Conformational Coverage by a Genetic Algorithm

Ovanes Mekenyan,*,† Dimitar Dimitrov,‡ Nina Nikolova,§ and Stoyan Karabunarliev†

Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov", 8010 Bourgas, Bulgaria, Institute of Water Problems, Bulgarian Academy of Sciences, 5000 Veliko Tarnovo, Bulgaria, and Central Laboratory for Parallel and Distributed Processing, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

Received February 16, 1999

A new approach for coverage of the conformational space by a limited number of conformers is proposed. Instead of using a systematic search whose time complexity increases exponentially with degrees of freedom, a genetic algorithm (GA) is employed to minimize 3D similarity among the conformers generated. This makes the problem computationally feasible even for large, flexible molecules. The 3D similarity of a pair of conformers is assumed to be reciprocal to the root-mean-square (rms) distance between identical atomic sites in an alignment providing its minimum. Thus, in contrast to traditional GA, the fitness of a conformer is not quantified individually but only in conjunction with the population it belongs to. The approach handles the following stereochemical and conformational degrees of freedom: rotation around acyclic single and double bonds, inversion of stereocenters, flip of free corners in saturated rings, and reflection of pyramids on the junction of two or three saturated rings. The latter two were particularly introduced to encompass the structural diversity of polycyclic structures. However, they generally affect valence angles and can be restricted up to a certain level of severity of such changes. Stereochemical modifications are totally/selectively disabled when the stereochemistry is exactly/partially specified on input. Three quality criteria, namely robustness, reproducibility, and coverage of the conformational space, are used to assess the performance of various GA experimental settings employed on four molecules with different numbers of conformational degrees of freedom. It was found that with the increase of the ratio between the number of parents and children, the reproducibility of GA runs increases whereas their robustness and coverage decrease. Force field optimization of conformers for each generation was found to improve significantly the reproducibility of results, at the cost of worse conformational coverage.

INTRODUTION

Quantitative structure−activity relationships (QSAR) often depend implicitly on the 3D molecular models adopted for the chemicals under study. This certainly applies to receptor-site mapping models dealing directly with molecular shapes and fields. Correlative QSARs may also be influenced indirectly when employing electronic quantum chemical descriptors that generally depend on the 3D structure. Typically QSARs rely on a single conformer to represent a chemical under study, while all others are neglected. In the best case, the representative conformer is the one of lowest potential energy for the isolated molecule or the one observed in the crystal phase. However, for flexible molecules such analysis is likely to fail to identify conformers that, albeit possibly less stable, have the required shape and electronic properties to interact with receptors. Indeed, the most stable conformers are often less likely to interact with a solvent or macromolecule.[1] At the macromolecular binding sites, conformational states can be populated which are substantially different than the isolated, lowest energy one or the crystal-phase one. This holds especially for enzyme-mediated reactions where enzyme-induced distortions in the direction of the transition state drive the molecules even out of the local potential energy minima.

To overcome this deficiency of conventional QSARs, new techniques which encompass molecular flexibility in the context of biological environment were introduced and employed. The basic assumption is that a molecule can interact as a variety of conformers, with solvation and binding interactions capable of compensating for the energy increase of higher conformational states. Thus, exploration of the conformational space of molecules has become an important issue in relating molecular structure to biological behavior. Recently, a method freely called "dynamic" was proposed for correlative QSARs, where descriptors reflect a subset of molecular conformers, rather than a single one.[2−6] By means of the so-called COREPA technique, conformational coverage was used to reveal areas in the descriptor space which are most populated by the conformers of the biologically active molecules and least populated by the inactive ones, simultaneously.[7,8] Pharmacophore mapping methods based on common functional groups[9] or molecular fields[10] also require exploration of the conformational space for flexible molecules.[11,12]

Usually conformational preferences are evaluated for the isolated molecule. In biological tissues, they strongly depend on the local environment imposed by the solvent and/or receptor and are explicitly taken into account in the theory of induced fit of ligands into receptors and allosteric effects

* Corresponding author.
† University "Prof. As. Zlatarov".
‡ Institute of Water Problems, Bulgarian Academy of Sciences.
§ Central Laboratory for Parallel and Distributed Processing, Bulgarian Academy of Sciences.

10.1021/ci990303g CCC: $18.00
Published on Web 07/15/1999

within protein systems.[13] In this respect, the notion of a continuum of accessible conformers under a potential-energy threshold instead of a discrete local minima set has been introduced[2,7,14−17] and employed in QSAR modeling and flexible database search.[11,18−20]

Conformer generation methods typically rely on a systematic search in the conformational space, followed by screening according to a user-defined requirement. Thus, one ultimately arrives at a reduced and manageable subset of conformers yet after the exhaustive generation. A commonly used deterministic, systematic search algorithm is the one included in the SYBYL molecular modeling package[21] (see also Dammkoehler et al.[22]) where all cyclic moieties are considered rigid. A more versatile system for exhaustive conformational search is the internal coordinate tree-searching procedure by Lipton and Still.[23] Proceeding from a local minimum in the conformational space, all possible combinations of rotamers at a certain torsion angle resolution are generated. Only those conformers which pass geometrical tests introduced to reject high-energy structures are retained. A similar approach has been used for exploring the orientational and conformational space of flexible ligands at macromolecular receptor sites.[24] In a further development of the tree-searching technique,[15] one initiates from the molecular topology and generates all conformers consistent with steric and stereochemical constraints. Of special note is that the tree-searching techniques resolve the conformational flexibility of cyclic saturated moieties, as opposed to other techniques where conformational degrees of freedom are restricted to rotations around noncyclic bonds. A practical drawback of all systematic algorithms is the fact that they inevitably encounter the problem of combinatorial explosion for large, flexible molecules: the number of possible solutions is exponential versus the number of degrees of freedom. One way to circumvent the combinatorial complexity of systematic algorithms is the stochastic approach. Chang et al.[25] describe a Monte Carlo methodology for conformational search and demonstrate its efficiency in finding low-energy conformers. A limitation of the stochastic approaches is its low efficiency when a large set of constraints needs to be imposed during the generation. Then, a large number of iterations per structure should be performed to fit the structure to those constraints.

Genetic algorithms (GAs) are a class of nondeterministic algorithms taking an intermediate position between both extremes: the systematic and stochastic methods. GAs provide nondeterministic solutions to the combinatorial optimization problem with constraints at a low computational cost. There have been several recent reports of the use of GAs for conformational search and analysis.[18,26−30] In general, GAs mimic the Darwinian evolution model according to which the process of natural selection tends to favor the survival of those children best adapted to the environment. Applied to conformational searching, GAs are trying to optimize generation of conformers with respect to steric and/or energetic criteria.[18,29−31] Most of the conformer searching GA implementations so far explore only the conformational subspace limited to rotatmers.[29,30] To account for flexibility of cyclic fragments, Payne and Glen[18] included flips of the so-called ring free corners which involve rotations around two bonds from a ring. 3D structural diversity generally includes variations of geometry of polycondensed saturated rings. It has been shown,[7,32] for instance, that sex steroids are conformationally flexible while retaining the stereochemistry of the natural enantiomer. Moreover, steroid conformer interconversions were assessed as kinetically and thermodynamically feasible.

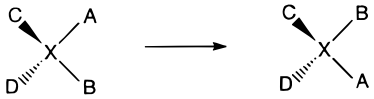
Applied to conformational search, GAs have been limited to enumeration of potential energy minima and identification of the lowest energy structure. Solutions have been reported for specific classes of molecules.[29,30] Herein we describe a GA approach aimed at optimal conformational coverage. It concentrates on the generation of a set of structurally diverse conformers that are energetically accessible under certain conditions and do not necessarily represent local potential energy minima in the conformational space. Genetic optimization is used to replace an exhaustive conformer search. The structural dissimilarity of conformers is defined in 3D Cartesian space, rather than as a Euclidian distance in the space of internal conformational variables. It is quantified by the so-called root-mean-square (rms) distance[33] between pairs of conformers. This is the average Cartesian distance between identical atomic sites of the two conformers in the mutual alignment providing its minimum. The GA is directed toward the attainment of a set of conformers exhibiting greatest rms distances among themselves. Thus, unlike conventional GAs, the fitness function does not pertain to each conformer separately but to the set of conformers jointly. To account for the structural flexibility of saturated polycyclic moieties, a new type of active variable is introduced.

METHOD

Background. According to the Darwinian theory, evolutionary preferences are given to individuals best adapted to the environment. In GAs, individuals are represented by chromosomes. Chromosomes are sequences of discrete quantities called genes which reflect the features liable to modification. The fitness of individuals is summarized by a functional that depends implicitly on the combination of features. In the area of molecular design, chromosomes determine a molecule or conformer, whereas fitness criteria typically address placement and orientation of functional groups, 3D charge distribution, molecular shape or electrostatic field, etc.

The GA typically begins from a random initial populations of size N_p (this is the size of the so-called permanent population). The permanent population is extended by N_c new individuals or children. Out of this extended population with $N_p + N_c$ individuals, N_p representatives are selected according to fitness criteria to form the next generation. The extension of a population, followed by its selective reduction to the permanent size N_p, forms a separate evolution step. The evolution is viewed as an iterative process, namely, this step is repeated until some ending criteria are reached. They typically involve a convergence test that requires no significant improvement with regard to fitness criteria over the last two or more successive generations. Basically, generation (or breeding) of children is attained by means of two operations called mutation and crossover. In the case of mutation, the parent chromosomes are modified by random modifications of some number of randomly selected genes. In the case of crossover, the child has at least two parents.

Scheme 1



Parts of their chromosomes are taken in a random manner and then concatenated to form the complete chromosome of the child.

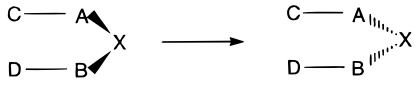
Structural Variables for Conformer Generation. Generally, the approach described herein handles four types of structural variables which change molecular conformation and/or configuration. Each such variable is encoded into a gene, and genes of the four different types are combined into chromosomes. Some of them reflect stereochemical features rather than conformational degrees of freedom. Such genes are totally or selectively disabled from modification depending on input. If stereochemical genes are left variable, the 3D structures generated involve, in the general case, different configurations. Hereafter, however, we will use the term conformer in a broader sense, as a 3D structure that has unique chromosome sequence among other ones.

Torsion Angles. Genes are attributed to all torsion angles associated with single and double bonds which are not terminal and do not participate in rings. Genes for different torsion angles have different cardinalities, i.e., different numbers of possible values. By default, the cardinality of a torsional gene is adjusted to the molecular force field adopted.[34,35] Namely, the cardinality is taken equal to the number of barriers (and minima) of the periodic torsional potential attributed to the particular bond. For a $C(sp^3)-C(sp^3)$ single bond, for instance, the period of the torsional potential is 120° and there are three isoenergetic minima. Thus, the actual torsion angle is converted into a gene by rounding it up with 120° resolution. The gene of cardinality 3, on its turn, fixes the angle of one of the three isoenergetic minima. For the sake of versatility, however, the cardinality of torsional genes may be changed selectively before generation starts. Binary genes attributed to noncyclic sp^2-sp^2 double bonds possibly reflect cis/trans isomers. Such genes are not liable to modifications when double-bond configurations are fixed on input.

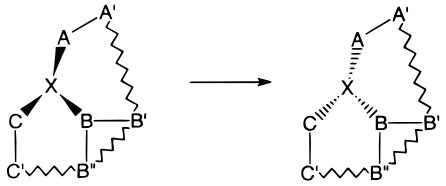
Stereocenter Parities. Genes are dedicated to all potential stereocenters that can be inverted without breaking a bond. As such, all sp^3 sites with at least three nonequivalent neighbors are taken. From them, only those whose deletion renders at least one pair of neighbors disjoint from the other one(s) are kept. The cardinality is 2 and reflects the parity of the stereosite. The latter is given by the sign of the mixed product of X−A, X−B, and X−C taken as vectors (see Scheme 1). Mutation of such a gene inverts the stereocenter. Inversion is attained geometrically by means of a 180° rotation of A and B and all attached moieties around the bisect of A−X−B. A, B, and X may participate in a common ring, but neither C nor D should belong to it. By default, inversions are enabled only for those stereocenters whose configuration is not provided on input.

Free Corners. Free corners are introduced by Payne and Glen[18] as moieties with a conformation degree of freedom to encompass the flexibility of saturated rings. Their genes are binary and fix two possible states. A free corner is sketched in Scheme 2. All sites shown participate in a common ring, and the free corner is the fragment A−X−B.

Scheme 2



Scheme 3



With exception of the C and D pair, no other pair of sites shown may participate in another common ring. With regard to the coordination of A and B, and the moieties attached to them outside the free corner, the flip can be viewed as two simultaneous ∼120° rotations around bonds C−A and D−B in opposite directions, respectively. With respect to site X and its incident moieties outside the free corner, the transformation can be viewed as a rotation around an axis through sites A and B. Hence, free-corner flips do not affect the stereochemical configuration. A free corner is perfect when it allows a flip which changes exclusively torsion angles. This is possible only when C−A and D−B are collinear. In the present approach, flips are selectively enabled, including imperfect free corners, as well. Their flips violate valence angles slightly, but do not change bond lengths. Selection of active free corners is achieved implicitly, by input of user-defined restrictions on the extent to which C−A and D−B may deviate from collinearity. Loosening this restrictions increases the diversity of conformers generated at the cost of strained structures. However, strain is usually relieved to a large extent when conformers undergo a force-field optimization.

Pyramids. Flipping of free corners is a heuristic solution that still does not completely resolve the flexibility of cyclic moieties. A further step in this direction proposed herein is the so-called reflection of pyramids (Scheme 3). Site X is in sp^3 hybridization and belongs to at least two condensed rings together with its first neigbors A, B, and C. All sites shown on the scheme belong to a single polycyclic fragment. Neighbors of X, A, B, and C, other than those shown, do not belong to it. Sites A, B, and C define the base plane of the pyramid. The reflection of the pyramid is its mirror reflection into the pyramid's base plane. The mirror reflection involves X and all moieties attached to X, A, B, or C, except those shown on the scheme. All mirror-reflected potential stereocenters including A, B, and C are inverted by the operations. If the parity of A, B, C, or X is fixed on input, the reflection of the particular pyramid is disabled. Reflections of pyramids are always performed prior to any other adjustments of structure to chromosomes. In this way, convertible stereo configurations are always intact with the corresponding chromosome in the final 3D structure.

In the general case, the reflection affects valence angles of sites A, B and C such as X−A−A′ and X−B−B′. These valence angles are not violated only if the cyclic second neighbors of X, namely A′, B′, B″, and C′, are exactly in the plane of the pyramid's base. To avoid generation of strained structures, active pyramids are identified on a complementary basis. Apart from the specific connectivity

of the molecular fragments involved, their 3D structures must approximately meet the requirement for coplanarity of bonds A−A′, B−B′, etc. For this purpose, a threshold for permissive coplanarity violation of pyramids is specified on input.

Structural Input. Force Field Optimization. Check for Degeneracy. The input structural information is introduced either by means of extended SMILES notations[36] or as a 3D molecular model in a special file format.[37] Standard SMILES[38] reflects only molecular connectivity in terms of an abstract chemical graph, whereas its extended version supports stereochemical notations, as well.[39] The actual generation starts anyway from an initial 3D molecular model. In the case of SMILES-based input, an automatic 3D molecular model builder is invoked. An approximate 3D model which complies to stereochemical specifications in SMILES is constructed heuristically. Upon direct 3D molecular input, all stereochemical features are implicitly fixed. Thereafter, a molecular force-field approach (called pseudo-molecular mechanics, or PMM) is employed[15] to minimize potential energy. The PMM force field includes additive empirical potentials of five types.[40]

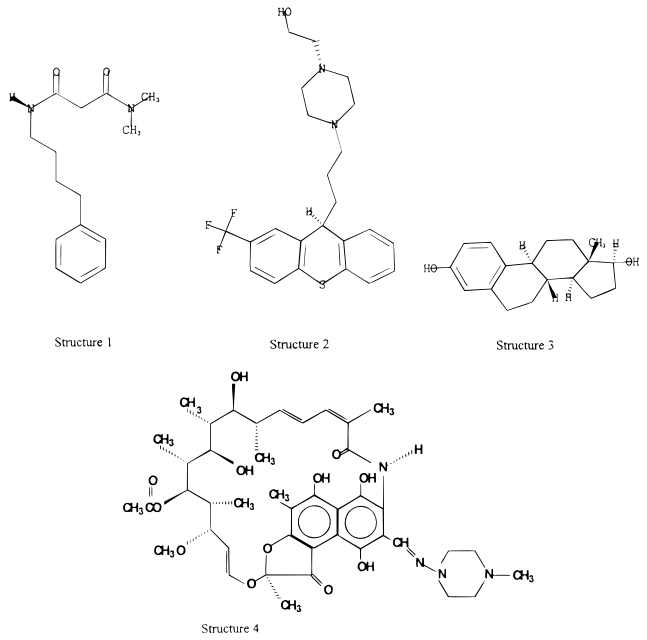
Before generation starts, all detected stereochemical features and conformational degrees of freedom are translated into chromosomes and listed, so that their genes can be selectively enabled/disabled for modification by the user. If 3D molecular input was used, all stereochemical features are initially disabled for modification. Otherwise, those detected, but not specified explicitly in terms of extended SMILES, are enabled initially. Conformational degrees of freedom are all enabled by default.

Within the software implementation, each conformer is represented by both its chromosomes and 3D structure. Genetic operations are on a chromosome level. The 3D model of a new conformer is constructed from its chromosomes. This is achieved by comparing the chromosomes with the ones of the source structure and then applying the corresponding elemental structural modifications to the latter. Optionally, the newly constructed crude 3D model undergoes PMM force field optimization that relieves possible strain from violated valence angles and close nonbonded contacts. The conformer may be yet rejected if its PMM potential energy exceeds some user-defined threshold.

If the chemical graph of the molecule has elements of symmetry, different chromosomes may actually reflect one and the same conformation. This is so because individual genes, and the order in which they appear in the chromosome, depend on the adopted numbering of atomic sites. Conformers that differ in chromosomes, but can be perfectly aligned in 3D, are called degenerate. Newly generated conformers are checked for degeneracy against the existing ones. This is achieved on the basis of the symmetry group Π of the chemical graph. The chromosome sequence is transformed according to all permutations $p \in \Pi$ of atomic sites, preserving the chemical graph invariant, and checked for identity with existing ones. The check for degeneracy is left optional, because conformers that can be closely aligned in 3D are very unlike to be preserved together in a permanent population with regard to the fitness criteria.

Generation of Zero Population. The initial (or zero) population is obtained from the source 3D molecular structure by random mutations of its chromosomes. To ensure equal probability for various modifications, both the set of genes changed and their new values are taken in a random manner. The mutated chromosomes are converted into a 3D model. The conformer may be still rejected from the population if found to be degenerate or too strained according the PMM force field. The source 3D structure is included in the initial population directly and its potential energy serves as a reference to evaluate the stability of other conformers.

Scheme 4



Fitness Criteria. For evaluation of its fitness, GAs normally score each individual. Because of the specific task herein, fitness criteria are defined in the context of a population by both individual and composite scores. A composite score pertains practically to any set of conformers. Scores are based on the so-called root-mean-square (rms) distance,[33] which is a dual relation of two chemically identical conformers i and j.

$$\mathrm{rms}_{ij} = \min_{p \in \Pi} \left\{ \sqrt{\frac{1}{m} \sum_{k=1}^{m} (r_{ik} - r_{ip_k})^2} \right\} \qquad (1)$$

Here, $\mathbf{r}$ are radius vectors, m is the number of atoms, and $p = \{p_1, p_2, ..., p_m\}$ are permutations of the numbers from 1 to m. Besides the trivial permutation with $p_k = k$, all other permutations from the symmetry group Π of the chemical graph are taken. The minimization for a specific $p \in \Pi$ is achieved numerically, by a linear search in the space of the rotational degrees of freedom of the molecule. The individual fitness score of the conformer i is the sum of its rms over all other individuals from the population S:

$$\mathrm{rms}(i) = \sum_{\substack{j \in S \\ j \neq i}} \mathrm{rms}_{ij} \qquad (2)$$

A set S of N conformers is scored by the so-called average dissimilarity (AD), which is the sum of the rms over all pairs, divided by the number of pairs:

$$\mathrm{rms}(S) = \frac{1}{N(N-1)} \sum_{\substack{i,j \in S \\ i \neq j}} \mathrm{rms}_{ij} \qquad (3)$$

Table 1. Settings of the Experiments for Structures **1**−**4**

(a) Structure **1**

parameter	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI
m/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9
E thresh	100	200	100	200	100	200	100	200	100	100	100	100	100	100	100	100
N_p	10	10	10	10	10	10	10	10	20	30	10	14	20	5	7	10
N_c	5	5	5	5	5	5	5	5	5	5	7	7	7	5	7	10
degeneracy	60	60	60	60	30	30	30	30	60	60	60	60	60	60	60	60

(b) Structure **2**

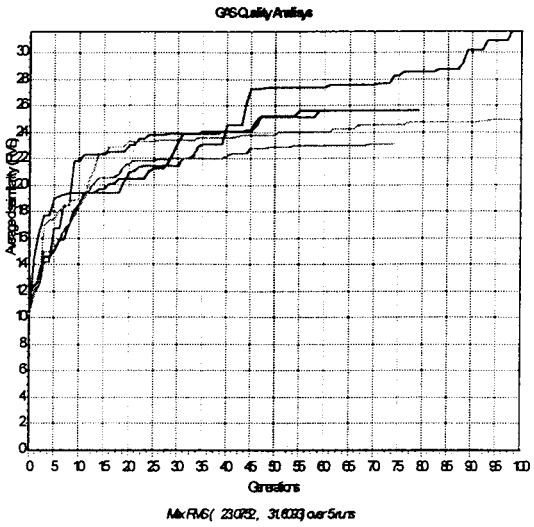
parameter	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI
m/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9
E thresh	100	200	100	200	100	200	100	200	100	100	100	100	100	100	100	100
N_p	10	10	10	10	10	10	10	10	20	30	10	14	20	5	7	10
N_c	5	5	5	5	5	5	5	5	5	5	7	7	7	5	7	10
degeneracy	60	60	60	60	30	30	30	30	60	60	60	60	60	60	60	60

(c) Structure **3**

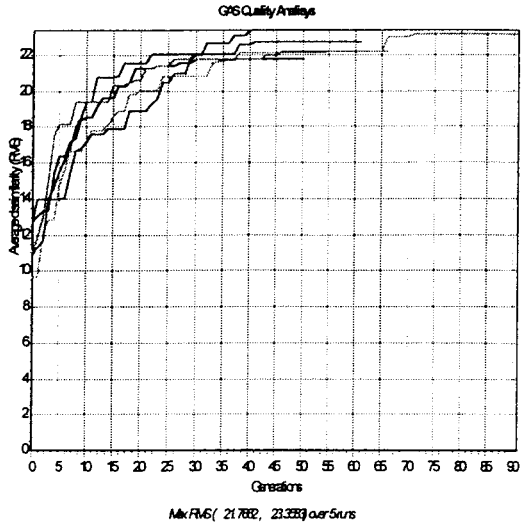
parameter	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
m/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9
E thresh	500	1000	500	1000	500	1000	500	1000	500	1000	500	2000	500	1000
N_p	3	3	3	3	3	3	3	3	4	4	4	4	4	4
N_c	1	1	1	1	2	2	2	2	1	1	1	1	2	2
degeneracy	30	30	30	30	30	30	30	30	30	30	30	30	30	30

(d) Structure **4**

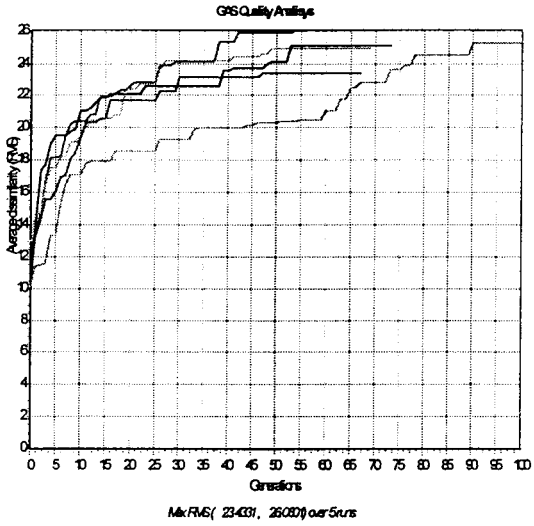
parameter	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
m/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.2/0.8	0.1/0.9	0.1/0.9
E thresh	500	1000	500	1000	500	1000	500	1000	1000	1000	1000	1000
N_p	5	5	5	5	5	5	5	5	7	7	10	10
N_c	3	3	3	3	3	3	3	3	3	3	3	5
degeneracy	60	60	60	60	30	30	30	30	60	60	60	60



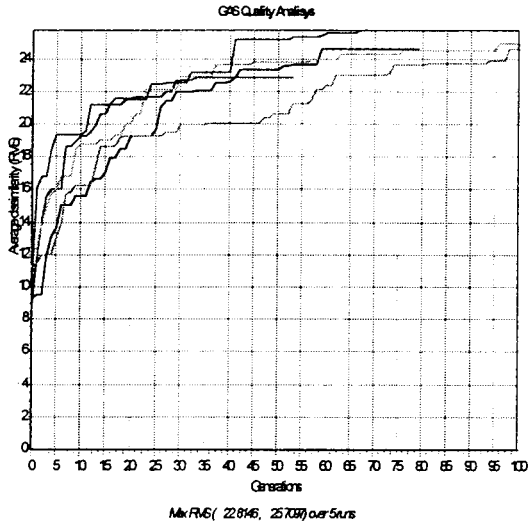
I



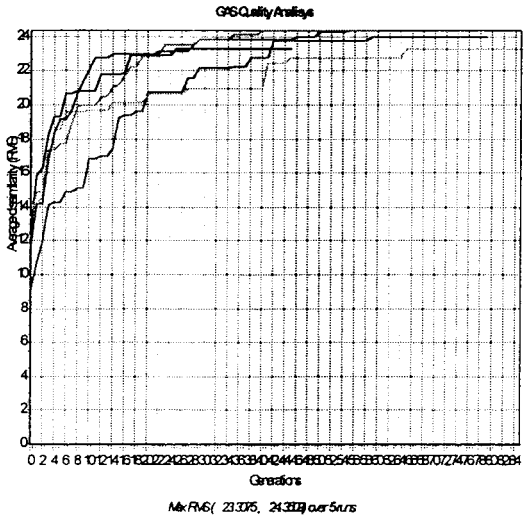
II



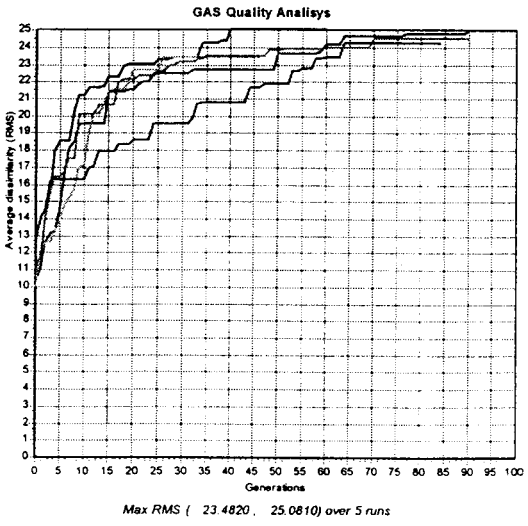
III



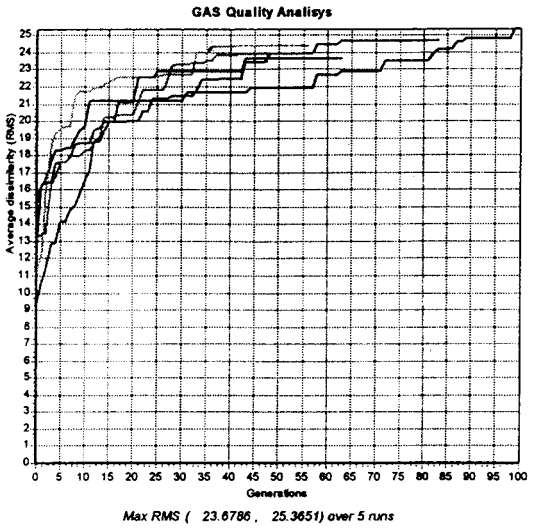
IV



V

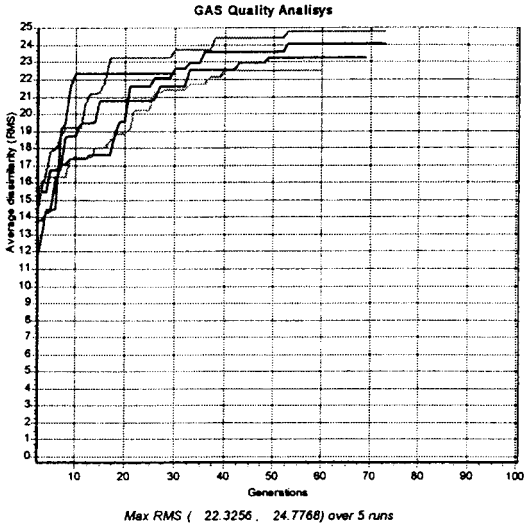


VI



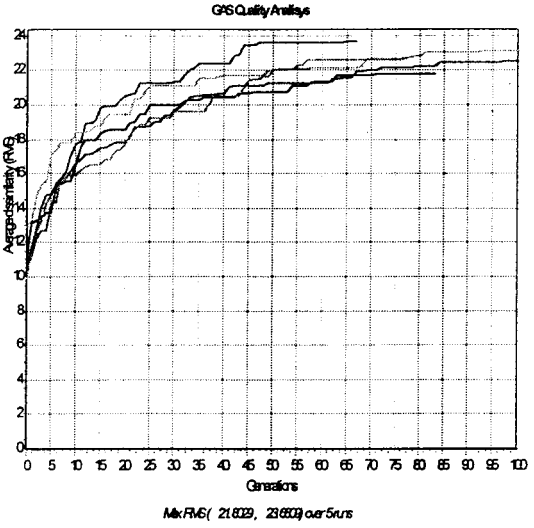
Max RMS (23.6786 , 25.3651) over 5 runs

VII



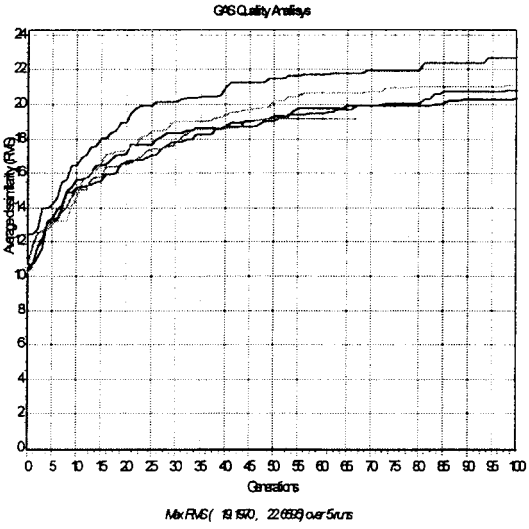
Max RMS (22.3256 , 24.7768) over 5 runs

VIII



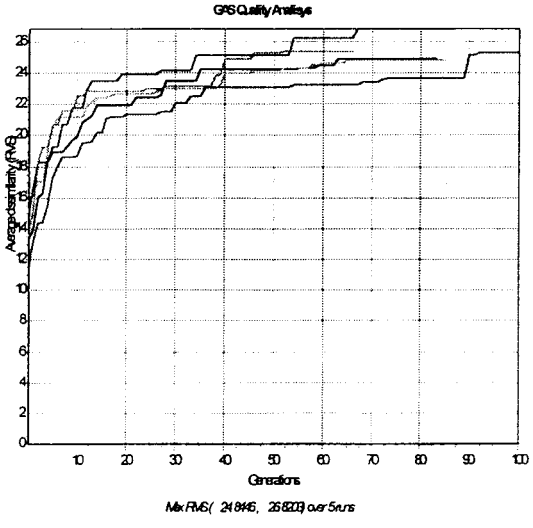
Max RMS(21.8029, 23.6609) over 5 runs

IX



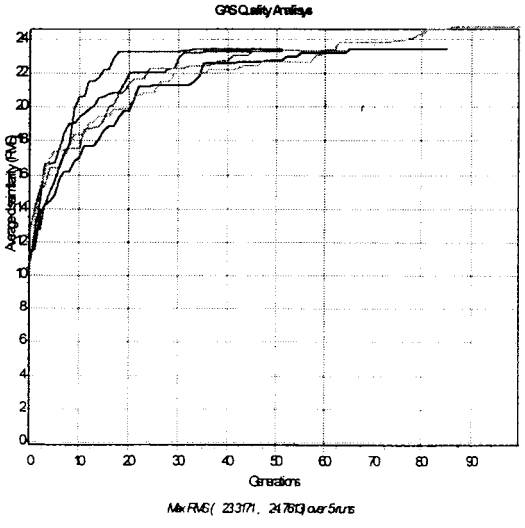
Max RMS(19.1940, 22.6589) over 5 runs

X



Max RMS(24.8945, 26.8203) over 5 runs

XI



Max RMS(23.3171, 24.7613) over 5 runs

XII

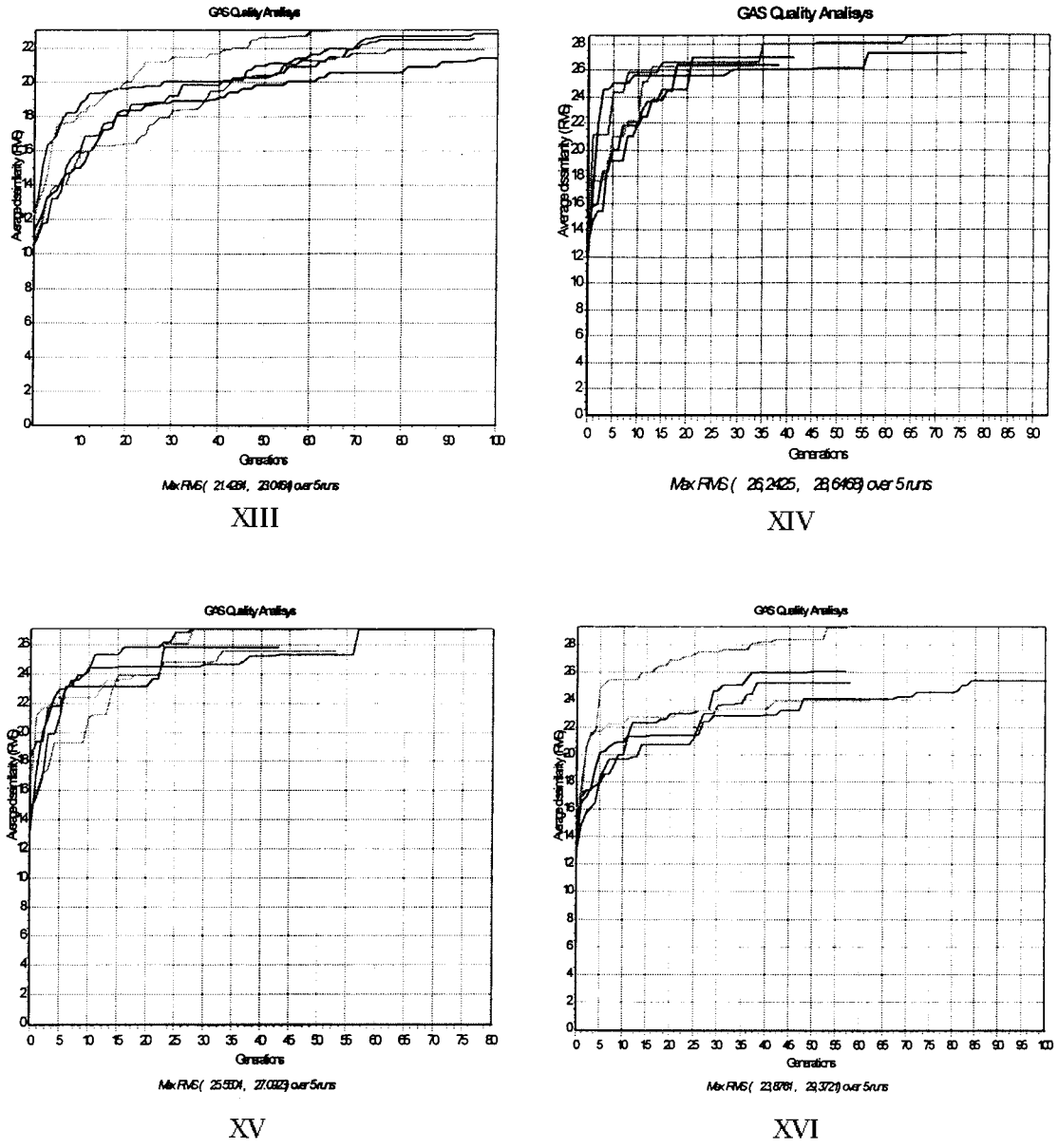
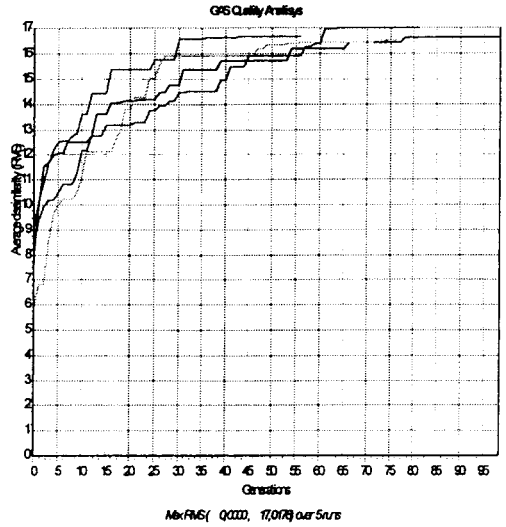


Figure 1. Average dissimilarity (rms) of the best populations during the GA optimization, associated with studied experiments (see Table 1a), for structure **1**.

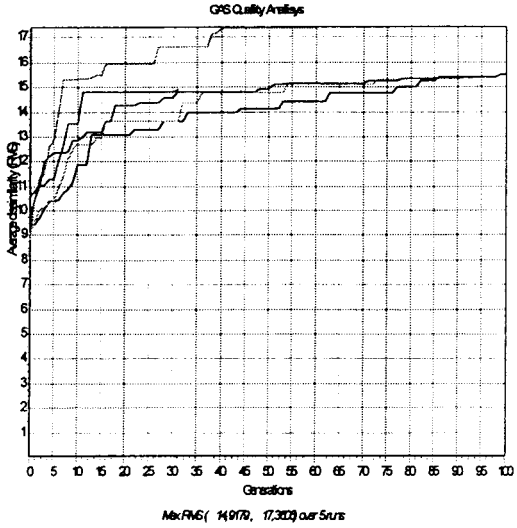
Generation of the Next Population. In the breeding of children from a permanent population of N_p individuals, both crossovers and mutations are employed. In a crossover operation, the chromosomes of two parents are cut at the same, randomly chosen point, and the parts are interchanged to produce the chromosomes of two children. This is the so-called one-point crossover. Practically, four-point crossovers are used, because the four different chromosomes are handled separately, and a one-point crossover is applied to each of them. Thus chromosomes inherited by a child are always from both parents, regardless of which type of structural variables they reflect. In contrast to the so-called roulette-wheel selection, the probability for conformers to be chosen as parents is not uniform. It is proportional to the individual rms score of the conformer within the permanent population it belongs to.

Prior to the actual crossover, the chromosomes of parents are subject to mutations. Mutations are usually needed to prevent the algorithm from stacking at a point where new extended populations approximately reproduce old ones. The probability of such mutations is given by the so-called gene mutation rate. The gene mutation rate is the probability with which genes of the parent chromosome are modified before participating in the actual crossover. A typical value of the gene mutation rate adopted by default is 5%.

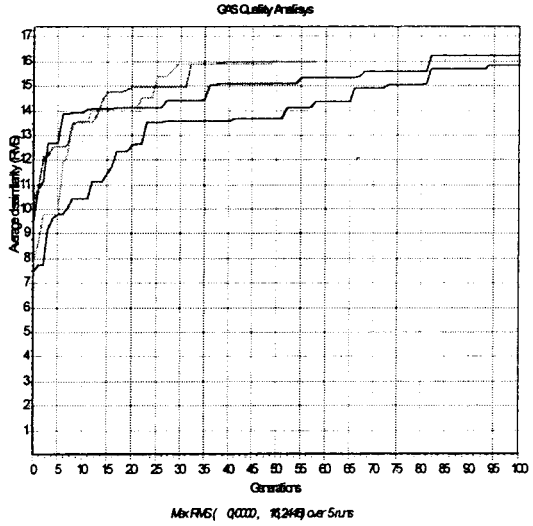
A new conformer obtained from genetic operations is not directly admitted as a child in the extended population. It might be rejected in advance if found degenerate in 3D with existing members or energetically inadmissible. The number of attempts to produce by genetic operations nondegenerate, low-energy conformers is further denoted as number of trials. New conformers are generated until the number of children accepted reaches N_c, or until the number of trials is exceeded. In the latter case, the whole process is aborted. Otherwise, the extended population is reduced from $N_p + N_c$ to N_p individuals without regard of origin to produce the next generation. The latter is chosen to be that subset S_p of N_p out of $N_p + N_c$ conformers which has the largest AD. The combinatorial problem of finding the subset S_p is solved exactly, by a trivial branch-and-bound algorithm.



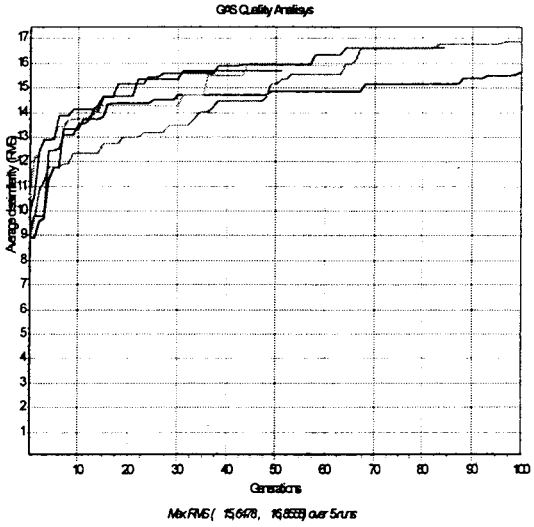
I



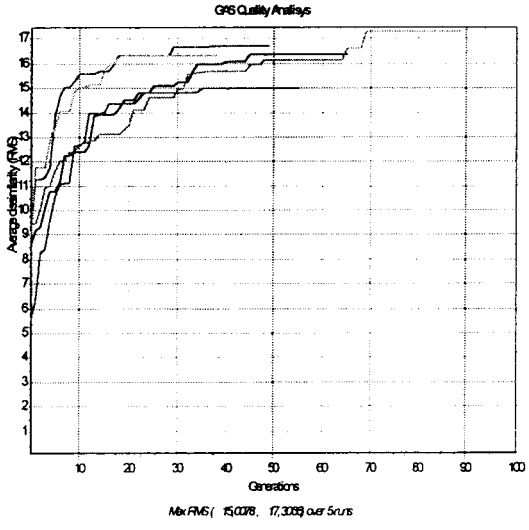
II



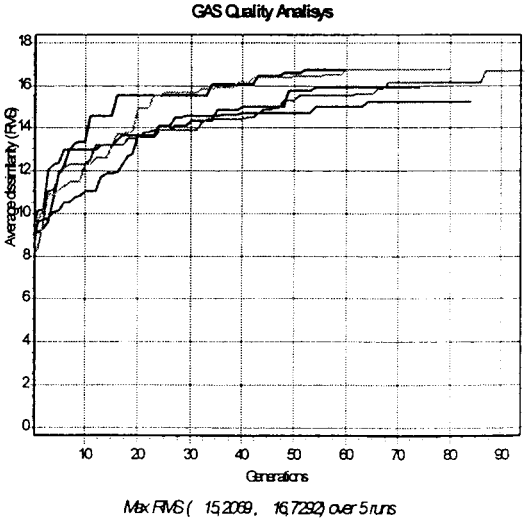
III



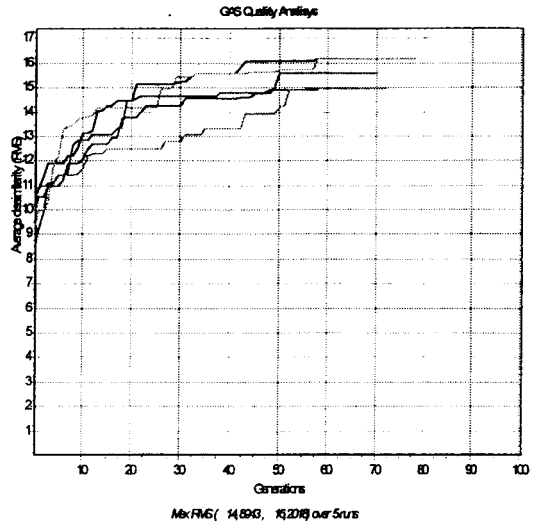
IV



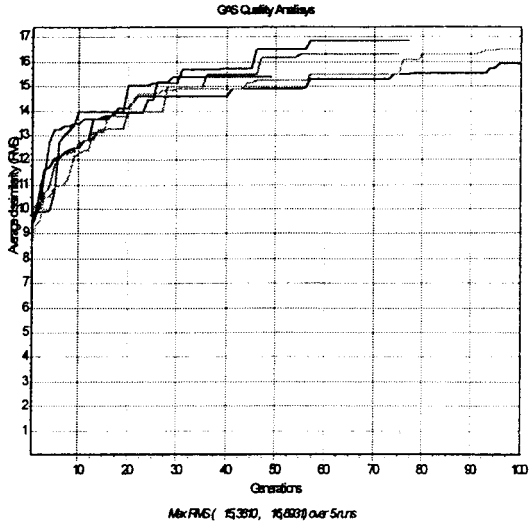
V



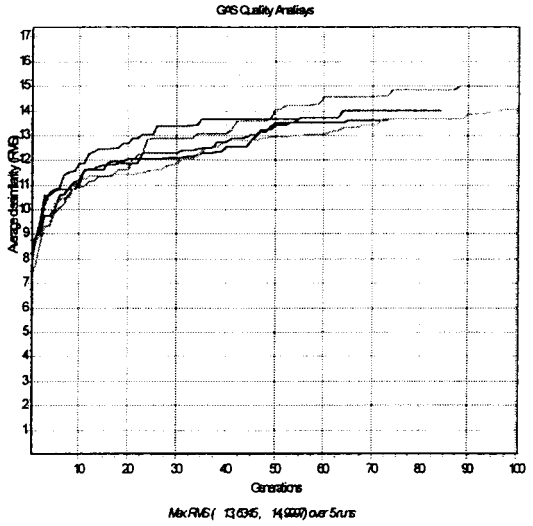
VI



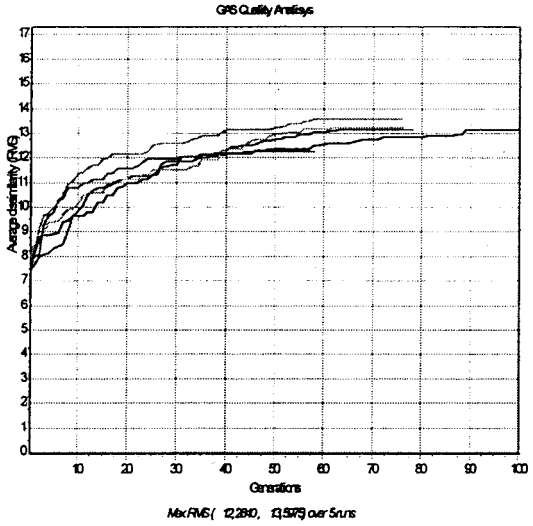
VII



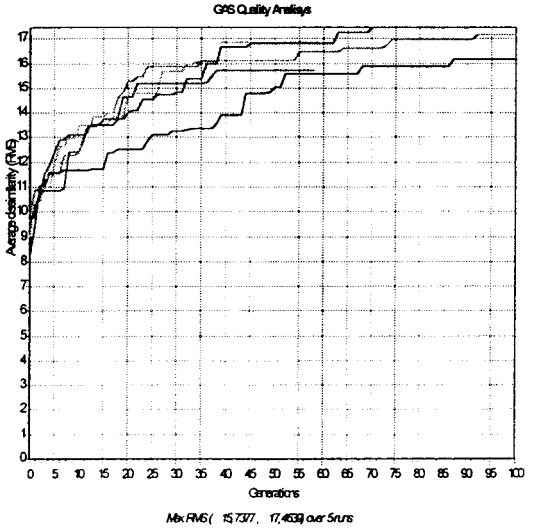
VIII



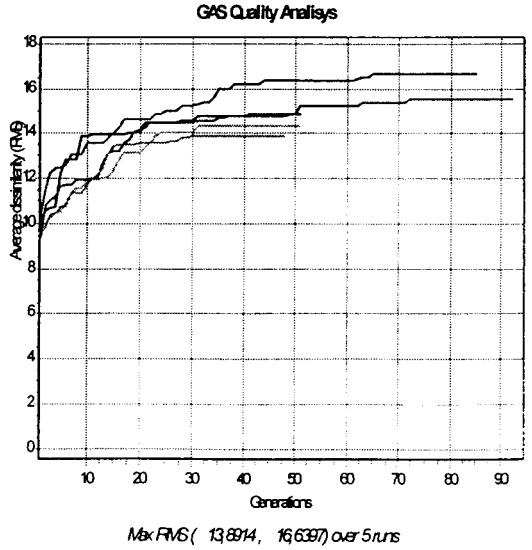
IX



X



XI



XII

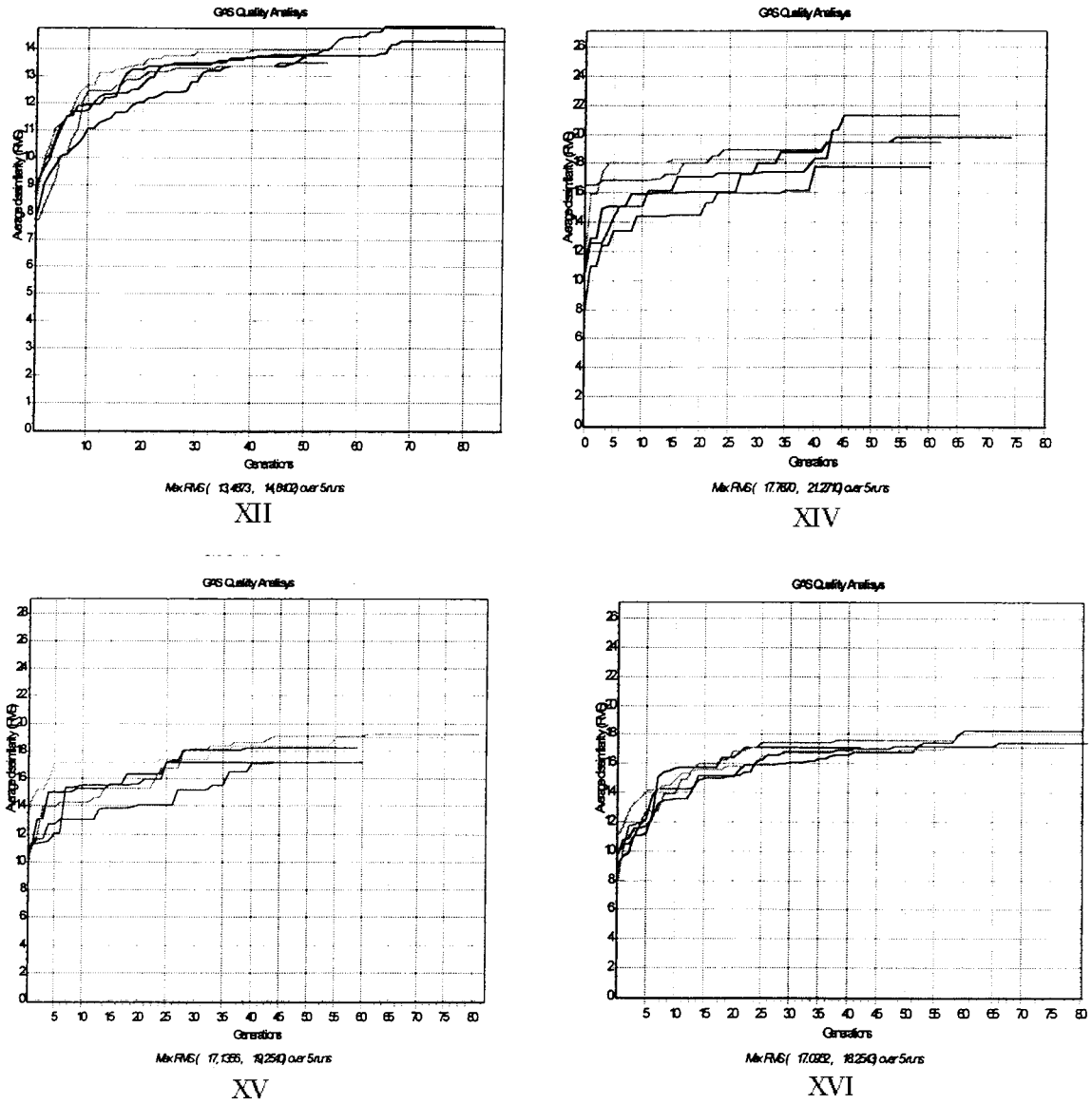


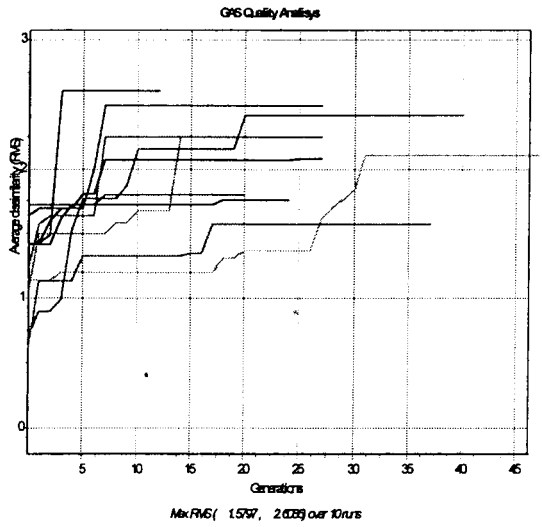
Figure 2. Average dissimilarity (rms) of the best populations during the GA optimization, associated with studied experiments (see Table 1b), for structure **2**.

Criteria for Ending. In view of the evolution step described, each new generation has AD at least equal to the one of the previous. Generally, AD increases in each step, both by chance and because more "fitted" parents with large individual scores are selected. In the implementation, three different ending criteria can be separately or jointly imposed. The first one is trivial and fixes a limit for the number of steps (iterations). The second one is a convergence test which requires that the AD increase over several successive steps drops below a user-defined threshold. Finally, the process can be terminated if AD exceeds a certain predefined limit.

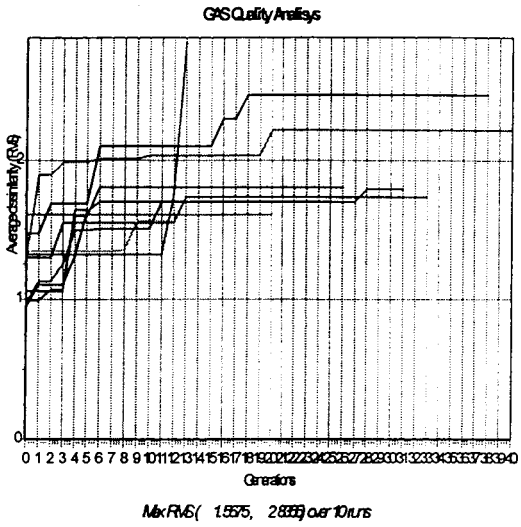
RESULTS

The quality of the GA is estimated by means of its robustness and reproducibility[41] and by the conformational coverage provided in terms of AD. In this respect, several GA runs were performed with test molecules (see Scheme 4) selected to differ in terms of conformational flexibility. The first one is *N*,*N*-dimethyl-*N*′-4-(phenylbutyl)malonamide (**1**), exemplifying a noncyclic moiety with conformational degrees of freedom (used also by Wehrens et al.[41]). The second one is flurpentixol (**2**), which is flexible both in its noncyclic and monocyclic parts (tested by Smellie et al.[16,17]). The third example is estradiol (**3**), which is a configurationally rigid molecule.[7,32] Finally, rifampicin (**4**) has a large flexible monocycle. The depictions in Scheme 4 are obtained from an automatic 2.5D model builder from extended SMILES, which is incorporated in the software.

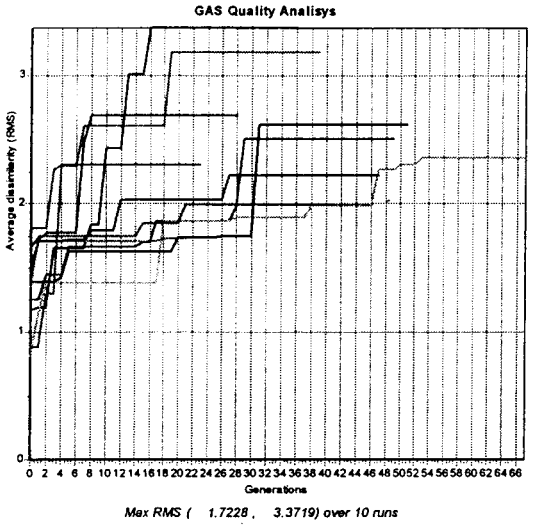
As shown in Table 1a−d, different control parameters are used in the process. These are the size of the permanent population N_p, the number of children N_c, the so-called degeneracy threshold, which is the resolution at which two torsion angles are considered essentially different, the mutation/crossover ratio (m/c ratio), and the potential energy threshold [E threshold (kJ/mol)]. Force field optimization of generated conformers was not used to save computational time. The number of trials to obtain nondegenerate, low-energy children was set to 100 and was exceeded in separate cases. The iteration limit was kept constant and equal to 100 for all experiments. Except for the several cases when the process ran out its iteration limit, generations ended according to the convergence test. The latter required average dis-



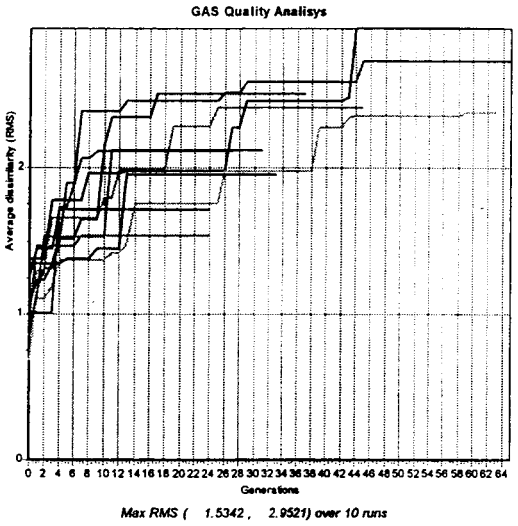
I



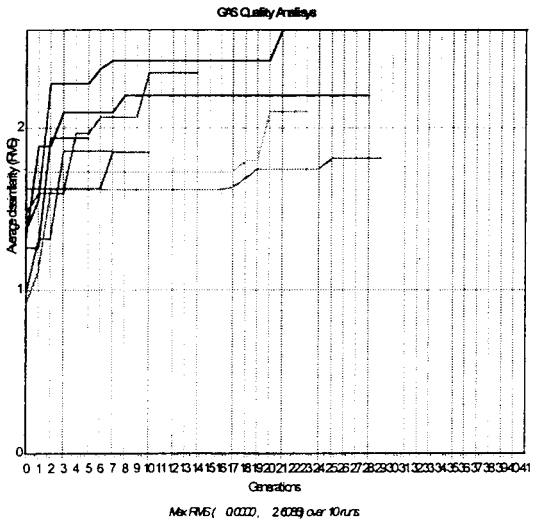
II



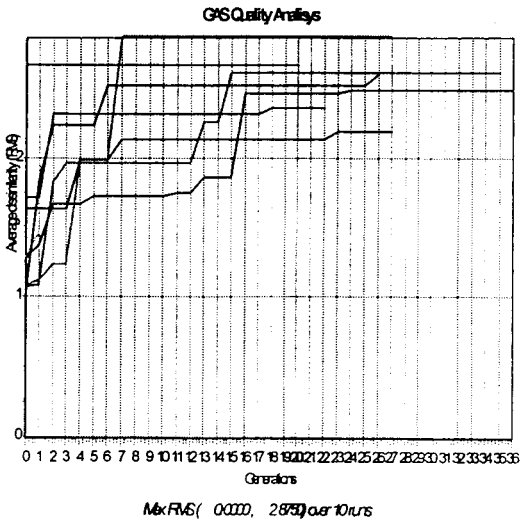
III



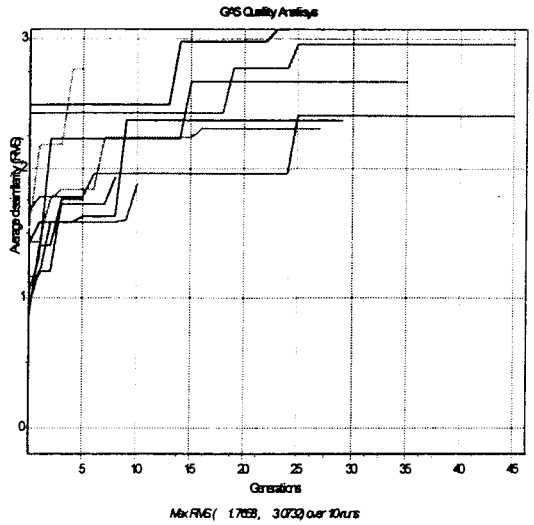
IV



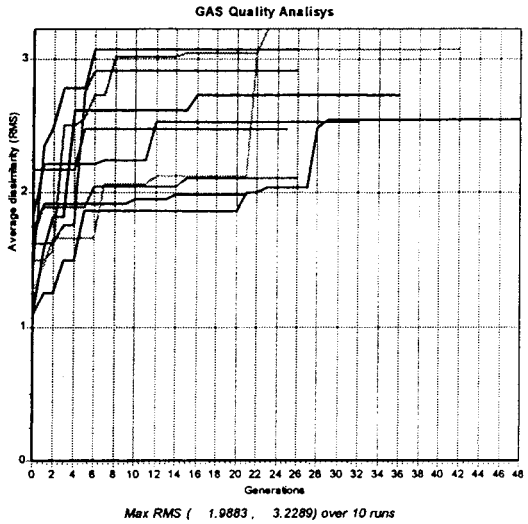
V



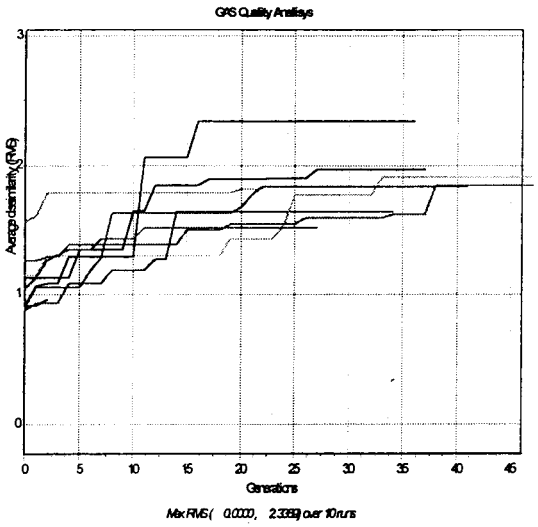
VI



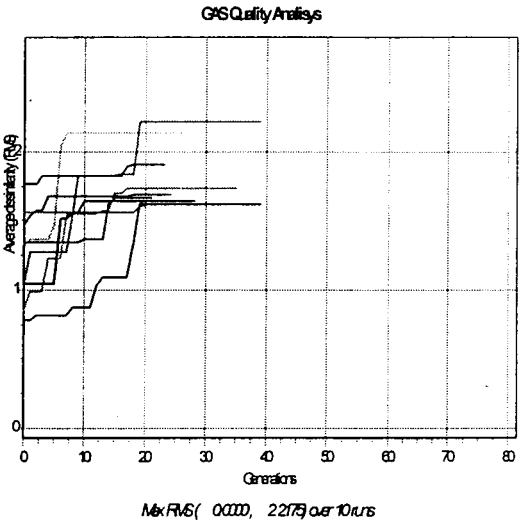
VII



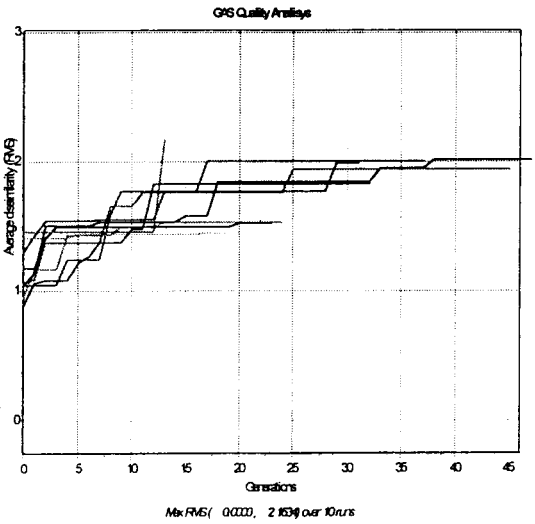
VIII



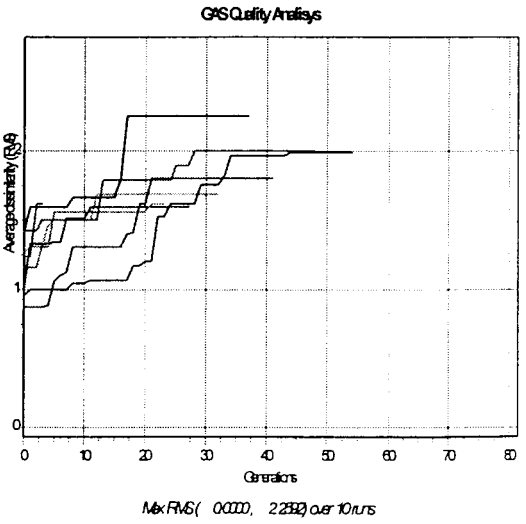
IX



X



XI



XII

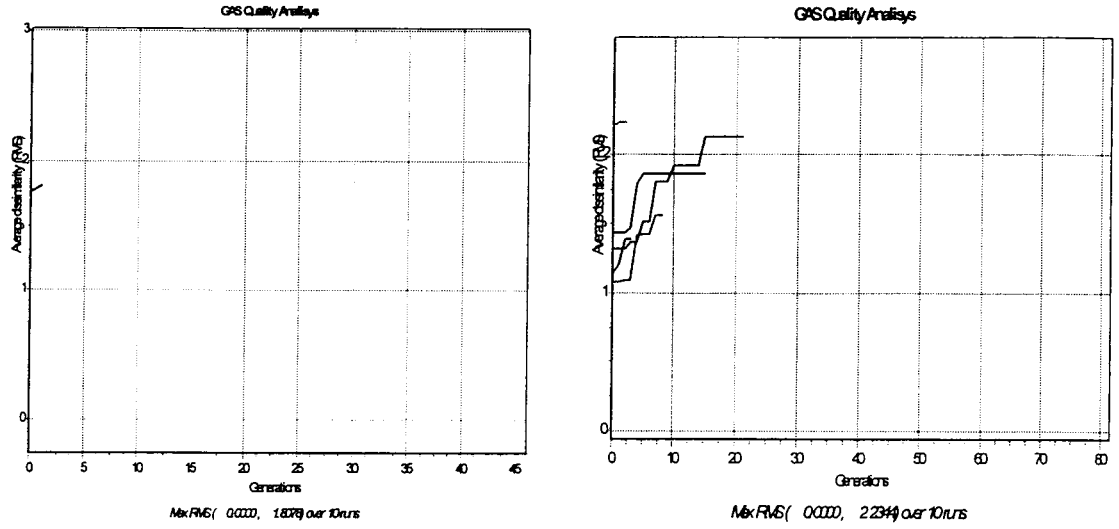


Figure 3. Average dissimilarity (rms) of the best populations during the GA optimization, associated with studied experiments (see Table 1c), for structure **3**.

similarity increase less than 0.1 Å over the last 20 generations. Parities of stereocenters were not preserved for **1**. For **2**, the parity of a single stereosite was kept constant, namely the C-atom of the B-ring incident to the side chain, as adopted by Wehrens et al.[41] and Smellie et al.[16,17] For rifampicin (**4**), conformers were restricted to the configuration of the active enantiomer.[42,43] Similarly, the generated conformers of **3** have the stereochemistry of the natural enantiomer.

As seen, the control parameters are set different for the four molecules handled. Higher potential energy thresholds were imposed for rifampicin and estradiol, allowing for more strained conformers. Because **3** and **4** are less flexible than **1** and **2**, the sizes of the permanent population N_p and the number of children N_c were set smaller.

Experiments involve five replicate runs for **1**, **2**, and **4** each, and 10 runs for **3**. In total, 360 runs were performed at 58 different parameter settings. For each experimental setting, the course of AD with successive generations is plotted in Figures 1−4 for molecules **1**−**4**, respectively. The plots serve as illustrations to the analysis of the results given hereafter.

DISCUSSIONS

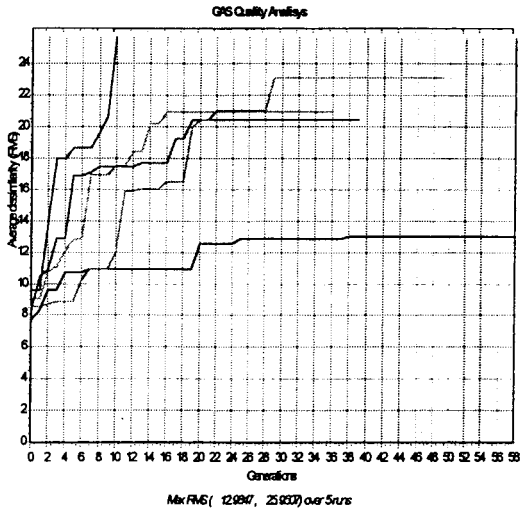
The robustness of the method was assessed by the so-called convergence rate. The latter is reciprocal to the number of generations required for AD to converge. As seen from Figures 3 and 4, convergence rates associated with strained molecules **3** and **4** vary considerably in different runs. Because of that, the effect of parameters on robustness was tested exclusively on molecules **1** and **2**. As a general trend, the robustness increases with the decrease of the N_p/N_c ratio. Thus, the lowest convergence rates are observed in experiments IX and X, where N_p/N_c equal 4 and 6 for **1** and **2**, respectively. Convergence is faster in experiments with $N_p/N_c = 2$, and especially for $N_p/N_c = 1$. Evidently, a larger N_c increases the probability that parents be replaced by children. Hence evolution is more dynamic and the chances for a rapid increase of AD are better.

In contrast to other GAs for conformer generation, our approach uses and maximizes a fitness function which directly reflects the extent of conformational coverage by a limited number of points. The range in which the fitness function AD varies at the end of replicate runs is given at the bottom of each AD/generation plot (Figures 1−4). Small range discrepancies of final ADs correspond to a better reproducibility of the GA evolution. Apparently, discrepancies are smaller for larger N_p/N_c. For instance, in experimental setting IX and X for molecule **1**, final ADs fall within 21.8 to 23.7 and 19.2 to 22.7 Å ranges, respectively. In fact, small final AD ranges were observed in other experiments, but the corresponding AD/generation curves did not follow a similar pattern in these cases. For molecule **2** again, smallest discrepancies are observed at $N_p/N_c = 4$ and $N_p/N_c = 6$ (experiments IX and X with final AD range of 13.6 to 15.0 and 12.4 to 13.6 Å). Similarly, the highest reproducibility for **3** and **4** was observed for large N_p/N_c ratios ($N_p/N_c = 4$ in experiment XI and $N_p/N_c = 3.3$ in experiment XI). The enhanced reproducibility at high N_p/N_c ratio can be related to the so-called better control of parents over children. Apparently, less dynamic evolutions tend to reproduce themselves better in replicate runs.

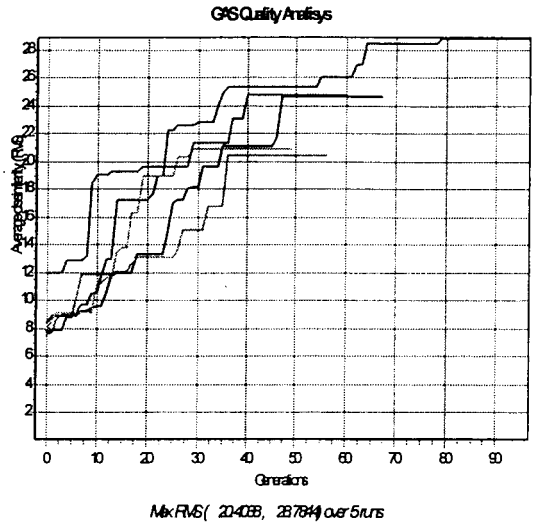
For certain experimental settings, evolution was terminated preliminary in some runs. This was caused by failure to produce nondegenerate conformers for initial or extended populations, meeting potential energy constraints. Such cases are typical for the relatively rigid molecules **3** (experiments V, VI, IX−XII) and **4** (experiments III and V). However, unsuccessful runs were observed for the more flexible molecule **2** (experiment I and III).

The obtained results showed that the coverage of conformational space decreases with the increase of the N_p/N_c ratio. Thus, the lowest AD values for molecules **1**−**4** were obtained for the highest values of the N_p/N_c ratio: see experiments IX and X for structure **1** (AD = 21.8 to 23.7 and 19.2 to 22.7 Å, respectively), experiments IX and X for **2** (AD = 13.6 to 15.0 and 12.3 to 13.6 Å, respectively), experiment IX for **3** (AD = 0.0 to 2.2 Å), and experiment XI for **4** (AD = 20.2 to 23.3 Å).

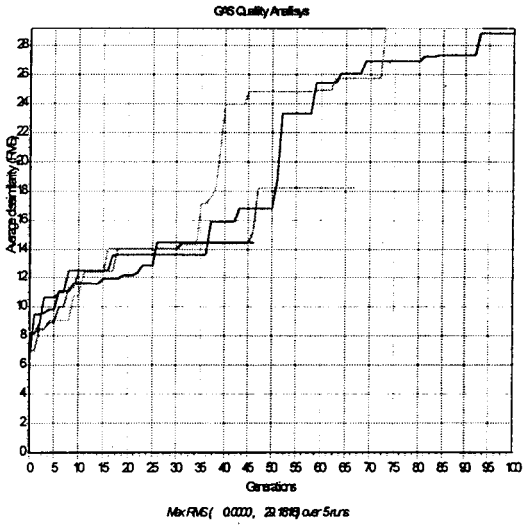
The results were obtained with potential energy constraints on nonoptimized 3D structures. Typically, conformers are



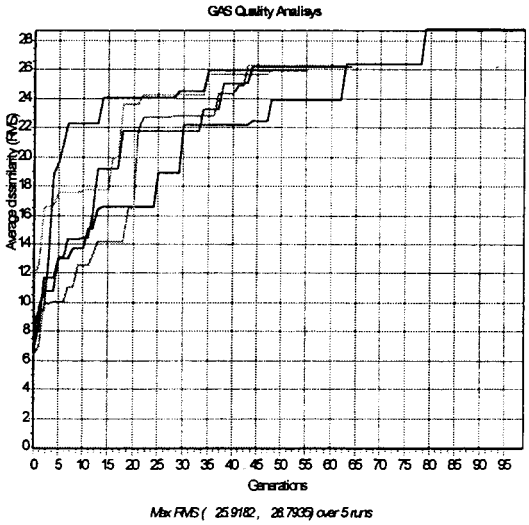
I



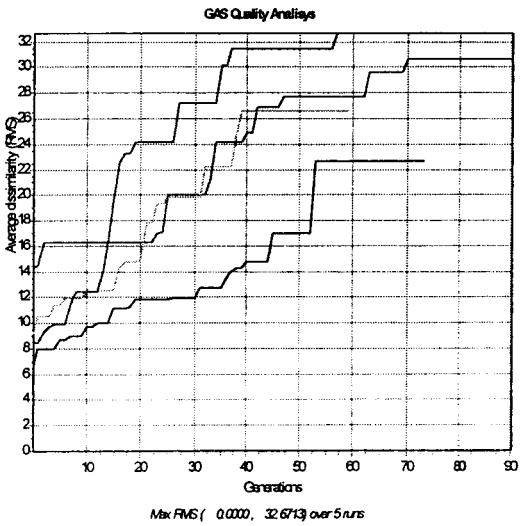
II



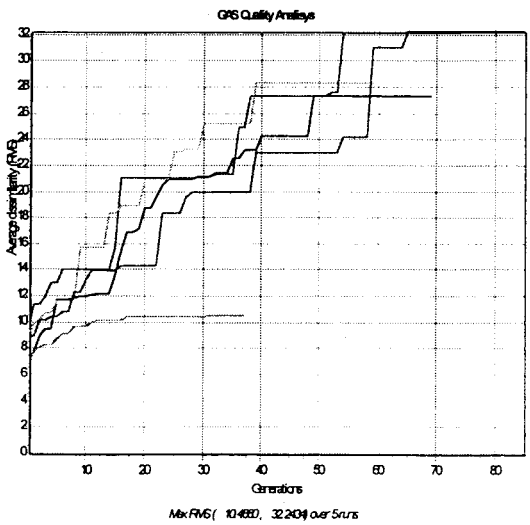
III



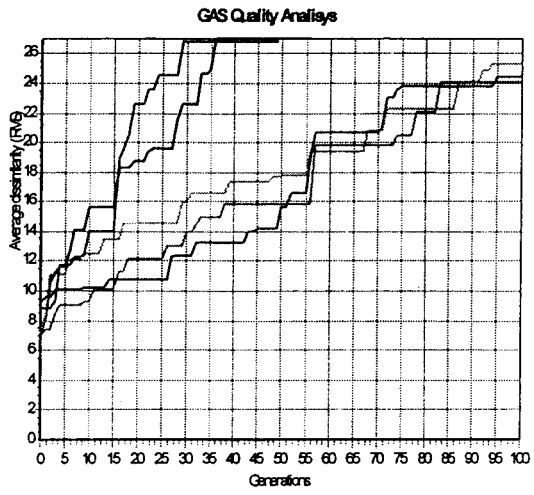
IV



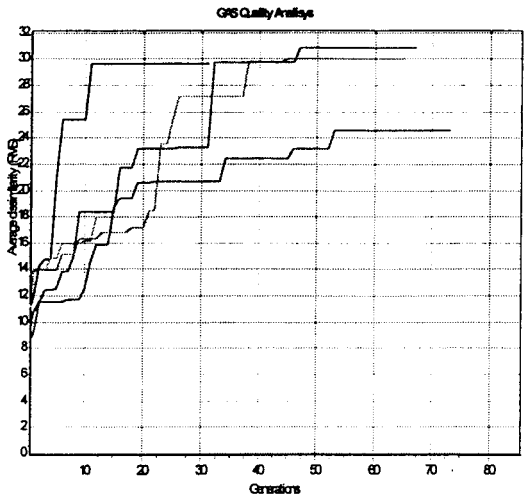
V



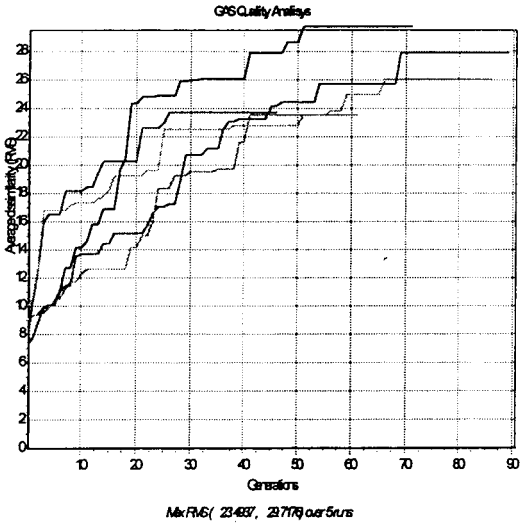
VI



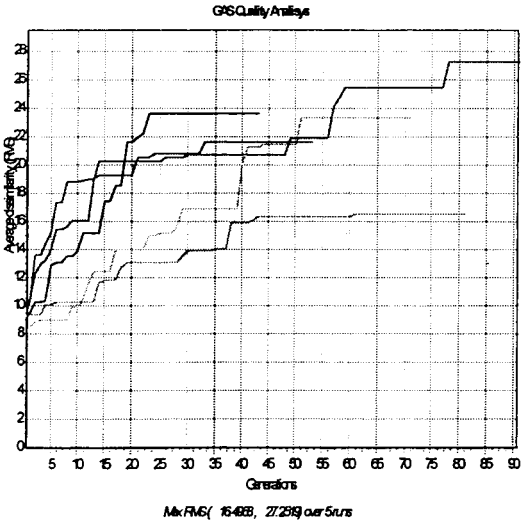
VII



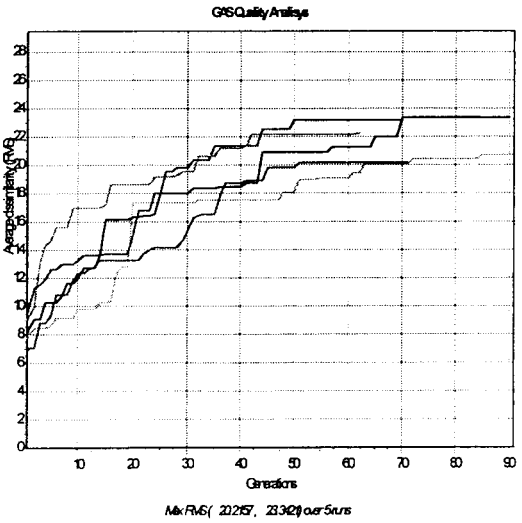
VIII



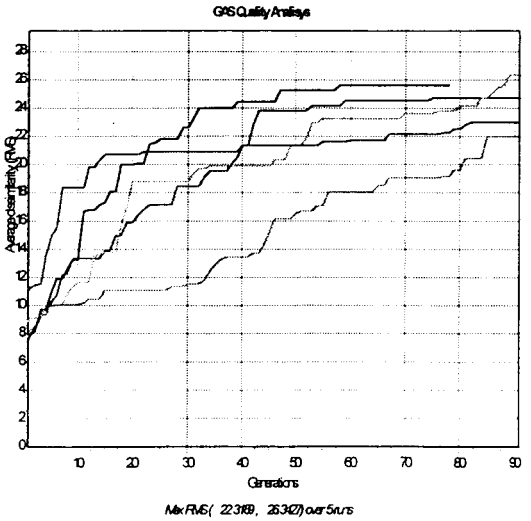
IX



X

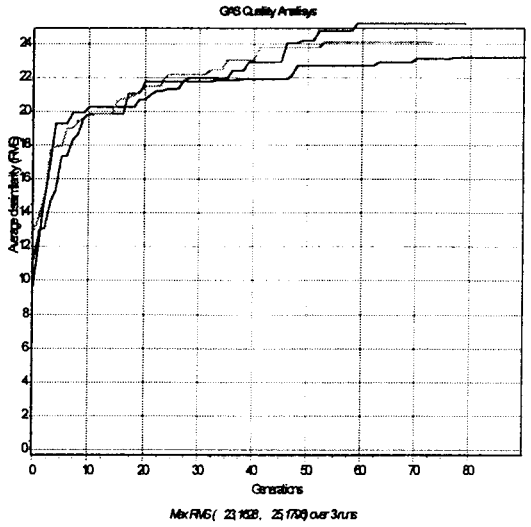


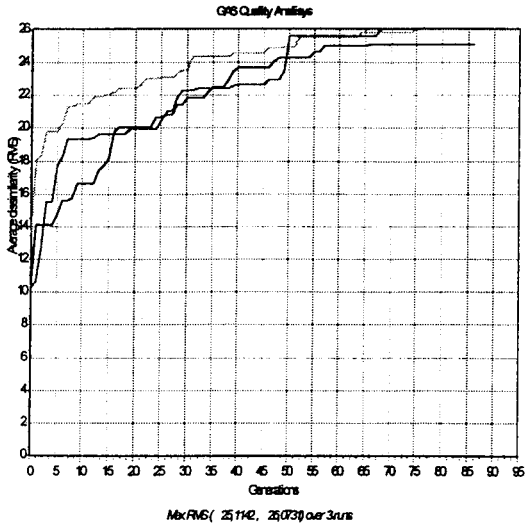
XI



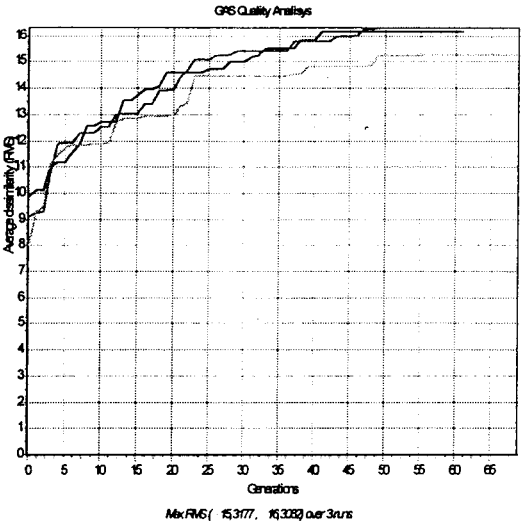
XII

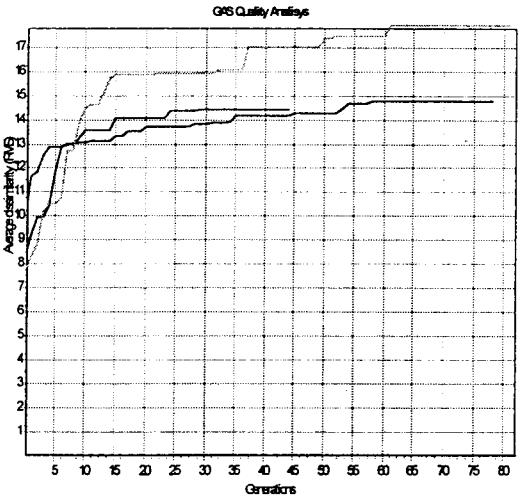
Figure 4. Average dissimilarity (rms) of the best populations during the GA optimization, associated with studied experiments (see Table 1d), for structure **4**.


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (23.1628, 25.1798) over 3 runs

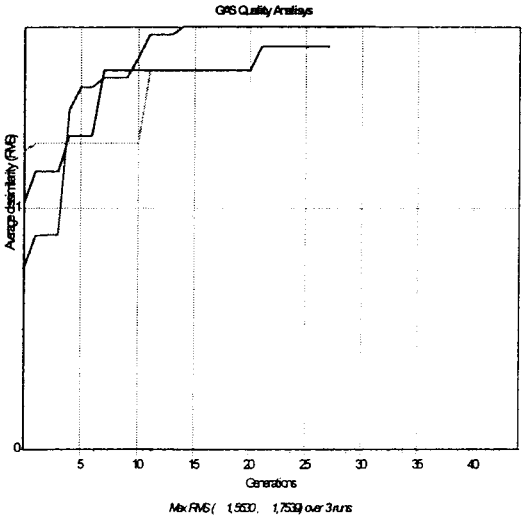

GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (25.1142, 26.0731) over 3 runs

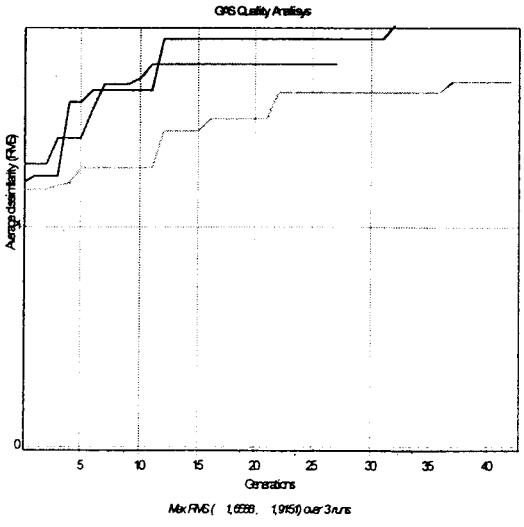
a


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (15.3177, 16.3082) over 3 runs


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (14.4108, 17.9375) over 3 runs

b


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (1.5530, 1.7534) over 3 runs


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (1.6588, 1.9151) over 3 runs

c

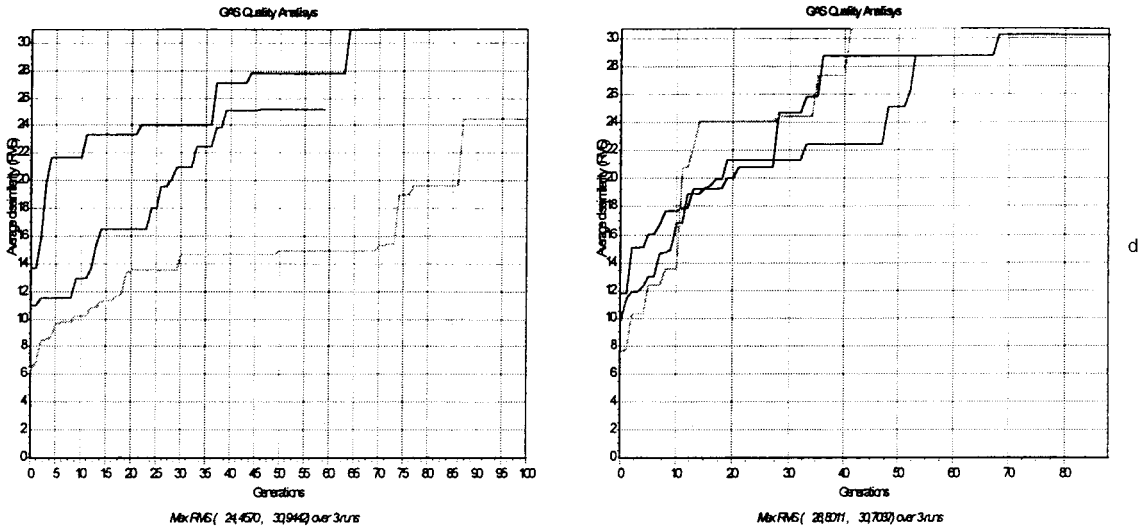


Figure 5. Average dissimilarity (rms) of the best populations across GA optimization, after PMM optimization and degeneracy screening. Single experiments were tested for each chemical: (a) experiment I for structure **1**; (b) experiment I for structure **2**; experiment X for structure **3**; and (d) experiment I for structure **4**.

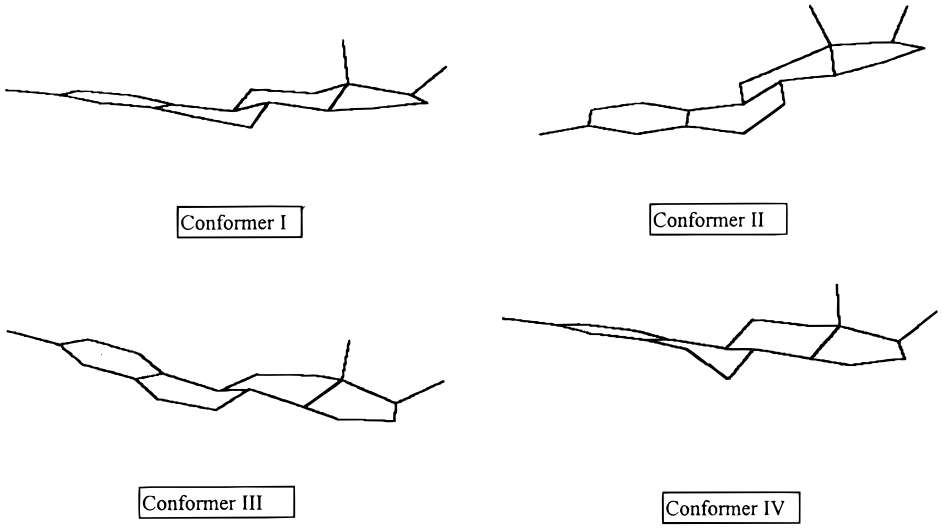


Figure 6. Generated conformers for estradiol.

submitted to potential energy mimimization on each step prior to that test or after the process terminates. Results obtained with force field oprimization are illustrated in Figure 5a−d: experimental setting I for molecules **1** (Figure 5a), **2** (Figure 5b), and **4** (Figure 5d), and experimental setting X for **3** (Figure 5c). The first and second plots in each figure are from runs performed without and with 3D degeneracy tests. As shown in Figure 5a−d, force field optimization significantly improves reproducibility but diminishes final AD values. Preliminary rejection of degenerate conformers tends to increase final AD values.

The GA runs on estradiol (natural enantiomer) under conditions X with force field optimization and degeneracy screening produced four different conformers (Figure 6). The comparison between these structures and the conformers obtained recently[7,32] by means of a systematic search[15] suggests that both approaches provide similar results in terms of conformational coverage. Beside the crystal-phase structure (conformer I) and conformers II and III, proposed by us[7,32] and Wiese and Brooks,[44] a new conformer IV is obtained. For the latter, the free corner of the B ring semichair is in the opposite direction with regard to the crystal phase structure. The kinetic feasibility of conformer IV, however, is still under question.

CONCLUSIONS

This work addresses the applicability of GA to the problem of conformational coverage. Instead of searching for a set of (meta)stable conformers, a small collection of conformers aimed at optimal coverage of the conformational space under potential energy constraints is generated. Like any other GA, the approach avoids the systematic search whose time complexity is the major obstacle in conformational analysis of large and flexible molecules. On the other hand, the nondeterministic character of the method is a common limitation of all GAs.

To quantify the (dis)similarity of conformers, the method uses a rather straightforward measure, the so-called rms distance. It has the advantage of being unambiguously defined in terms of Cartesian atomic coordinates. Initially defined as a dual relation comparing a pair of conformers, the rms distance is generalized in terms of AD to reflect the 3D dissimilarity of a set of conformers, independently of

their number and chemical nature. Unlike other GAs, which score and optimize individuals rather than populations, the method is directed toward obtaining an optimal set of individuals.

In handling the flexibility of molecules, the method operates with an extended set of conformational degrees of freedom. Namely, in parallel to rotatable bonds, flips of free corners and pyramids in cyclic fragments are considered. Thus cyclic fragments are no longer taken as rigid like in typical conformation analysis software. Along with conformations, the method encompasses the complete diversity of stereochemical configurations. Depending on demands, stereochemistry can be totally or selectively preserved during the run.

The performance of the approach is assessed by tests on four molecules which differ significantly in terms of flexibility. The effect of control parameters on the robustness and reproducibility of the GA and on the final AD as a measure of conformational coverage is evaluated. Robustness is measured by the convergence rate with which the final, maximal AD is reached over generations. Reproducibility is assessed by the variation of final ADs from replicate runs. It was found that with the increase of the ratio between parents and children, N_p/N_c, the reproducibility of the runs increases, whereas robustness and coverage in terms of AD decrease. The trends were observed for all molecules used in the tests. We suggest that the effect of N_p/N_c on robustness and reproducibility is common for all GAs. Loosely speaking, the larger number of children diversifies populations and renders the evolution process more dynamic and less determined.

Force field optimization was found to improve reproducibility at the cost of diminished final AD values. As expected, explicit rejection of 3D-degenerate structures was found to increase AD under all other equal conditions.

ACKNOWLEDGMENT

This research was supported, in part, by a U.S. EPA Cooperative Agreement (CR822306-01-0) with the Bourgas University "As. Zlatarov". The authors thank Dr. Steve Bradbury, Dr. Gilman Veith, and Dr. Julian Ivanov for valuable discussions.

REFERENCES AND NOTES

(1) Eliel, E. L. Chemistry in Three Dimensions. In *Chemical Structures*; Warr, W. A., Ed.; Springer: Berlin, Germany, 1993; Vol. 1, p 1.
(2) Mekenyan, O. G.; Ivanov, J. M.; Veith, G. D.; Bradbury, S. P. DYNAMIC QSAR: A New Search For Active Conformations and Significant Stereoelectronic Indices. *Quant. Struct.-Act. Relat.* **1994**, *13*, 302−307.
(3) Mekenyan, O. G.; Schultz, T. W.; Veith, G. D.; Kamenska, V. B. "Dynamic" QSAR for Semicarbazide-Induced Mortality in Frog Embryo. *J. Appl. Toxicol.* **1996**, *16*, 355−363.
(4) Mekenyan, O. G.; Veith, G. D.; Call, D. J.; Ankley, G. T. A QSAR Evaluation of Ah Receptor Binding of Halogenated Aromatic Xenobiotics. *Environ. Health. Perspect.* **1996**, *104*, 1302−1309.
(5) Veith, G. D.; Mekenyan, O. G.; Ankley, G. T.; Call, D. J. QSAR Evaluation of α-Terthienyl Phototoxicity. *Environ. Sci. Technol.* **1995**, *29*, 1267−1272.
(6) Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. Quantitative Structure−Activity Relationships for Polychlorinated Hydroxybiphenyl Estrogen Receptor Binding Affinity: An assessment of conformational flexibility. *Environ. Chem. Toxicol.* **1996**, *15*, 1945−1954.
(7) Mekenyan, O. G.; Ivanov, J. M.; Karabunarliev, S. H.; Bradbury, S. P.; Ankley, G. T.; Karcher, W. COREPA: A New Approach for the Elucidation of COmmon Reactivity PAtterns of Chemicals. I. Stereoelectronic requirements for androgen receptor binding. *Environ. Sci. Technol.* **1997**, *31*, 3702−3711.
(8) Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. The Role of Ligand Flexibility in Predicting Biological Activity: Structure−Activity Relationships for Aryl Hydrocarbon, Estrogen and Androgen Receptor Binding Affinity. *Environ. Sci. Technol.* **1997**, *17*, 115−25.
(9) Marshall, G. R. Binding Site Modeling of Unknown Receptors. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H. Ed.; Escom: Leiden, 1993; p 80.
(10) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(11) Handschuh, S.; Wagener, M.; Gasteiger, J. Superposition of Three-Dimensional Chemical Structures Allowing for Conformational Flexibility by a Hybrid Methodol. *J. Chem. Inf. Comput. Sci.* **1997**, *38*, 2220−232.
(12) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.
(13) Ferhst, A. *Enzyme Structure and Mechanisms*, 2nd ed.; Freeman: New York, 1985; p 263, 331, 342.
(14) Mekenyan, O. G.; Nikolova, N.; Karabunarliev, S. H.; Bradbury, S. P.; Ankley, G. D.; Hansen, B. New Developments in a Hazard Identification Algorithm For Hormone Receptor Ligands. *Quant. Struct.-Act. Relat.* In press.
(15) Ivanov, J. M.; Karabunarliev, S. H.; Mekenyan, O. G. 3DGEN: A System for an Exhaustive 3D Molecular Design. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 234−243.
(16) Smellie, A.; Kahn, S. D.; Teig, L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285−294.
(17) Smellie, A.; Kahn, S. D.; Teig, L. Analysis of Conformational Coverage. 2. Application of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295−304.
(18) Payne, A. W. R.; Glen, R. C. Molecular recognition using a binary genetic search algorithm. *J. Mol. Graphics.* **1993**, *11*, 74−91.
(19) Clark, D. E.; Jones, G.; Willet, P. 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.
(20) Jones, G.; Willet, P.; Glen, R. C. Genetic Algorithms for Chemical Structure Handling and Molecular Recognition. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London, 1996; p 211.
(21) SYBYL. Tripos Associates Inc., St. Louis, MO.
(22) 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.
(23) Lipton, M.; Still, W. C. The Multiple Minimum Problem in Molecular Modeling. Tree Searching Internal Coordinate Conformational Space. *J. Comput. Chem.* **1988**, *9*, 343−355.
(24) Leach, A. R.; Kuntz, I. D. Conformational Analysis of Flexible Ligands in Macromolecular Receptor Sites. *J. Comput. Chem.* **1992**, *13*, 730−748.
(25) Chang, G.; Guida, W. C.; Still, W. C. An Internal Cordinate Monte Carlo Method for Searching Conformational Space. *J. Am. Chem. Soc.* **1989**, *111*, 4379−4386.
(26) Goldberg, D. E. *Genetic Algorithms in Search Optimization and Machine Learning*; Addison-Wesley: Wokingam, MA, 1989.
(27) Davis, L. *Handbook of Genetic Algorithms;* Van Nostrand Reinhold: New York, 1991; p 385.
(28) Lucasius, C. B.; Blommers, M. J. J.; Buydens, L. M. C.; Kateman, G. A genetic algorithm for conformational analysis of DNA. In *Handbook of Genetic Algorithms*; Davis, L., Ed.; Van Nostrand Reinhold: New York, 1991; p 251.
(29) Blommers, J. J. M.; Lucasius, C. B.; Kateman, G.; Kaptein, R. Conformational Analysis of a Dinucleotide Photodimer with the Aid of the Genetic Algorithm. *Biopolymers* **1992**, *32*, 45−52.
(30) Nair, N.; Goodman, J. M. Genetic Algorithms in Conformational Analysis. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 317−320.
(31) Legrand, S.; Merz, K. The Application of the Genetic Algorithm to Conformational Search. *FASEB J* **1992**, *6A*, 132.
(32) Ivanov, J. M.; Mekenyan, O. G.; Bradbury, S. P.; Schuurmann, G. A Kinetic Analysis of the Conformational Flexibility of Steroids. *Quantum Struct.-Act. Relat.* **1998**, *17*, 437−449.
(33) Sippl, M. J.; Stegbuchner, H. Superposition of Three-Dimensionsl Objects: A Fast and Numerically Stable Algorithm for the Calculation of the Matrix of Optimal Rotation. *Comput. Chem.* **1991**, *15*, 73−78.
(34) Davies, E. K. and Murrall, N. W. How Accurate a Force Field Need Be? *Comput. Chem.* **1989**, *13* (2), 149−156.
(35) White, D. N. J. Molecular Mechanics Calculations. *Spec. Rep. Chem. Soc.* **1987**, *6*, 38−63.

(36) Karabunarliev, S., Nikolova, N.; Nikolov, N.; Mekenyan, O. G. Rule Interpreter: A Chemical Language that Implements Decision Rules Base on Molecular Structure. Submitted).

(37) Mekenyan, O. G.; Karabunarliev, S. H.; Ivanov, J. M.; Dimitrov, D. N. A New Development of the OASIS Computer System. *Comput. Chem.* **1994**, *18*, 173−187.

(38) Weininger, D. SMILES, a Chemical Language and Information System. Part 1. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36

(39) Among other features not pertinent to this paper, extended SMILES supports qualifiers for stereochemical components. These qualifiers are keywords enclosed in curly brackets attached to the corresponding atom or bond entry. Valid stereochemical atom qualifiers are {p+} and {p−}, which denote positive and negative parity of the corresponding site. Bond qualifiers fixing the coordination of a chemical bond are: {t}, *trans (E)*; {c}, *cis (Z)*; {g}, gauche; {g+}, gauche (clockwise); and {g−} gauche (counterclockwise).

(40) The basic form and parametrization of PMM was taken from the Chem-X force field.[34,35] The following potential energy terms are included: $(k_{BL}/2)(x - x_{BL})^2$, x = bond length; $(k_{VA}/2)(x - x_{VA})^2$, x = valence angle; $k_{TA}\{1 - \cos[n_{TA}(x - x_{TA})]\}$, x = torsion angle; $k_{OP}/2\sin^2(x)$, x = out-of-plane angle of sp^2 sites; $k_{LJ}[(x/x_{LJ})^{12} - 2(x/x_{LJ})^6 + 1]$, x = nonbonded distance. Quantities with subscript are parameters that depend on the type of the atoms and bonds involved.

(41) Wehrens, R.; Pretsch, E.; Buydens, M. C. Quality Criteria of Genetic Algorithms for Structure Optimization. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 151−157.

(42) Arora, K. Correlation of Structure and Activity in Ansamycins: Structure, Conformation, and Interactions of Antibiotic Rifamycin S. *J. Med. Chem.* **1985**, *28*, 1099−1102.

(43) Cellai, L.; Cerrini, S.; Segre, A.; Battistoni, C.; Cossu, G.; Mattogno, G.; Brufani, M.; Marchi, E. Study of Structure−Activity Relationships in 4-Deoxypyrido[1′,2′-1,2]imidazo[5,4-c]rifamycin SV Derivatives by Electron Soectroscopy for Chemical Analysis and ^{1}H NMR. *Mol. Pharmacol.* **1985**, *27*, 103−108.

(44) Wiese, T.; Brooks, S. C. Molecular Modeling of Steroidal Estrogens: Novel Conformations and their Role in Biological Activity. *J. Steroid Biochem. Mol. Biol.* **1994**, *50*, 61−72.

CI990303G