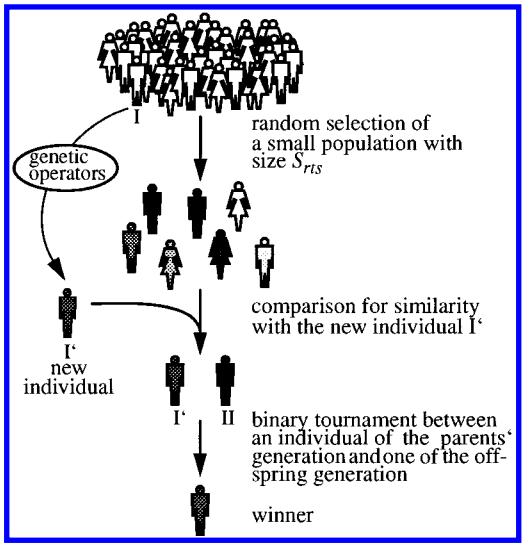

Figure 6. Schematic presentation of the restricted tournament selection. Tournaments are held only between similar individuals. This guarantees the preservation of a variety in genetic information.

Thus, restricted tournament selection (Figure 6) is based on the concept of local competition. The winners of each tournament are moved into the next generation. An element I is randomly chosen from the basic population and changed by the operators of the GA into a new element I′. For each I′ a small population with an optional member size S_{rts} is selected from the basic population. The individual II that is most similar to I′ among the chosen individuals is saved. I′ has then to compete with II for a place in the new population. This form of binary tournament restricts an individual from competing with individuals too different to it. Hence, the variety in the information is maintained. A further advantage of the described mechanism of RTS is that it enables a so-called continuous selection. A continuous selection allows individuals from different generations (e.g., II and I′ in Figure 6) to compete with each other.

2.6. The Operators. 2.6.1. Crossover. The crossover operator exchanges random parts of two individuals, e.g., partial substructures, and combines partial solutions of the MCSS search problem in a new and potentially better way. The crossover operator is the main mechanism for improvement during optimization. The crossover operator developed in our case is a two-point permutation crossover operator.[35] As a two-point crossover was implemented, two parts of equal length are randomly chosen in the match list of two parental individuals (Figure 7). Each partial list is copied to the tail of the other. After this first step, double references are introduced that have to be deleted later: If an atom of molecule I appears twice in the match list, the corresponding original match pair has to be replaced by the new one that was copied to the tail (e.g., atoms 3, 4, or 5 in Figure 7). Any double references remaining after this process are

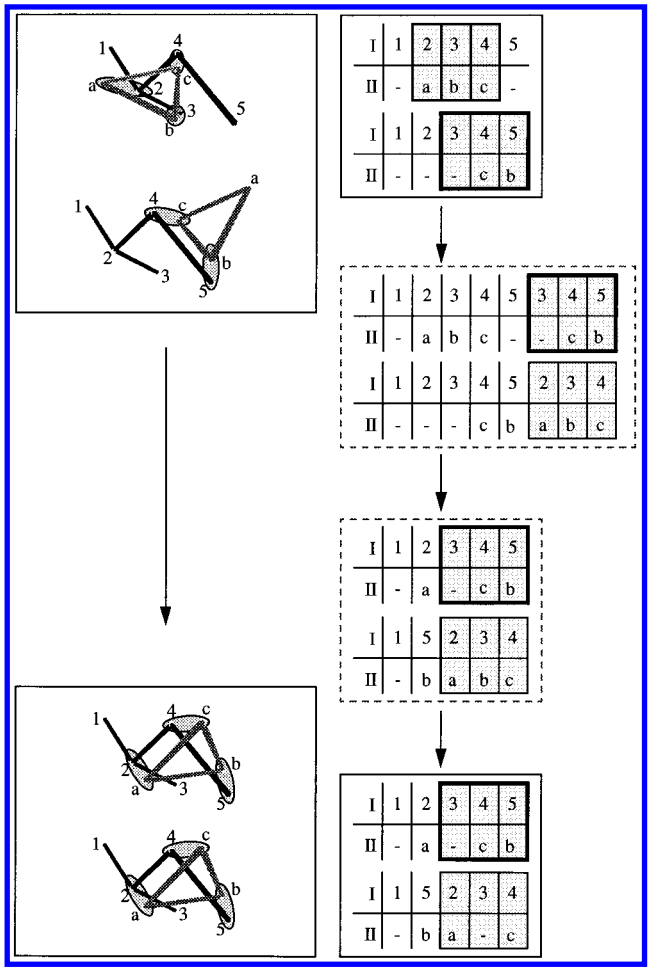

Figure 7. The mechanism of the *crossover* operator for random exchange of information between two individuals of the population. Two parts of equal length in both bit strings are chosen. Each partial list is copied to the tail of the other. Match pairs which appear twice have to be deleted in the original part. Any double references that remain after this procedure have to be replaced by randomly chosen ones not currently in the match list.

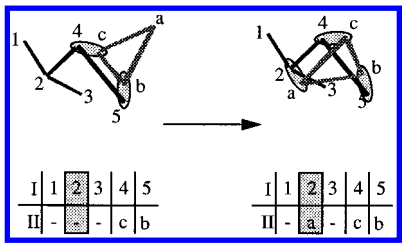

Figure 8. The mechanism of the *mutation* operator for randomly exchanging a single atom pair assignment. One match pair has to be chosen randomly and replaced by another not currently existing in the original match list.

replaced by randomly chosen ones conforming to the constraints. This procedure ensures that the match lists always have the same length and that each atom is referenced only once. In a similar manner, the crossover operator working on the representation of the torsion angles exchanges two parts of two randomly chosen strings of torsion angles. This leads to new conformations for which the geometric fit has to be assessed.

2.6.2. Mutation. The mutation operator randomly changes a pair of atoms in the match list (Figure 8). One boundary condition must be taken into account: None of the atoms is allowed to appear twice. Hence, each atom of structure I has to remain in the match list, and only a corresponding atom of structure II can be changed to an atom index not yet existing in the information string. These constraints are also valid when changing a match pair which has no corresponding atom in structure II (Figure 8).

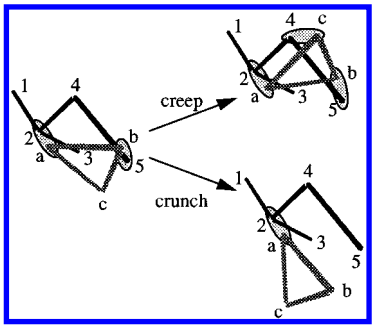

Figure 9. The mechanism of the target oriented *creep* and *crunch* operator. Creep leads to a larger substructure, whereas crunch leads to a smaller one.

The mutation operator is a random process and changes are not made to intentionally achieve the goal of a larger substructure. A matching atom pair may be lost in the entire population during the optimization process, and, therefore, certain solutions can become inaccessible. Mutation is a mechanism protecting against this irreversible loss of genetic information.

Mutation in the data structure representing the torsion angles changes one bit of a binary coded torsion angle string. As mentioned in section 2.2 a Gray-coded representation of the torsional angles was chosen. Therefore, the mutation operator which leads to a change of only one bit of the chromosome does not cause large changes in the torsion angles.

2.6.3. Creep and Crunch. In addition to the operators *crossover* and *mutation* which work nondeterministically and are based on the mechanics of genetics and evolution, it was found helpful to develop operators that are tailored to the specific MCSS search problem at hand.[36] These operators do not act stochastically like the genetic operators *crossover* and *mutation* but make use of knowledge specific to the problem to be solved. Hence, they are called "knowledge-augmented operators".[24]

To improve the efficiency of the standard GA two new operators were developed: The *creep* and the *crunch* operator. The purpose of the *creep* operator is to refine solutions found by the other operators (Figure 9). This can be achieved by adding a matching pair of atoms to the match list while obeying restrictions imposed by the spatial arrangement of the atoms. The new matching atom pair should not cause a large increase in the rms value of the original match. Therefore, two existing match pairs are selected randomly. The distances to these two match pairs to every atom not yet contained in the matchlist of molecule I and molecule II are calculated. Then, atoms of molecule I and molecule II are searched which have similar distances to the two match pairs. One of these atom pairs are chosen to build a new match pair. In this way, the creep operator leads to a "hill climbing" mechanism in the GA.

The *crunch* operator (Figure 9) is the second newly implemented operator. It acts as an antagonist to the creep operator in reducing the size of the substructure, while the creep operator increases it. The goal of the crunch operator

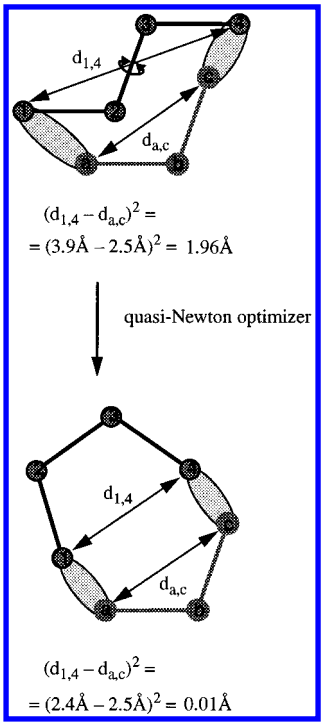

Figure 10. The mechanism of the directed-tweak technique using a quasi-Newton optimizer. The goal is to minimize the difference of corresponding atom distances ($d_{1,4}$ and $d_{a,c}$) in both structures by changing torsion angles.

is to eliminate match pairs which are responsible for bad geometric distance parameters.

First, an atom pair of the original match list is chosen. Then, a second atom pair in the match list is searched for which the differences of the distances of the corresponding first atoms (atoms of the first molecule) and second atoms (atoms of the second molecule) exceed a given tolerance value. Finally, a randomly selected new atom of the second molecule displaces the original one in the chosen match pair. This operation should help the search to avoid becoming trapped in local minima during the optimization process.

2.7. Directed Tweak. The directed tweak method reported by T. Hurst[13] was implemented in our procedure based on a quasi-Newton optimizer. The objective was to combine nondeterministic genetic mechanisms with a numerical optimizer in order to improve potential solutions. After each generation, the fitness of each individual has to be determined in order to evaluate the winner of the restricted tournament selection. Before the fitness is calculated, the geometric fit of each individual or superposition is improved by the directed tweak method. The selection is then based on these values. However, the next generation consists of the original individuals before applying the directed tweak to avoid loss of genetic information or premature convergence.

The technique makes use of a quasi-Newton optimizer (Figure 10) to minimize differences in the conformations. The squared differences of the distances of corresponding atom pairs (e.g., 1,4 and a,c in Figure 10) are used to minimize the differences in the geometry of the superimposed structures by changing torsion angles. The improved individuals thus obtained are taken to determine the geometric fitness as a starting condition for the selection. At the end of each cycle through all generations the final solutions are

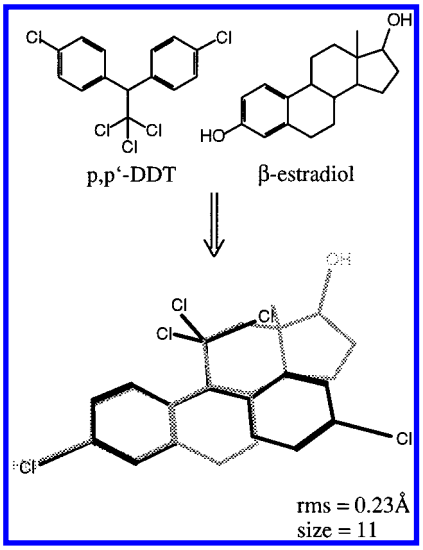

Figure 11. The structural formulas of β-estradiol and p,p'-DDT and the most frequent superposition of the three-dimensional structures of β-estradiol (grey) and p,p'-DDT (black) among 40 GA runs.

presented after the improvement of the geometric fit by the directed tweak procedure. The obtained superpositions are not limited to low-energy conformations. In contrast to the genetic algorithm published by Jones et al.,[18,20] no van der Waals energy check of the conformations is implemented. This allows one to find conformations of ligands that correspond to those found in the binding of a ligand to a receptor that do not correspond to low-energy conformations in the free state.

3. RESULTS

3.1. Restricted Tournament Selection. To determine an optimal size S_{rts} of the subpopulation which must be chosen for similarity comparison (first step in Figure 6), the calculation of the three-dimensional MCSS between β-estradiol and p,p'-DDT (Figure 11), the two neurotoxins saxitoxin and tetrodotoxin (Figure 12), and two angiotensin II antagonists of the AT1 receptor, Losartan and L-158,809 (see later in Figures 16 and 19) were evaluated.

The superposition of Losartan and L-158,809 will be explored in more detail later (see 3.4). Saxitoxin is a paralytic shellfish poison produced by the California sea mussel *mytilus californianus*, and tetrodotoxin is a toxin from the ovaries and liver of species of *tetraodontidae*, particularly the globe fish.[33] This example was chosen because it had been investigated with other superposition and similarity perception techniques.[34]

Physicochemical properties of atoms, e.g., ranges of partial atomic charges, can be chosen as superposition restrictions. Hydrogen binding properties which are very important for receptor ligand interactions and are mainly based on dipole−dipole interactions can be described indirectly by variables such as atom electronegativity, number of free electrons, number of binding hydrogens, and the atom number. The atoms to be overlaid must conform to these constraints. Other atom properties, such as distinguishing between aromatic and nonaromatic ring atoms, or ring and non-ring atoms can also be selected as mapping conditions. In the superpositions presented here, it was decided that atoms should have the same atomic number because this is most illustrative.

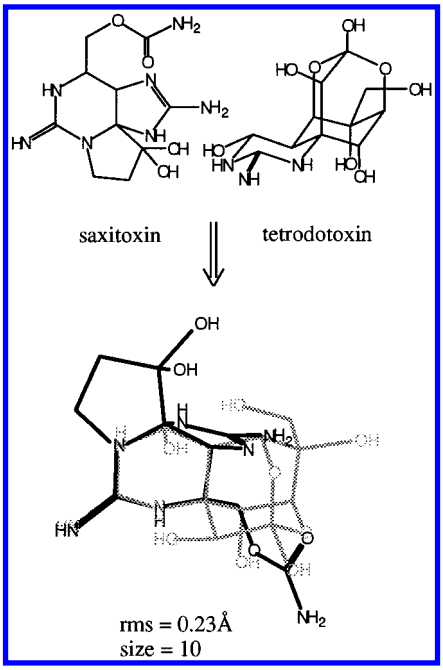

Figure 12. The structural formulas of tetrodotoxin and saxitoxin and the most frequent superposition of the three-dimensional structures of tetrodotoxin (grey) and saxitoxin (black) among 40 GA runs.

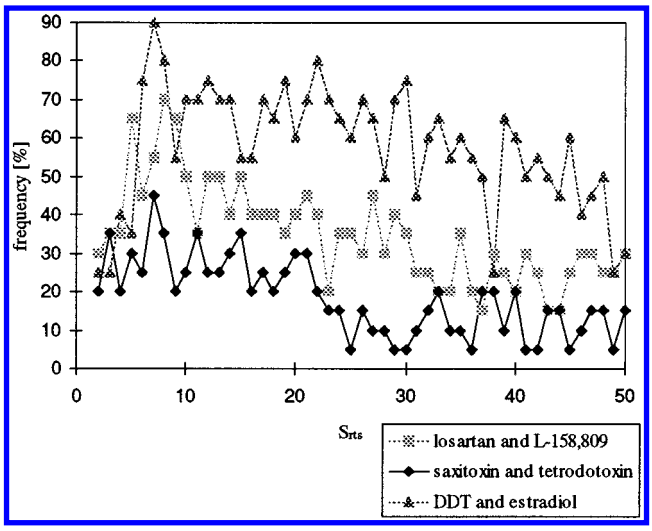

Figure 13. The frequency of achieving relevant solutions in several superpositions correlated with S_{rts}, the size of the randomly selected subpopulation in the first step of the restricted tournament selection. The superposition between β-estradiol and p,p'-DDT, tetrodotoxin and saxitoxin, and Losartan and L-158,809 were investigated. S_{rts} values of 7 and 8 are taken to be the best ones.

The size S_{rts} of the randomly chosen subpopulation was varied from 2 to 50 in steps of 1. The size i of the entire population was 100.

The following requirements were set in order to consider a superposition as an acceptable solution: for β-estradiol and p,p'-DDT a size of 11 atoms and an *rms* error smaller than 0.3 Å, for Losartan and L-158,809 a substructure size of 22 atoms and an *rms* error smaller than 0.6 Å, and for saxitoxin and tetrodotoxin substructures of a size of 10 atoms and an *rms* error smaller than 0.6 Å. The frequency of the superpositions obeying these constraints attained among 40 GA runs was determined. It is always necessary to make more than one GA run because not every GA run leads to the same solutions. The solution considered to be the best one is that reached most frequently. Figure 13 shows the relationship between this frequency of finding acceptable superpositions and the current S_{rts} in 40 GA runs. Based on the results shown in Figure 13 a value of S_{rts} of 7 or 8 can be considered as sufficient. It is true that higher S_{rts} values also cause optimal solutions, but they lead to quite an increase in computation times. A doubling of S_{rts} (from 10 to 20) leads to a 20% increase in computation time. Hence, further superpositions will be calculated with an S_{rts} of 8 in a population of 100 individuals.

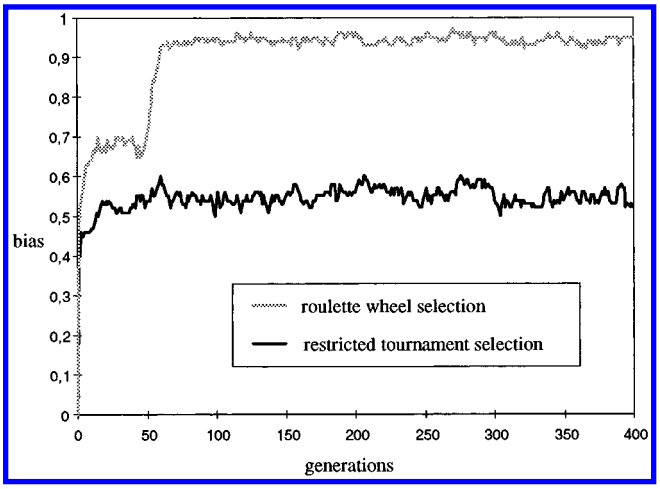

Figure 14. The average bias of the superposition of β-estradiol with p,p'-DDT as dependent on the number of the generations during optimization. The gray curve shows the roulette wheel selection and the black one the restricted tournament selection. Restricted tournament selection keeps a higher variety in genetic information.

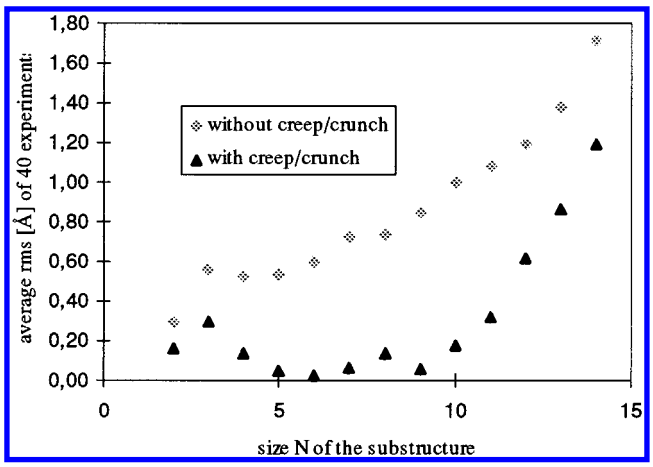

Figure 15. The comparison between GA runs with the creep and crunch operators and without the creep and crunch operators. The superposition of β-estradiol and p,p'-DDT was investigated. The plot shows the average rms for one substructure size in 40 GA runs. The correlation between the rms value and the size N of the substructures shows that applying the creep and crunch operators offers superpositions which have a much better geometric fit.

To investigate the preservation of the variety of information while passing generations from one cycle into the next one, two examples were investigated: one with *restricted tournament selection* and the other one with *roulette wheel selection*. The superposition of β-estradiol with one of its mimetics p,p'-DDT (Figure 11) was chosen as a specific problem.

The bias b (eq 4) was used as a measure for monitoring the convergence of the algorithm. An atom $A_{2,i}$ of structure

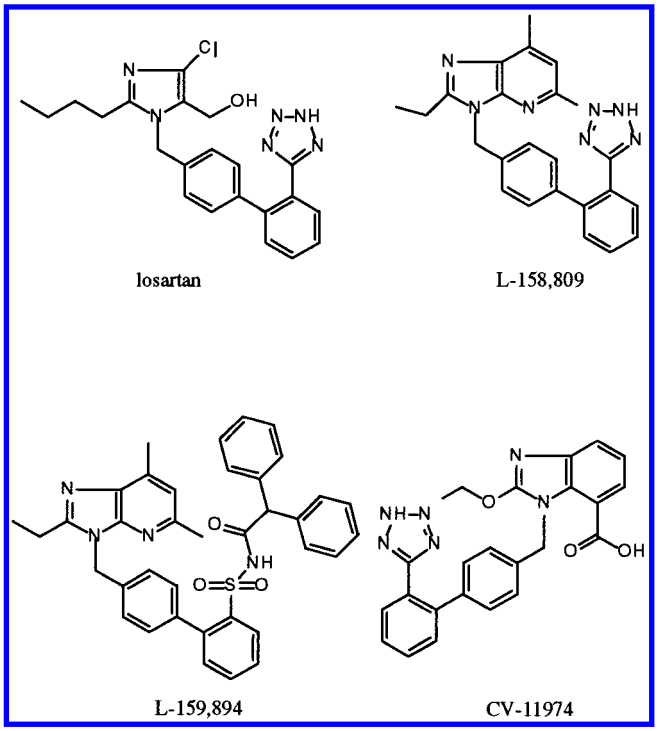

Figure 16. The structural formulas of four angiotensinII antagonists.

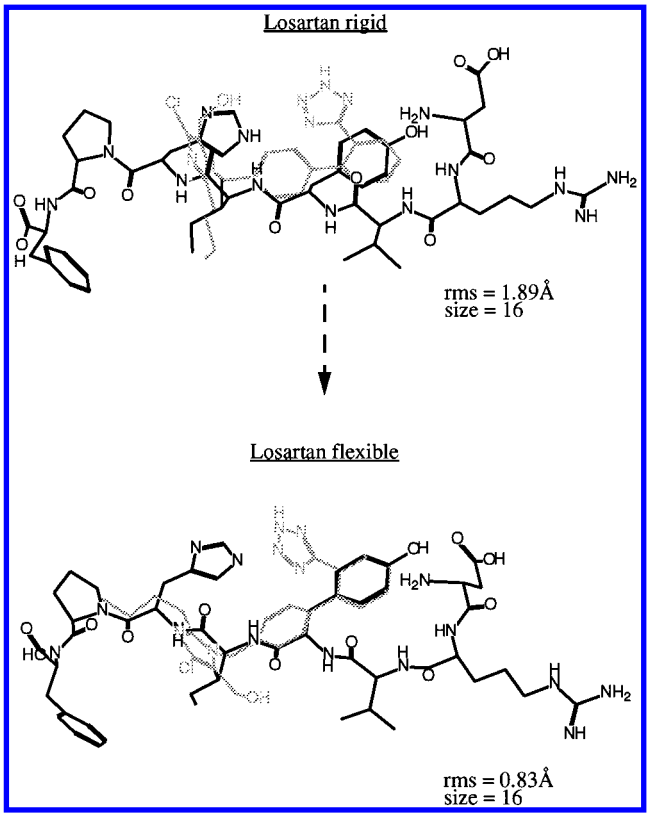

Figure 17. A comparison between the rigid and the flexible overlay of angiotensinII (black) and Losartan (grey). AngII was kept rigid as well as Losartan in the upper part of the figure. In the alignment shown in the bottom part, the conformation of Losartan was changed automatically by the program in a query-directed manner. Substructures of the same size and different rms values are chosen for the comparison. This shows the improved geometric fit for substructures of same size by using the hybrid method.

II (M_2) is most frequently matched to an atom $A_{1,i}$ of structure I (M1). The bias then measures the probability for finding

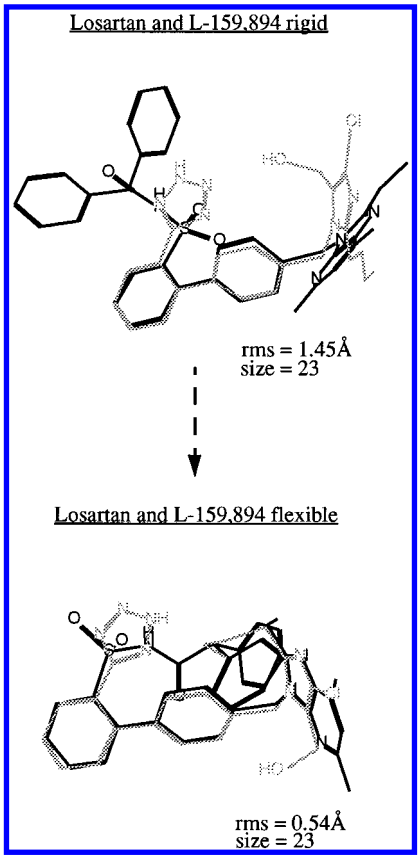

Figure 18. A comparison between the rigid and the flexible overlay of Losartan (grey) and L-159,894 (black). Both structures were regarded as rigid in the upper presentation and as flexible in the lower presentation.

this specific match pair in one individual of the whole population. The presented bias is the average value of all atoms $A_{1,i}$ and individuals.

$$b = \frac{1}{n_1}\sum_{i}^{n_1}\frac{\text{maxfreq}(A_{1,i}, M_2)}{I} \qquad (4)$$

with n_1 = number of atoms in molecule I (M_1), $A_{1,i}$, $A_{2,j}$ = an atom in M_1, respectively M_2, maxfreq ($A_{1,i}$, M_2) = the number of the atom $A_{2,j}$ of M_2, which is mapped on an atom $A_{1,i}$ of M_1 most frequently, and I = size of the population.

A bias of 0.75 means that a certain match pair appears with an average probability of 75% in each individual. The highest value the bias can attain is 1.0. A bias of 1.0 means that in each individual of a generation a specific atom of M_1 is matched always onto one and the same atom of M_2. Therefore, this same match pair would occur in all individuals.

Figure 14 presents the relationship between the average bias obtained in 40 GA runs of the superposition of β-estradiol and *p,p′*-DDT and the 400 generations in a roulette wheel selection and in a restricted tournament selection (RTS). During the entire optimization process the bias of the population of the RTS is always lower than the bias of the population of the roulette wheel selection. Hence, a higher information variety is guaranteed while passing all generations, and the probability for finding an optimal solution is strongly increased.

3.2. Determination of Operator Probabilities. The operators crossover, mutation, and creep and crunch are

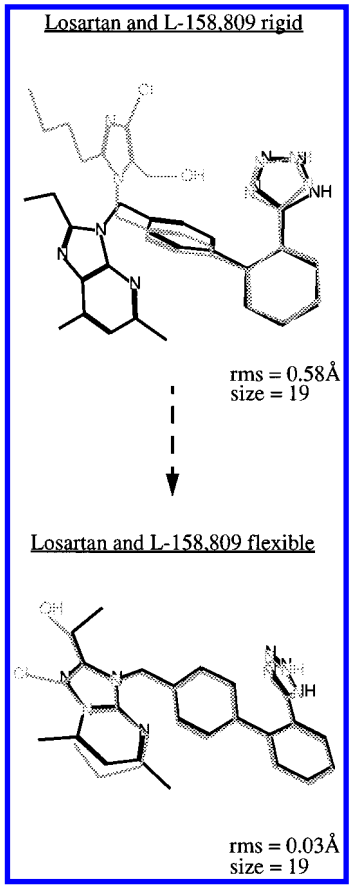

Figure 19. A comparison between the rigid and the flexible overlay of Losartan (grey) and L-158,809 (black). Both structures were regarded as rigid in the upper presentation and as flexible in the lower presentation.

Table 1. The Best Probabilities for the Employment of the Genetic Operators

p_{mut}	p_{cross}	p_{tormut}	$p_{torcross}$	p_{creep}	p_{crunch}
0.4	0.6	0.7	0.4	0.5	0.3
0.5	0.9	0.9	0.9	0.7	0.5
1.0	0.8	1.0	0.8	0.7	0.9

applied with a probability that has to be selected. The best probabilities were determined in several series of tests again by determining the three-dimensional MCSS of β-estradiol and *p*,*p*′-DDT. Optimized values for the probability of application of mutation and crossover to the match lists (p_{mut}, p_{cross}), mutation and crossover to the torsion angles (p_{tormut}, $p_{torcross}$) and creep and crunch (p_{creep}, p_{crunch}), given in Table 1 were found.

The best combination of probability values was determined by calculating the frequency of finding an optimal solution (size of the substructure $N = 11$ and rms ≤ 0.2) in 40 GA runs. The higher this frequency the better the combination of the probabilities of applying the operators. Based on the results shown in Table 1, all further calculations were made with a p_{mut}/p_{cross}-combination of 0.4/0.6, a $p_{tormut}/p_{torcross}$-combination of 0.7/0.4, and a p_{creep}/p_{crunch}-combination of 0.5/0.3.

3.3. Creep and Crunch Operators. Test series showed that applying the creep and crunch operators led to better solutions or found optimal solutions much more often. Again, the superposition of β-estradiol with *p*,*p*′-DDT has been used to show this. Forty GA runs were made with a p_{creep}/p_{crunch} of 0.0/0.0 and the default values for the probabilities mentioned in 3.2 and shown in Table 1. Another 40 GA runs were made with a p_{creep}/p_{crunch} of 0.5/0.3 and with the default employment probabilities given in Table 1. The comparison between the GA runs without the creep and crunch operators and those including them is shown in Figure 15.

It can be seen that using the creep and crunch operators leads to much better solutions: substructures of all sizes have better geometric fits (lower rms values) in comparison to those calculated without the creep and crunch operators.

3.4. Application of the Method to Angiotensin II Antagonists. Angiotensin II (AngII) antagonists are of high medical importance because they promise to inhibit the last step in the renin−angiotensin system (RAS), the binding of AngII to the AT_1 or AT_2 receptor. This system was shown to be a key element in blood pressure regulation and electrolyte/fluid homeostasis.[37] The former clinical approach to the treatment of hypertension and congestive heart failure was to block the conversion of AngI to AngII by angiotensin converting enzyme (ACE) inhibitors. But ACE is a nonspecific protease and is also responsible for the degradation of bradykinin as well as other peptides such as enkephalins.[38] Therefore, ACE inhibitors lead to several side effects, which are mainly attributed to bradykinin potentiation.[39] Large efforts are being made to search for new methods for blocking the RAS. This has led to attempts to inhibit angiotensin II binding to the AT1 or AT2 receptor, the last step in the reaction cascade. Takeda Chemical Industries[40] was built on the concept that nonpeptide AngII antagonists would lack the disadvantages of peptide AngII receptor antagonists, in particular rapid cleavage on oral dosing. The compounds developed by Takeda Chemical Industries were further investigated by DuPont and culminated in the discovery of Losartan (DuP 753, Figure 16), which is the prototype of a new class of potent, orally active, nonpeptide AngII receptor antagonists.[40]

The structures of the AngII receptor antagonists (Figure 16) and the nomenclature used are taken from the publication of Wexler et al.[40] The three-dimensional structures of these molecules and the peptide angiotensin II (Asp-Arg-Val-Tyr-Ile-His-Pro-Phe) were generated by the 3D structure generator CORINA.[1,2] It should be emphasized again that any single conformation of a molecule is sufficient for initiating the superposition method presented here.

As Pareto optimality is used (see 2.4) not only one superposition per GA run is obtained but also a set of superpositions with increasing sizes and optimized geometric fits is obtained. Every superposition presented here is taken from the Pareto solution set that was obtained most frequently in the GA runs made.

In the following discussion it will be shown that the hybrid method has a higher effectiveness and reaches better superpositions, as compared to a rigid superposition of 3D structures.[12] Four examples, the superposition of AngII and Losartan (Figure 17), of Losartan and L-158,809 (Figure 18), of Losartan and L-159,894 (Figure 19), and of Losartan and CV-11974 will be used to demonstrate this.

Superpositions between the octapeptide angiotensin II Asp-Arg-Val-Tyr-Ile-His-Pro-Phe and Losartan are shown in Figure 17. The superposition presented in the upper half of Figure 17 is generated by keeping both structures rigid during

Table 2. Parameters of a GA Run

parameter	value
rts selpar	8
size of the population	100
no. of generations	1000
no. of experiments	40

the optimization process. In the calculation of the superposition shown in the lower part of Figure 17, the 3D structure of the octapeptide was kept rigid, while the structure of the small molecule was adjusted by the optimization method to fit the octapeptide conformation. This is an option of the program where one molecule is taken as a template. The octapeptide is regarded as a template, and the conformation of the antagonist is changed to approximate the template as well as possible.

The parameters used in these runs of the genetic algorithm were the default values of the probabilities for applying the operators as shown in Table 1. Additionally, the parameter values presented in Table 2 were used.

The number of generations had to be 1000 in order to achieve convergence in the superposition of such large molecules. The average time for the entire run was 2 h on a SGI origin 200 (single R10000, 180 MHz). Using more than 1000 generations did not result in better solutions. The solutions shown in Figure 17 of the overlay of AngII and Losartan were taken from the Pareto sets that were reached most frequently during the 40 GA runs. In Table 3 the entire Pareto set for the flexible superposition (bottom of Figure 17) is shown. For each size of the substructure (number of atoms *N* of the substructure) one optimized superposition is obtained.

The superposition with $N = 16$ was chosen for further discussion because it is the superposition with the highest size and an rms value smaller than 1.0 Å.

The C-terminal segment of AngII is regarded as the structure element responsible for receptor affinity, and it can be seen that the small molecules mimic the interactions made by this part of the peptide.[40] Only the flexible superposition presented in the lower part of Figure 17 leads to a substructure aimed at the C-terminus of AngII. In comparison with the corresponding superposition of Wexler et al. the main part of the common substructure also comprises the Pro-His-Ile-Tyr residues, the C-terminus. The rigid overlay in the upper part of Figure 17 has an rms value too high for the given size of 16 atoms of the substructure. Thus, in this case the geometric fit is so far from the one that is obtained by the flexible superposition that it should not be used for any further similarity evaluation.

In the overlay of Losartan with L-158,809 (Figure 18), with L-159,894 (Figure 19) and with CV-11974 (Figure 20), the superpositions presented in the upper part of the figures were calculated keeping both structures rigid during the superposition process. In the flexible overlay shown in the lower part of Figures 18−20, both molecules were considered as flexible, and their torsional angles were changed during

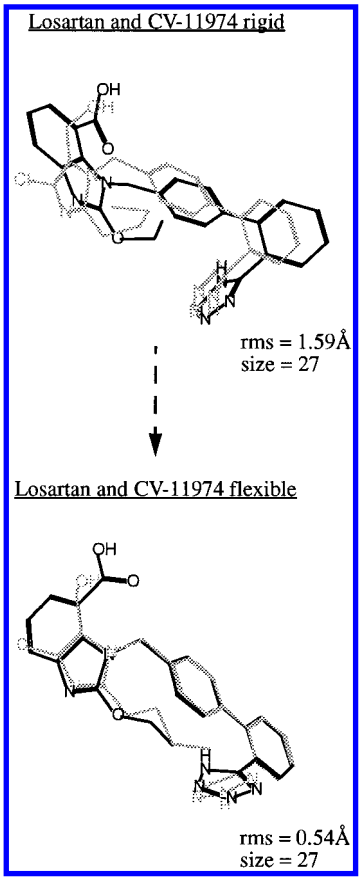

Figure 20. A comparison between the rigid and the flexible overlay of Losartan (grey) and CV-11947 (black). Both structures were regarded as rigid in the upper presentation and as flexible in the lower presentation.

overlay. Therefore, the conformations of both molecules were adapted to each other during superposition.

The superpositions shown in Figures 18−20 were calculated with 400 generations. The average execution time of one GA run regarding the structures as flexible was 7 min 20 s on a SGI origin 200 (single R10000, 180 MHz). The average run time for a rigid superposition is 43 s. Thus, consideration of conformational flexibility leads to about a 10-fold increase in computation time. Again, the 40 GA runs reveal several different Pareto solutions, but only the most frequent superpositions are presented here.

Comparison of the rigid overlays with the flexible ones (Figures 17−20) shows the increased geometric fit of the superpositions (much lower rms values), when comparing substructures of the same size, by using the flexible hybrid method instead of a rigid genetic algorithms. In the flexible superpositions, in addition to the two aromatic rings, the alkyl side chains of the imidazole rings and the imidazole rings of Losartan were also aligned in corresponding regions of L-158,809, L-159,894 or CV-11974. These results agree with the SAR studies of Lin et al.[41] that point out that the lipophilic and basic sites are responsible for high affinity binding. Using the flexible algorithm the superpositions are

Table 3. The Entire Pareto Set of a Superposition of AngII and Losartan

size of the substructure	8	9	10	11	12	13	14	15	16
rms [Å]	1.34	1.10	1.23	1.00	0.90	1.10	0.71	1.11	0.83
size of the substructure	17	18	19	20	21	22	23	24	25
rms [Å]	0.99	1.15	1.22	1.34	1.28	1.49	1.54	1.85	1.89

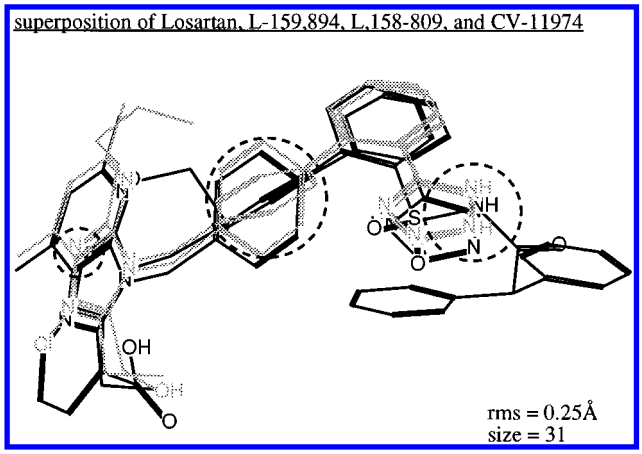

Figure 21. The superposition of Losartan (grey), L-159,894 (black), L-158,809 (grey), and CV-11974 (black). The common substructure contains the recently reported pharmacophoric patterns: two aromatic rings, a basic site in the form of the imidazole rings, and a hydrophobic site. The circles of broken lines indicate the pharmacophoric points in equivalence with Figure 18 of ref 20 (Jones et al. "A genetic algorithm for flexible molecular overlay and pharmacophoric elucidation").

not limited to those between low-energy conformations. Hence, the probability of finding the conformation that the structure obtains during receptor binding is highly increased. This can thus serve as a basis for the understanding of ligand receptor interactions.

The program also offers the possibility of overlaying more than two molecules. The molecules are then not aligned in a concerted process but in a stepwise manner. The third structure is added to the set obtained in a superposition of molecules (both of them being kept flexible). Then, the fourth molecule is superimposed onto the results obtained from these three molecules, etc. An unlimited number of molecules can be handled this way. The process is repeated with each conceivable sequence of molecules considered. As an example, the structures of four AngII antagonists, Losartan, L-159,894, L-158,809, and CV-11974 (Figure 16), were aligned. The results obtained in this example were not dependent on the order of feeding the molecules into the superposition process. The presented superposition is the one of the Pareto set obtained most frequently in 40 GA runs (Figure 21). Thus, the stepwise calculation is here equivalent to a concerted process.

Again the maximum common 3D substructure contains the two aromatic rings and the imidazole rings and thus comprises the hydrophobic site and the basic site so important for receptor binding. These fragments correspond to the pharmacophore elements reported for AngII antagonists.[41] Jones et al.[20] (Figure 18 in ref 20) presents an equivalent overlay of a set of angiotensinII antagonists. Two of them (Losartan and L-158-809) are part of our set of molecules. Jones et al. point out that a basic nitrogen and an aromatic ring are responsible for receptor affinity. These two pharmacophoric points emerge from the superposition shown in Figure 21 as the nitrogen atom of an imidazole ring, a structure element contained in all molecules, and a benzene ring, also contained in all four molecules. These two points are indicated by circles with broken lines. A third superposition is indicated which results from a tetrazole ring in Losartan, L-158,809, and CV-11974 and a sulfonamide in L-159.894. It is interesting to note that this site has an NH group of appreciable acidity pointing out a third pharmacophoric site.

4. DISCUSSION

Jones, Willett, and Glen[18−20] have also developed a genetic algorithm for the flexible molecular overlay of 3D structures. Their approach has many features in common with our work, and a comparison of the two methods is therefore appropriate.

The differences start with the representation of chemical structures. Jones et al. represent the structures by features that consist of hydrogen-bond donor protons, lone pairs on acceptor atoms (as defined by SYBYL), and rings. We, however, have decided to represent molecules at the atomic level, putting each atom of a molecule into the data structure. This prevents any bias from creeping into the data structure, and it also gives the program the possibility to perceive ring atoms and chain atoms as equivalent (e.g., both being hydrophobic groups) in certain situations.

Clearly, we have to pay a price for this: the chromosomes become much longer because we include all atoms of the molecules to be compared. This by itself must lead to a strong increase in computation times.

Furthermore, our data structure does not prohibit the use of pharmacophoric points. A simple preprocessing step to determine interesting features in the molecules under study would enable us to base the superposition on these features. This can be useful in some applications, as those features, like pharmacophoric points, represent additional information about the kind of interactions frequently observed when a ligand binds to a receptor.

To keep computation times as low as possible we have decided, as a first attempt, not to check for van der Waals interactions during the superposition of molecules. To our surprise, we have not yet found in our studies any particular problems of atoms of a molecule bumping into each other. We assume that such conformations are generated for pairs of molecules that are both highly flexible and differ drastically in size. If this is not the case, superpositions of structures without van der Waals clashes will have a far higher fitness than those that exhibit severe van der Waals interactions. This is a very interesting result that clearly needs more extensive investigations which are pursued by us.

In the approach of Jones et al., the survival of individuals in populations is controlled by the fitness values calculated from the fitness function. This fitness function includes a variety of factors such as van der Waals energy, volume integral, and similarity scores obtained from features that the superimposed molecules have in common. The fitness score was obtained by calculating a weighted sum from these various factors. This balancing act has to account for conflicting influences, and a heuristic solution is proposed for this problem.

We, on the other hand, do not try to come up with an artificial solution to the problem that a larger maximum common substructure by definition must have a larger deviation in the coordinates of the superimposed atoms. Rather, we keep these conflicting criteria open by explicitly showing—and calculating—the Pareto optima for each number of superimposed atoms. Thus, it is still in the hands of the user whether he/she wants to define a limit for the

standard deviation in the coordinates of the superimposed atoms or give a number for the atoms that have to be superimposed (or work with default values).

The selection of individuals for the next generation of the genetic algorithm is based in the work of Jones et al.[20] on the procedure of roulette-wheel selection, selecting members of the new population on the basis of their fitness value. We have chosen a restricted tournament selection. It is hard to compare the results of these two different selection methods. In any case, our decision, not to work with a fitness function forced us to abandon a roulette-wheel selection. The restricted tournament selection is known to avoid premature loss of genetic information, and it apparently works fine in our approach.

Jones et al. only work with two genetic operators, crossover and mutation, that are also part of our approach. However, we have also included two operators, creep and crunch, that have been specifically coined for the problem at hand, the search for the maximum common 3D substructure.

A major feature of our approach is that we have combined optimization by a genetic algorithm with a classical optimization method, the directed tweak method.[13]

As far as the results are concerned, both Jones et al.[20] and we have studied angiotensin II receptor antagonists. The datasets of molecules are different, but two of the six molecules of Jones et al. (Figure 17 in ref 20) are identical to the molecules of our dataset (Figure 20), thus somehow allowing a comparison. Visual inspection of Figure 18 of ref 20 with our results (Figures 19 and 21) indicates quite an agreement in the general structure of the superposition (see also 3.4).

Overall, it is quite difficult to compare the effectiveness and results of the two approaches to the flexible superposition of molecules by a genetic algorithm, the method of Jones et al.[20] and our approach. This would need studies on the same datasets by the two groups, an endeavor that seems very worthwhile.

5. CONCLUSIONS

The program described here overlays and aligns structures independent of the initially chosen conformation. Therefore, only one conformation per structure is necessary, and, thus, the program can work even when only one conformation of a compound is stored in a database. The automatic finding of structural similarities has its particular efficiency in a hybrid method, the combination of a genetic algorithm and a numerical optimization method, the directed-tweak method. The genetic algorithm process leads to an optimization of the assignment of the atoms in the form of match lists. An optimization of the geometric fit by adapting the conformations of molecules to each other is obtained by the combination of the genetic algorithm with the directed-tweak technique. The genetic algorithm is further improved by two additional operators which are goal-directed, tailored to the superposition problem: the creep and crunch operators. The applied restricted tournament selection avoids the loss of variety in genetic information during the optimization process. The problems studied here have to optimize several conflicting criteria; therefore, always a set of so-called Pareto solutions is obtained at the end of each GA run.

During optimization of the superposition, the conformations of both structures were adapted to each other. The program also offers as another possibility a template method which allows one to consider one structure as rigid and then adapts the conformation of the other molecule onto this template. Furthermore, an unlimited number of structures can be treated. These methods were applied to a set of angiotensin II antagonists. The superpositions presented are the most frequent ones among 40 GA runs and were compared to recently reported SAR studies on angiotensin II antagonists. The obtained substructure elements correspond to the known pharmacophore elements of AngII antagonists.

ACKNOWLEDGMENT

We thank the Bundesminister für Bildung, Wissenschaft, Forschung und Technologie (bmb+f) for the support of our work by the program "Molekulare Bioinformatik" through the project 01 IB 305 and the "Fonds der Chemischen Industrie" for a scholarship (Promotionsstipendium) to Sandra Handschuh.

REFERENCES AND NOTES

(1) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(2) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(3) Sheridan, R. P.; Nilakantan, R.; Rusinko, III, A.; Baumann, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255−260.
(4) Greene, J.; Kahn, S. D.; Savoi, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.
(5) Martin, Y. C.; Danaher, E. B.; May, C. S.; Weininger, D. MENTHOR, a Database System for the Storage and Retrieval of Three-Dimensional Molecular Structures and Associated Data Searchable by Substructural, Biological, Physical or Geometric Properties. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 15−29.
(6) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An Integrated Tool for Computer-Assisted Molecular Design and Pharmacophoric Pattern Recognition from Geometric, Steric and Substructure Searching of Three-Dimensional Molecular Structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225−251.
(7) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity using Inter-Atomic Distances. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 455−474.
(8) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.
(9) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.
(10) Lauri, G.; Bartlett, P. A. CAVEAT: A Program to Facilitate the Design of Organic Molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51−66.
(11) Fisanick, W.; Cross, K. P.; Rusinko, III, A. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.
(12) Wagener, M.; Gasteiger, J. Die Bestimmung gröβter deckungsgleicher Teilstrukturen mit einem genetischen Algorithmus: Anwendung in der Syntheseplanung und zur strukturellen Analyse biologischer Aktivität. *Angew. Chem.* **1994**, *106*, 1245−1248. The Determination of Maximum Common Substructures by a Genetic Algorithm: Application in Synthesis Design and for the Structural Analysis of Biological Activity. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 1189−1192.
(13) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.
(14) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation. Research Studies Press: Letchworth, U.K., 1988.

(15) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Constrained Search of Conformational Hyperspace. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3-21.
(16) Payne, A. W. R.; Glen, R. C. Molecular Recognition Using a Binary Genetic Search Algorithm. *J. Mol Graphics* **1993**, *11*, 74−91.
(17) Fontain, E. Application of Genetic Algorithms in the Field of Constitutional Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 748−752.
(18) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(19) Jones, G.; Willett, P.; Glen, R. C., Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking, *J. Mol. Biol.* **1997**, *276*, 727−748.
(20) Jones, G.; Willett, P.; Glen, R. C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.
(21) Wild, D. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures. Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 159−167.
(22) Thorner, D. A.; Wild, D. J.; Willett, P.; Wright M. Similarity Searching of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 900−908.
(23) Clark, D. E.; Jones, G.; Willett, P.; Kenny, P. W.; Glen, R. C. Pharmacophoric pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational−Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197−206.
(24) Goldberg, D. E. *Genetic Algorithms in Search Optimization and Machine Learning*; Addison-Wesley Publishing Company: New York, 1989.
(25) Fonseca, C. M.; Fleming, P. J. Genetic Algorithm for Multiobjective Optimization: Formulation, Discussion and Generalization. In *Proceedings of the 5th International Conference on Genetic Algorithms*; Forrest, S., Ed.; Morgan Kaufmann Publishers: San Mateo, CA, 1993; pp 416−423.
(26) Jones, G. Genetic and Evolutionary Algorithms. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, UK, in print.
(27) Treasurywala, A. M. Genetic Algorithms. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, UK, in print.
(28) Venkatasubramanian, V.; Sundaram, A. Genetic Algorithms: Introduction and Applications. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, UK, in print.
(29) Harik, G. R. Finding Multimodal Solutions Using Restricted Tournament Selection. In *Proceedings of the 6th International Conference on Genetic Algorithms*; Eshelman, L. J., Ed.; Morgan Kaufmann Publishers: San Francisco, CA, 1995; pp 24−31.
(30) Brown, R. D.; Jones, G.; Willett, P. Matching Two-Dimensional Chemical Graphs Using Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 63−70.
(31) DeJong, K. A. An Analysis of the Behaviour of a Class of Genetic Adaptive Systems. Dissertation, University Michigan, 1975.
(32) Goldberg, D. E.; Deb, K. An Investigation Of Niche and Species Formation in Genetic Function Optimization. In *Proceedings of the 3rd International Conference on Genetic Algorithms*; Schaffer, D., Ed.; Morgan Kaufmann Publishers: San Mateo, CA, 1989; pp 42−50.
(33) *The Merck Index*; Budavari, S., O'Neil, M. J., Smith, A., Eds.; Merck & Co., Inc.: Whitehouse Station, NJ, 1996.
(34) Dean, P. M. Molecular Recognition: The Measurement and Search for Molecular Similarity in Ligand−Receptor Interaction. In *Concepts and Application of Molecular Similarity*; Johnson, A. M., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990; pp 211−238.
(35) Oliver, I. M.; Smith, D. J.; Holland, J. R. C. A Study of Permutation Crossover Operators on the Travelling Salesman Problem. *Genetic Algorithms and their Applications:* In *Proceedings of the 2nd International Conference on Genetic Algorithms*; Grefenstette, J. J., Ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1987; pp 224−230.
(36) Davis, L. Adapting Operator Probabilities in Genetic Algorithms. In *Proceedings of the 3rd International Conference on Genetic Algorithms*; Schaffer, D., Ed.; Morgan Kaufmann Publishers: San Mateo, CA, 1989; pp 61−69.
(37) Sealy, J. E.; Laragh, J. H. The Renin-Angiotensin-Aldosterone System for Normal Regulation of Blood Pressure and Sodium and Potassium Homeostasis. In *Hypertension: Pathophysiology, Diagnosis and Management*; Laragh, J. H., Brenner, B. M., Eds.; Raven Press: New York, 1990; pp 1287−1317.
(38) Böhm, H. J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign*, Spektrum Akademischer Verlag: Heidelberg, 1996.
(39) Levens, N. R.; de Gasparo, M.; Wood, J. M.; Bottari, S. P. Could the Pharmacological Differences Observed between Angiotensin II Antagonists and Inhibitors of Angiotensin Converting Enzyme be Clinically Beneficial? *Pharmacol. Toxicol.* **1990**, *71*, 241−249.
(40) Wexler, R. R.; Greenlee, W. J.; Irvin, J. D.; Goldberg, M. R.; Prendergast, K.; Smith R. D.; Timmermans, P. B. M. W. M. Nonpeptide Angiotensin II Receptor Antagonists: The Next Generation in Antihypertensive Therapy. *J. Med. Chem.* **1996**, *39*, 625−656.
(41) Lin, H.-S.; Rampersaud, A. A.; Zimmerman, K.; Steinberg, M. I.; Boyd, D. B. Nonpeptide Angiotensin II Receptor Antagonists: Synthetic and Computational Chemistry of N-[[4-[2(2H-Tetrazol-5-yl)-1-cycloalken-1-yl]phenyl]methyl]imidazole Derivatives and Their in Vitro Activity. *J. Med. Chem.* **1992**, *35*, 2658−2667.

CI970438R