

Conformational Boosting

Dimitris K. Agrafiotis,^{A,D} Alan Gibbs,^A Fangqiang Zhu,^A
Sergei Izrailev,^{A,B} and Eric Martin^C

^A Johnson & Johnson Pharmaceutical Research & Development, 665 Stockton Drive, Exton, PA 19341, USA.

^B Current address: Fortress Investment Group, 1345 Avenue of the Americas, New York, NY 10105, USA.

^C Chiron Corporation, 4560 Horton Street, Emeryville, CA 94608, USA.

^D Corresponding author. Email: dagrafio@prdus.jnj.com

Stochastic proximity embedding (SPE) is a novel self-organizing algorithm for sampling conformational space using geometric constraints derived from the molecular connectivity table. Here, we describe a simple heuristic that can be used in conjunction with SPE to bias the conformational search towards more extended or compact conformations, and thus greatly expand the range of geometries sampled during the search. 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. The approach is compared to several popular conformational sampling techniques using a reference set of 59 bioactive ligands extracted from the Protein Data Bank, and is shown to be significantly more effective in sampling the full range of molecular radii, with the exception of the *Catalyst* program, which was equally effective.

Manuscript received: 21 June 2006.

Final version: 21 October 2006.

Most organic molecules of non-trivial size can assume a multitude of different three-dimensional (3D) conformations. Identifying which of these conformations are relatively stable and likely to be populated at room temperature has been the subject of countless studies in the computational chemistry literature.^[1] This problem is particularly critical in computer-assisted drug design. Recent studies of crystal structures of protein–ligand complexes have shown that bioactive conformations tend to be more extended than random ones,^[2] and may lie several kcal mol⁻¹ higher in energy than their respective global minima.^[3] Since the bioactive conformation of a ligand also depends on the geometry of its host, it is imperative that the search for conformational minima casts a wide net over the potential energy surface. Several applications depend critically on the diversity of conformations sampled during the search, including protein docking, pharmacophore modelling, 3D database searching, and 3D quantitative structure–activity relationship (QSAR) methods, to name a few.

Conformation generation algorithms fall into two broad categories—deterministic, which exhaustively enumerate all possible torsions at certain discrete intervals, and stochastic, which use a random element to explore the molecule's conformational space. 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.^[4–7] For flexible molecules, stochastic

methods designed to sample low energy conformations represent a viable alternative.^[8–10] A common problem with many stochastic techniques, however, particularly those based on molecular dynamics,^[11–14] is that they spend a considerable amount of time generating and minimizing many transitional conformations along the trajectories connecting the local minima.

Recently, 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,^[15,16] and extended the method further to the problem of conformational sampling^[17] using a distance geometry formalism.^[18,19] SPE attempts to generate conformations that satisfy a set of interatomic distance constraints. These constraints are derived from the molecular connectivity table, and are encoded in the form of upper (u_{ij}) and lower (l_{ij}) bounds for every possible interatomic distance d_{ij} (such that $l_{ij} \leq d_{ij} \leq u_{ij}$). Distance constraints are usually supplemented by volume constraints to enforce planarity of conjugate systems and correct chirality of stereocenters. By generating coordinates that satisfy these constraints, one should, in theory, be able to sample the entire conformational space. In a recent study, distance geometry was shown to identify conformations that were missed by alternative systematic search methods.^[20]

SPE is a very efficient algorithm for solving these constraints. The method starts from a random initial configuration and gradually refines it by repeatedly selecting an individual constraint at random, and updating the respective

atomic coordinates towards satisfying that specific constraint. This procedure is performed repeatedly until a reasonable conformation is obtained. A detailed description of the algorithm can be found in ref. [17].

Here, we describe a simple boosting strategy that can be used in conjunction with SPE to bias the search towards more extended or more compact conformations, and thus extend the range of geometries sampled during the search. The method generates increasingly extended conformations through a series of embeddings, each seeded on the result of the previous one. In the first iteration, a normal SPE embedding is performed as described above, generating a reasonable conformation c_1 . The lower bounds of all atom pairs $\{l_{ij}\}$ are then replaced by the actual interatomic distances $\{d_{ij}\}$ in conformation c_1 , and used along with the unchanged upper bounds $\{u_{ij}\}$ to perform a second embedding to generate another conformation, c_2 . This process is repeated for a prescribed number of iterations. The lower bounds are then restored to their original default values, and a new sequence of embeddings is performed using a different random number seed. Because the distance constraints in any iteration are always equal to or greater than those in the previous iterations, successively more extended conformations should be generated. An analogous procedure can be used to generate increasingly compact conformations. More details can be found in ref. [21]. We should note that the strategy described above is not related to the boosting approaches employed in the context of machine learning using predictor ensembles.

The method is here compared to seven other conformational sampling programs using 59 ligands^[2] extracted from the Protein Data Bank,^[22,23] containing 3–23 rotatable bonds. These techniques included classical *SPE* (without boosting), *Rubicon*, *Omega*, *Macromodel*, two variants of *Catalyst*, and two variants of *MOE*. The specific parameters used for each one are as follows:

SPE – A total of 10000 conformers were generated for each molecule, using the default parameters described in ref. [17].

Boosting SPE – A total of 10000 conformations were generated for each molecule, using the same defaults. 6250 of these conformations were generated by tightening the lower bounds (i.e. towards more extended conformations), using 1250 independent trials of five boosted conformers per trial. The remaining 3750 conformations were generated by tightening the upper bounds (i.e. towards more compact conformations), using 1250 trials of only three boosted conformers per trial.

Rubicon – *Rubicon*^[24] uses a distance geometry method very similar in principle to *SPE*. Like *SPE*, *Rubicon* sets upper and lower bounds for distances and volumes, but uses a different algorithm to find atomic coordinates that satisfy those bounds. Rather than selecting random starting coordinates, *Rubicon* selects a random set of distances within the bounds and uses the ‘metric matrix’ algorithm^[25] to generate approximate 3D coordinates, followed by conjugate gradient minimization to refine these coordinates so they satisfy the original bounds. A total

of 10000 conformers were generated for each molecule, using one trial per conformation, no hydrogen atoms, 1–4 bump checking enabled, redundancy checking disabled, and default values for all the remaining parameters.

Catalyst – Two sets of conformers were generated using the BEST conformational sampling option in *Catalyst* ver. 4.10. The first set used *Catalyst*’s ‘poling’ technique, which introduces an artificial potential during the sampling to repel similar conformers, thereby promoting conformational variation. 500 conformers were requested for each molecule, though the actual number produced ranged from 43 to 456. The second set of conformers was obtained with poling disabled. In this case, 10000 conformers were requested, though for some molecules fewer conformers were produced. Duplicate checking was disabled, and all other parameters were kept at their default values.

Omega – *Omega* divides each molecule into component fragments, which may contain up to five contiguous rotatable bonds. A library of predefined angles is used to generate conformations for each fragment, which are then assembled to construct the conformations of the whole molecule using a depth-first, divide-and-conquer approach, driven by the fragment energies. 10000 conformers were requested for each molecule, but the actual number generated varied. Default values were used for all the remaining parameters.

Macromodel – A total of 10000 conformers were generated for each molecule using the serial torsional/low-mode conformational sampling method in *Macromodel* ver. 9.0016. This is a hybrid technique that combines broad Monte Carlo sampling of torsional space^[26] with local low-frequency eigenvector sampling in the vicinity of the current conformation.^[27] Since minimization could not be decoupled from the actual search, a single cycle of truncated Newton conjugate gradient^[28] minimization was performed on each conformer using the *MMFF94s* force field (i.e. the resulting conformers were effectively not minimized). No energy cutoff was applied to discard unreasonable conformations. Default values were used for all other parameters.

MOE – Two sets of conformers were generated. The first was obtained using *MOE*’s stochastic conformational search, a variant of Ferguson’s Random Incremental Pulse Search method.^[29] 10000 conformations were requested for each molecule. The resulting geometries were minimized using *MMFF94s* with the distance-dependent dielectric disabled. All minimized conformations with energy values above 500 kcal mol⁻¹ from the global minimum were discarded. 90 successive failed attempts at generating new conformations before termination was applied in order to maximize the number of conformers produced within the specified energy window. All other parameters were set to their default values. The second set of conformers was generated using a fragment-based approach. This algorithm breaks the molecule up into overlapping fragments, retrieves pre-computed conformations of those fragments, and re-assembles them by rigid body superimposition. In this work, only fragments with strain energy less than

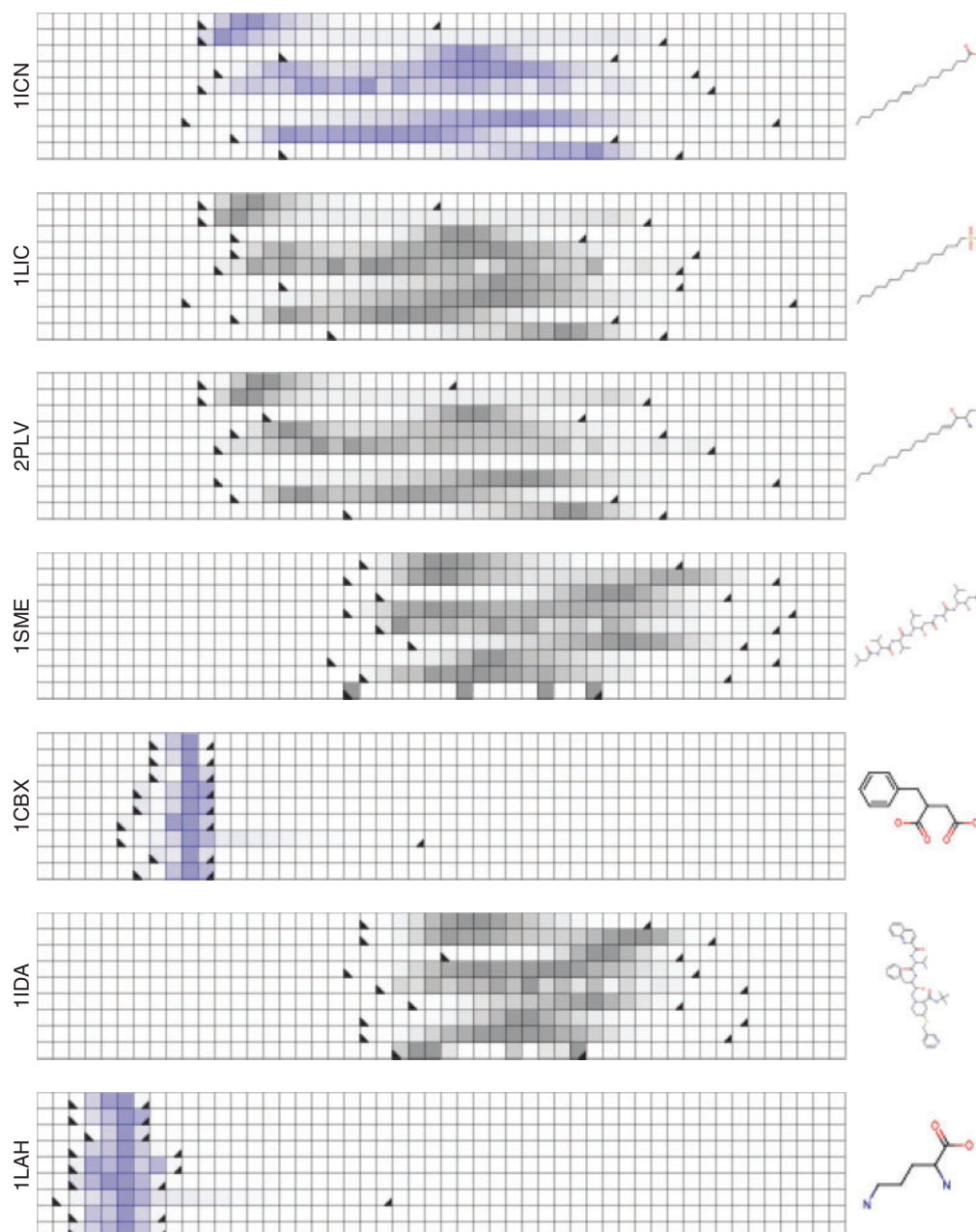


Fig. 1. Distribution of radius of gyration of the conformations generated by each method for each molecule. All rows are scaled identically from 1.5 to 8.5, which covers the full range of molecular radii encountered across all conformers, all molecules, and all methods. The intensity of the shade is proportional to the fraction of conformations that fall within each bin. The black triangles indicate the bins containing the minimum and maximum radius for that particular method and molecule. Within each molecule panel, the methods appear, from top to bottom: *SPE*, boosting *SPE*, *Rubicon*, *Catalyst*, poling *Catalyst*, *Omega*, *Macromodel*, *MOE* stochastic, and *MOE* systematic.

5 kcal mol^{-1} were considered. Once again, 10000 conformations were requested for each molecule. All other parameters were set to their default values.

The results are summarized in Fig. 1. The Figure contains a series of panels, one for each molecule, which show the distribution of the radius of gyration (a measure of extend- edness) for all the conformers obtained by each method. The histograms are identically scaled from 1.5 to 8.5, which covers the full range of radii encountered across all conformers, all molecules, and all methods.

SPE tends to produce significantly more compact conformations compared to the other methods, including its close relative, *Rubicon*. This is not entirely surprising if one considers the way in which *SPE* refines the atomic coordinates. Because the embedding starts from a random initial configuration, many topologically neighboring atoms have to approach each other from opposite directions. A typical *SPE* optimization proceeds through an initial contraction phase where atoms cross each other to get close to their neighbors, followed by an expansion phase where the coordinates are relaxed to satisfy all the distance constraints. There is

no intrinsic force in *SPE* driving the system to an extended geometry, as there are typically many geometrically feasible compact conformations that are more easily accessible from these intermediate 'imploded' states. The algorithm will generate extended conformations only if these are enforced by the distance constraints. Boosting is an effective strategy to achieve this goal.

Interestingly, *Rubicon* does not suffer from this problem, perhaps because it starts from the nearest 'linear' 3D embedding of initial random distances, rather than from random initial coordinates. Also, *Rubicon's* conjugate gradient minimization works in two phases—it first minimizes bounds violations in four dimensions (which allows atoms to pass through each other), and then minimizes the fourth dimension towards zero, collapsing the conformation into three dimensions. The use of a fourth dimension increases the effective volume where atoms can move, causing the molecule to expand more freely. It will be interesting to see if this approach would work equally well with the *SPE* stochastic minimizer (work in progress). In general, the two methods explore different, complementary regions of conformational space. The range of radii sampled by either of them is relatively narrow, with *SPE* offering a slight advantage except for long, linear molecules such as 1ICN, 1LIC, and 2PLV, and to a lesser extent 1SME.

SPE's affinity for compact geometries is compensated by the use of boosting, which greatly expands the range of conformations sampled, and covers the broadest spectrum of geometric sizes of all the methods examined, except *Catalyst*, which we found to be equally effective. Interestingly, the heatmaps in Fig. 1 show a characteristic bimodal distribution for every molecule of considerable flexibility. This is not an intrinsic limitation of our method, but rather an artifact of our decision to explore the effect of boosting in both directions (towards more extended as well as more compact conformations). As it turns out, the latter is unnecessary, since *SPE* already shows a strong bias towards compact conformations. The bimodal effect is not observed when boosting is used only towards more extended conformations (results not shown).

Of all the other methods, *Catalyst* emerges as the most competitive to boosting *SPE*. It covers an equally broad range of geometric sizes, and identifies comparable minima and maxima in terms of extendedness for virtually every molecule. Poling works as expected, producing just as broad of a sampling but with significantly fewer conformations. *Omega* shows a strong preference for extended conformations, in contrast to *MOE's* stochastic search, which is comparable to conventional *SPE*. *MOE's* systematic search yields greater diversity, with a notable preference for extended conformations. (We should note, parenthetically, that we had considerable difficulty getting *Omega* to work for many of these molecules, and failed to generate any conformers for two ordinary compounds, 1ICN and 2PLV.)

Perhaps the most striking feature of Fig. 1 is the substantially wider range of radii sampled by *Macromodel*. Closer inspection reveals that unlike conformations generated by

other methods the raw conformations produced by *Macromodel* are severely distorted (results not shown), most likely as a result of low-mode sampling. Indeed, *Macromodel* relies primarily on energy minimization to generate chemically sensible geometries. It is unclear whether this form of sampling offers any significant advantages over much simpler and faster procedures, such as Saunders' stochastic search, particularly in light of the significant computational overhead associated with computing the eigenvectors.

We have chosen to generate on the order of 10000 conformations by each method to avoid the possibility of the results suffering from undersampling. However, it may be possible to generate a comparable set of conformations by applying the boosting heuristic with fewer conformers generated in each iteration. An even smaller number of conformations actually tested by computationally intensive applications such as docking and 3D-QSAR could be obtained from a large number of generated conformers, for example, by selecting a smaller diverse set of conformations covering the whole range of radii of gyration.

Although the current analysis is based on raw, non-minimized conformations, it is clear that our boosting strategy greatly increases the sampling capacity of *SPE*. Boosting is a general strategy that is likely to benefit any distance geometry method. It can be easily implemented, and provides a way of sampling regions of conformational space that may not be easily accessible by alternative methods. Since bioactive conformations tend to be extended and often fall outside the range sampled by an unbiased search, this heuristic significantly improves the chances of finding such conformations. This conclusion is also supported by the fact that for *SPE* with boosting, conformations in later iterations were found to have lower root-mean-squared deviation from the bioactive conformations in the Protein Data Bank crystal structures.^[21] Our ultimate plan is to complete this analysis with geometries minimized using the same force field and minimization algorithm, but as of this writing these calculations are still in progress. A more detailed report is forthcoming.^[30]

Accessory Publication

A fuller version of Fig. 1, featuring all 59 bioactive ligands, is available from the author or, until December 2011, the *Australian Journal of Chemistry*.

Acknowledgments

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

References

- [1] A. R. Leach, in *Reviews in Computational Chemistry* (Eds K. B. Lipkowitz, D. B. Boyd) **1991**, Vol. 2 (VCH: New York, NY).
- [2] D. J. Diller, K. M. Merz, Jr, *J. Comput. Aid. Mol. Des.* **2002**, *16*, 105.

- [3] J. Kirchmair, C. Laggner, G. Wolber, T. Langer, *J. Chem. Inf. Model.* **2005**, *45*, 422. doi:10.1021/CI049753L
- [4] M. Lipton, W. C. Still, *J. Comput. Chem.* **1988**, *9*, 343. doi:10.1002/JCC.540090409
- [5] R. E. Bruccoleri, M. Karplus, *Biopolymers* **1987**, *26*, 137. doi:10.1002/BIP.360260114
- [6] R. E. Bruccoleri, M. Karplus, *Macromolecules* **1985**, *18*, 2767. doi:10.1021/MA00154A069
- [7] N. Go, H. A. Scheraga, *Macromolecules* **1970**, *3*, 178. doi:10.1021/MA60014A012
- [8] M. Saunders, *J. Am. Chem. Soc.* **1987**, *109*, 3150. doi:10.1021/JA00244A051
- [9] D. M. Ferguson, D. J. Raber, *J. Am. Chem. Soc.* **1989**, *111*, 4371. doi:10.1021/JA00194A034
- [10] G. Chang, W. C. Guida, W. C. Still, *J. Am. Chem. Soc.* **1989**, *111*, 4379. doi:10.1021/JA00194A035
- [11] P. Auffinger, G. Wipff, *J. Comput. Chem.* **1990**, *11*, 190. doi:10.1002/JCC.540110103
- [12] R. E. Bruccoleri, M. Karplus, *Biopolymers* **1990**, *29*, 1847. doi:10.1002/BIP.360291415
- [13] Z. Li, H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6611. doi:10.1073/PNAS.84.19.6611
- [14] W. L. Jorgensen, J. Tirado-Rives, *J. Phys. Chem.* **1996**, *100*, 14508. doi:10.1021/JP960880X
- [15] D. K. Agrafiotis, H. Xu, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 15869. doi:10.1073/PNAS.242424399
- [16] D. K. Agrafiotis, *J. Comput. Chem.* **2003**, *24*, 1215. doi:10.1002/JCC.10234
- [17] H. Xu, S. Izrailev, D. K. Agrafiotis, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1186. doi:10.1021/CI0340557
- [18] G. M. Crippen, *J. Comput. Phys.* **1978**, *26*, 449. doi:10.1016/0021-9991(78)90081-5
- [19] D. C. Spellmeyer, A. K. Wong, M. J. Bower, J. M. Blaney, *J. Mol. Graphics Modell.* **1997**, *15*, 18. doi:10.1016/S1093-3263(97)00014-4
- [20] E. J. Martin, T. J. Hoeffel, *J. Mol. Graphics Modell.* **2000**, *18*, 383. doi:10.1016/S1093-3263(00)00064-4
- [21] S. Izrailev, F. Zhu, D. K. Agrafiotis, *J. Comput. Chem.* **2006**, *27*, 1962.
- [22] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235. doi:10.1093/NAR/28.1.235
- [23] H. M. Berman, T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, J. Westbrook, *Nat. Struct. Biol.* **2000**, *7*, 957. doi:10.1038/80734
- [24] D. Weininger, *Rubicon ver. 4.9 2005* (Daylight Chemical Information Systems: Irvine, CA).
- [25] G. M. Crippen, T. F. Havel, *Acta Crystallogr. A* **1978**, *34*, 282. doi:10.1107/S0567739478000522
- [26] G. Chang, W. C. Guida, W. C. Still, *J. Am. Chem. Soc.* **1989**, *111*, 4379. doi:10.1021/JA00194A035
- [27] I. Kolossvary, W. C. Guida, *J. Am. Chem. Soc.* **1996**, *118*, 5011. doi:10.1021/JA952478M
- [28] J. W. Ponder, F. M. Richards, *J. Comput. Chem.* **1987**, *8*, 1016. doi:10.1002/JCC.540080710
- [29] D. M. Ferguson, D. J. Raber, *J. Am. Chem. Soc.* **1989**, *111*, 4371. doi:10.1021/JA00194A034
- [30] D. K. Agrafiotis, A. Gibbs, F. Zhu, S. Izrailev, E. Martin, unpublished.