

Self-Organizing Superimposition Algorithm for Conformational Sampling

FANGQIANG ZHU, DIMITRIS K. AGRAFIOTIS

*Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive,
Exton, Pennsylvania 19341*

Received 25 August 2006; Accepted 18 October 2006

DOI 10.1002/jcc.20622

Published online 13 February 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: A novel self-organizing algorithm for conformational sampling is introduced, in which precomputed conformations of rigid fragments are used as templates to enforce the desired geometry. Starting from completely random coordinates, the algorithm repeatedly superimposes the templates to adjust the positions of the atoms, thereby gradually refining the conformation of the molecule. Combined with pair-wise adjustments of the atoms to resolve steric clashes, conformations that satisfy all geometric constraints can be generated from this procedure. The algorithm is demonstrated to achieve good performance and promises potential applications on more challenging modeling problems.

© 2007 Wiley Periodicals, Inc. *J Comput Chem* 28: 1234–1239, 2007

Key words: self-organizing superimposition; self-organization; self-organizing; conformational sampling; conformational analysis; distance geometry; molecular fragment; superimposition; stochastic proximity embedding

Introduction

Conformational sampling, the generation of low-energy three-dimensional (3D) conformations of organic molecules, is a problem of central importance in molecular modeling and computer-aided drug design. Several applications depend critically on the diversity and quality of the generated conformations, including protein docking, pharmacophore modeling, 3D database searching, and 3D-QSAR, to name a few. A variety of algorithms have been devoted to the conformational sampling problem.¹ For molecules with limited conformational flexibility, systematic search^{2,3} over all possible torsions at certain discrete intervals is both practical and effective. Stochastic methods^{4,5} such as Monte Carlo and simulated annealing represent an alternative approach that produces a sequence of conformations, each generated by randomly perturbing the previous one. Many variants of the systematic and stochastic methods have been proposed and implemented in different computer programs.^{6–8}

A more recent technique, known as stochastic proximity embedding (SPE), attempts to generate conformations through a process of self-organization.⁹ In this method, the coordinates of the atoms are assigned random initial values and are then gradually refined toward a feasible geometry. SPE randomly selects a pair of atoms at each step and adjusts their coordinates so that the distance between them falls within a certain allowed range, which is typically derived from the covalent structure of the molecule. It was demonstrated that by repeating this procedure, the coordinates could converge toward a feasible geometry.⁹ Unlike perturbation-based stochastic methods, a new conforma-

tion generated by SPE is not derived from previously sampled conformations, but represents a completely random cast on the low-energy surface in the conformational space. Therefore, the sampling will not be limited within certain regions in the conformational space even if only a small number of conformations are generated. SPE was shown to compare favorably to a conventional distance geometry algorithm and emerged as a promising technique for conformational sampling.

The conformation of a molecule containing N atoms is represented by $3N$ coordinates. However, the actual degrees of freedom for low-energy conformations are far fewer than $3N$ due to geometric constraints, such as those on bond lengths and angles. As a result, a flexible molecule can be decomposed into rigid fragments, each containing a small number of atoms. A fragment normally has a unique geometry as determined by the geometric constraints. For instance, a phenyl group can be taken as a rigid fragment that must assume the geometry of a regular hexagon. In SPE, no special effort is made to explicitly identify or handle fragments, and the correct geometries for the fragments are enforced by the normal pair-wise adjustments. In contrast, in some other conformation generation methods, the geometries of all common fragments are precomputed and stored in a library. The fragment conformations for a given molecule can then be directly retrieved from the library during the sampling, thereby speeding up the process. In this article, we propose a new method, the self-organizing superimposition (SOS) algorithm, which takes

Correspondence to: F. Zhu; e-mail: fzhu2@prdus.jnj.com

advantage of the precomputed fragment conformations and achieves a significant improvement in performance over SPE.

Methods

In this section, we first describe the scheme to decompose a molecule into overlapping templates. Then we introduce the template fitting operation used in our algorithm to enforce the desired geometry of a template. Following a simple illustration, we provide the details of the self-organizing superimposition (SOS) algorithm. We then discuss the strategy for properly assigning weights to template atoms in order to accelerate convergence. Finally, the implementation and testing of the algorithm are briefly described.

The Reference Templates

A chemical molecule normally consists of rigid fragments connected together by bonds (typically single σ bonds) that can be considered freely rotatable. In our algorithm, the conformation of each fragment is retrieved from a precomputed library and used to form a reference template. In addition to the fragment atoms, the template also includes the atoms that are directly attached to the fragment through rotatable bonds. Because of the constraints on bond lengths and angles, the positions of these attached atoms can be determined by the coordinates of the fragment atoms in the reference template. Therefore, the two atoms in a rotatable bond connecting two fragments will be included in both templates. This ensures that the geometric constraints involving these attached atoms are represented in the templates and are enforced in the subsequent conformation generation. As an example, for the simple molecule shown in Figure 1a with only one rotatable bond, the two reference templates are shown in Figure 1b, with the bond (between atoms no. 6 and no. 7) represented in both templates.

The Template Fitting Operation

The conformation of a reference template represents the ideal geometry that the corresponding template atoms should adopt in the generated conformation. In our algorithm, this is enforced through the template fitting operation, as described below, where *coord1* refers to the coordinates of the reference template and *coord2* to the coordinates of the corresponding atoms in the molecule that belong to the template.

fit(coord1, coord2)

- ```
{
 1. Using coord2 as reference, perform a rigid-body alignment to superimpose coord1 on top of coord2, such that the RMSD between coord1 and coord2 is minimized.
 2. Copy coord1 to coord2.
}
```

In the first step, a rigid-body transformation (by translation and rotation) is performed on the reference template, which essentially places it in the closest position to the corresponding template atoms in the molecule. This can be done using the standard rigid-body RMSD superimposition technique. Then in the second step, the coordinates of those template atoms in the

molecule are replaced by the new coordinates of the reference template. In essence, the fitting operation applies the smallest possible adjustment to the template atoms in the molecule to achieve the geometry of the reference template.

### Iterative Refinement

While a fitting operation enforces the correct geometry for one template, it also alters the geometries of its neighboring templates in the molecule, due to the movement of the atoms they have in common. Through iterations of template fitting operations, however, a feasible molecular conformation can be obtained in which all templates approach their reference geometries. This procedure is illustrated in Figure 1. Starting from completely random coordinates (step 0 in Fig. 1c), by repeatedly performing fitting operations on the two templates alternately, an almost perfect conformation (step 20b) for the molecule is obtained in 20 cycles.

### The Self-Organizing Superimposition Algorithm

When all templates in the molecule approach their ideal (reference) geometries, the molecular conformation will satisfy all of the constraints on bond lengths, angles, planarity, etc. A feasible conformation, however, also requires the absence of steric clashes, which is not enforced by fitting operations. Therefore, pair-wise adjustments, the elementary refinement step in the SPE algorithm,<sup>9</sup> are also adopted in our algorithm to resolve steric clashes. The combination of template and pair-wise refinements will drive the coordinates toward satisfying all geometric constraints. Our self-organizing superimposition (SOS) algorithm is outlined below.

1. Read in the molecule and decompose it into overlapping templates.
2. Retrieve the conformation of each reference template from the library.
3. Determine the upper bounds  $\{u_{ij}\}$  and lower bounds  $\{l_{ij}\}$  for each pair of atoms  $i$  and  $j$  in the molecule.
4. Assign random coordinates to all atoms in the molecule.

Repeat for  $n_c$  times

- ```
{
  For each template  $t$  in the molecule do
  {
    Repeat for  $n_p$  times
    {
      5. Perform a pair-wise adjustment: randomly select a pair of atoms  $i$  and  $j$  that are not in the same template, and measure the distance  $d$  between them. If  $d < l_{ij}$ , then move the two atoms away from each other so that their new distance is  $l_{ij}$ ; if  $d > u_{ij}$ , then move them toward each other to a new distance of  $u_{ij}$ .
    }
    6. Perform a fitting operation on template  $t$ .
  }
}
```

Steps 1–3 of the algorithm represent the initialization phase when a new molecule is read in. As described in earlier sections, the molecule is decomposed into fragments, and a template is

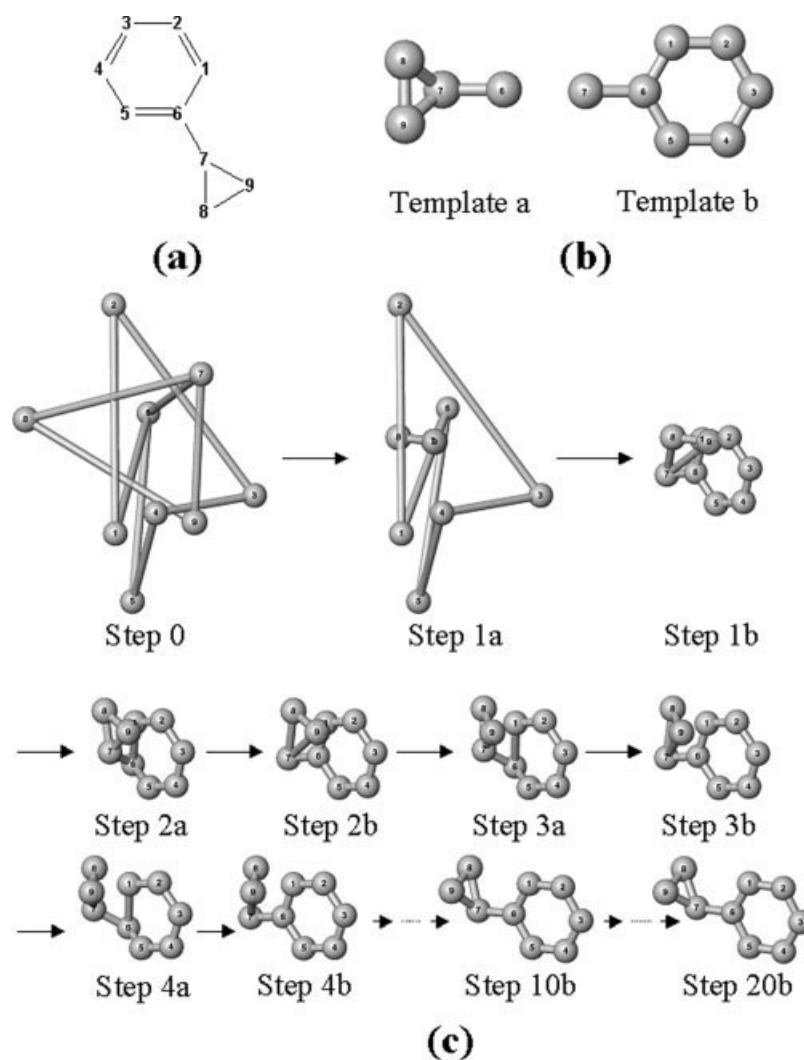


Figure 1. An illustration of the self-organizing superimposition process for a simple molecule. (a) The chemical structure of the molecule. (b) The geometries of the two reference templates for the molecule. Note that the atoms (no. 6 and no. 7) in the connecting bond are represented in both templates. (c) Some snapshots during the refinement process. Starting from random coordinates at step 0, each following iteration consists of two steps of fitting operations on templates a and b, respectively.

initialized for each fragment with its attached atoms. The reference geometries for the templates are obtained from a pre-computed library. Moreover, to enable the pair-wise adjustments, the lower bounds $\{l_{ij}\}$ and upper bounds $\{u_{ij}\}$ for every interatomic distance are determined in step 3. The sum of the VDW radii for two atoms (except for those connected through three or fewer bonds) dictates the minimum distance between them without causing a steric clash and is therefore taken as the lower bound for the atom pair. The Floyd-Warshall algorithm is employed to determine the shortest path between each pair of atoms based on the topology of the molecular graph, which is then taken as the upper distance bound for the pair. Since all of the upper bounds will be automatically satisfied as all templates approach ideal geometries, the direct enforcement of the upper

bounds is not necessitated by the algorithm, but rather serves to achieve faster convergence and improve the efficiency of the algorithm.

After random assignment of the atomic coordinates in step 4, the self-organizing process starts, and proceeds for n_c cycles (iterations). In each cycle, a number of pair-wise adjustments (step 5) are performed, and every template undergoes a fitting operation (step 6) as described in earlier sections. In each pair-wise adjustment, two atoms in different templates are randomly selected and their distance is measured. If the distance is not within the corresponding lower and upper bounds, the two atoms will be moved toward or away from each other to satisfy those bounds. In our algorithm, the pair-wise adjustments are inter-laced with the fitting operations such that each template fit is

accompanied by n_p pair-wise refinement steps. Therefore, the total number of the pair-wise refinements during the self-organizing procedure is $n_c N_t n_p$, where N_t denotes the number of templates in the molecule.

Improving the Performance

As mentioned earlier, if two templates are connected by a rotatable bond, the two atoms in the connecting bond are represented in both templates. Although a template adopts the exact reference geometry right after a fitting operation on it, the geometry will be distorted when another fit is subsequently performed on one of its neighboring templates and displaces the connecting bond (see Fig. 1). Therefore, in order to expedite convergence, it would be desirable to minimize the displacement of the connecting bonds during a template fit and hence reduce the distortion to the geometries of the neighboring templates. This can be achieved by assigning a higher weight to the atoms in the connecting bonds and performing a weighted superimposition in the fitting operation. Since atoms with higher weights are fitted more closely in the weighted superimposition, the positions of the connecting bonds will be better preserved during the fit. For the simple molecule shown in Figure 1, if the two atoms (no. 6 and no. 7) in the connecting bond are assigned a weight of 10 while all other atoms a weight of 1, it takes only five cycles to converge to an almost perfect geometry as compared to 20 cycles required for the unweighted fitting operations shown in the figure.

If a template is connected to two or more other templates, even when the atoms in the connecting bonds are assigned high weight as described above, they are still likely to undergo large displacements, because these connecting bonds may not have been positioned consistently with respect to each other and hence may not superimpose well against their counterparts in the reference geometry. In such general cases, a hierarchical scheme described below could be used to expedite convergence. If the connectivity between the templates is acyclic, the system can be viewed as a tree structure in which each node (representing a template) except the root is connected to a single parent and to any number of children. In a template, we denote the bond connecting to its parent template as an “upward bond”. In our scheme, during a template fit, we assign a high weight only to the two atoms in the upward bond, while all other atoms, including those in the bonds connecting the template to its children, are given a lower weight. In this scheme, the position of the upward bond will be preserved with a higher priority during the fit, which will prevent a large distortion to the parent template. Consequently, once the templates at higher levels in the tree obtain good geometries, they will not be significantly disturbed by the subsequent fitting operations on lower level templates. During the iterative refinement, therefore, convergence will be first achieved at the highest level and then propagated to the templates at lower levels.

If a sufficiently high weight is assigned to the atoms in the upward bond in each template, as described above, it takes only a few cycles of template fits to achieve convergence. When pair-wise adjustments are incorporated, however, the weight should not be assigned arbitrarily high, because if those atoms are too immobile, the ability of the algorithm to resolve steric clashes

will be compromised. Through experimentation, we found that a weight of 5 for the atoms in the upward bonds and a weight of 1 for all other atoms is a reasonable choice for most molecules.

Implementation and Testing

The self-organizing superimposition algorithm was implemented in the C++ programming language, and tested on four representative molecules shown in Figure 2, which were also used in some other conformational sampling studies.^{9,11,12} For each molecule, all single bonds except those in small rings were treated as rotatable, and the fragments were identified after removing these rotatable bonds from the molecule. The 3D conformations of the fragments were retrieved from a fragment library. After adding the attached atoms (see “The Reference Templates” Section), these conformations were used as the reference templates. The self-organizing process consisted of 50 cycles ($n_c = 50$) as described in “The Self-Organizing Superimposition Algorithm” Section. The number of pair-wise adjustments following each

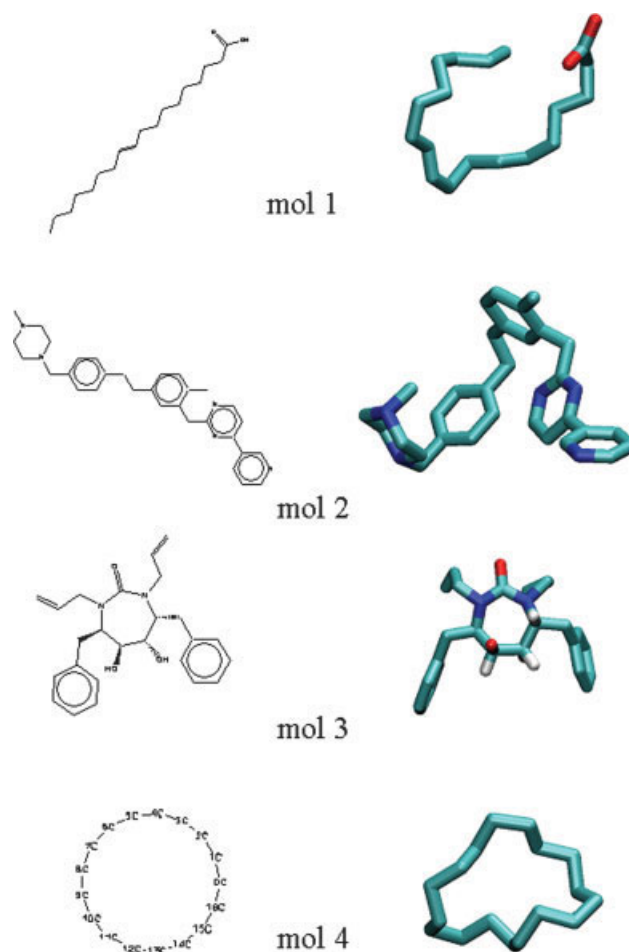


Figure 2. The chemical structures and 3D conformations generated by the self-organizing superimposition (SOS) algorithm for the four molecules under test. The molecular images were rendered using VMD.¹⁰

template fit, n_p , was determined such that the total pair-wise steps during the cycles, $n_c N_t n_p$, is at least $2N^2$, where N_t and N denote the numbers of templates and atoms in the molecule, respectively. The hierarchical scheme for weight assignment, as described in “Improving the Performance” Section, was adopted, which assigned a weight of 5 to the atoms in the upward bond in a template and a weight of 1 to all other atoms. For each molecule under test, 10,000 conformations were generated. For the purpose of comparison, we also used the SPE algorithm with default parameters⁹ to generate the same number of conformations for each molecule. All tests were run on an Intellistation workstation equipped with a 3.8 GHz Intel Xeon processor.

Results and Discussion

Both the SOS and SPE⁹ algorithms adopt self-organizing schemes to refine the coordinates toward satisfying the geometric constraints. Naturally, more cycles of refinements will yield conformations of better quality, but in practice one needs to balance quality and computing time. Two relevant criteria for evaluating the performance of such algorithms are how well the geometric constraints are satisfied in the resulting conformations and how much computing time is needed to generate the results. To quantify the former criterion, we examined the bond and angle deviations in the generated conformations. For each conformation, we measured the actual length of every bond and computed its deviation from the ideal bond length. The root mean square and the maximum of such deviations over all bonds in the molecule were then calculated. The mean and maximum deviations for angles were obtained similarly for each conformation. These values serve as indicators of the quality of the generated conformations.

The average values for these deviations over the 10,000 generated conformations, along with the computing times, are provided in Table 1 for the four molecules under test. One can see from the table that in all cases the SOS algorithm took less time and generated conformations of higher quality (smaller bond and angle deviations) than the SPE algorithm⁹ did, indicating that the former achieved a faster convergence rate. The gain in performance is more prominent for mol2 and mol3, which consist of templates of relatively larger sizes. For the cycloheptadecane molecule (mol4), its macrocyclic structure would pose a challenge to systematic methods due to problems associated with ring closures. In contrast, both of the self-organizing algorithms handled it well without any difficulty. In the SOS algorithm, since all of the bonds in the cycloheptadecane were treated as rotatable, the molecule was decomposed into 17 fragments, each containing only one atom, and each template consisting of a single fragment atom and two attached atoms only. This would be a most unfavorable case for the SOS algorithm in comparison to SPE, given that each template fit enforces only two bond lengths and one angle. The weight assignment scheme described earlier had no effect in this case either. Nevertheless, the SOS algorithm still clearly outperformed SPE on this molecule, although by a smaller margin than in other cases.

Some randomly chosen 3D conformations generated by the SOS algorithm are shown in Figure 2, which can be visually confirmed to have sensible geometries. Although Table 1 indi-

Table 1. Performance Benchmarks for the Self-organizing Superimposition (SOS) and the SPE⁹ Algorithms on the Four Molecules (See Fig. 2) Under Test.

ID	Algorithm	Time (s)	Bond deviation (Å)		Angle deviation (°)	
			Mean	Max	Mean	Max
mol1	Iterative superimposition	0.0047	0.005	0.018	0.5	1.6
	SPE	0.0116	0.025	0.083	6.3	23.1
mol2	Iterative superimposition	0.0038	0.016	0.079	1.5	6.8
	SPE	0.0396	0.045	0.154	6.8	21.6
mol3	Iterative superimposition	0.0038	0.021	0.095	1.8	8.6
	SPE	0.0290	0.082	0.333	10.4	36.9
mol4	Iterative superimposition	0.0047	0.006	0.017	0.6	1.5
	SPE	0.0070	0.010	0.026	1.6	3.9

Each algorithm was used to generate 10,000 conformations, and the computing times per conformation are shown in the table. For each conformation, the root mean square and maximum deviations (with respect to the ideal values) over all bond lengths and angles in the molecule were calculated as explained in the text, and the averages of these deviations over the 10,000 conformations are listed in the table.

icates that the generated conformations still exhibit some deviations with respect to the ideal (reference) geometric parameters, the deviations are not severe if one considers the consistency of these reference parameters over different sets of rules. For example, in some rule-based parameter set widely used in conformational sampling programs¹³ including SPE,⁹ all σ bonds between two carbon atoms have a length of 1.50 Å, and all angles centered on an sp^3 carbon are 109.5°; in contrast, in some protein force field, the lengths of this type of bonds vary from 1.49 Å to 1.54 Å, and the angles from 107.0° to 115.0°, depending on the specific atom types. In light of this, the geometric deviations in the conformations generated by our SOS algorithm are comparable to the variance between different reference parameter sets. When more accurate geometries are required, rather than running more cycles of refinement in our algorithm, one should perform a standard energy minimization on the generated conformations using a high-quality force field.

Rigid-body superimposition has been used in some software to assemble fragments into a 3D molecular conformation.⁶ In this study, we demonstrated that this technique can also be effectively applied in the context of self-organization starting from random coordinates. In our SOS algorithm, since each template fitting operation simultaneously enforces a number of geometric constraints, a relatively small number of steps are required to achieve convergence. In comparison, the elementary step in the SPE algorithm,⁹ a pair-wise adjustment, needs much less computation than a template fit, but enforces only one distance constraint at a time, and consequently many more steps are required for convergence. Taken together, as our test results indicate, the SOS algorithm represents a significant improvement in performance over SPE. On the other hand, unlike SPE, the

SOS algorithm requires a precomputed fragment library as input. Fortunately, this is not a difficult requirement, as quite a few conformational sampling programs also utilize fragment libraries and provide routines to handle them.^{6,7} We analyzed the J&JPRD corporate chemical database containing about 1.5 million compounds, and found that they could be decomposed into fewer than 20,000 unique fragments. The geometries of these fragments can be easily generated using SPE, followed by energy minimization. Interestingly, building a fragment library for large biomolecules such as proteins is actually much more trivial, given the fact that all proteins can be decomposed into merely a few dozens of unique fragments.

In SPE and some other distance geometry¹⁴ algorithms, correct chirality is achieved by enforcing volume constraints.⁹ During the generation of every conformation, SPE needs to apply a number of adjustments on the signed volume of each stereocenter.⁹ In our SOS algorithm, on the other hand, since the local chirality in each reference template will be preserved in the generated conformations, chirality needs to be handled only in the initialization phase. Once the reference templates are initialized properly, the correct chirality will be naturally enforced in all subsequently generated conformations without requiring any additional effort.

When a systematic search is feasible and affordable, it would be the most effective approach for sampling the conformations of a molecule, due to its exhaustive nature. In comparison, the self-organizing methods, such as SOS and SPE,⁹ are more general approaches with wider applicability and better scalability, and are especially well suited for molecules with many rotatable bonds or with macrocycles, when a straightforward exhaustive torsion search is not feasible. In addition, some systematic or stochastic approaches require a good 3D conformation as input, which is then subjected to torsional driving or perturbations to generate more conformations. In many programs, this initial conformation is generated by distance geometry.¹⁴ The SPE and SOS algorithms are also well applicable for this purpose since they do not require an existing 3D conformation as input.

More importantly, the self-organizing scheme allows the flexibility to apply various external constraints (e.g., distance constraints derived from NOEs). In this study, the pair-wise adjustments were mixed with template fitting operations to enforce

both types of constraints during the course of the refinement. It was also demonstrated in a recent study that the distance constraints in SPE can be modified to bias the sampling toward more extended or more compact conformations.¹⁵ Such biasing heuristics are also perfectly applicable in our SOS algorithm. The ability to enforce external constraints is especially important for the sampling of large molecules, where additional constraints usually exist that can restrict the otherwise enormous conformational space. Although this study is only focused on the conformational sampling of small molecules, the self-organizing approach and the efficient superimposition algorithm presented here are also expected to show great potential in more complex and challenging applications, such as NMR structure determination for proteins and other protein modeling problems.

References

1. Leach, A. R. In *Reviews in Computational Chemistry*, Vol. 2; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991, pp. 1–55.
2. Lipton, M.; Still, W. C. *J Comput Chem* 1988, 9, 343.
3. Bruccoleri, R. E.; Karplus, M. *Biopolymers* 1987, 26, 137.
4. Saunders, M. *J Am Chem Soc* 1987, 109, 3150.
5. Ferguson, D. M.; Raber, D. J. *J Am Chem Soc* 1989, 111, 4371.
6. MOE. www.chemcomp.com.
7. Omega. www.eyesopen.com.
8. Accelrys Software Inc., Catalyst, Release 4.10, Accelrys Software Inc.: San Diego, 2005.
9. Xu, H.; Izrailev, S.; Agrafiotis, D. K. *J Chem Info Comput Sci* 2003, 43, 1186; Agrafiotis, D. K.; Xu, H. *Proc Natl Acad Sci USA* 2002, 99, 15869.
10. Humphrey, W.; Dalke, A.; Schulten, K. *J Mol Graphics* 1996, 14, 33.
11. Diller, D. J.; Merz, K. M., Jr. *J Comput Aid Mol Des* 2002, 16, 105.
12. Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. *J Med Chem* 2006, 46, 1848.
13. Rubicon, 1997. www.daylight.com.
14. Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. *J Mol Graphics Modell* 1997, 15, 18.
15. Izrailev, S.; Zhu, F.; Agrafiotis, D. K. *J Comput Chem* 2006, 27, 1962.