

A Distance Geometry Heuristic for Expanding the Range of Geometries Sampled During Conformational Search

SERGEI IZRAILEV, FANGQIANG ZHU, DIMITRIS K. AGRAFIOTIS

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

Received 10 February 2006; Revised 19 April 2006; Accepted 24 April 2006

DOI 10.1002/jcc.20506

Published online 10 October 2006 in Wiley InterScience (www.interscience.wiley.com).

Abstract: A recent study of crystal structures of protein–ligand complexes has shown that bioactive conformations tend to be more extended than random ones (Diller and Merz, *J. Comput. Aid. Mol. Des.* 2002, 16, 105–112). Existing conformational sampling techniques produce molecular conformations with a distribution of geometric sizes that may not cover that of the bioactive conformations. Here, we describe a simple heuristic for biasing the conformational search toward more extended or compact conformations, while maintaining excellent sampling. The method uses a boosting strategy to generate a series of conformations, each of which is at least as extended (or compact) as the previous one. We demonstrate that this method significantly expands the range of geometric sizes generated during the search and thus increases the efficiency of sampling bioactive conformations.

© 2006 Wiley Periodicals, Inc. *J Comput Chem* 27: 1962–1969, 2006

Key words: conformational analysis; bioactive conformation; stochastic proximity embedding; distance geometry; boosting

Introduction

The generation of low-energy three-dimensional (3D) conformations of small organic molecules is a theme of central importance in computational drug design. Conformation generation algorithms fall into two broad categories: deterministic methods, which exhaustively enumerate all possible torsions at certain discrete intervals, and stochastic methods, which use a random element to explore the molecule's conformational space.¹ Although systematic search can be very effective for molecules with limited conformational flexibility, the exponential growth of the search space with the number of rotatable bonds, as well as problems associated with ring closures, limit its utility as a general conformational sampling technique.^{2–5} For flexible molecules, stochastic methods designed to sample low energy conformations represent a viable alternative. In its simplest form, a stochastic method randomly perturbs the current conformation of the molecule, minimizes it in energy, and repeats the process to generate a sequence of minimized conformations.^{6–8} Standard simulation techniques, such as molecular dynamics and Monte Carlo methods, have been used to generate an ensemble of conformations that lie in the low energy regions on the potential energy surface.^{9–12} All of these methods generate conformations in a continuous trajectory, in that each trial conformation is derived from the preceding one by a relatively small change. Because of this continuity, a large number of conformations are

generated between the important low energy ones, and a considerable amount of computer time is spent on the calculation and minimization of potential energies for these transitional conformations.

The conformations generated by conformational search are typically used as starting points for biomolecular docking, pharmacophore modeling, 3D database searching, 3D-QSAR, and other molecular modeling techniques. Most of these techniques attempt to model the interaction of small molecules with a biological target, typically a protein. Although a small molecule may adopt many different conformations in solution, it generally assumes a distinct conformation (usually referred to as “bioactive conformation”) when bound to a protein. Questions have been raised as to whether these bioactive conformations are similar to the conformations found in solution. This is of particular relevance to the problem of conformational search, because most of the sampling algorithms generate conformations in free space, in the absence of a target protein.

On the basis of a survey of 65 crystal structures of protein–ligand complexes, it was recently determined that bioactive conformations are usually considerably more extended (less compact) than randomly generated conformations in free space, as measured by a descriptor of molecular size such as the radius of gyra-

Correspondence to: D. K. Agrafiotis; e-mail: dagrafio@prds.jnj.com

tion (R_g).¹³ A recent report based on an examination of 510 crystal structures concluded that bioactive conformations have significantly higher energies than the corresponding energy minima.¹⁴ These studies cast doubt on the efficiency of conventional searching methods in locating bioactive conformations, and suggest that it may be desirable to bias the search to improve the sampling of conformational space near the bioactive conformation.¹³

Previously, we introduced a self-organizing algorithm called stochastic proximity embedding (SPE) for producing coordinates in a low-dimensional space that best preserve a set of distance constraints,¹⁵ and extended the method further to the problem of conformational sampling using a distance geometry formalism.¹⁶ In this article, we describe a simple boosting strategy that can be used in conjunction with SPE to bias the search toward more extended or more compact conformations. This method provides an effective means to sample all possible degrees of compactness for any given molecule.

Methods

Brief Overview of the SPE Conformational Sampling Algorithm

The SPE conformational sampling algorithm¹⁶ is based on distance geometry.^{17,18} Distance geometry has been well established over the past 20 years as a general and powerful method for generating molecular conformations.^{19,20} It has been successfully applied to a wide range of problems including conformational analysis,¹⁸ protein structure prediction,²¹ and ligand docking,²² and is the method of choice for NMR structure determination.^{23–25} Distance geometry generates molecular conformations that satisfy a set of geometric constraints. The connectivity and covalent structure of a molecule dictate that the distance between any pair of atoms (bonded and nonbonded alike) falls between certain lower and upper bounds. For bonded atoms (1,2 distances), the lower and upper bounds are identical and are set to the corresponding bond lengths. For atoms bonded to a third, common atom (1,3 distances), the bounds are also identical and can be derived by converting standard bond lengths and angles to specific distances. Similar calculations can be performed to derive specific distances for 1,4 nonbonded atoms in nonrotatable bonds. 1,4 distances in freely rotatable bonds can assume a minimum value defined by the syn form and a maximum value defined by the anti form. For longer topological distances (1,5 and beyond), the lower and upper bounds can be set to the sum of the van der Waals radii and the sum of the bond lengths along the shortest path connecting the two atoms, respectively. Collectively, these distance constraints are sufficient to define all possible 3D geometries attainable by a given molecule. By generating coordinates that satisfy these constraints, one should, in theory, be able to sample the entire conformational space.¹⁹ In a recent study, distance geometry was shown to identify conformations that were missed by alternative systematic search methods.²⁶

SPE is a very efficient algorithm for solving these distance constraints. It starts from a random initial configuration and uses a self-organizing scheme to rapidly refine the atomic positions

Table 1. PDB Codes, Molecule IDs, and the Number of Rotatable Bonds of the Molecules in the Data Set.

PDB code	Mol. ID	No. rot. bonds	PDB code	Mol. ID	No. rot. bonds
1apt	0	18	1hri	35	9
1b0h	1	15	1htf	36	12
1b1h	2	16	1hvr	37	8
1b2h	3	16	1hyt	38	5
1b3h	4	15	1icn	39	15
1b4h	5	15	1ida	40	15
1b5h	6	14	1lah	41	4
1b6h	7	15	1lic	42	15
1b7h	8	16	1lmo	43	6
1cbx	9	5	1lna	44	8
1etr	10	10	1lst	45	5
1hvi	11	19	1lyb	46	22
1hvj	12	19	1mcr	47	5
1hvk	13	19	1pgp	48	7
1hvl	14	19	1poc	49	23
1stp	15	5	1ppc	50	9
1tmn	16	13	1pph	51	7
1tni	17	4	1rne	52	21
1aaq	18	17	1sme	53	22
1apu	19	7	2cgr	54	9
1atl	20	9	2cmd	55	5
1baf	21	7	2lgs	56	4
1dwd	22	9	2plv	57	15
1eap	23	10	2r07	58	8
1eed	24	19	2sim	59	5
1epo	25	15	2yhx	60	3
1eta	26	5	3cpa	61	5
1ets	27	9	3tmn	62	6
1ett	28	7	3tpi	63	6
1fkg	29	10	4dfr	64	9
1glp	30	10	4phv	65	12
1glq	31	13	5p2p	66	20
1hef	32	19	8gch	67	7
1hfc	33	9			

so as to satisfy all the input constraints.¹⁶ The algorithm was described in detail in Ref. 16. The core of the algorithm is summarized below:

- Establishing distance constraints.* In this step, distance bounds matrices $\{l_{ij}\}$ and $\{u_{ij}\}$ are populated, where elements l_{ij} and u_{ij} denote the lower and upper bounds for the distance between atoms i and j , respectively. $\{l_{ij}\}$ and $\{u_{ij}\}$ are determined from the connectivity table and geometric parameters such as bond lengths, angles, and atomic radii. The interatomic distances r_{ij} in any feasible conformation must satisfy the distance constraints $l_{ij} \leq r_{ij} \leq u_{ij}$.
- Initialization.* In this step, the atoms are randomly placed in a 3D box.
- Self-organization.* A pair of atoms i and j are randomly selected, and their distance d_{ij} is computed. If the corresponding constraint $l_{ij} < d_{ij} \leq u_{ij}$ is not satisfied, the coordinates

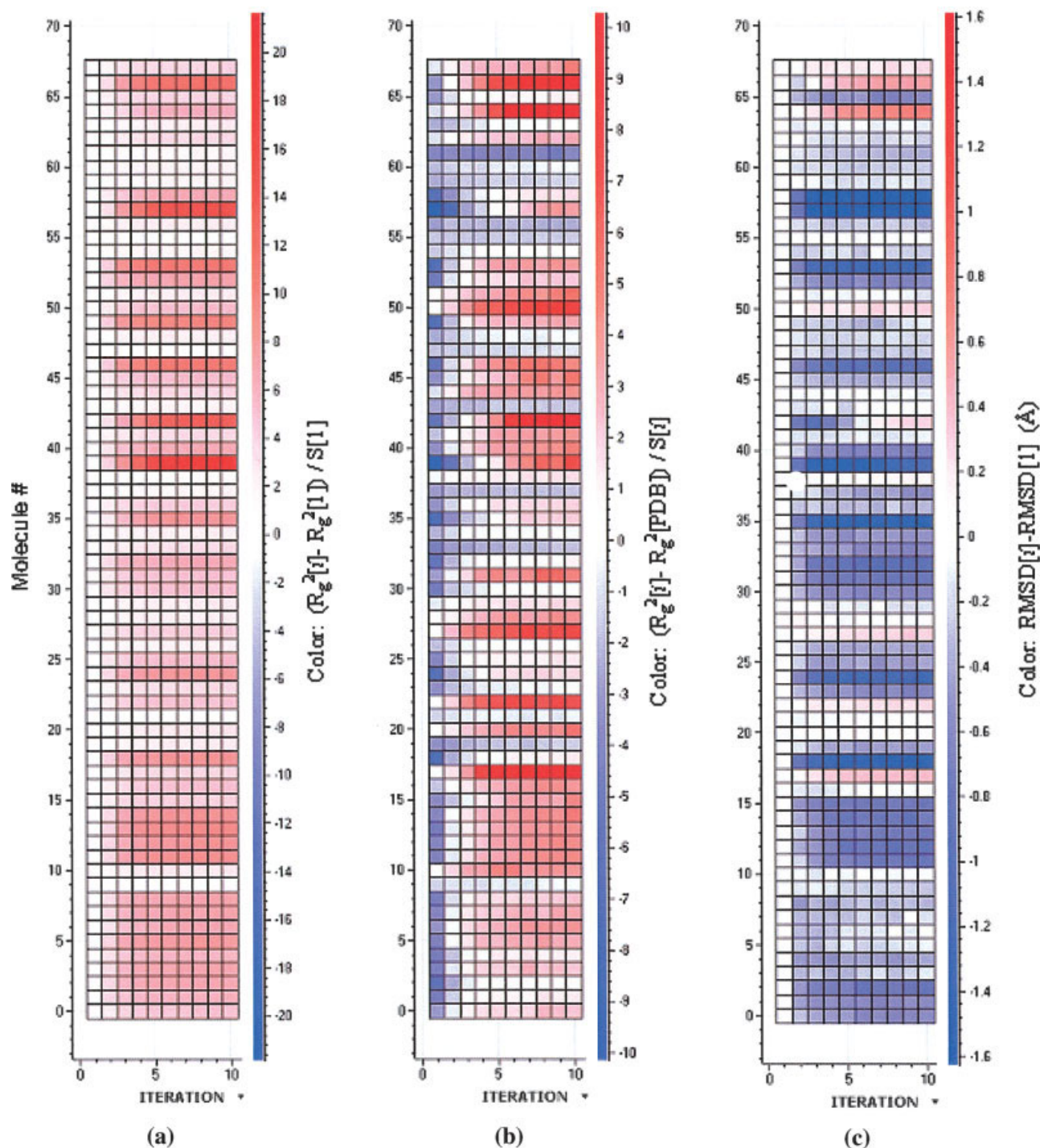


Figure 1. (a) Color map of $[\overline{R_g^2}(i) - \overline{R_g^2}(1)] / \text{Std}(1)$ for each molecule in iteration i , where $\overline{R_g^2}(i)$ and $\text{Std}(i)$ denote the mean and the standard deviation of the square radius of gyration, respectively, as explained in the text; (b) Color map of $[\overline{R_g^2}(i) - R_g^2(0)] / \text{Std}(i)$, where $R_g^2(0)$ denotes the square radius of gyration of the bioactive conformation, as found in the PDB structures. The color of each square on the maps represents the corresponding value for a specific molecule and iteration. (c) Color map of $\text{rmsd}(i) - \text{rmsd}(1)$, where $\text{rmsd}(i)$ denotes the average RMSD of the conformations generated at the i -th iteration with the bioactive conformation, as explained in the text. The color of each square on the map represents the corresponding value (in Å) for a specific molecule and iteration.

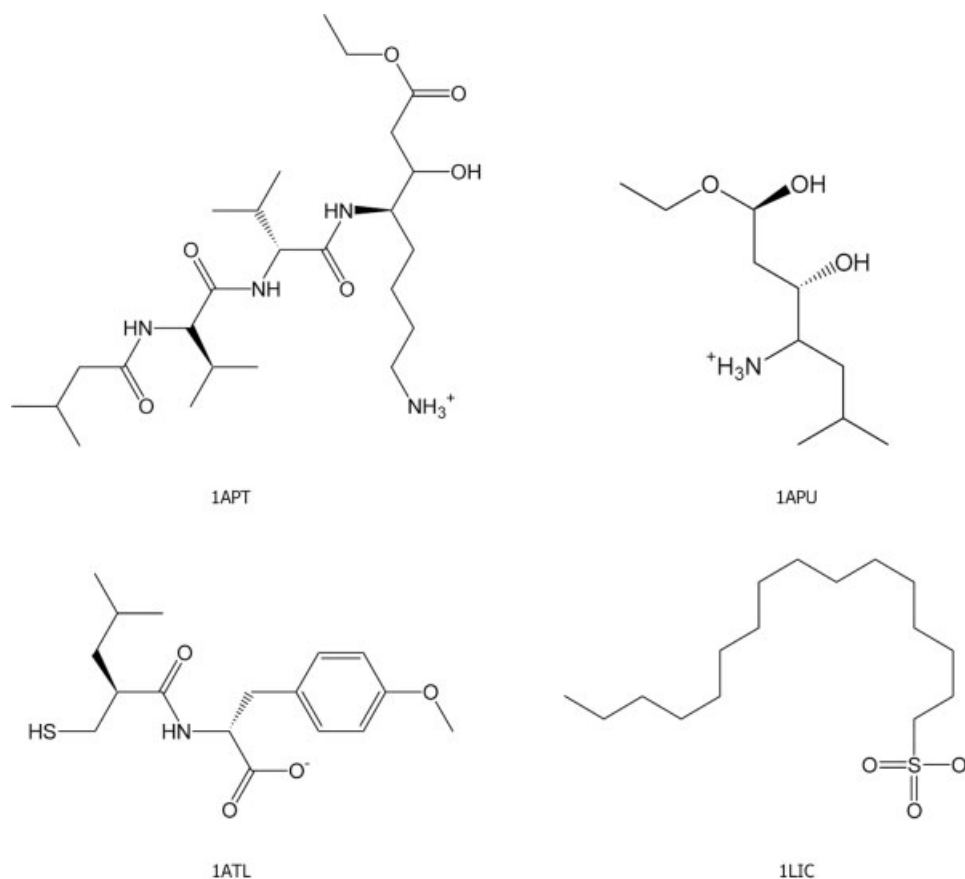


Figure 2. Chemical structures and the PDB codes of the four molecules used as examples.

of both atoms are updated to optimize the distance. This procedure is performed repeatedly until a reasonable conformation is obtained.

For the sake of simplicity, we did not cover some aspects of the algorithm earlier. For example, the algorithm uses some learning rates to control the self-organization process. Also, like in other well-known distance geometry methods, volume constraints are used to enforce chirality and planarity. We refer the reader to Ref. 16 for a complete description, validation, and performance benchmark of the algorithm, as well as a comparison to other distance geometry methods.

Biasing Heuristic

The SPE conformational sampling algorithm generates conformations that satisfy a set of distance constraints as specified by the lower and upper bounds $\{l_{ij}\}$ and $\{u_{ij}\}$. This suggests a simple method to bias the extendedness of the generated conformations. For example, by increasing the values of the lower bounds, one could enforce an increase of the interatomic distances, thereby generating more extended conformations. Such artificial modification, however, may be problematic, because when

the distance bounds are tightened, the existence of a geometry satisfying these constraints is not always guaranteed.

Here, we propose a heuristic that adopts a safeguarded strategy to update the bounds. For the sake of conciseness, we describe only the procedure to generate more extended conformations below, but note that more compact conformations can also be generated in a very similar fashion.

In our heuristic, increasingly extended conformations are generated through iterations. In the first iteration, a normal SPE procedure is performed as described in the previous section, generating a reasonable conformation. We then measure the actual distances between each atom pair in this conformation, and use these values to update the lower bound matrix $\{l_{ij}\}$. In the second iteration, the updated $\{l_{ij}\}$ and the unchanged upper bound matrix $\{u_{ij}\}$ are used as the distance constraints, and the initialization and self-organization steps (steps (b) and (c) in the previous section) of SPE are performed. Similarly, in the beginning of any further iteration, the lower bounds are updated by the actual interatomic distances in the conformation generated in the preceding iteration.

We note that in our heuristic, the lower bounds of the distance constraints in any iteration are always equal to or greater than those in the previous iterations. As a result, the interatomic distances in the generated conformations should always be non-

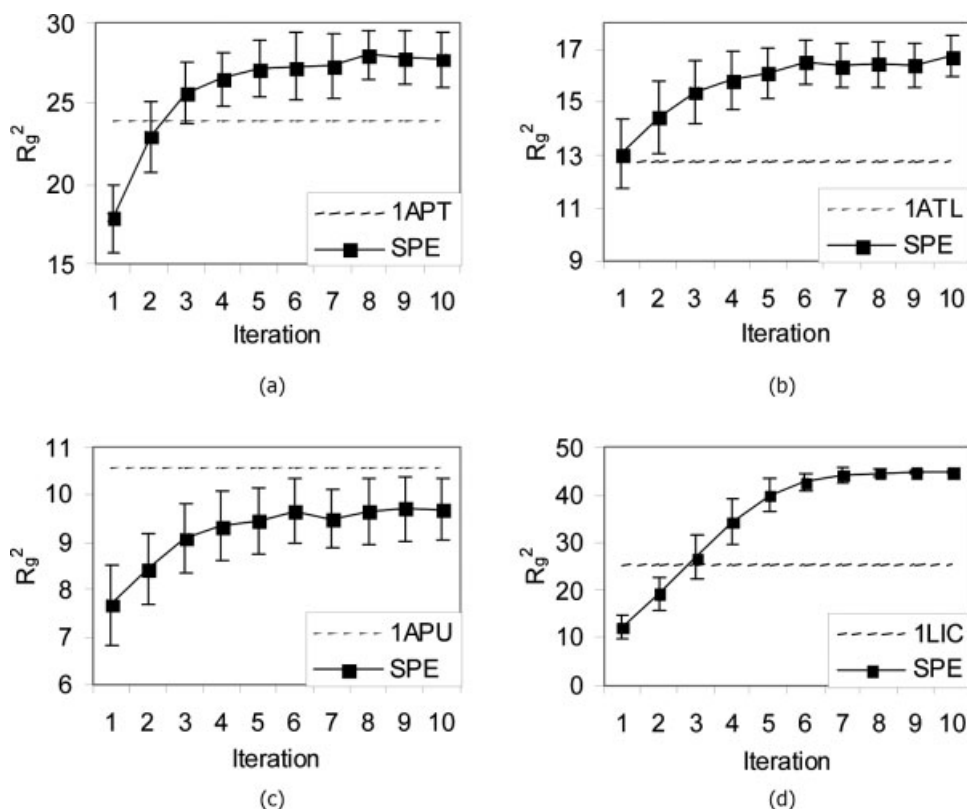


Figure 3. $\bar{R}_g^2(i)$ and $\text{Std}(i)$ for the conformations generated at each iteration for the four example molecules. $R_g^2(0)$ for the bioactive conformation is presented by the dotted line in each graph.

decreasing over the iterations, and more and more extended conformations should be generated. Moreover, our heuristic will never yield a set of distance constraints that are impossible to satisfy, because there already exists a conformation (i.e., the one generated in the preceding iteration) that satisfies them. Therefore, the conformational space defined by the distance constraints will shrink, but not vanish over the iterations, thus effectively biasing the SPE sampling toward more extended conformations.

Testing of the Heuristic

The heuristic was tested on a data set of 68 small molecule conformations extracted from the Protein Data Bank (PDB),^{27,28} which was a slightly updated version of the set used by Diller and Merz.¹³ The molecules contain between 3 and 23 rotatable bonds, as presented in Table 1. For each molecule, a conformation was generated by the SPE algorithm, using standard geometric constraints.²⁹ This conformation was labeled as iteration 1. Then, for each atom pair i and j with their actual distance r_{ij} in this conformation satisfying $l_{ij} \leq r_{ij} < u_{ij}$, we updated l_{ij} to the value of r_{ij} . Another conformation was generated using the new set of constraints, and labeled as iteration 2, and so on, until 10 iterations were performed. This process was repeated 100 times to produce a total of 1000 conformations for each molecule, 100 conformations at each of the 10 iterations. Note that each conformation was generated from a random set of initial coordinates, using a set of

geometric constraints that depended only on a single conformation generated at the previous iteration. The square radius of gyration R_g^2 , defined as the mean square distance between the atoms and the center of mass of the molecule, was computed for each conformation, as well as the root mean square deviation (RMSD) of each conformation from the bioactive conformation found in the PDB structure. All calculations were carried out without explicit hydrogens, and no energy minimization was performed.

The algorithm was implemented in the C++ programming language. All calculations were performed on an IBM Intellistation workstation running Windows XP Professional and equipped with two 3.2 GHz Xeon processors and 2048 Mb of RAM.

Results and Discussion

For each molecule, the mean and the standard deviation of the square radius of gyration over the 100 conformations generated in the i th iteration, denoted as $\bar{R}_g^2(i)$ and $\text{Std}(i)$, respectively, were calculated. Since a normal SPE sampling without bias is performed in iteration 1, $\bar{R}_g^2(1)$ and $\text{Std}(1)$ characterize the “natural” distribution of compactness for the molecule. The difference between $\bar{R}_g^2(i)$ and $\bar{R}_g^2(1)$, scaled by $\text{Std}(1)$ is visualized in Figure 1a. As expected, $\bar{R}_g^2(i)$ increases with the iteration number i for every single molecule, indicating that as the number of boosts increases, more and more extended conformations

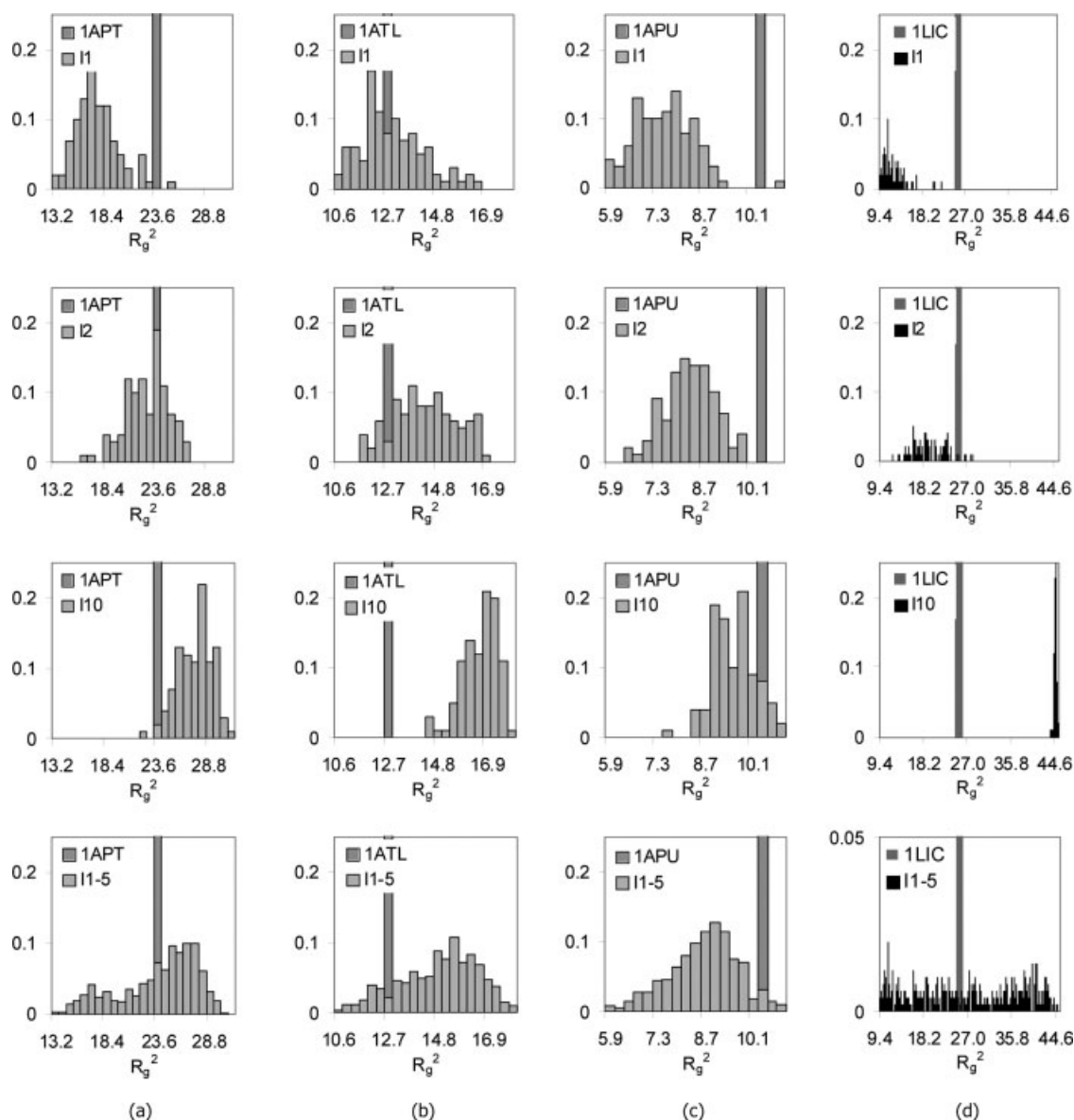


Figure 4. The distributions of the square radius of gyration R_g^2 of the four example molecules' conformations generated at iteration 1 (labeled I1), iteration 2 (labeled I2), iteration 10 (labeled I10), and iterations 1 through 5 combined (labeled I1–5). Each plot displays the value of R_g^2 for the bioactive conformation, labeled with the respective PDB code. The y-axis is the fraction of the conformations found with the given R_g^2 .

are generated. Let $R_g^2(0)$ denote the square radius of gyration of the bioactive conformation found in the crystal structure. The plot of the differences between $\overline{R_g^2}(i)$ and $R_g^2(0)$ scaled by $\text{Std}(i)$, as shown in Figure 1b, reveals that while the conformations of most molecules generated in iteration 1 are significantly more compact than the bioactive conformations, the conformations generated in iteration 10 are significantly more extended. Furthermore, with few exceptions, one can find for each molecule one or more iterations in which the $\overline{R_g^2}(i)$ closely matches $R_g^2(0)$, indicating that the desired region of compactness has been sampled in these iterations.

Since RMSD is a measure of the similarity between two conformations, we calculated this value for each generated conformation and the corresponding bioactive conformation in the crystal structure, after superimposing the two. In this calculation, all equivalent mappings of the molecule due to its internal symmetry were enumerated, and the RMSD was taken from the mapping with the best superimposition. The mean of the RMSD values in iteration i , $\overline{\text{rmsd}}(i)$, over the 100 generated conformations was computed, and the difference between $\overline{\text{rmsd}}(i)$ and $\overline{\text{rmsd}}(1)$ was visualized in Figure 1c for each molecule. As shown in the figure, most molecules have $\overline{\text{rmsd}}(i)$ values in later

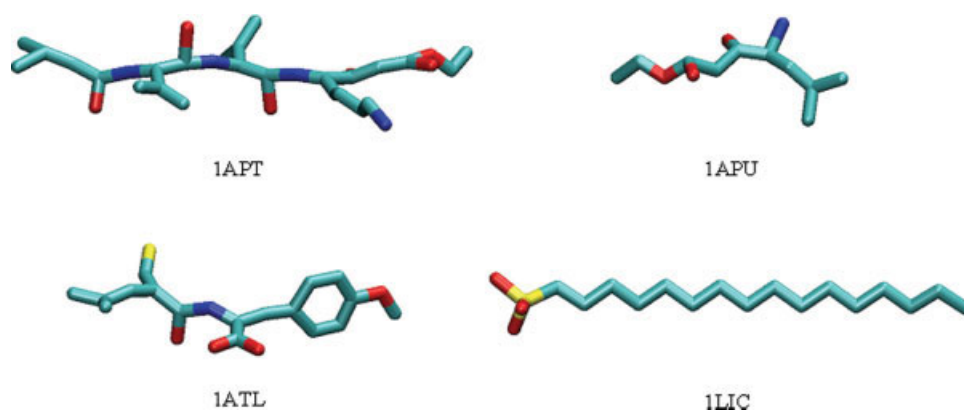


Figure 5. Conformations of the four molecules (see Fig. 2) generated from iteration 10 of the sampling. The images were rendered using VMD.³⁰ [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

iterations ($i > 1$) smaller than $\overline{\text{rmsd}}(1)$, which indicate that the boosting heuristic improves the efficiency of sampling the bioactive conformations.

We now take a closer look at four exemplary molecules from our dataset, whose chemical structures are given in Figure 2. The mean and standard deviation of R_g^2 in each iteration for these molecules are plotted in Figure 3, showing that $\overline{R_g^2}(i)$ generally increases as the number of boosts increases. Furthermore, although the $\overline{R_g^2}(1)$ values (except for 1ATL) in iteration 1 are well below $R_g^2(0)$ (as indicated by the dotted lines in the figure), within a few iterations, the $\overline{R_g^2}(i)$ values either exceed $R_g^2(0)$ (1APT and 1LIC), or approach it much more closely (1APU). The distributions of R_g^2 in iterations 1, 2, 10 are shown in the first three rows of Figure 4, respectively. From that figure, one can discern a shift of the distribution toward higher R_g^2 values as the iterations increase, and, consequently, an enhanced sampling of the region where the bioactive $R_g^2(0)$ lies. As a practical question for our heuristic, one may wonder how many iterations should be performed to achieve a reasonable sampling. From Figure 3, we see that the $\overline{R_g^2}(i)$ curves generally become flat and saturated after about five iterations. Moreover, in all cases the distributions of R_g^2 over all conformations generated in the first five iterations, as shown in the last row of Figure 4, have reasonable overlap with $R_g^2(0)$. This suggests that in practice, performing five iterations may be sufficient to sample the bioactive conformation.

The three-dimensional structures of a randomly chosen conformation generated in iteration 10 for each of the four molecules depicted in Figure 5 visually confirm that they are all highly extended. In particular, the conformation of 1LIC is a fully extended chain, which represents the most extended conformation possible under the intrinsic distance constraints. Figure 3d also shows that $\overline{R_g^2}(i)$ for this particular molecule reaches a plateau in the last few iterations and that the standard deviation becomes very small, which indicates that almost all of the sampled conformations have converged to the maximally extended conformation. Most importantly, all the conformations depicted in Figure 5 are geometrically sensible even though they were derived from purely geometrical considerations without any energy minimization.

To find a bioactive conformation, the conformational space with the correct compactness (characterized by R_g^2) must be sufficiently sampled. This can be a very difficult task for unbiased sampling, especially when the molecule is highly flexible and can assume conformations with a wide range of compactness. For example, as shown in Figure 3d, the $R_g^2(0)$ of 1LIC lies more than three standard deviations above the mean of the “natural” distribution in iteration 1, which means that roughly only 3 out of 1000 generated conformations would have a similar or higher R_g^2 value. The unbiased search is therefore highly inefficient. On the other hand, we have demonstrated that a few iterations of biased sampling extend the explored region of compactness to cover the bioactive conformation, therefore significantly enhancing the efficiency of the search.

For some branched molecules, the conformation may be extended in multiple directions. While a single series of iterations in our heuristic would extend the conformations in one direction, other extending directions can be explored by repeating the process multiple times, each starting from a random initial geometry, as described in this study.

In our current heuristic, the distance bounds of all atom pairs in the molecule are updated at the end of an iteration. This is a simple and general strategy that can be automatically applied to all molecules, and it is shown to work well for the molecules tested in this study. Alternatively, with specific knowledge of the molecules, one may also choose to update the bounds of only a subset of atom pairs, e.g., those at the extremity of the molecules, or those between functional groups of interest. In fact, doing so can improve the effectiveness of the heuristic for some molecules, such as those with macrocycles, where stretching along one direction is possible only when interatomic distances in the orthogonal direction are allowed to decrease.

The radius of gyration of the bioactive conformation is usually unknown when one performs a conformational search. Although the majority of the bioactive conformations under study tend to be extended, there are some molecules that adopt more compact bioactive conformations as compared to random. We note that our heuristic is also capable of biasing toward generating more compact conformations (with smaller R_g^2 values).

Therefore, it may be advisable to bias the sampling in both directions, to ensure that the compactness of the bioactive conformation is well covered. Finally, although this heuristic is based on the SPE algorithm, it is perfectly applicable to other distance geometry methods as well. We believe that other sampling algorithms may also benefit from modifications that bias the search to obtain conformations with a wider range of compactness.

Conclusions

In this article, we propose a simple heuristic to bias conformational sampling toward more extended or more compact conformations. The heuristic can be easily implemented, and provides a method of sampling the full range of compactness of molecular conformations during the search. Since the compactness of a bioactive conformation often lies outside the range sampled by an unbiased search, this heuristic significantly improves the chances of finding bioactive conformations.

Acknowledgments

We thank Dr. David J. Diller of Pharmacoepa, Inc. for providing the data set and Dr. Huafeng Xu of D. E. Shaw and Co. for his earlier work on SPE and for his critical review of this manuscript.

References

1. Leach, A. R. A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: New York, 1991; Vol. 2.
2. Lipton, M.; Still, W. C. *J Comput Chem* 1988, 9, 343.
3. Bruccoleri, R. E.; Karplus, M. *Biopolymers* 1987, 26, 137.
4. Bruccoleri, R. E.; Karplus, M. *Macromolecules* 1985, 18, 2767.
5. Go, N.; Scheraga, H. A. *Macromolecules* 1970, 3, 178.
6. Saunders, M. *J Am Chem Soc* 1987, 109, 3150.
7. Ferguson, D. M.; Raber, D. J. *J Am Chem Soc* 1989, 111, 4371.
8. Chang, G.; Guida, W. C.; Still, W. C. *J Am Chem Soc* 1989, 111, 4379.
9. Auffinger, P.; Wipff, G. *J Comput Chem* 1990, 11, 190.
10. Bruccoleri, R. E.; Karplus, M. *Biopolymers* 1990, 29, 1847.
11. Li, Z.; Scheraga, H. A. *Proc Natl Acad Sci USA* 1987, 84, 6611.
12. Jorgensen, W. L.; Tirado-Rives, J. *J Phys Chem* 1996, 100, 14508.
13. Diller, D. J.; Merz, K. M., Jr. *J Med Chem* 2002, 16, 105.
14. Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. *J Chem Inf Model* 2005, 45, 422.
15. Agrafiotis, D. K.; Xu, H. *Proc Natl Acad Sci USA* 2002, 99, 15869.
16. Xu, H.; Izrailev, S.; Agrafiotis, D. K. *J Chem Inf Comput Sci* 2003, 43, 1186.
17. Crippen, G. M. *J Comput Phys* 1978, 26, 449.
18. Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. *J Mol Graph Model* 1997, 15, 18.
19. Blaney, J. M.; Dixon, J. S. *Perspect Drug Discov Design* 1993, 1, 301.
20. Feuston, B. P.; Miller, M. D.; Culberson, J. C.; Nachbar, R. B.; Kearsley, S. K. *J Chem Inf Comput Sci* 2001, 41, 754.
21. Huang, E. S.; Samudrala, R.; Ponder, J. W. *Prot Sci* 1998, 7, 1998.
22. Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. *Proteins* 1993, 17, 266.
23. Havel, T. F.; Wüthrich, K. *J Mol Biol* 1985, 182, 281.
24. Kuszewski, J.; Nilges, M.; Brünger, A. T. *J Biomol NMR* 1992, 2, 33.
25. Mumenthaler, C.; Braun, W. *J Mol Biol* 1995, 254, 465.
26. Martin, E. J.; Hoeffel, T. J. *J Mol Graph Model* 2000, 18, 383.
27. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res* 2000, 28, 235.
28. Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. *Nat Struct Biol* 2000, 7, 957.
29. Rubicon, 1997. www.daylight.com.
30. Humphrey, W.; Dalke, A.; Schulten, K. *J Mol Graph* 1996, 14, 33.