A New Approach to Design Virtual Combinatorial Library with Genetic Algorithm Based on 3D Grid Property

DongXiang Liu, HuaLiang Jiang, KaiXian Chen,* and RuYun Ji

Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 294 TaiYuan Road,
Shanghai 200031, P. R. China

Received January 2, 1997

Combinatorial chemistry is a type of synthetic strategy which systematically connects a set of "building blocks" of varying structures to yield a large array of diverse molecular entities. Since the molecular diversity of combinatorial library used for the two stages of drug discovery process which are exploring lead compounds and performing structural modification on lead compounds is different, choosing appropriate fragments as the building blocks becomes a critical step for combinatorial chemistry. The number of available chemical fragments is often so great that it is impossible to use all of the fragments in one synthetic experiment. An effective method is much needed to select building blocks from the enormous chemical fragments to synthesize a combinatorial library with desired molecular diversity. In this article, we proposed a novel idea to define molecular diversity based on three-dimensional (3D) grid force field properties including electric potential and steric potential parameters and demonstrated how to apply genetic algorithm to choose a subset of fragment for the construction of a virtual combinatorial library. In the first example, we tried to construct a virtual library for benzodiazepine derivatives with maximum molecular diversity. In the second example, a virtual library is constructed for structural modification of (−)-huperzine A based on the 3D grid properties. By analyzing fragments frequency in the target library, we can understand which fragments are appropriate as building blocks. Furthermore, we can predict the biological activity of library compounds with 3D-QSAR model or by evaluating the binding energy of library compounds with the target receptors sites.

INTRODUCTION

The experimental process of drug discovery has been influenced by the development of combinatorial chemistry in recent years. Some researchers have successfully applied combinatorial chemistry technique to discover new active compounds as HIV inhibitors[1] and μ-opioid receptor agonists.[2] The main difference between combinatorial chemistry and traditional synthetic strategy is that using the former method thousands or even billions compounds can be obtained in a reaction vessel simultaneously, while with traditional methods compounds can be only synthesized one by one. Moreover, library compounds can be also assayed directly with some appropriate pharmaceutical methodology without any further separation process. It is apparent that the advantages of combinatorial chemistry arise from its parallel synthesis and high throughput screening method. For a long time, research on combinatorial chemistry has been focused on developing more effective synthesis strategy and more accurate pharmaceutical assay methods. It is now generally accepted that a successful combinatorial chemistry work can not only construct a combinatorial library which contains a large amount of diverse structures but also can guarantee appropriate molecular diversity in the library.

Combinatorial chemistry technique can be used in the two different stages of drug discovery process: finding lead compounds and structural modification on lead compounds. However, the requirement for molecular diversity of combinatorial library is different for these two stages. If the structure specificity of expected bioactive compound and the active sites of its target receptors is yet unknown, chemical structures in the library should be mostly diverse from each other in order to find a lead compound. Otherwise, if we will perform structural modification on lead compounds, chemical structures in the designed library should be some similar with the specific lead compound. Such ideas imply that choosing building blocks is very important for constructing a combinatorial library with desired molecular diversity.

Since the purpose of constructing a virtual library is to ensure molecular diversity, the final library compounds should well distribute in the universe of organic chemical structures or in some region around the lead compound. However, the number of chemical fragments that can be used as building blocks is usually so big that it is impractical to combine every fragment at each position of the "core" molecules one at a time. Namely, to construct a virtual combinatorial library with desired molecular diversity, an effective combinatorial optimization algorithm must be applied to explore the chemical structure space trying to identify the subspace which will fit the molecular diversity criteria. Simulated annealing and genetic algorithm are the two kinds of very useful combinatorial optimization methods. Some researchers have already used simulated annealing and topological index based molecular similarity metrics to design target combinatorial chemical library.[3] Sheridan and Kearsley used genetic algorithm to suggest combinatorial libraries.[4] Gillet et al. have also proposed ways to generate structurally diverse libraries and put upper and lower limits on the diversity for a library with genetic algorithm.[5]

S0095-2338(97)00086-3 CCC: $15.00
Published on Web 03/04/1998

Considering the figurative evolution mechanism and flexible genetic protocols, we chose genetic algorithm to construct virtual combinatorial libraries based on 3D grid property.

In this article, we wish to present our study to construct a virtual combinatorial library for broad screening and for structure modification of lead compounds, representively. The first example is to design a benzodiazepine derivatives library with maximum molecular diversity. The second example is to select fragments as building blocks for constructing combinatorial library of (−)-huperzine A, which is a potent reversible inhibitor of acetylcholinesterase (AChE) that lacks potentially complicating muscarinic effects.[6] Though it seems currently impossible to synthesize a (−)-huperzine combinatorial library with a combinatorial organic synthetic (COS) method, the simulation result would be very useful for traditional structural modification work on (−)-huperzine A. The most important aspect of our idea is that it is practical for all organic compounds also including those compounds available by combinatorial synthetic methodology.

METHODS

Determination of active conformation of drug compounds is very important in conventional 3D-QSAR analysis. The direct way to determine the active conformation is to resolve the crystal structure of drug complex with the receptor. But in many cases, we are not clear about the drug's target and cannot obtain the receptors in crystalline formation, so we usually apply a lot of calculation methods such as conformational search to determine the active conformation or take the ground-state conformation as the active conformation of the compound. Because our molecular diversity is based on the 3D force field property including the electric potential and steric potential properties, determination of active conformation is also very critical to our work. Four different ways of selecting the compound conformation are proposed, which are picking one at random, averaging a number of compound conformations, boltzmann weight a number of conformations and using a 3D template. In this paper, we used a 3D template as the active conformation.

Genetic Algorithm for Combinatorial Library. Genetic algorithm is a kind of stochastic optimization method that has been inspired by Darwinian evolutionary principles.[7] The distinctive aspect of genetic algorithm is that it investigates many possible solutions simultaneously, each of which explores different regions in the chemical structure universe. The first step of genetic algorithm is to create a chromosome pool of *N* individuals. Each individual encodes the same number of genome, which represents a fragment. The fitness score of each individual in this generation is determined by a user-specified fitness function. The fitness function can be varied depending on the purpose of our constructing the virtual combinatorial library. If the virtual combinatorial library will be used for exploring lead compounds aimed at a specific bioactivity, the fitness function can be defined as the molecular dissimilarity. Otherwise, if our purpose is to do structural modification on a lead compound, the fitness function should be able not only to identify the molecular similarity of library compounds with the lead compound but also to describe the molecular diversity of the combinatorial library. The next step is the reproduction process. Two individuals are selected probabilistically on the basis of their fitness scores and serve as parents. Three kinds of reproduction mechanisms, inheritance, mutation and crossover, are performed on the parent compounds to generate a new library from the old one. Next, the fitness score of each member of the new generated chromosome pool is evaluated, and the reproduction cycles continue until the target fitness score is achieved.

All chromosomes in the pool will be scored by a fitness function, and their surviving probability into the next generation depends on their fitness score. The intrinsic random property involved in the three genetic manipulations ensures that "eugenic" solutions will be inherited in the next generation. Maybe the final combinatorial library is not always guaranteed to have the maximum molecular diversity, but for most large chemical structure universe it is sure that the better combinatorial library will be obtained much faster with a well organized genetic algorithm than system search methodology.

Generation of Virtual Compounds. Our current program is designed for organic compound library including peptoid and small organic molecules. The substituent sites of core molecules on which fragments can be attached are connected with a putative atom ("Du" atom). The number of the Du atoms on the core molecules is not limited. Also there is only one putative atom on the fragments which labels the position where a covalent bond can be formed with the core molecule. The program randomly reads in a fragment and attaches it to a Du atom labeled site of the core molecule. When all Du atoms of core molecule have been replaced by fragments, a virtual compound is "synthesized" and is later saved in Sybyl MOL2 format. Then the structure of virtual compounds will be optimized with molecular dynamics (MD) method and molecular mechanics (MM2) method in Sybyl 6.1 environment. In the optimization process, any geometry constraints which are beneficial to keep the active conformation of library compounds can be applied. Afterward, the optimized structure of virtual compounds is aligned onto the template structure and put into a database which represents a virtual combinatorial library.

During the generation of virtual compounds, it is necessary to trace the fragments information of each library compound in case two of the same compounds would exist in a library or a newly generated compound which is the parent library compound member would be also included into the new combinatorial library. We set the size of a virtual library to be 100. When 100 compounds have been generated into the database, a virtual combinatorial library has been constructed.

Molecular Dissimilarity. After we have constructed a virtual combinatorial library, the next step is to analyze molecular diversity of the library and molecular dissimilarity of library compounds. In principle, chemical structures can be characterized by various physicochemical parameters including partition coefficient, molar refractivity, and molecular volume as well as quantum mechanical quantities such as HOMO and LUMO energies and electrostatic potentials. Topological and topographical descriptors based on 2D and 3D representation of chemical structures can also be used to calculate molecular dissimilarity of compounds and molecular diversity of combinatorial library. For the consideration of interaction mechanism between ligand and receptors, we have developed a novel method to define molecular dissimilarity and molecular diversity.

As we know, Comparative Molecular Field Analysis (CoMFA) is a useful QSAR approach whose models have shown unprecedented accuracy in prediction.[8] In CoMFA, molecules are represented and compared by their steric and electrostatic fields sampled at the intersections of one or more lattices (or grids or boxes) spanning a three-dimensional region. Thus each CoMFA descriptor contains the magnitudes of either the steric or electrostatic field exerted by the atoms in the tabulated molecules on a probe atom located at a point in Cardesian space.

There are a lot of atom probes: H_2O, Carbonyl-O, Aromat.C, NH^{4+}, and Csp^3. These probes can be considered as a representative selection among the variety of the main interaction modes with amino acids in order to consider possible interactions of the molecule with a putative receptor. In our study, we selected Csp^3 with +1 charge as the probe atom. The interaction energy matrix (i.e., constituted by CoMFA descriptors) obtained were submitted to PCA (SYBYL QSAR factor analysis options: no factor rotation; Nipals factor analysis algorithm; minimum σ value: 0.00; same weight for all columns: no scaling). We used 28 principle components which were extracted from CoMFA matrix to describe the molecular dissimilarity

$$\mathrm{MS}_{\mathrm{ij}} = \sqrt{\sum_{k=1}^{m} (\mathrm{FFP}_{\mathrm{i}k} - \mathrm{FFP}_{\mathrm{j}k})^2} \qquad (1)$$

where $\mathrm{MS}_{\mathrm{ij}}$ is the molecular dissimilarity between compound i and compound j, m is the principal components number, and $FFP_{\mathrm{i},k}$ and $FFP_{\mathrm{j},k}$ are the kth principal component of compound i and compound j from the CoMFA matrix, respectively. The bigger the $\mathrm{MS}_{\mathrm{ij}}$ is, the more different compound i will be from compound j. If $\mathrm{MS}_{\mathrm{ij}}$ equals to zero, it means that compound i is the same as compound j.

According to formula 1, we can define the molecular dissimilarity between compound i and compound j. However, it is not adequate except that we can define the overall molecular dissimilarity of one compound with other compounds in the library. One way that we used to define the overall molecular dissimilarity of one compound with the other compounds in the library is to calculate the average dissimilarity values of this compound over other library compounds

$$\mathrm{MS}_{\mathrm{i}} = \frac{\sum_{j=1}^{n} \mathrm{MS}_{\mathrm{ij}}}{n} \qquad (2)$$

where MS_{i} is the overall molecular dissimilarity of compound i with other compounds in the library, n is the number of compounds in the library. Similar to $\mathrm{MS}_{\mathrm{ij},}$ molecule with large MS_{i} is more diverse from the other molecules in the library.

Molecular Diversity of Library. Molecular diversity is an important index to estimate the quality of a combinatorial library. The value of molecular diversity is closely related with every molecule in the library. From formula 2, we can see that even if the CoMFA descriptors of one compound in the library have little variation, MS_{i} of all library compounds would also be different. This implies that MS_{i} can be used to define the molecular diversity of a combinatorial library, the calculation formula of molecular diversity for a library as eq 3

$$\mathrm{MD} = \frac{\sum_{i=1}^{n} \mathrm{MS}_{\mathrm{i}}}{n} \qquad (3)$$

where n is the number of library compounds.

If we can design a combinatorial library of which MS_{i} of each library compound locates in the top region, i.e., each library compound is as more different from the other library compounds as possible under particular constraints such as the common scaffold being similar to the lead compound, etc., molecular diversity of the combinatorial library should be the maximum. The library compounds would well distribute in the universe of organic chemical molecule.

Scoring Library Compounds. In genetic algorithm, fitness function is used to score each individual in the pool. Higher score means that the compound has more chances to be inherited into the descendant generation. The molecular diversity of the combinatorial library required for broad screening and structural modification of lead compounds is a little different. In the first case, the principle of designing virtual library is to maximize the molecular dissimilarity. If library compounds are most diverse, the library would be the best. In the second case, our purpose is to modify the structure of lead compounds to improve its bioactivity or decrease its side effect and toxicity, the structure specificity of designed library compounds should have some similarity with the original lead compounds which means that not only the molecular diversity of combinatorial library but also the molecular similarity of library compounds with the lead compound should be considered. Because of these reasons, fitness functions corresponding to the above two cases should be constructed differently.

For the first case, we use MS_{i} to evaluate the fitness score of library compounds

$$\mathrm{Score}_{\mathrm{i}} = \mathrm{MS}_{\mathrm{i}} \qquad (4)$$

For the second case, we use the ratio of MS_{i} with $\mathrm{MS}_{\mathrm{it}}$ to evaluate the fitness score of library compounds (as formula 5)

$$\mathrm{Score}_{\mathrm{i}} = \frac{\mathrm{MS}_{\mathrm{i}}}{\mathrm{MS}_{\mathrm{it}}} \qquad (5)$$

where $\mathrm{MS}_{\mathrm{it}}$ is the molecular dissimilarity of library compounds with the lead compound, t labels the lead compound. From formula 5, we can see that larger $\mathrm{Score}_{\mathrm{i}}$ comes from larger MS_{i} and smaller $\mathrm{MS}_{\mathrm{it}}$. Virtual compounds only with larger MS_{i} or only with smaller $\mathrm{MS}_{\mathrm{it}}$ are not sure to be inherited into the next generation with high probability. Only those compounds which are not only very similar with the lead compound but also very different from the other library compounds would have a high score. Using such fitness function, we can construct a virtual combinatorial library satisfying the two conditions: large molecular diversity and large molecular similarity with the lead compound.

Implementation of Genetic Algorithm. We encoded the genetic algorithm with Sybyl Program Language (SPL) [9] and

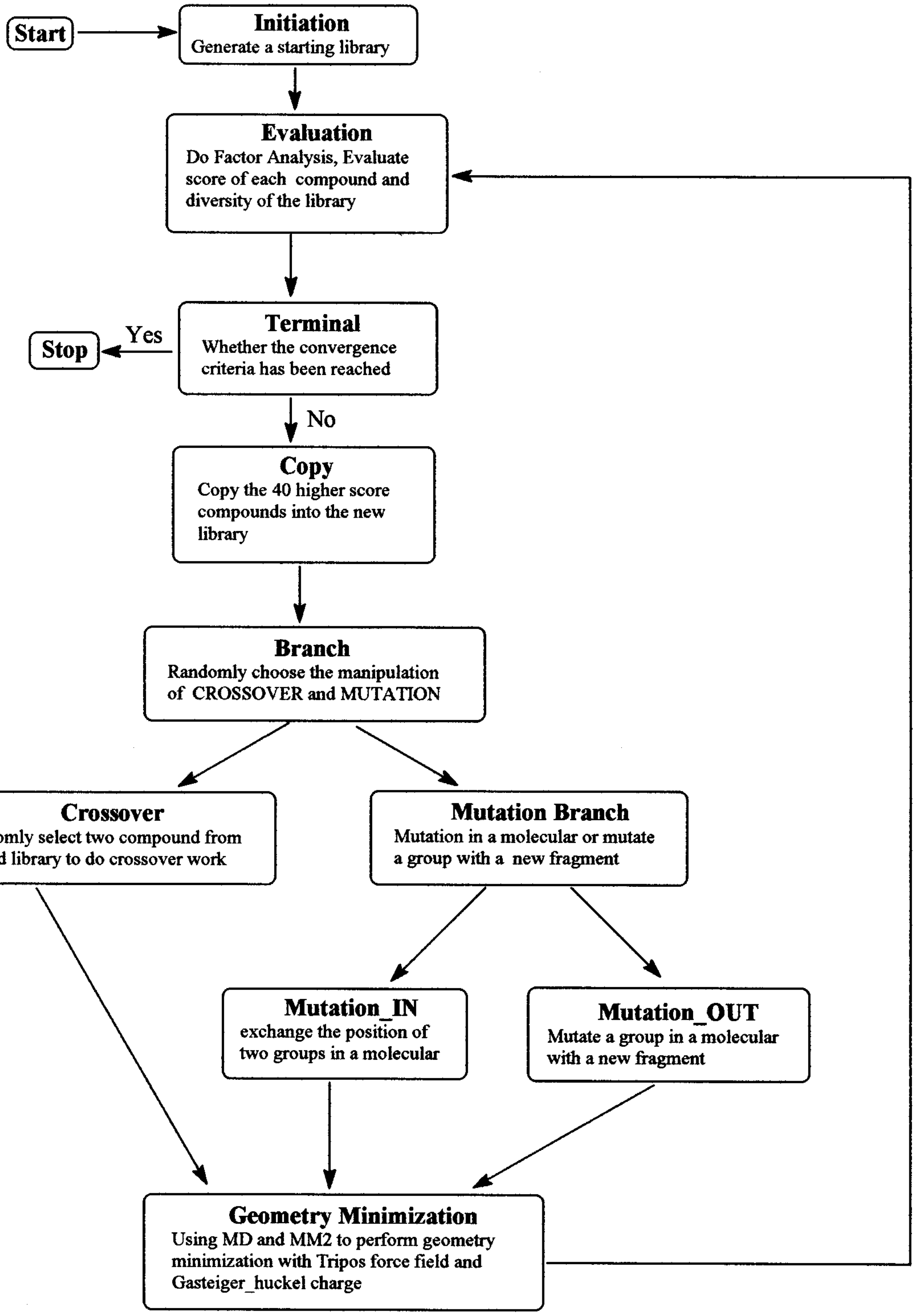
Start
Initiation
Generate a starting library
Evaluation
Do Factor Analysis, Evaluate score of each compound and diversity of the library
Terminal
Whether the convergence criteria has been reached
Yes
Stop
No
Copy
Copy the 40 higher score compounds into the new library
Branch
Randomly choose the manipulation of CROSSOVER and MUTATION
Crossover
Randomly select two compound from the old library to do crossover work
Mutation Branch
Mutation in a molecular or mutate a group with a new fragment
Mutation_IN
exchange the position of two groups in a molecular
Mutation_OUT
Mutate a group in a molecular with a new fragment
Geometry Minimization
Using MD and MM2 to perform geometry minimization with Tripos force field and Gasteiger_huckel charge

Figure 1. Schematic illustration of constructing target virtual combinatorial library with a genetic algorithm.

C++ language. The program scheme is represented in Figure 1. The population size and maximum cycles of genetic generation are read from a control file. A database file named “fragment.mdb” contains the fragments with Du atom labeling the connecting site. And the virtual combinatorial library is represented as a database file.

First, we generate an initial population of molecules. All library compounds are aligned onto a given template compound. To locate the connecting sites of the core molecule, we set up an array buffer to record the atom ID which is connected to Du atom on the core molecule. By this array, it is convenient to split the combined fragments from library compounds and perform crossover and mutation manipulation. Meanwhile, we also set up another array buffer to record fragments in each virtual compounds avoiding duplicate compounds in the same library. When a new library is generated, the information in this array buffer will be written into a trace file “trace.dat” for analysis of fragment frequency. Next, we used an appropriate fitness function to evaluate the score of each library compound. If the maximum cycles of genetic generation is reached, the genetic generation will stop. Otherwise, a new library will be generated from the old one by three genetic manipulations which are demonstrated in Figure 2.

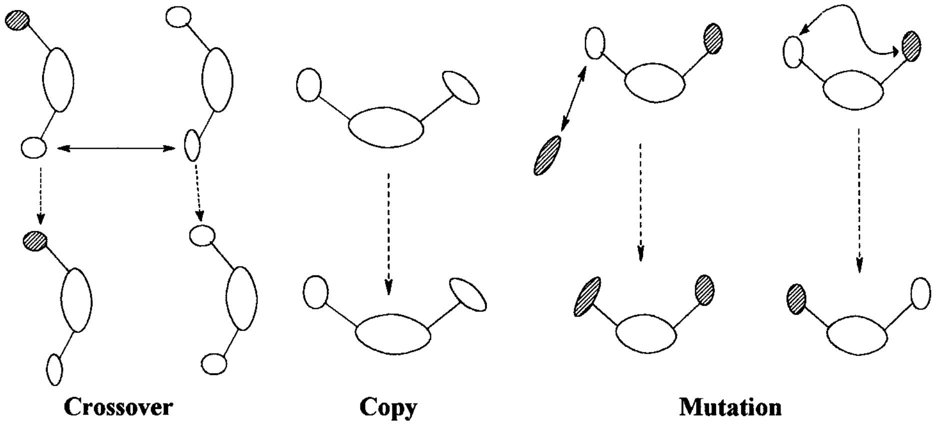

Figure 2. Three kinds of genetic manipulation: crossover, copy, and mutation.

The first step of generation is to copy the 40 compounds of higher score into the new library intactly. Then, crossover or mutation is randomly selected according to a Gaussian function which is driven by the computer random number generator. Two different compounds are selected from the old library randomly as the parents. If crossover manipulation is chosen, fragments linked onto the randomly selected connection site of the core molecule will be exchanged between the two parent compounds to produce two descendant compounds. If mutation manipulation is chosen, two alternative protocols for mutation will be further decided. The first is Mutation_In which will exchange two fragments at two different connection sites of a molecule. The second is Mutation_Out which will replace a randomly selected fragment in the parent compound with a new fragment also randomly selected from the fragment database. After new compounds are generated, molecular dynamics and molecular mechanics method will be used to optimize the structures with geometry constraints if necessary. Finally, these new compounds will be aligned onto the template structure. After the new library is prepared, we analyze the CoMFA matrix of library compounds with Nipals algorithm in Sybyl QSAR module to extract 28 principal components for scoring library compounds and identifying molecular diversity of the virtual combinatorial library. Scores of library compounds and molecular diversity of library are traced by another data file "score.dat".

To ensure the molecular diversity of combinatorial library, we have taken two effective means. One is to avoid duplicate compounds in the same library. Another is to take molecular diversity into account when constructing the fitness function as formulas 4 and 5. As long as enough genetic generation has been performed, the final virtual library will have the maximum molecular diversity.

Of course, there is no single standard method by which a genetic algorithm generates a new population from the previous one based on the fitness scores. It is problem_related. Genetic manipulation can surely be performed on the several best compounds to generate the new library. However, considering the molecular diversity, we included all library compounds for genetic manipulation.

Analysis of the Top Candidates. When the virtual combinatorial library has been constructed, we should further analyze the fragments frequency in the library to determine which fragments would be appropriate for building blocks. We can either use the final virtual library or include those libraries at the equilibrium state of genetic generation to do frequency analysis. High frequent fragments mean that they are important for the molecular diversity. And they should be considered as building blocks for experimental combinatorial synthesis. However, it is not to say that only these fragments could be applied as building blocks. On the contrary, other fragments possessing the necessary atoms or groups for the bioactivity should also be included in the building block set even though they may have low frequency in the analysis. This is because our main purpose of constructing the virtual combinatorial library is to guarantee the molecular diversity of the library with consideration of molecular dissimilarity (or molecular similarity).

According to the maximum score of library compounds traced in "score.dat" data file, we can determine that at which step the genetic generation reaches the equilibrium state. The fragments frequency involved in the final virtual library or those libraries at the equilibrium state can be calculated by the following formula

$$f_i = \frac{N_i}{\sum_{t=1}^{m} N_t} \qquad (6)$$

where N_i and m are the number of occurrences for fragment i in the library (or libraries) compounds and the number of fragments in the fragment set, respectively.

Further, library compounds could be screened by QASR model and/or its target receptors. If the conformation of the template structure on which the library compounds are aligned is in consistence with that of compounds for 3D-QSAR analysis, the predictive activity of library compounds can be calculated directly with 3D-QSAR model. If the structure of target receptor and the active sites have already been elucidated, the binding energy of library compounds with the target receptor can also be evaluated. With the 3D-QSAR prediction combined with the binding energy evaluation of library compounds with the receptor, we can decide the high active library compounds as the drug candidates.

RESULTS

To test our method and program, we have demonstrated two examples: one is benzodiazepine derivatives library, which is expected to have the maximum molecular diversity;

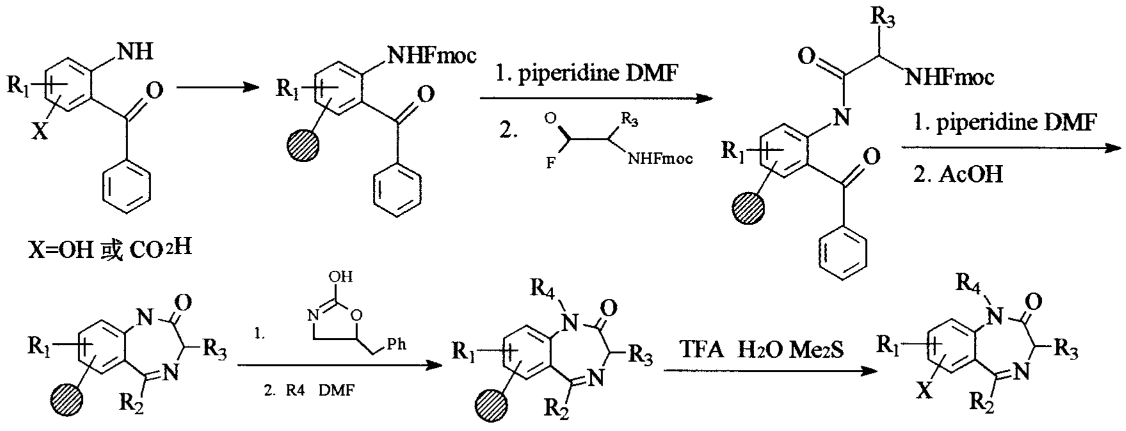

Figure 3. Ellman's solid-phase synthesis of benzodiazepine.

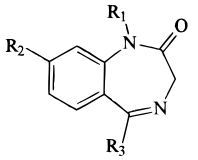

Figure 4. Structure of benzodiazepine, R_1, R_2, and R_3 are the connection sites which fragments can be linked on.

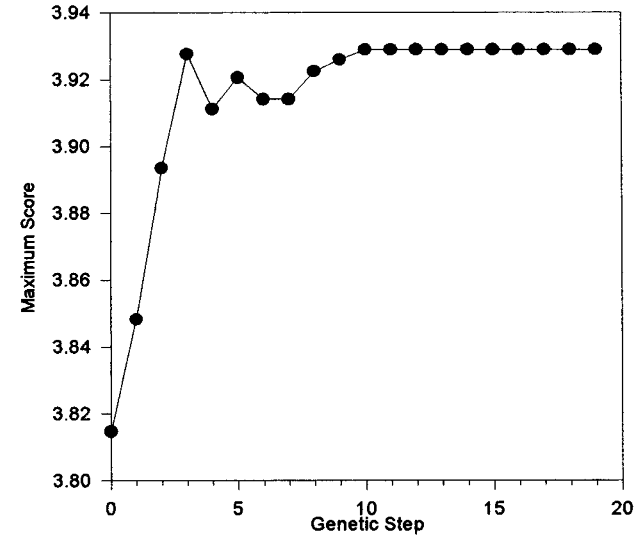

Figure 5. Maximum score of virtual benzodiazepine library compounds along the genetic step.

the other is (−)-huperzine A analogues library, which is designed for the structural modification of (−)-huperzine A.

Benzodiazepine Derivatives Library. Nonoligomeric molecules, which are nonpeptidic in nature and are below 600−700 in molecular weight, have become the major focus of library synthesis efforts for the development of medicinal agents. In one of the first articles to address the synthesis and evolution of small molecule combinatorial libraries, Bunin and Ellman reported the solid-phase synthesis of 1,4-benzodiazepine derivatives.[10] The original benzodiazepine synthesis sequence was based upon the combination of three different building block sets: 2-aminobenzophenones, amino acids, and alkylating agents. The synthesis scheme is shown in Figure 3. To increase the diversity of 1,4-benzodiazepine-2-ones available through solid-phase synthesis, Plunkett and Ellman utilized the Stille coupling reaction to synthesize a variety of 2-aminoaryl ketones on solid support.[11] Bunin and co-workers have reported the preparation of a library of 11 200 discrete 1,4-benzodiazepines from 20 acid chlorides, 35 amino acids, and 16 alkylating agents.[12] The acid chlorides were selected from a set of over 300 commercially available compounds that are compatible with the synthesis sequence using a similarity group procedure developed by Dr. Steven Muskal to select as diverse a set as possible. This library is currently being screened by a number of industrial and academic collaborators.

To demonstrate the ability of our algorithm, we selected 72 chemical structures as fragments which are not limited within 2-aminobenzophenones, amino acids, acid chlorides, and alkylating agents. The structure of benzodiazepines and the predefined connection sites R_1, R_2, and R_3 are shown in Figure 4. The fragments that we used for construction of virtual benzodiazepine library are included in the Supporting Information.

To synthetic chemists, three connection sites on the core are different. There should be three different sets of fragments corresponding to the three connection sites, respectively. This can be easily implemented in our program. Since our study is just to explore a virtual benzodiazepine library with maximum molecular diversity, we use the same fragment sets including 72 chemical structures for the three connection sets. In practice, they would be different.

The number of library compounds is also important for the quality of combinatorial library. More library compounds imply that the designed combinatorial library samples consist of more of the universe of organic chemical compounds. Actually, the number of virtual library compounds should be as same as that of the experiment library. For example, if we select N fragments as building blocks and these fragments can be linked on M connection sites of the core, there should be about N^M compounds in the virtual library. A bigger virtual combinatorial library will give a more accurate statistic result. But at the same time, it will also burden the computational resources simultaneously. In this article, we set population size to be 100 arbitrarily though it can be easily adjusted in our program.

The fitness function for maximizing molecular diversity is shown in formula 4. The maximum score of libraries compounds along the genetic step is shown in Figure 5, which indicates that the maximum score remains almost unchanged after generation 10. Figure 6, Figure 7, and

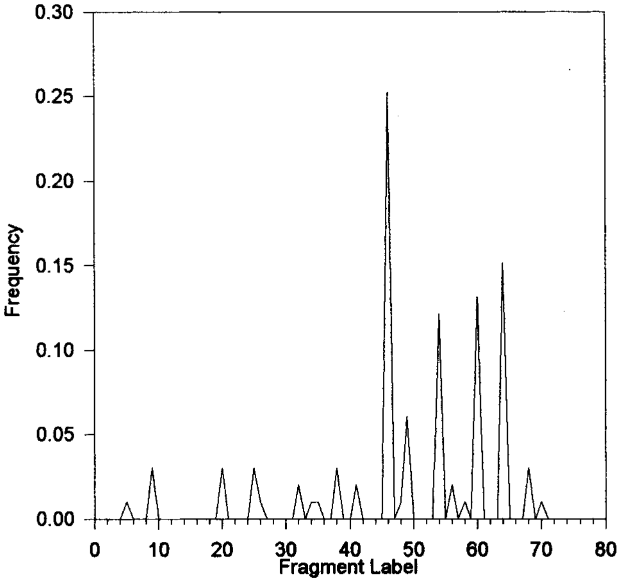

Figure 6. Fragments frequency at connection site R_1.

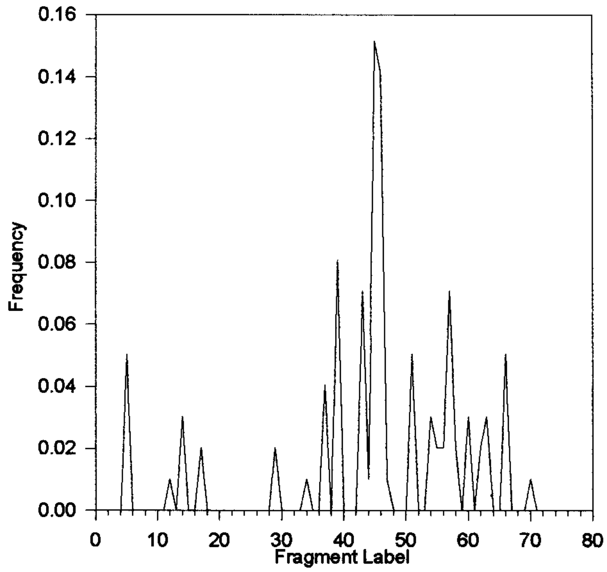

Figure 7. Fragments frequency at connection site R_2.

Figure 8 show the fragments frequency at connection sites R_1, R_2, and R_3, respectively. Figure 9 shows the fragments frequency at connection sites of R_1, R_2, and R_3 in the final virtual library.

According to Figure 6, the fragments with high frequency at R_1 connection site are fragments 46, 48, 54, 60, and 64. And the fragments with high frequency at R_2 connection site are fragments 5, 39, 43, 45, 51, 57, and 66 which are shown in Figure 7. From Figure 8, we know that those fragments with high frequency at connection site R_3 are fragments 3, 32, 33, 34, 44, 45, 51, 52, 55, 57, 60, and 63. The fragments occurring on connection sites R_1, R_2, and R_3 with high frequency are fragments 39, 44, 45, 46, 51, 54, 57, 60, and 64 according to Figure 9.

It is instructive to look at the structure of highly frequent fragments. Figure 10 shows those highly frequent fragments in the final virtual benzodiazepine library according to Figure 9.

(−)-Huperzine A Analogues Library. In the past several years, there have been a lot of structural modification works on huperzine A.[13−15] The pharmaceutical assay results showed that it seemed there are little chances to improve its AChE inhibitory activity except introducing methyl group at C-10 position of (−)-huperzine A. Several months ago, Mia L. Raves et al.[16] elucidated the crystal structure of

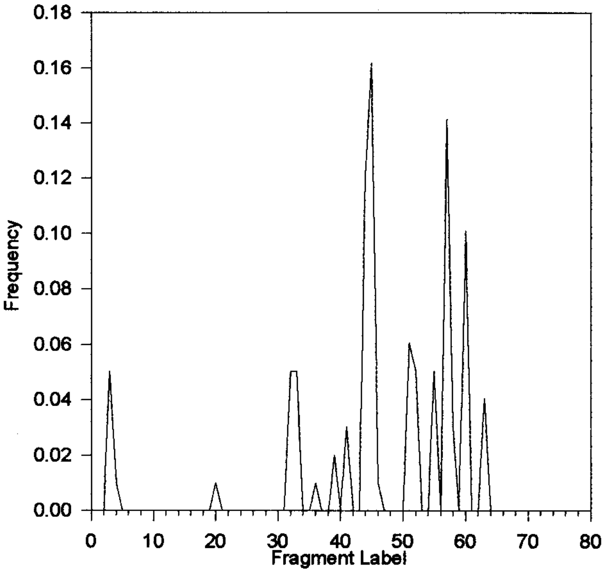

Figure 8. Fragments frequency at connection site R_3.

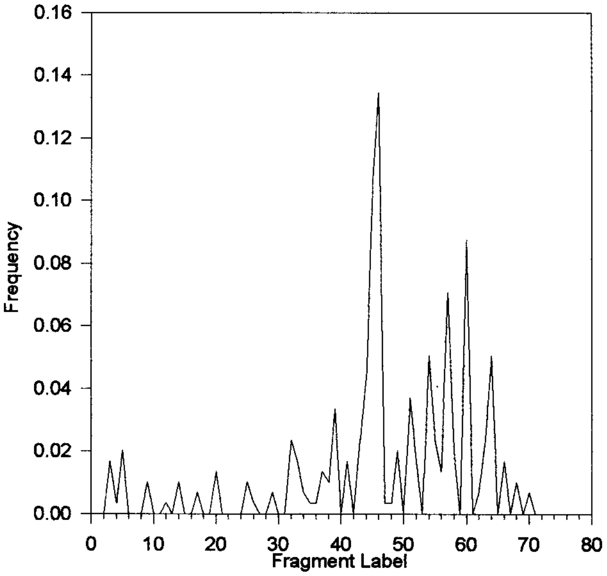

Figure 9. Fragments frequency at connection sites R_1, R_2, and R_3.

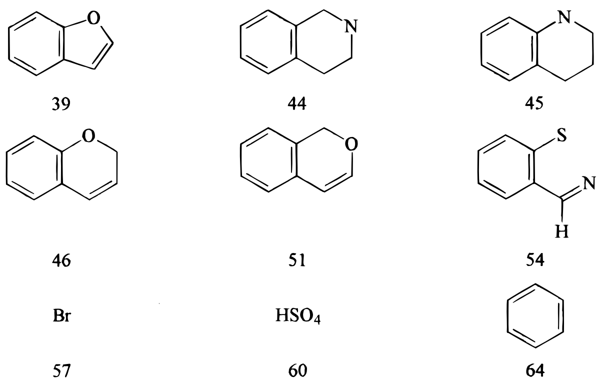

Figure 10. Structures of high frequent fragments at connection sites R_1, R_2, and R_3.

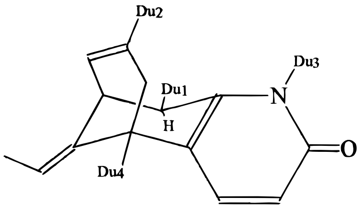

Figure 11. Structure of (−)-huperzine A, atom Du represents the substitute positions that building blocks will be added on.

—CH_3	—CH_2CH_3	—CH_2OH
F_1	F_2	F_3
—OH	—CH_2CH_2OH	—CH_2NH_2
F_4	F_5	F_6
—NH_2	—COOH	—H
F_7	F_8	F_9

Figure 12. The structure representation of building blocks used for construction of virtual (−)-huperzine A analogue library and their abbreviation labels.

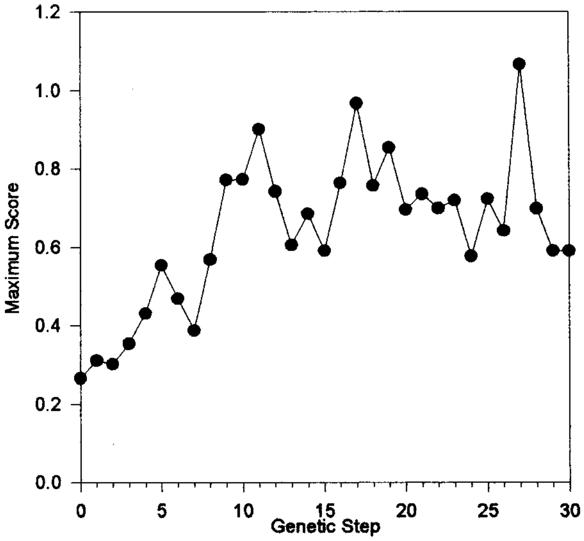

Figure 13. Maximum score of virtual (−)-huperzine A library compounds along the genetic step.

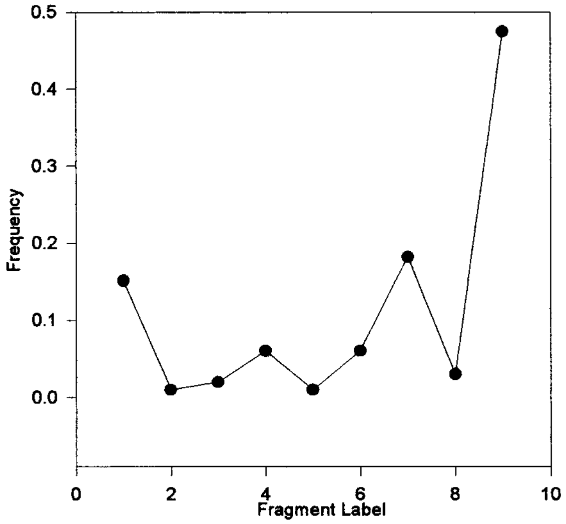

Figure 14. The occurrence probability of fragments at position one.

complex of (−)-huperzine A and AChE at 2.5 Å resolution. This provided us important information to understand the ligand−protein interaction of AChE inhibitors with AChE. And it also makes it practical to do structure-based drug design of AChE inhibitors.

The structure of (−)-huperzine A and the assigned substituent positions are shown in Figure 11. From the structure, we see that it is hard to synthesize the (−)-huperzine A library with combinatorial synthesis methodology. The reason that we determined to choose this scaffold is that it is a good example because the 3D structure of its target is known. Though we probably cannot synthesize this library, fragments frequency and screen result can give us useful clues as how to design (−)-huperzine A analogues which have certain molecular similarity with (−)-huperzine A.

The fragments set is shown in Figure 12. With formula 5 as the fitness score, we utilized a genetic algorithm to construct a (−)-huperzine A virtual library. We used the coordinate of (−)-huperzine A in the complex of huperzine A and AChE as the structure orientation of (−)-huperzine A analogues in the virtual library when docked in AChE to evaluate their binding energy with the receptor.

The maximum score of library compounds along the genetic step is shown in Figure 13. From the trend of the curve, we get to know that the average value of maximum score at the equilibrium state is 0.73. After generation 9, the maximum score has reached the average value. In the following generations, the maximum score fluctuates around the average value. Inspecting the molecular structure of compounds in the library, we find that (−)-huperzine A was already included in the 40 higher score compounds in generation 9 and was inherited thereafter.

The frequency of fragments that occurred at position Du1, Du2, Du3, and Du4 were presented in Figures 14−17. From these results, we see that CH_3, NH_2, and H most likely appear at position Du1; fragment CH_3 and NH_2 are likely to appear at position Du2; fragment H most probably appears at position Du3; fragments of CH_3, OH, and NH_2 are likely to appear at position Du4. This result implies that it is not beneficial to perform structure modification at position Du3 and position Du1, Du2, and Du4 are the suggestive substituent sites of the core molecule.

Moreover, we have evaluated the binding energy of library compounds with AChE receptor. The energy points distribution is shown in Figure 18. The dashed line in the figure is the binding energy of (−)-huperzine A with AChE. The points whose value is 1000.0 are those compounds whose binding energy is so huge that they could not be highly active AChE inhibitor. Points in Figure 18 can be assigned into three regions: big binding energy, modest binding energy, and low binding energy. It is very clear that most library compounds have a low binding energy which is near that of (−)-huperzine A.

Whether the designed virtual library is a good one with appropriate molecular diversity, it should be verified by experiment. In fact, a library compound including fragments F_1, F_1, F_9, and F_7 at position Du1, Du2, Du3, and Du4 (compound 16) is a potential AChE inhibitor. The pharmaceutical screening result of this compound (IC50=3.5 × 10^{-8} M) demonstrates its higher bioactivity than (−)-huperzine A.[17] This verified and tested our program indirectly.

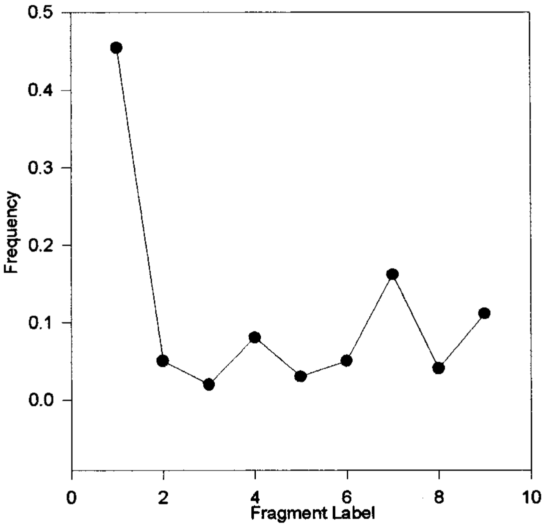

Figure 15. The occurrence probability of fragments at position two.

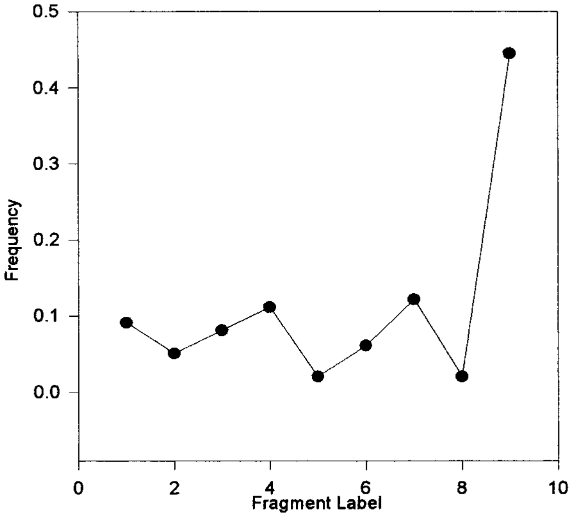

Figure 16. The occurrence probability of fragments at position three.

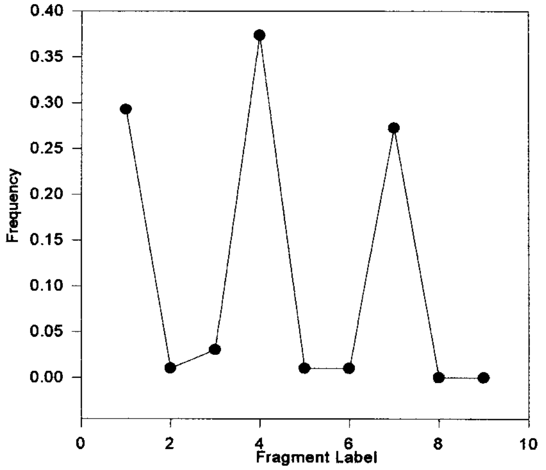

Figure 17. The occurrence probability of fragments at position four.

DISCUSSION

We have described our idea of using genetic algorithm to design virtual combinatorial library and demonstrated its application in designing combinatorial library for broad screening and for structural modification. The simulation result tells us that a virtual library can be constructed with a modest amount of computational resources given the molecular scaffold and the fragments set. From the process of genetic generation, we can analyze the fragments frequency within the high score library compounds. High frequent fragments could be taken as building blocks of a combinatorial library. Since our molecular dissimilarity and molecular diversity are based on the CoMFA matrix, determining active conformation of the template compound is very important. This precursor work should be completed before the genetic generation. The geometry parameters of the template compound could be used as geometry constraints when performing structure optimization on virtual library compounds. Therefore, as long as the correct conformation could be determined, we could correctly identify the molecular dissimilarity of library compounds and molecular diversity of a combinatorial library. Our idea of describing molecular dissimilarity and molecular diversity relates closely with the structure specificity of the active site of targeted receptors. Different classes of compounds with the same biological function may activate different residues involved in the same active site of the receptor. Therefore, in designing virtual combinatorial library for structural modification, there is something worthy of considering carefully in selecting the lead compound. If the lead compound has the same scaffold as that of library compounds, we can directly superimpose the library compounds onto the lead compound. Otherwise, the structure orientation of library compounds relative to the lead compound should be carefully examined. If there are common atoms or groups on the scaffold of library compounds and the lead compound which binds the same sites of the target receptor, such atoms or groups may be used as the overlap points. Thus, the way to superimpose library compounds onto the lead compound is problem-specific. And there are also a lot of ways such as 2D-QSAR, 3D-QSAR, cluster analysis, and binding energy evaluation to predict the bioactivity of library compounds.

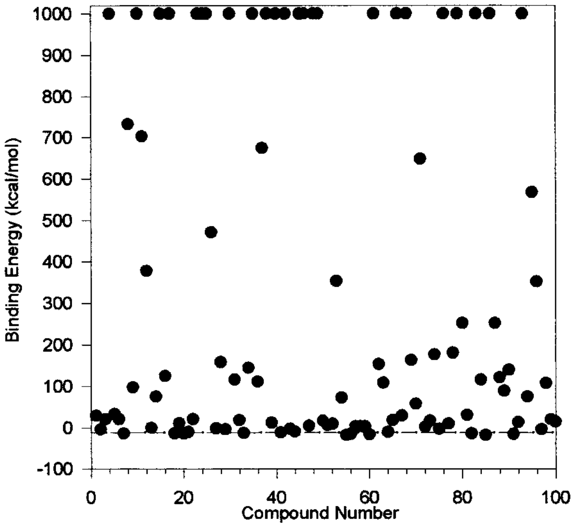

Figure 18. The distribution of binding energy of library compounds. The dashed line is the energy of (−)-huperzine A.

There are three kinds of genetic manipulations. Protocols for choosing population size, for making selection, and for fixing the relative frequency of mutation and crossover have been discussed in some research articles. High frequent crossover will make the genetic mechanism to sample more potent solution space. The number of virtual library compounds should be the expected number of our experimental combinatorial library. How will the population size influence the result of genetic algorithm? It should be further investigated in the future work.

The fitness function in this article mainly concerns on the molecular dissimilarity. One can of course think out a number of variations. The fitness function can be any function of molecular dissimilarity.

We have shown small organic molecules as examples, but this assembly method and molecular dissimilarity description can also apply to any type of molecules. The fragments can be also any type of chemical structures. From a practical sense, these fragments should be reactive and can be attached on the connection sites with chemical synthesis methods. Therefore, after fragments frequency analysis, the high score fragments should be further screened by synthesis chemists to decide which reactive fragments can be used as building blocks in combinatorial synthesis.

Compared with other methods to design combinatorial library based on 2D descriptors, our idea of describe molecular dissimilarity based on 3D grid property have some obvious advantages. The activity of designed library compounds can be easily predicted with 3D-QSAR model. And the binding energy of library compounds with the target receptor can also be easily evaluated with computational methods. The predominant feature of CoMFA method suggests that our idea of constructing virtual combinatorial library should be a successful one.

ACKNOWLEDGMENT

We wish to present our grateful thanks to Prof. Joel L. Sussman, Prof. Israel Silman, and Mr. Mia L. Raves for providing the (−)-huperzine A and AChE complex crystal coordinate and their research article.

Supporting Information Available: Fragments used for construction of virtual benzodiazepine library. See any current masthead page for Web access instructions.

REFERENCES AND NOTES

(1) Gary, T. W.; Sam, L. Synthetic Chemical Diversity Solid Phase: Synthesis of Libraries of C2 Symmetric Inhibitors of HIV Protease Containing Diamino Diol and Diamino Alcohol Cores. *J. Med. Chem.* **1995**, *38*, 2995−3002.

(2) Zuckermann, R. N.; Martin, E. J.; Spellmeyer, D. C.; Stauber, G. B.; Shoemaker, K. R.; Kerr, J. M.; Figliozzi, G. M.; Goff, D. A.; Siani, M. A.; Simon, R. J.; Banville, S. C.; Brown, E. G.; Wang, L.; Richter, L. S.; Moos, W. H. Discovery of Nanomolar Ligands for 7-Transmembrane G.-Protein-Coupled Receptors From a Diverse N−(Substituted) glycine Peptoid Library. *J. Med. Chem.* **1994**, *37*, 2678−2685.

(3) Weifan, Z.; Sung, J. C.; Tropsha, A. FOCUS-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries Using Simulated Annealing and Topological Index Based Molecular Similarity Metrics. *Molecular Graphics and Modeling Society-Electronic Conference (MGMS.-EC 1)*, **1996**.

(4) Sheridan, R. P.; Kearsley, S. K. Using Genetic Algorithm to Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.

(5) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.

(6) Wang, Y. E.; Yue, D. X.; Tang, X. C.; Hanin, I. Huperzine A.-A possible lead structure in the treatment of Alzheimer's disease. *Acta Med. Chem.* **1992**, 7, 175−205.

(7) Goldberg, D. E. *Genetic Algorithm in Search. Optimization, and Machine Learning*; Addison-Wesley: New York, 1989.

(8) 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.

(9) SYBYL programming language (SPL). [Computer programming language for SYBYL software]. St. Louis, MO, Tripos Associates, 1994.

(10) Bunin, B. A.; Ellman, J. A. A General and Expedient Method for the Solid-Phase Synthesis of 1,4-Benzodiazepine Derivatives. *J. Am. Chem. Soc.* **1992**, *114*, 10997−10998.

(11) Plunkett, M. J.; Ellman, J. A. A Silicon-Based Linker for Traceless Solid-Phase Synthesis. *J. Org. Chem.* **1995**, *60*, 6006−6007.

(12) Bunin, B. A.; Plunkett, M. J.; Ellman, J. A. *Methods Enzymol.* **1996**.

(13) Kozikowski, A. P.; YanXia, E.; Reddy, R.; Tuckmantel, W.; Hanin, I.; Tang, X. C. Synthesis of Huperzine A and Its Analogues and Their Anticholinesterase Activity. *J. Org. Chem.* **1991**, *56*, 4636−4645.

(14) Lamiani, G.; Li-Qiang, S.; Kozikowski, A. P. Palladium-Catalyzed Route to Huperzine A and its Analogues and Their Anticholinesterase Activity. *J. Org. Chem.* **1993**, *58*, 7660−7669.

(15) Kozikowski, A. P.; Miller, C. P.; Yamada, F.; Yuan-Ping, Pan. Delineating the Pharmacophoric Elements of Huperzine A: Importance of the Unsaturated three Carbon Bridge to its AChE Inhibiting Activity. *J. Med. Chem.* **1993**, *34*, 3399−3402.

(16) Raves, M. L.; Harel, M.; Yuan-Ping, Pang; Silman, I.; Kozikowski, A. P.; Sussman, J. L. Structure of acetylcholinesterase complexed with the nootropic alkaloid, (−)-huperzine A. *Nature Struct. Biol.* **1997**, *4* (1), 57−63.

(17) Vrland, G. WHAT IF A molecular modelling and drug design program. *J. Mol. Graph.* **1990**, *8*, 52−56.

CI970086O