Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm

Mikko J. Vainio* and Mark S. Johnson

Structural Bioinformatics Laboratory, Department of Biochemistry and Pharmacy, Åbo Akademi University, Tykistökatu 6A (BioCity), FI-20520 Turku, Finland

Received December 21, 2006

The task of generating a nonredundant set of low-energy conformations for small molecules is of fundamental importance for many molecular modeling and drug-design methodologies. Several approaches to conformer generation have been published. Exhaustive searches suffer from the exponential growth of the search space with increasing degrees of conformational freedom (number of rotatable bonds). Stochastic algorithms do not suffer as much from the exponential increase of search space and provide a good coverage of the energy minima. Here, the use of a multiobjective genetic algorithm in the generation of conformer ensembles is investigated. Distance geometry is used to generate an initial conformer, which is then subject to geometric modifications encoded by the individuals of the genetic algorithm. The geometric modifications apply to torsion angles about rotatable bonds, stereochemistry of double bonds and tetrahedral chiral centers, and ring conformations. The geometric diversity of the evolving conformer ensemble is preserved by a fitness-sharing mechanism based on the root-mean-square distance of the atomic coordinates. Molecular symmetry is taken into account in the distance calculation. The geometric modifications introduce strain into the structures. The strain is relaxed using an MMFF94-like force field in a postprocessing step that also removes conformational duplicates and structures whose strain energy remains above a predefined window from the minimum energy value found in the set. The implementation, called Balloon, is available free of charge on the Internet (http://www.abo.fi/~mivainio/balloon/).

1. INTRODUCTION

The majority of the methods within the domain of computer-assisted drug design make the simplification that molecules are static objects. This simplification may be too harsh for applications that are based on the three-dimensional (3D) properties of molecules. Examples of such applications are molecular docking[1] and pharmacophore construction and search methods.[2,3] The ability to account for the conformational flexibility of molecules is of specific importance for those methods that quantitatively predict some measurable real-life property of a molecular system. The observed value of such a property is usually an average over the values for different microscopic states of the system. In this average, the contribution of each microscopic state is weighted by the probability of occurrence for that state. The probabilities are given by the Boltzmann equation, according to which high-energy states have a very small probability of occurrence. Thus, high-energy states contribute very little to the observed average value of the measured quantity. Recently, quantitative structure−activity relationship methods that account for molecular flexibility have been introduced.[4−6] These methods, and the ones mentioned earlier, require the enumeration of a set of conformers for each processed compound. The set of conformers should reflect the distribution of microscopic states observed in reality. Usually the enumerated conformers are required to have a potential energy value within a given tolerance from the global minimum for that compound. The conformers should also be geometrically as distinct from each other as possible (within the given energy window) in order to reflect the flexibility of the system; duplicate conformations do not provide new information. Thus, the task of generating conformer ensembles can be described as a multimodal optimization problem, i.e., one where the objective space has multiple optima. The use of additional criteria besides the potential energy may also be of interest when, e.g., exploring the conformational flexibility of a molecule against a pharmacophore model. Therefore, the ensemble generation task may also be formulated as a multiobjective optimization problem. Genetic algorithms[7] (GAs) can be used to solve such tradeoff optimization problems with multiple objectives.

In the workflow of a multiobjective GA (MOGA) (see also the pseudocode listing below), a population of possible solutions (individuals) is first generated randomly. A solution is usually presented as an array (chromosome) of numbers (genes) to be used as parameters for the functions that are being optimized. Each solution is assigned an array of numerical fitness values that estimate the goodness of the solution, one value per criterion. A solution is said to dominate another solution if it is equally good in all objective criteria but better in one or more. The population can then be sorted according to the number of other individuals that dominate a given solution (degree of domination or Pareto rank).[8] Solutions with the same degree of domination are said to belong to the same Pareto front. A set of new solutions is generated by combining the genes of two individuals selected randomly (with bias toward the least dominated, i.e., better solutions) from the population in an operation called crossover. The combination of two solutions introduces a large move in the search space (exploration).

* Corresponding author phone: +358-2-215 4600; e-mail: mikko.vainio@abo.fi.

10.1021/ci6005646 CCC: $37.00
Published on Web 09/25/2007

The newly generated offspring solutions are subject to mutation operations that introduce small changes to the genes. The mutations search the neighborhood of the existing good solutions (exploitation). The new population is evaluated for fitness, and the process starts a new cycle (generation). Many GAs copy a small number of the best parent solutions directly into the next generation without crossover or mutation (elitism).

Genetic algorithms, although relatively easy to implement, involve a number of control parameters that maintain a balance between the runtime and the quality of the obtained results. The selection of parents for reproduction is biased toward the least dominated (better) solutions. The bias introduces evolutionary pressure (survival of the fittest). The magnitude of the selection bias affects the speed at which the solutions become very similar to each other (convergence). A too large bias leads to premature convergence, where the search space is no longer efficiently explored and the global optima may not be found. A too small bias causes the solutions to jump around the search space and failure to converge near the optima within the runtime. In addition, the probability of mutation affects the diversity of the solutions and thus the convergence properties of the algorithm: a large mutation probability may be too disruptive for any found good solution to be maintained in the population, while an insufficiently low mutation probability results in poor local sampling of the search space. The number of solutions in a population affects the convergence because diversity is easier to maintain in a larger set of solutions.

Rules of thumb exist for setting reasonable values for the control parameters; Djurdjevic et al. studied the relative effects of algorithm design and control parameter settings on the performance of GAs for ab initio protein structure prediction.[9] The sometimes complex interplay of the control variables boils down to the diversity of the genetic material maintained in the evolving population. In general, the more diverse the solutions in the evolving population are, the closer to the global optimum the GA converges. Diversity is commonly maintained by niching, a technique where the probability of solutions in crowded regions of the objective (or parameter) space to be selected for reproduction is penalized in favor of the less crowded (more unique) solutions.

Several examples of the use of a single-objective GA for conformational analysis and generation of conformer ensembles exist in the literature, e.g., refs 10−20 (for a broad review on evolutionary algorithms in drug design see ref 21). MOGA has been applied to, e.g., pharmacophore generation[22,23] and QSAR.[24] To our knowledge only a few reports of the use of a MOGA for conformational analysis exists to date: ref 25 is the latest paper in a series of studies where MOGA was applied to protein structure prediction. Here, we report a MOGA (see Chart 1) for the generation of conformer ensembles for drug-sized organic molecules. The algorithm is implemented in a computer program named Balloon (describing the expansion of a structure from a single 2D representation to an ensemble of 3D models). The implementation features phenotype based niching and selection in a novel manner for GAs for conformational analysis. The design of these algorithmic elements follows that of the Niched Pareto-optimal Genetic Algorithm (NPGA)[26] and its successor NPGA2.[27] The population handling mechanism used in this work derives from the Nondominated Sorting Genetic Algorithm (NSGA-II).[28]

Chart 1. Algorithm 1: The Implemented MOGA

Require: topology of a molecule
generate template conformation
fill population with N_0 random individuals
evaluate energy of the conformers encoded by the individuals
sort the population according to Pareto rank
generation ← 0
while not(Termination()) **do**
  select parents for reproduction and create N_0 offspring
  make mutations to the genes of the offspring
  evaluate energy of the conformers encoded by the offspring
  add the offspring to the population (size grows)
  sort the population according to Pareto rank
  discard geometrically redundant conformers
  shrink the population back to size max(N_0, size of the Pareto front)
  generation ← generation + 1
end while
discard geometrically redundant conformers
optimize remaining conformers
discard high energy conformers
discard conformers with incorrect stereochemistry
discard geometrically redundant conformers
output remaining conformers

Most of the GA implementations for conformational analysis mentioned above encode conformers as an array of torsion angle values applied to a template structure. Our algorithm does the same and uses additional structural modifications that enable more fine-grained sampling of the conformational space than the mere use of torsional rotations would allow for. The geometric modifications build upon and extend those used in existing GAs.

The aim of this study was to develop a MOGA that produces conformer ensembles that are both low in potential energy and geometrically dissimilar. As a proof-of-concept test, the algorithm is run on a set of drug-sized organic molecules, and the geometric diversity and energy of the generated conformers is assessed. The results obtained on the set of molecules indicate that valid design decisions were made. The program is available free of charge from www.abo.fi/~mivainio/balloon.

2. METHODS

2.1. Generation of Template Conformation. The input structure may or may not contain 3D atomic coordinates. A SMILES string[29] is an example of the latter case. Initial atomic coordinates for such topology-only input are generated using stochastic proximity embedding[30,31] or, should that fail, using metric matrix embedding[32] (for a review on the distance geometry method see ref 33). The initial coordinates usually violate the distance and signed volume bounds. The violation is minimized by the conjugate gradient method[34] using a sum of the bound violations as the objective function. The initial atomic coordinates are produced in four-dimensional space. The fourth dimension is required in the initial optimization in order to ensure correct configuration of stereochemical centers—incorrect chiralities can flip around as the atoms can "pass through" each other in 3D without imposing high proximity penalty values due to difference in the fourth coordinate value. After the optimization has terminated, the fourth dimension is discarded, and another optimization pass is made in 3D, first against the

distance geometry bounds and then using the conformational energy as calculated by an in-house reimplementation of the MMFF94[35] molecular force field as the objective function.

Distance geometry can produce unrealistic structures, e.g., intersecting rings (analogous to a link in a chain) that are impossible to get rid of using gradient based optimization methods. Therefore, an option is provided to apply the downhill simplex minimization algorithm of Nelder and Mead[34,36] prior to conjugate gradient optimization in order to "shake" the structure away from local minima. The resulting structure will have reasonable bond lengths and valence angles provided that the optimization is allowed to iterate until convergence. The bond lengths and valence angles remain intact (with exceptions in aliphatic rings as described below) during the generation of the conformer ensemble.

The configuration of the stereochemical centers of the initial conformer are verified to conform to the values specified in the input. Any discrepancies are resolved by iterative application of the geometric modifications for stereocenters (described below) and short MMFF94 energy minimizations (conjugate gradient) and by mirror reflection of the whole structure if necessary. If any discrepancy remains, the structure is output to a separate file, an error is reported, and the structure is not processed further.

The distance geometry method was chosen for the generation of the template structure because it is general in terms of covered chemistry, the resulting geometry does not depend on the input sequence as has been observed for several conformer generator programs,[37] and the used algorithms have been published in detail and are therefore straightforward to implement. Distance geometry can, however, be very time-consuming, and it is provided only in order to ensure a valid startpoint for the GA. Initial 3D coordinates produced by some fast rule-based method, such as the one implemented in Corina,[38,39] should be used whenever possible.

2.2. Genome. The molecular conformations are encoded relative to the initial structure. The genome object resembles those used in previously reported GAs for conformational analysis:[11,18,19] it consists of four chromosomes that encode different structural modifications applied to the template conformation, two chromosomes that encode the order of execution of two of the modifications as described below. The resulting atomic coordinates, or the phenotype, are stored within the genome object for later use during the calculation of the distance between genomes.

2.2.1. Torsion Angles. Djurdjevic et al.[9] studied the relative effects of algorithm design and control parameter settings on the performance of GAs for protein structure prediction. Their results encouraged the use of the real-valued encoding of torsion angles. Here, the first chromosome is an array of floating point values that describe the values of torsion angles of rotatable bonds, i.e., acyclic single bonds connecting nonterminal atoms. The torsion angles assume values within the half-open interval $[-\pi, \pi)$. The molecule is divided into a set of rigid fragments that are connected with rotatable bonds, and the set is ordered as a tree data structure rooted at the largest fragment that remains immobile. The geometric transformations required to rotate about a bond, namely rotation and translation, are additive, which allows the torsion angles to be adjusted in linear time with respect to the number of atoms by recursively traversing the tree and moving the atoms according to the net transformation associated with the rotatable bonds between each fragment and the root.[40]

The crossover operator produces two copies of the parent solutions, one copy of each. The chromosomes of the offspring are then reorganized. A uniform crossover operation is used for both the torsion angle and global transformation arrays: the gene values at each locus are swapped between the two offspring with some probability, usually between 0.1 and 0.2. In general, high crossover rates can be overly disruptive for any found good solutions to be maintained in the population.

A recent study found an improvement in the rate of convergence of a conformer generator GA upon the introduction of biased torsional mutations.[41] Torsion angle values that result in low-energy geometries were identified based on a set of random conformations generated in a preprocessing step. Torsion angle mutations were then biased toward these favorable values. It was acknowledged that the conformational energy is not linearly dependent on the torsion angle values, but the relation is of a combinatorial nature. Consequently, Gaussian perturbation of the favored torsion angles was used in order to allow exploration of unfavorable ranges of angle values. Parent et al.[20] found no advantage in using a similar biased torsion mutation scheme over using a flat distribution of random angle values. Because of the controversy associated with these previous reports, we implemented a 2-fold torsion angle mutation operation that allows for both large exploratory moves and small changes that help to sample the vicinity of the good angle values that have already been found. The mutation operator selects a random locus from the array of torsion angle values. The selected locus is either assigned a random value in the range $[-\pi, \pi)$, drawn from a uniform distribution, or the current value is perturbed by a small random amount in order to sample the nearby solution space. These two alternatives have equal probability of occurrence. An analogous 2-fold torsion mutation scheme was adopted by Cutello et al.,[25] who used scaling functions for the torsion mutation rates so that local sampling becomes pronounced as evolution proceeds. We did not implement any scaling of the mutation rates.

2.2.2. Chiral Inversion. The stereochemistry of a compound is not always completely defined in the input, but the conformer generation algorithm is requested to sample the other possible stereochemical configurations, too. The chiral inversion operation is provided for that purpose. The inversions of tetrahedral and double bond stereochemical centers are encoded into the second chromosome, which is an integer array with one element per chiral center. The element values corresponding to those stereogenic centers that cannot be inverted without breaking a bond are set to zero as well as those corresponding to atoms whose chirality is defined in the input. Values at other array positions are set equal to one, which indicates no inversion relative to the initial conformation. A value of -1 triggers inversion in the evaluation. The chromosome is mutated by swapping the sign of the gene value at a randomly selected locus.

Let X be a tetrahedral chiral atom connected to four neighboring atoms: A−X(−B)(−C)(−D). The inversion is achieved by a rotation by angle π about the axis defined by atom X and the midpoint between two of its neighboring atoms, say, A and B. The rotation is applied to atoms A and B and the subtrees rooted under them. Inversion by rotation

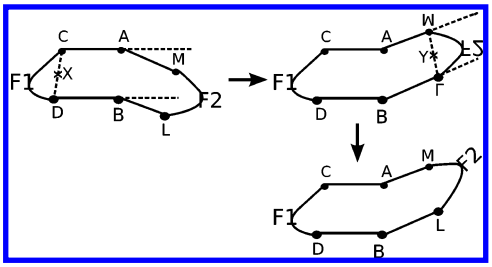

Figure 1. The flip-of-fragments operation. The fragment F2 is subject to two reflections that invert the valence angles C−A−M and D−B−L. See the text for details.

is possible even when atoms A, X, and B are part of the same ring but impossible if either of atoms C or D is also a member of that same ring system. Chiral centers at a singular junction of two rings (X is the only atom in both rings) can be inverted. Chiral centers at ring junctions that are comprised of more than one atom are handled with the "inversion of pyramids" operation as described below.

Inversion of acyclic stereogenic double bonds is achieved analogous to the tetrahedral case, except the axis of rotation is the double bond.

A molecule can be partitioned into a tree of rigid fragments according to stereogenic centers, analogous to fragmentation with respect to rotatable bonds. Thus, the geometric transformations required to invert the centers can be done in linear time with respect to the number of atoms in the molecule.

2.2.3. Flip of Fragments. The conformation of aliphatic rings is modified by applying the "flip of fragments" operation,[19] which is a generalization of the "flip of free corners" operation introduced in ref 11. The flip-of-fragments procedure is described in ref 19 and only shortly reviewed here.

The flip-of-fragments operation is defined for all pairs of sp^3-hybridized nonvicinal ring atoms A and B whose removal would break the molecular graph into two "disconnected fragments that are incident to both atoms".[19] This condition can be checked for a pair of atoms A and B using the following algorithm:

(1) Use breadth-first search in the molecular graph to find the set *S* of all different paths from A to B. (In particular, paths not leading to B are not in *S*. The atoms in these paths must be rotated about the C−A bond as described below in order to keep valence angles intact.) If A or B is encountered twice in a path during the search, then the pair of atoms does not give rise to a flip-of-fragments operation.

(2) Merge all paths that have at least one atom in common (excluding A and B, of course). Replace the two coinciding paths with the merged path in *S*.

(3) A and B define a flip-of-fragments if the size of $S = 2$. The flip-of-fragments operation involves the pair of atoms A and B, the smaller fragment F2, and two atoms from the larger fragment, C and D. Atom C is directly bonded to A, and D is directly bonded to B (Figure 1). The fragment F2 is reflected twice, first with respect to the plane defined by atoms A and B, and the midpoint of atoms C and D, denoted X. The first reflection converts fragment F2 to its mirror image. The second reflection with respect to the plane defined by atoms A and B, and the midpoint of atoms M and L, denoted Y, restores the original "image" of fragment F2. The flip-of-fragments operation becomes equal to the flip-of-free corner operation when there is only one atom in fragment F2 (i.e., M = L). The reflections retain the bond lengths intact, but the valence angles might change if the connecting bonds A−C and B−D, or A−M and B−L, are not parallel to each other. Therefore, a parallelity threshold is used for the angle between the directions of the A−C and B−D bonds (a default value of 30° was used in this study). The flip is performed only if the angle between the bond directions is below this threshold.

The number of flip-of-fragments operations for a given cycle increases in proportion to the second power of the number of atoms in the ring. The memory requirements may therefore become overwhelming for large cyclic structures, e.g., cyclic peptides, if all possible flip-of-fragments operations are taken into account. Balloon uses an upper limit for the size for rings to be treated as flexible in order to avoid exhaustion of memory upon processing of large cyclic structures.

Atom A may have neighbor atoms that are not members of the ring system. When the fragment F2 is reflected, the valence angles between the atom M and the neighbors change. In order to restore the valence angles, the neighbors and the subtrees rooted at those (i.e., the paths not leading to B as found by the breadth-first search in step 1 above) must be rotated about the C−A bond with the same amount as the (atom-in-F1)−C−A−M torsion angle changed upon the reflections. An analogous procedure must be applied to the nonring neighbors of atom B; only the subtrees to be rotated are not known as a side-product of the breadth-first search but must be found by a separate search step.

The flip-of-fragments operations are encoded by an array of bits that has one element per operation. The value of the element is used to determine whether to execute the corresponding operation or not. The mutation operator toggles a randomly selected bit on or off.

The flip-of-fragments operation uses the coordinates of the ring atoms as reference points. Therefore, the order of execution of flips defined for a ring affects the resulting geometry. The order of execution of the flip operations was randomized in the original formulation.[19] The randomization detaches the genotype from the phenotype (the atomic coordinates), which can be seen as the effect of environment on the development of an individual. However, determination of the optimal order of execution is a permutation problem, and GAs can be used to solve permutation problems as well. Here, an additional chromosome is used to encode the order of execution of the flip operations. The permutation chromosome is an array of integers, in which each element value is a unique index to the flip-of-fragment operations. The flip-of-fragment operations are executed in the order given by the permutation chromosome.

The permutation chromosome requires a specialized crossover operation that ensures the uniqueness of the element values and preserves the relative order of the elements as closely as possible. We implemented a position-based crossover operator[42] that has the desired properties. The mutation operation for the permutation chromosome swaps the values of two randomly selected genes.

2.2.4. Inversion of Pyramids. The flip-of-fragments operation described above does not apply to atoms at ring junctions. The "reflection of pyramids" operation[18] was introduced in order to modify the geometry of sp^3-hybridized atoms at the junctions of two (or more) fused aliphatic rings.

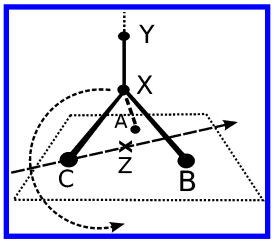

Figure 2. The inversion of pyramids operation. Atom X and the fragment rooted under it are subject to rotation by an angle π about the dashed axis going through Z and C. The axis lies in the plane defined by A, B, and C. See text for details.

Let X denote such an atom (Figure 2). Atoms A, B, and C are sp^3-hybridized neighbors of X so that all are members of the same polycyclic system. Let Y denote a fourth moiety that is connected to X. If Y exists, it may not be a member of the same ring system as X and its other neighbors. Atoms X, A, B, and C define a pyramid, the base plane of which is defined by A, B, and C. In the original procedure, atom X and the moiety Y were reflected with respect to the plane of the base of the pyramid.[18] Since an odd number (one) of reflections was performed, all chiral centers involving any of the moving atoms, including those in Y, were inverted. The configuration of centers for which the configuration was defined in the input was restored in a later step. Our encoding scheme does not use absolute stereodescriptors (required in order to assess whether a center was inverted in a reflection or not), hence the configurations would not be restored later. Therefore, we introduce a reflectionless procedure that inverts a pyramidal center: a rotation of the X−Y moiety by an angle π about an axis defined in the base plane of the pyramid (Figure 2). The axis must pass through the projection point of X on the plane, denoted Z, but its direction can be chosen arbitrarily (atom C is used to define the direction in Figure 2).

The use of a rotation instead of a reflection abolishes the problematic inversion of chiral centers possibly present in the moiety Y. If any of the atoms X, A, B, or C are chiral, then their stereochemical configuration will still be inverted. Inversion is not allowed for those pyramidal centers for which the stereoconfiguration is defined in the input.

The torsion angle about the X−Y bond (if any) may change upon inversion of the pyramid. The torsional rotations are encoded as absolute values and applied after the pyramidal inversions, which restores the correct torsion angle and orientation of moiety Y.

Another issue with both the reflection and inversion of pyramids is that the valence angle values may change. If the base plane atoms A, B, and C are coplanar with their ring system neighbors other than X, the valence angles can be restored using the following correction operation applied to the nonring neighbors of the base plane atoms: rotate about a bond L−M between a base plane atom M ($\in$ {A,B,C}) and its vicinal ring atom L ($\neq$ X) by an angle obtained as the difference of a torsion angle K−L−M−X (K ($\neq$ M) is any neighbor of L) before and after the rotation of atom X.

If atoms A, B, and C are not coplanar with their ring system neighbors other than X, the valence angles involving A, B, and C will still change upon the operation. This leads to somewhat strained structures. This strain is relaxed in the postprocessing phase (see below).

The modifications to the pyramidal centers are encoded in the fourth chromosome, which is an array of integers. One element is allocated for each pyramidal center, and the sign tells whether to invert or not. Values corresponding to fixed pyramidal centers are set to zero. Mutation in the pyramidal inversion chromosome swaps the sign of the value at a randomly selected locus (zero leaves a pyramidal center untouched).

Pyramidal centers frequently occur next to each other. Therefore, the execution order of the inversions has an effect on the resulting geometry. This permutation problem is tackled by using the same procedure as with the flip-of-fragments operation, namely an additional chromosome that encodes the order of execution of the pyramidal inversions.

2.3. Genetic Operators. *2.3.1. Mutations.* Mutations to different chromosomes of the genome each have their own implementation and adjustable mutation probabilities as described above. A common trait of the mutation operations is the random selection of the locus of mutation.

The mutation probabilities should be kept relatively low, typically below 0.1.[9] The default setting in Balloon is to use a probability of 0.05 for all mutations.

2.3.2. Crossover. A uniform crossover operator is used for all but the permutation chromosomes, for which the position-based crossover is used. The default probability for entering the crossover process is 0.9, and the default probability of performing a crossover at any given locus is 0.2.

2.4. Objective Functions. Potential energy values calculated using the MMFF94-like molecular force field are used as objectives. The total energy value contains contributions from valence angle bending and stretch-bending terms of the force field. As pointed out before, the valence angles might change because of the flip-of-fragments and pyramidal inversion operations. Relaxation of the angle strain is not feasible during the evolution of the system due to the overwhelming CPU time required. Therefore, the use of the total energy as an objective function would likely hinder the exploration of the conformational flexibility of saturated rings.

The energy related to rotation about bonds is accounted for by the torsional terms of the MMFF94 potential energy function.[35] Since torsion angles are subject to geometric modifications, the torsional energy is an obvious choice for an objective function. The van der Waals energy term measures the nonbonded steric interactions between atoms and increases exponentially when atoms are in close proximity, thus penalizing evolving structures for steric clashes. The van der Waals energy term is used as a second objective function. These two objective functions can, however, conflict with each other in certain situations.[25]

The number of the van der Waals interaction pairs scales with the second power of the number of atoms in the structure, while distant pairs of atoms make a very small contribution to the energy. Most molecular mechanics software apply a distance cutoff to remove distant interactions from consideration. This approximation results in a remarkable speed-up in the processing of large structures. Balloon employs a distance cutoff in the form of a simple polynomial switching function that, when multiplied with the van der Waals interaction energy, brings the energy smoothly to zero at the cutoff distance.

The electrostatic interaction potential is decidedly not used as an objective function. Previous studies have indicated that inclusion of the electrostatic energy term as an objective function may lead to the formation of intramolecular interactions (hydrogen bonds and salt bridges), which causes packed conformations that would not be observed for a solvated molecule.[43−45]

The two objective functions consider only the internal geometry of the molecule. Other objective functions related to the estimation of biological activity, such as pharmacophore matching or interaction energy with a binding site of a receptor, could be included without modifications to the existing parts of the algorithm.

2.5. Selection. Individuals are picked for reproduction in a process called selection. The selection routine used in NPGA2[27] (NSGA-II uses a comparable approach) was used as a basis for the development of the tournament selection routine used in this study. Solutions are selected randomly from the population, and the solution with the lowest degree of domination wins the tournament. The number of individuals selected for the tournament can be used to adjust the selection pressure: the greater the tournament size, the smaller the probability for a strongly dominated solution to reproduce.

Because the conformational potential energy landscape of a molecule usually has multiple relevant minima to be found, one needs a mechanism to promote geometric diversity in the population of evolving conformers. One such mechanism is to use the niche count *m*, a number calculated from the distances between individuals, to break ties (two or more solutions with the same degree of domination) in the tournament selection. The individual with the lowest niche count wins the tie. NPGA2 and NSGA-II use a distance calculated in the objective space for determining the niche count. In our case a distance metric defined in the objective space (the MMFF94-like potential energy) is not feasible because conformers with equal energy may have distinctly different geometries. Instead, the atomic coordinates (the phenotype encoded by the genes) are defined in a metric space in which unambiguous distances can be readily calculated. The niche count for conformer *i* is obtained from the commonly used formula

$$m_i = \sum_{j \neq i}^{N} \begin{cases} \left(1 - \dfrac{d_{ij}}{\sigma_{\text{share}}}\right) & |d_{ij} < \sigma_{\text{share}} \\ 0 & |d_{ij} \geq \sigma_{\text{share}} \end{cases} \qquad (1)$$

where σ_{share} is the niche radius (1.5 Å in this study), *j* runs over all conformers in the population, and the distance value d_{ij} is the root-mean-square deviation (rmsd) of an optimal superimposition of conformers *i* and *j* calculated in a least-squares sense[46] using the coordinates of non-hydrogen atoms.

The requirement of defining a value for σ_{share} can be considered a weakness in conjunction with problems where the scale and range of variation of the distance metric can be arbitrary, but the scale of the rmsd over atomic coordinates is readily comprehensible to a scientist who has a modest amount of experience in molecular modeling.

Calculation of the rmsd between conformers of a molecule with a degree of topological symmetry requires special attention. Obviously, the rmsd between two perfectly superimposable conformers can become very high if one conformer has undergone a symmetry operation and the coordinates of a "wrong" pair of symmetrically equivalent atoms are used in the calculation of the rmsd. A brute-force solution to the problem is to calculate the rmsd for all possible automorphisms (isomorphisms of a graph against itself, or combinations of topological symmetry operations) of the molecular graph and take the lowest rmsd value as the distance.

The number of automorphisms can become very large for graphs with a high degree of symmetry. Fortunately, the number can be reduced drastically by not considering hydrogen atoms. The position of a hydrogen depends on the position of the heavy atom the hydrogen is bonded to. Omitting hydrogens does not therefore affect the order in which the automorphisms appear when ranked according to increasing rmsd but only the scale is affected. Thus, the backtracking search algorithm[47] used to perceive the automorphisms is applied only to the set of non-hydrogen atoms. The automorphisms are perceived and cached before the evolutionary cycle starts, and only the coordinates of atoms in the stored automorphisms are used in the calculation of the rmsd.

The number of individuals taken into the tournament can be used to adjust the selection pressure in situations where the whole population has the same Pareto rank: the greater the tournament size, the smaller the probability for a crowded solution to reproduce. Because the less crowded solutions in the current generation will reproduce more than the crowded ones, the phenotype space of the next generation will be crowded at regions that are uninhabited in the current population. Balloon uses a default tournament size of two solutions.

2.6. Population Handling and Elitism. Elitism has been shown to improve the convergence properties of evolutionary algorithms especially in cases where the fitness landscape is multimodal.[48] Two of the model algorithms for this study, NPGA and NPGA2, do not have any mechanism for elitism: they use tournament selection to select individuals from the current population and the offspring of those individuals fill the next generation. There is no guarantee for a superior solution to be carried on to the next generation without modifications to its genetic material.

NSGA-II implements elitism in an explicit manner: the current population (*N* individuals) is merged with the offspring (also *N* individuals), and the augmented population (of size 2*N*) is sorted according to the degree of domination. The *N* individuals for the next generation are then selected from the sorted and augmented population in a manner that ensures the next generation will contain the best solutions found so far: starting from the nondominated Pareto front, all individuals of the front are added to the population of the next generation provided that the size of the growing population does not exceed *N* as a consequence of the addition. This is done successively for fronts with increasing rank. When a front that cannot be accommodated within the *N* individuals is encountered, the individuals in the front are sorted according to a metric defined in the objective space (the authors of NSGA-II call the metric *crowding distance*), and the least crowded individuals are selected to fill the remaining slots in the population.

The GA implemented in this study uses a merge-populations-and-reduce procedure that resembles the one in NSGA-II. There are, however, four differences. First, crowded solutions are deleted from the population: the rmsd is measured between all pairs of solutions as described above, and if the rmsd falls below a tolerance d_{crowd} specified by the user (by default 0.5 Å), then the conformer with higher Pareto rank (or higher niche count according to eq 1 if the ranks are equal) is deleted. Consequently, the population size N can decrease.

Second, the population is allowed to grow in order to accommodate all remaining individuals that have zero Pareto rank (the nondominated front) if necessary. Consequently, the population size N can increase.

Third, the comparison of objective function values in the determination of dominance employs a tolerance value: two energy values are considered equal if they differ by less than an energy threshold E_t value calculated from a linear energy window function E_w and the current population size N using

$$E_t = E_w \cdot \min\left(0.5, \frac{N_0}{2N}\right) \quad (2)$$

where N_0 is the initial population size as given by the user, and

$$E_w = E_0 + k \cdot N_{rb} \quad (3)$$

where N_{rb} is the number of rotatable bonds in the molecule, the (user-definable) constant E_0 takes by default a value of 10 kcal mol^{-1} and the (user-definable) slope k 0.5 kcal mol^{-1} per rotatable bond. The use of a scaling function for the energy window (eq 3) was inspired by the dependence of the energy difference between the bioactive conformation and the closest local minimum and the number of rotatable bonds observed by Perola et al.[44] Using a threshold in the energy comparisons causes all conformations that are within the energy threshold from the so-far lowest energy conformer to reside in the nondominated front. Because the population size is adjusted to hold the entire nondominated front, a large energy threshold tends to increase the population size. A large population size, in turn, tends to decrease the energy threshold according to eq 2. These opposing tendencies combined with the removal of crowded conformers lead to self-adjusting behavior so that the population size does not grow infinitely. In case the removal of crowded solutions causes the population size to decrease below the user-defined value N_0, random conformers are added to the population in order to increase its diversity. Thus, the population size adapts to the flexibility and the shape of the potential energy hypersurface of the compound.

Fourth, the phenotype based niche count is used as the metric for selecting individuals from the borderline front (of rank > 0) instead of the crowding distance. Thus, geometric diversity of the population affects the evolutionary pressure at two phases of the algorithm: the tournament selection and the nondominated sorting.

The removal of crowded solutions can cause the solution with the absolute minimum energy to be dropped out from the next generation if it resides in a crowded region of the phenotype space. Despite this, a reduction in the energy values takes place (see Results).

2.7. Termination Criteria. Evolution has no well-defined natural endpoint. As a consequence, there are no natural termination criteria for a GA run in general. Some problems might provide a nonambiguous termination criterion, e.g., convergence to a zero fitness value in the minimization of bound violations in distance geometry where the objective function value is always ≥ 0. Two simple termination conditions are the maximum allowed number of generations and the maximum allowed runtime. We have implemented both of these criteria.

2.8. Population Postprocessing. The final population is pruned in order to conform to the low-energy window as given by eq 3: the geometry of the conformers is optimized using the conjugate gradient method against the MMFF94-like potential energy, excluding the electrostatic term in order to avoid the formation of intramolecular hydrogen bonds. The torsional component of rotatable bonds is also excluded from the *gradient* in order to prevent the driving of the torsion angles down the energy well, which would result in only a few conformers and decrease the coverage of the torsional space within the allowed energy window. The gradient arising from the nonrotatable (including ring) bonds is taken into account in the gradient. All atoms are allowed to move, which allows the angle strain introduced by the flip-of-fragments and the pyramidal inversions to relax. Conformers whose energy is not within the energy window or whose stereochemistry does not conform to that specified in the input are discarded.

The remaining conformers are again checked for geometric redundancy based on the rmsd as described above. The geometric transformations (rotation and translation) required for the optimal superimposition are obtained as a side product of the calculation of the rmsd. The transformations are applied in order to superimpose the conformers on the minimum energy conformer, which facilitates visual comparison of the resulting structures.

2.9. Test Run. The applications of a conformer generation algorithm are usually related to the modeling of biomolecular recognition. Conformer generation algorithms are therefore typically tested by a comparison of the generated ensemble and the protein-bound ligand coordinates observed in an X-ray crystal structure stored in the Protein Data Bank[49] (PDB)[43−45,50−54] or against single-molecule X-ray crystal structures in the Cambridge Structural Database.[55,56] A test run operating mode was implemented in Balloon in order to perform such comparisons. The input structures are treated as topology-only (2D) when in test run mode: the input atomic coordinates (i.e., those of the crystal structure) are stored aside in a separate data structure, and the coordinate values are explicitly set to zero prior to the generation of the template conformation upon which the individuals of the GA operate. The zeroing of initial coordinates removes any bias introduced by the input geometry. After the evolutionary cycle has terminated, the rmsd values of the optimal superimposition between each conformer in the final ensemble and the crystal structure coordinates are calculated in the symmetry-aware manner described above. The minimum of the optimal rmsd and the used CPU time are recorded for each ensemble.

The CCDC/Astex test set of protein−ligand complex structures,[57] augmented with the recently published Astex Diverse Set,[58] was chosen for the test runs. The CCDC/Astex

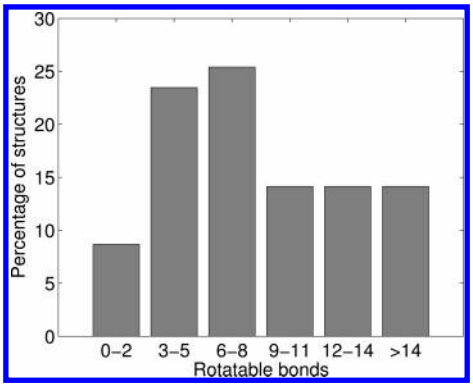

Figure 3. The distribution of rotatable bonds in the combined set of the 222 "clean" structures from the CCDC/Astex set and the 85 compounds in the Astex Diverse set.

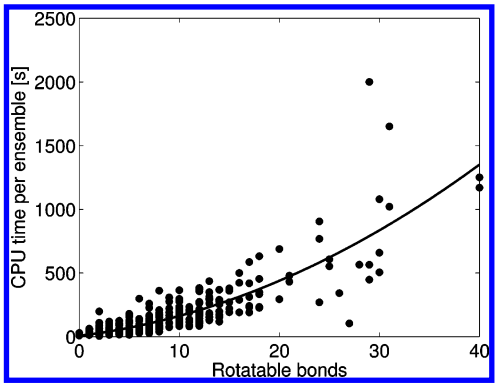

Figure 4. The CPU time used per conformer ensemble as a function of the number of rotatable bonds. The observed quadratic trend is plotted as a solid line.

set consists of 305 entries selected from the PDB and was originally used as a benchmark test for docking algorithms. Because the full CCDC/Astex set contains some rather low-resolution and information-deficient structures, we used the "clean" set of 224 ligands listed in ref 57 excluding the ligand in 1mbi (imidazole) that has no conformational degrees of freedom. In the original PDB entry 6rsa the ligand (uridine-2′,3′-vanadate) contains a vanadium atom, which was replaced by a phosphorus in the CCDC/Astex test set structure. Because the software used in this study were unable to handle vanadium, 6rsa was discarded from the set. The test sets were converted to the SD file format, and the formal atomic charges were manually assigned based on the connectivity of the atoms. The combined set of ligands contains 311 compounds. Figure 3 presents the distribution of rotatable bonds for the combined set. For other property distributions of the CCDC/Astex set and the Astex Diverse Set the reader is directed to ref 58.

Balloon (version 0.6.0) was run in a test run operating mode for 300 generations on each of the 311 ligand structures of the test set using an initial population size of 20 conformers. The maximum allowed CPU time for the GA was set to 60 000 s per structure, a value large enough to ensure that even for structures with a high degree of conformational freedom the GA run terminates due to a maximum number of elapsed generations instead of by exceeding a time limit. Because the set of ligands does not contain structures with very large rings, no upper limit was used for the size of rings to be treated as flexible. The default probabilities were used for the genetic operators. The number of maximum allowed initial conjugate gradient geometry optimization steps was set to 1000, and the convergence criterion based on the gradient root-mean-square (rms) was set to 0.1 kcal mol^{-1} $Å^{-1}$. The number of allowed iterations is enough for producing reasonable geometries, although the gradient rms remained above 0.1 in some cases. Downhill simplex minimization was not used in the generation of the template geometry. The allowed number of conjugate gradient iteration steps for the postprocessing phase was set to 100 (the default value).

In order to compare program performance, single conformers were generated for the test set structures using Corina (version 3.2) and conformer ensembles using MacroModel (version 9.5)[59] in the "Serial torsional/Low-mode sampling" mode[60] with "Distinguish enantiomers" enabled, starting from the crystal structure coordinates. Otherwise the default parameters were used for both Corina and MacroModel. The minimum rmsd values between the crystal structure coordinates and the generated conformer models were calculated as described above. Processing times were extracted from the respective log files. Balloon and MacroModel were run on a 2.4 GHz Intel Xeon CPU. Corina was run on an UltraSPARC IV CPU.

3. RESULTS

The obtained processing times for Balloon are shown in Figure 4 as a function of the number of rotatable bonds found in the structure: a quadratic trend is observed (plotted as a solid line in Figure 4). The majority of the structures (0.74%) are processed in less than 200 s.

The GA requires force field potential energy values as a basis for fitness evaluation. The MMFF94 force field is not completely parametrized for all chemical elements, such as boron that occurs in one compound of the test set, the ligand in the PDB entry 1vgc (L-1-(4-chlorophenyl)-2-(acetamido)-ethane boronic acid). No MMFF94 atom type can either be assigned for the atoms in the thiodiimine group in the ligand in 1cps (S-(2-carboxy-3-phenylpropyl)thiodiimine-S-methane) and for the 11-nitrogen in the ligand in 3hvt (11-cyclopropyl-4-methyl-5,11-dihydro-6H-dipyrido[3,2-b:2′,3′-e][1,4]diazepin-6-one). Such atoms are assigned the wildcard atom type that has zero force field parameter values in our implementation of MMFF94. Thus, potential energy calculations on structures that contain these insufficiently parametrized elements or functional groups may result in unrealistic values and distorted geometry. The resulting structures for the ligand in 1vgc indeed have collapsed geometry for the boronic acid group.

The number of conformers retained after the postprocessing step increases with the increasing number of rotatable bonds in the structure as shown in Figure 5 ($R = 0.74$). The number of generated conformers was on average 14 ± 11 over all ensembles when the initial population size N_0 was 20 conformers. The average minimum rmsd was 1.1 ± 0.7 Å over all ensembles. The rmsd of superimposition for the conformer closest to the experimental structure is shown as a function of the number of rotatable bonds in Figure 6. The correlation is approximately linear over the observed range ($R = 0.71$). The distribution of the rmsd values is shown in Figure 7. The majority of the ensembles contain at least one conformer within 2 Å from the bioactive conformation.

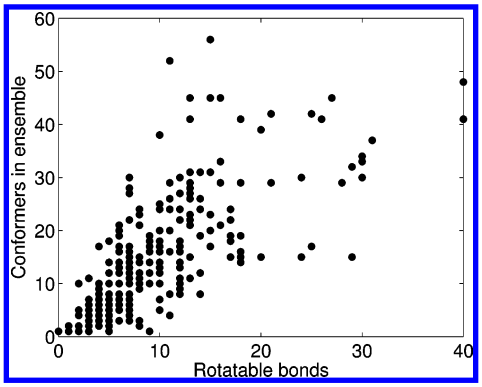

Figure 5. The number of generated conformers per ensemble as a function of the number of rotatable bonds.

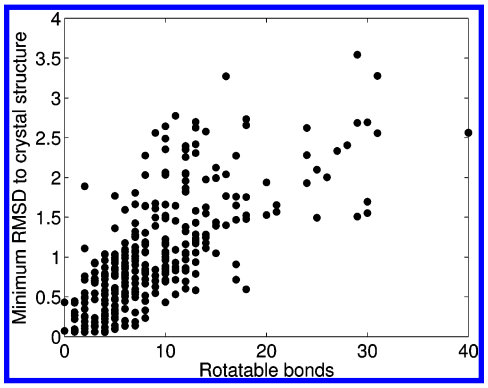

Figure 6. The minimum rmsd [Å] of optimal superimposition of generated conformers and the experimental protein-bound conformation taken from the PDB as a function of the number of rotatable bonds.

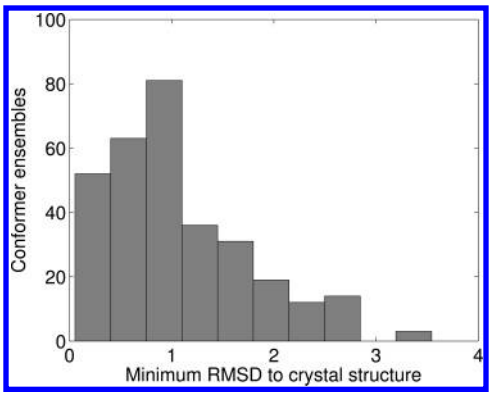

Figure 7. Distribution of the minimum rmsd [Å] of optimal superimposition of the generated conformers and the experimental protein-bound conformation. See also Table 1.

The ligand in 1hos ((2-phenyl-1-carbobenzyloxyvalylamino)ethylphosphinic acid) had the highest rmsd of all the studied structures, 3.54 Å, when the number of rotatable bonds is 29 and the number of generated conformers 32. Another run was made on the ligand in 1hos using the same settings as above but an increased initial population size N_0 of 50 conformers. The population size is plotted in Figure 8 as a function of elapsed generations. The size of the nondominated front exceeds the initial population size at generation 42, after which the population size varies between 50 and 145 conformers depending on the size of the nondominated front. The final population contains 117 conformers of which 15 crowded conformers are discarded, leaving 102 conformers in the output with a minimum rmsd of 2.75 Å to the crystal structure, an improvement of 0.79

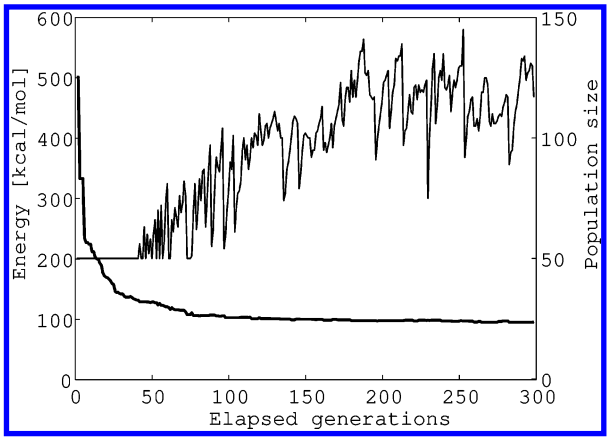

Figure 8. The minimum of the sum of the torsional and van der Waals energies (the objective functions) and the population size as a function of elapsed generations for the rerun of the ligand in 1hos (see text). The minimum energy in the population is represented by the thick line, and the population size is represented by the thin line.

Å over the run using N_0 of 20 conformers. MacroModel generated 607 conformers for the ligand in 1hos with a minimum rmsd of 1.13 Å. The conformer generated by Corina had an rmsd of 4.81 Å with the crystal structure.

The tolerance allowed in the objective value comparisons of the nondominated sorting procedure (see population handling above) might at first feel too loose for any reduction to be achieved in the conformational energies. The energy window will, however, shift toward lower values during the course of the evolutionary cycle because the crossover and mutations produce geometries that happen to be of lower energy than the current lowest, as shown in Figure 8 for the rerun of the ligand in 1hos. The conformer with the absolute minimum energy can be lost if it happens to reside in a highly populated region of the phenotype space, but another very similar conformer will survive. The conjugate gradient energy minimization, done in the postprocessing step, will drive the conformations toward the closest minimum. It is therefore important that the ensemble contains at least one conformer close to each of the relevant low-energy conformations, while preserving the absolute minimum energy conformer in the population is of secondary interest. The average energy (MMFF94 potential excluding the electrostatic term) of the final of conformers for the ligand in 1hos was 133 ± 2 kcal/mol, and the minimum energy was 127.4 kcal/mol. No conformers were discarded in the postprocessing step because of high energy.

The ligand in 1fki ((21S)-4,4-dimethyl-6,19-dioxa-1-azabicyclo[19.4.0]pentacosane-2,3,7,20-tetrone) stands out with an rmsd of 1.11 Å and only two rotatable bonds. The structure, depicted in Figure 9 with the 10 produced conformers, is a polycycle with 21 atoms in the larger ring. A total of 95 flip-of-fragment operators are defined for the structure when no ring size restrictions are applied (see Methods), and therefore the ligand has a total of 97 degrees of structural freedom, which explains the seemingly high rmsd value. The conformer produced by Corina has an rmsd of 2.29 Å with the crystal structure. MacroModel produced 557 conformers with a minimum rmsd of 0.55 Å.

The results above were obtained using an rmsd of 0.5 Å between two conformers as a threshold for crowdedness (d_{crowd}). The value of the threshold can be expected to affect

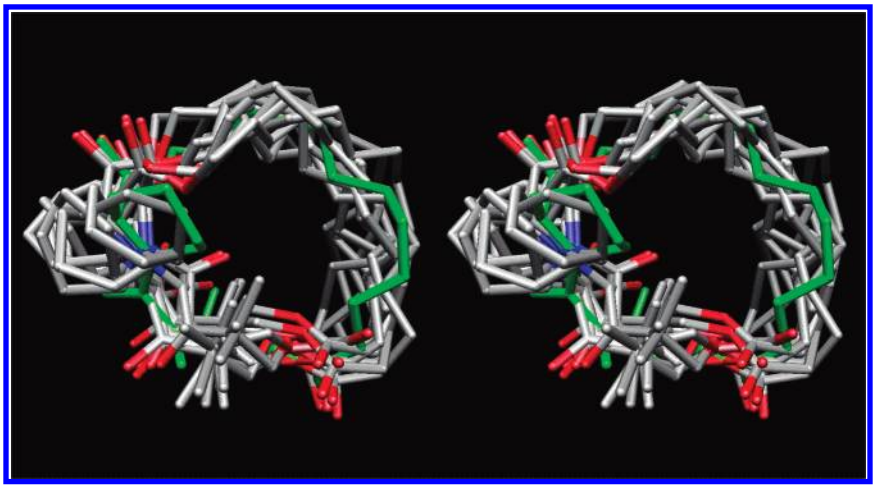

Figure 9. A wall-eyed stereoview of the conformers produced for the ligand in 1fki. The crystal structure conformation is shown in green. Hydrogens are suppressed for clarity. The figure was created using the BODIL molecular modeling environment.[61]

Table 1. Distribution of the Minimum rmsd to Crystal Structure for the Generated Conformer Ensembles[d]

	cumulative percentage of ensembles below rmsd							
	<0.1 Å	<0.5 Å	<1 Å	<1.5 Å	<2 Å	<3 Å	t_{CPU}[a] [s]	NOC[b]
$d_{crowd} = 0.6$ Å	3.2	19.0	58.5	76.5	90.0	98.4	163 ± 194	13 ± 9
$d_{crowd} = 0.5$ Å	3.5	21.9	55.9	76.5	87.8	99.0	172 ± 225	14 ± 11
$d_{crowd} = 0.4$ Å	2.6	23.5	55.6	76.2	90.0	98.4	187 ± 239	15 ± 11
$d_{crowd} = 0.3$ Å	3.2	23.2	56.3	77.5	88.4	98.7	202 ± 270	18 ± 13
$N_0 = 50$	3.9	31.2	65.9	82.3	93.2	99.7	532 ± 965	36 ± 31
$N_{gen} = 600$	3.2	21.9	57.2	79.7	89.7	97.7	229 ± 372	15 ± 1
MacroModel[c]	6.2	41.4	66.4	82.4	89.3	98.7	1612	202 ± 239
Corina	4.5	14.5	33.4	59.2	73.0	92.6	0.019	1

[a] Average CPU time per ensemble. [b] Average number of generated conformers. [c] Processed 307 structures; skipped ligands in 1apt, 1apu, 1cps, and 1vgc. [d] The first colum indicates altered settings for Balloon. Other parameters were kept at the values given in the text.

the resolution at which the GA samples the conformational space and the size of the final population because conformers get discarded for being too close to one another. Table 1 lists the percentages of the generated conformer ensembles binned according to the minimum rmsd to crystal structure for four runs each using a different value for d_{crowd}. Other parameters were as given above. No significant changes were observed in the minimum rmsd to crystal structure over the studied range of d_{crowd} values. The average number of generated conformers and the average CPU time per ensemble increase with decreasing d_{crowd}.

Table 1 also lists statistics for a run with an increased initial population size N_0 and a run with an increased number of allowed generations N_{gen}. The minimum rmsd to crystal structure improves when a larger population is used. Increase in the number of allowed generations has little effect on the accuracy, which indicates that the GA converges on average before 300 generations have elapsed.

4. DISCUSSION

A previous study on commercial conformer generation software showed a correlation between the number of produced conformers and the minimum rmsd of superimposition to the bioactive conformation.[45] The finding simply follows the laws of probability: the more times one rolls the dice, the more probable it is to get any given result. The number of conformers generated in the previous study varied from 30 to 300 models per ensemble, and the corresponding minimum rmsd averaged from 1.115 to 0.941 Å, respectively.[45] In this study, MacroModel produced an average of 202 conformers per structure and achieved an average rmsd of 0.9 ± 0.8 Å. Balloon achieved an average of 1.1 Å on 14 conformers. (For reference, a conformer with an rmsd greater than 2 Å is not regarded as representative of the bioactive conformation in ref 45. For the smallest ligands 2 Å is already a substantial deviation.) The difference in accuracy between Balloon and the commercial software, ~0.2 Å, is small considering the average 20-fold difference in the number of generated conformers. When the initial population size is increased to 50 conformers, Balloon produces ensembles with a minimum rmsd of 0.9 ± 0.6 Å, comparable to commercial software. The single conformers from Corina averaged to 1.5 ± 0.9 Å.

The use of a force field optimization step is a source of error for the prediction of the bioactive geometry: ligands do not always bind to the target protein in a minimum energy conformation as predicted by a force field.[44,62,63] Flexible ligands with hydrophobic groups usually adopt a globular conformation in solution but can "unfold" when in contact with a hydrophobic binding site of a receptor.[44] An example of such a situation is the ligand in 1qbr (3-[[(4R,5S,6S,7R)-5,6-dihydroxy-2-oxo-4,7-bis(phenylmethyl)-3-[[3-(1,3-thiazol-2-ylcarbamoyl)phenyl]methyl]-1,3-diazepan-1-yl]methyl]-N-(1,3-thiazol-2-yl)benzamide), an HIV protease inhibitor, where the generated conformation closest to that in the crystal structure differs by an rmsd of 3.27 Å. The ligand is composed of two thiazolylbenzamides and two benzyl moieties connected symmetrically to a 7-membered cyclic

urea core.[64] In the generated conformers the phenyl moeities of the thiazolylbenzamides and benzyl groups tend to pack against each other as would be expected for the hydrophobic groups in the solvated state. The distance between the thiazoles range from 10 to 17 Å. In the protein-bound conformation the thiazole rings of the residues are far apart (20 Å), while the phenyl rings are accommodated by the hydrophobic pockets of the enzyme.[64] Because of the unfolding of hydrophobic ligands upon binding, a conformer generator utilizing the van der Waals potential of a force field cannot be expected to find the bioactive conformation in cases such as the ligand in 1qbr without the incorporation of the receptor structure into the conformational analysis (known as the molecular docking method) or tweaking the form of the potential function. The latter modification may be difficult to justify on a physical basis.

The generated conformer ensemble is usually input to a downstream software tool that might have substantial runtime requirements per each input structure. It is therefore desirable to restrict the number of produced conformers to some threshold value that is large enough to capture (some of) the flexibility of the structure but still small enough to allow the downstream tool to be used on the ensemble. Because the size of the search space increases significantly upon increasing structural complexity, the population size should increase accordingly in order to ensure sufficient sampling. Mekenyan et al. scaled the population size in their GA according to the flexibility of the compound.[19] Balloon uses a constant initial population size, but the population is allowed to grow in order to accommodate the nondominated conformers, and the tolerance for domination is adjusted by the flexibility of the compound (eq 3), which achieves the same goal as scaling the population size explicitly. The number of generated conformers for Balloon, seen in Figure 5, does not increase as heavily with the number of rotatable bonds in the structure as for the commercial software reported in Figure 2 of ref 45 or for MacroModel as observed in this study. With Balloon the parameters of eq 3 can be adjusted in order to achieve larger ensembles at a lower number of rotatable bonds than with the settings used in this study.

The use of an initial population size of 20 conformers is reflected in the results: the minimum rmsd tends to exceed the 2.0 Å limit when the number of rotatable bonds approaches 20 (Figure 6), and simultaneously the number of generated conformers exceeds the initial population size (Figure 5). The rmsd improves upon using a larger population size (Table 1), in line with earlier studies. An investigation on the optimal initial population size and other control parameters for the implemented GA is a combinatorial optimization problem and beyond the scope of this paper but a worthwhile future direction to take. The work of Djurdjevic et al.[9] provides an excellent basis and point of reference for such a study.

Geometry optimization is known to be the rate-limiting step in stochastic conformer analysis algorithms.[52] The performance of Balloon is dependent on the performance of the used force field, both time-wise and with regard to the quality of produced geometries. According to our experience, the absolute scale of the CPU time used per ensemble depends largely on the number of allowed iteration steps and the strictness of the termination criteria for strain relaxation in the postprocessing step. While the processing time can probably be reduced by further optimizations of the source code, a stochastic conformer generation method will hardly ever be as fast as a deterministic construction method because of the need of postoptimization of the generated structures. Despite being run on different processor types, the timing results in Table 1 indicate that Corina is by far the fastest program of the three, using on average 19 ms per conformer, whereas Balloon uses 12.3 s and MacroModel 8.1 s per conformer in the final ensemble.

The quality of the produced geometries depends on the level of theory used in the energy calculations. Although druglike molecules seldom contain boron or other "exotic" elements, the force field should also be able to produce reasonable geometries for systems that do not fall into the category of druglike molecules. The MMFF94-like force field used in Balloon is fairly general but not complete in terms of the covered chemistry. Mekenyan et al. used semiempirical ab initio methods for post-GA structure optimization in their conformer ensemble generator.[19] While quantum chemical methods produce accurate geometries, the computational cost is high. Balloon does not presently make use of parallel processing on multiple CPUs, which is needed for conformational expansion of large virtual molecular libraries within a reasonable time frame especially when quantum chemical energy calculations are involved. Therefore, the use of a force field is a compromise between speed and accuracy (and generality to some extent): force fields provide sufficient precision for most molecular modeling and computational drug-design applications.

5. CONCLUSIONS

A multiobjective GA for generating ensembles of molecular conformers was designed and implemented. The method combines elements of published GAs for conformer generation and includes modifications and additions that expand the conformational space available for sampling. The goal was to design a method that can produce low-energy conformers that are geometrically distinct from each other. Because an average of 14 conformers passed the rmsd filter applied in the postprocessing step, we can state that a reasonable degree of geometric diversity is preserved. Since the conformers also passed the low-energy filter, it is evident that the algorithm achieves the result it was designed to obtain.

ACKNOWLEDGMENT

We thank Susanna Repo and J. Santeri Puranen for their critical reading of this manuscript. The Academy of Finland and Sigrid Jusélius Foundation are acknowledged for their financial support. The Structural Bioinformatics Laboratory belongs to the Center of Excellence in Cell Stress of Åbo Akademi University.

REFERENCES AND NOTES

(1) Kontoyianni, M.; McClellan, L.; Sokol, G. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(2) *Pharmacophore Perception, Development, and Use in Drug Design*, Vol. 2 of *IUL Biotechnology Series*; Güner, O. F., Ed.; International University Line: La Jolla, CA, U.S.A., 2000.

(3) *Pharmacophores and Pharmacophore Searches*, Vol. 32 of *Methods and Principles in Medicinal Chemistry*; Langer, T., Hoffmann, R. D., Eds.; Wiley-VHC Verlag GmbH & Co. KGaA: Weinheim, Germany, 2006.
(4) Mekenyan, O.; Nikolova, N.; Schmieder, P.; Veith, G. COREPA-M: A Multi-Dimensional Formulation of COREPA. *QSAR Comb. Sci.* **2004**, *23*, 5−18.
(5) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526−1539.
(6) Vainio, M. J.; Johnson, M. S. McQSAR: A Multiconformational Quantitative Structure−Activity Relationship Engine Driven by Genetic Algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 1953−1961.
(7) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
(8) Fonseca, C. M.; Fleming, P. J. Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. In *Genetic Algorithms: Proceedings of the Fifth International Conference*; Forrest, S., Ed.; Morgan Kaufmann: 1993; pp 416−423.
(9) Djurdjevic, D. P.; Biggs, M. J. Ab Initio Protein Fold Prediction Using Evolutionary Algorithms: Influence of Design and Control Parameters on Performance. *J. Comput. Chem.* **2006**, *27*, 1177−1195.
(10) Blommers, M. J. J.; 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.
(11) Payne, A. W. R.; Glen, R. C. Molecular Recognition Using a Binary Genetic Search Algorithm. *J. Mol. Graphics* **1993**, *11*, 74−91.
(12) McGarrah, D.; Judson, R. Analysis of the Genetic Algorithm Method of Molecular Conformation Determination. *J. Comput. Chem.* **1993**, *14*, 1385−1395.
(13) Judson, R.; Jaeger, E.; Treasurywala, A.; Peterson, M. Conformational Searching Methods for Small Molecules. II. Genetic Algorithm Approach. *J. Comput. Chem.* **1993**, *14*, 1407−1414.
(14) Clark, D.; Jones, G.; Willet, P.; Kenny, P.; Glen, R. 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.
(15) Glen, R. C.; Payne, A. W. R. A Genetic Algorithm for the Automated Generation of Molecules within Constraints. *J. Comput.-Aided Mol. Des.* **1995**, *V9*, 181−202.
(16) Nair, N.; Goodman, J. Genetic Algorithms in Conformational Analysis. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 317−320.
(17) Keser, M.; Stupp, S. I. A Genetic Algorithm for Conformational Search of Organic Molecules: Implications for Materials Chemistry. *Comput. Chem.* **1998**, *22*, 345−351.
(18) Mekenyan, O.; Dimitrov, D.; Nikolova, N.; Karabunarliev, S. Conformational Coverage by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 997−1016.
(19) Mekenyan, O.; Pavlov, T.; Grancharov, V.; Todorov, M.; Schmieder, P.; Veith, G. 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm. *J. Chem. Inf. Model.* **2005**, *45*, 283−292.
(20) Parent, B.; Kokosy, A.; Horvath, D. Optimized Evolutionary Strategies in Conformational Sampling. *Soft Comput.* **2007**, *11*, 63−79.
(21) Lameijer, E.-W.; Bäck, T.; Kok, J. N.; Ijzerman, A. P. Evolutionary Algorithms in Drug Design. *Nat. Comput.* **2005**, *4*, 177−243.
(22) Cottrell, S. J.; Gillet, V. J.; Taylor, R.; Wilton, D. J. Generation of Multiple Pharmacophore Hypotheses Using Multiobjective Optimisation Techniques. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 665−682.
(23) Cottrell, S. J.; Gillet, V. J.; Taylor, R. Incorporating partial Matches within Multiobjective Pharmacophore Identification. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 735−749.
(24) Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. S. Multiobjective Optimization in Quantitative Structure-Activity Relationships: Deriving Accurate and Interpretable QSARs. *J. Med. Chem.* **2002**, *45*, 5069−5080.
(25) Cutello, V.; Narzisi, G.; Nicosia, G. A Multi-Objective Evolutionary Approach to the Protein Structure Prediction Problem. *J. R. Soc. Interface* **2006**, *3*, 139−151.
(26) Horn, J.; Nafpliotis, N.; Goldberg, D. E. A Niched Pareto Genetic Algorithm for Multiobjective Optimization. In *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*; IEEE Service Center: Piscataway, NJ, 1994; Vol. 1, pp 82−87.
(27) Erickson, M.; Mayer, A.; Horn, J. The Niched Pareto Genetic Algorithm 2 Applied to the Design of Groundwater Remediation Systems. In *EMO '01: Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization*; Springer-Verlag: London, U.K., 2001; pp 681−695.
(28) Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE T. Evolut. Comput.* **2002**, *6*, 182−197.
(29) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(30) Rassokhin, D. N.; Agrafiotis, D. K. A Modified Update Rule for Stochastic Proximity Embedding. *J. Mol. Graphics Modell.* **2003**, *22*, 133−140.
(31) Xu, H.; Izrailev, S.; Agrafiotis, D. Conformational Sampling by Self-Organization. *J. Chem. Inf. Model.* **2003**, *43*, 1186−1191.
(32) Kuszewski, J.; Nilges, M.; Brünger, A. T. Sampling and Efficiency of Metric Matrix Distance Geometry: A Novel Partial Metrization Algorithm. *J. Biomol. NMR* **1992**, *2*, 33−56.
(33) Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. Conformational Analysis Using Distance Geometry Methods. *J. Mol. Graphics Modell.* **1997**, *15*, 18−36.
(34) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C: the Art of Scientific Computing*, 2nd ed.; Cambridge University Press: 1992.
(35) Halgren, T. A. Merck Molecular Force Field. 1. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(36) Nelder, J.; Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **1965**, *7*, 308−313.
(37) Carta, G.; Onnis, V.; Knox, A. J. S.; Fayne, D.; Lloyd, D. G. Permuting Input for More Effective Sampling of 3D Conformer Space. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 179−190.
(38) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(39) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to 3-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(40) Choi, V. On Updating Torsion Angles of Molecular Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 438−444.
(41) Strizhev, A.; Abrahamian, E.; Choi, S.; Leonard, J.; Wolohan, P.; Clark, R. The Effects of Biasing Torsional Mutations in a Conformational GA. *J. Chem. Inf. Model.* **2006**, *46*, 1862−1870.
(42) Syswerda, G. In *Handbook of Genetic Algorithms*; Davis, L., Ed.; Van Nostrand Reinhold Co.: 1991; Chapter Schedule Optimization using Genetic Algorithms, pp 332−349.
(43) Boström, J. Reproducing the Conformations of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.
(44) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.
(45) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848−1861.
(46) Kearsley, S. K. Structural Comparisons Using Restrained Inhomogeneous Transformations. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1989**, *45*, 628−635.
(47) Krissinel, E. B.; Henrick, K. Common Subgraph Isomorphism Detection by Backtracking Search. *Software Pract. Exper.* **2004**, *34*, 591−607.
(48) Zitzler, E.; Deb, K.; Thiele, L. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evol. Comput.* **2000**, *8*, 173−195.
(49) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(50) Ricketts, E. M.; Bradshaw, J.; Hann, M.; Hayes, F.; Tanna, N.; Ricketts, D. M. Comparison of Conformations of Small-Molecule Structures from the Protein Data-Bank with Those Generated by Concord, Cobra, ChemDBS-3D, and Converter and Those Extracted from the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 905−925.
(51) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of OMEGA with Respect to Retrieving Bioactive Conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449−462.
(52) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational Analysis by Intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10−20.
(53) Good, A. C.; Cheney, D. L. Analysis and Optimization of Structure-Based Virtual Screening Protocols (1): Exploration of Ligand Conformational Sampling Techniques. *J. Mol. Graphics Modell.* **2003**, *22*, 23−30.
(54) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A Distance Geometry Heuristic for Expanding the Range of Geometries Sampled During Conformational Search. *J. Comput. Chem.* **2006**, *27*, 1962−1969.
(55) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380−388.

(56) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic 3-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(57) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein-Ligand Interaction. *Proteins* **2002**, *49*, 457−471.
(58) Hartshorn, M.; Verdonk, M.; Chessari, G.; Brewerton, S.; Mooij, W.; Mortenson, P.; Murray, C. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(59) *MacroModel*, *version 9.5*; Schrödinger, LLC: New York, 2007.
(60) Kolossváry, I.; Guida, W. C. Low-Mode Conformational Search Elucidated: Application to $C_{39}H_{80}$ and Flexible Docking of 9-Deazaguanine Inhibitors into PNP. *J. Comput. Chem.* **1999**, *20*, 1671−1684.
(61) Lehtonen, J. V.; Still, D.-J.; Rantanen, V.-V.; Ekholm, J.; Björklund, D.; Iftikhar, Z.; Huhtala, M.; Repo, S.; Jussila, A.; Jaakkola, J.; Pentikäinen, O.; Nyrönen, T.; Salminen, T.; Gyllenberg, M.; Johnson, M. S. BODIL: A Molecular Modeling Environment for Structure-Function Analysis and Drug Design. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 401−419.
(62) Boström, J.; Norrby, P.-O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383−396.
(63) Tirado-Rives, J.; Jorgensen, W. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein-Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880−5884.
(64) Jadhav, P. K.; Ala, P.; Woerner, F. J.; Chang, C. H.; Garber, S. S.; Anton, E. D.; Bacheler, L. T. Cyclic Urea Amides: HIV-1 Protease Inhibitors with Low Nanomolar Potency against Both Wild Type and Protease Inhibitor Resistant Mutants of HIV. *J. Med. Chem.* **1997**, *40*, 181−191.

CI6005646