A Method for Automatic Generation of Novel Chemical Structures and Its Potential Applications to Drug Discovery

RAMASWAMY NILAKANTAN,* NORMAN BAUMAN,* and R. VENKATARAGHAVAN

Medical Research Division, Lederle Laboratories, Pearl River, New York 10965

Received April 15, 1991

A novel method for generation of chemical structures of potential pharmaceutical interest is presented. Structures are generated by random combination of known fragments and selected by statistical topological techniques. The power of the method lies in the great profusion of candidates generated together with the extremely high selectivity imposed by the techniques of selection.

INTRODUCTION

We present a novel approach to the generation of chemical structural candidates for the discovery phase of medicinal chemistry research. We have devised a simple algorithm for rapid automatic generation of chemical structures on the computer. The method assembles chemical structures by the random fusion of chemical fragments. We then select a small fraction of these compounds using statistical structure–activity models. The purpose of this paper is to show that

(i) It is possible to generate reasonable chemical structures very rapidly

(ii) These structures can be made to resemble 'real' compounds as measured by various gross topological indexes, and at the same time be novel

(iii) These structures can be used to augment databases of 'real' compounds by using statistical methods of selection

(iv) This method is a viable automated idea generator which can be used to assist the medicinal chemists' imagination

In particular, we wish to emphasize the fact that undirected assembly enables us to construct very large numbers of structural candidates very rapidly; such large numbers also make 'low yield' statistical selection techniques effective.

In an earlier publication[1] we had outlined the basic philosophy of this technique. This paper presents the method in detail with some examples.

METHOD

Preliminary Construction of a Database of Fragments. The first step consists of assembling a database of different fragments. These include rings and ring systems, acyclic fragments, functional groups, etc. For the present work, we derived this set of fragments from our own proprietary database of about 200000 compounds. We then attach a statistical weight to each fragment depending on its frequency of occurrence in the database. These weights can be modified to alter the type of structures produced by the program. The database of fragments can be augmented or altered as desired, but it is usually held fixed.

Our procedure consists of the following three major steps:

1. Generation of Structures
2. Selection
3. Archival

These steps are described in greater detail below.

1. Generation of Structures. We outline below the algorithm for structure assembly:

(i) A truncated gaussian random number generator is used to select a minimum molecular weight. We use a mean of 150 and a standard deviation of 50 as our gaussian parameters and cutoffs of 100 and 400 at the low and high ends, respectively.

(ii) A fragment is chosen from the database with probability proportional to its statistical weight.

(iii) Step ii is repeated to select successive fragments. As each new fragment is selected, free sites are picked randomly from it and the molecule built so far, and the two are joined by elimination of hydrogens. Care is taken to see that valence rules are not violated. Fragments are added until the molecular weight exceeds the number chosen in step i. At this stage, the molecule is assigned a temporary identification number and stored in a highly compressed binary file.

This procedure can be fine-tuned by adjusting the statistical weights, gaussian function parameters, etc. until the gross characteristics of the generated compounds closely mimic those in the database whence the fragments were generated, or adjusted otherwise as desired.

The random selection of sites for fusion of fragments often produces structures that are chemically unstable or very unusual. We overcome this problem as shown in the selection step.

2. Selection. The purpose of this step is to use structure–activity models to eliminate a vast majority of the randomly generated structures and keep only a highly selected few. We use two topological methods developed in our laboratory for such selection: (a) similarity probe and (b) trend vector analysis. These have been described in detail in earlier publications;[2,3] here we give only a brief outline and describe their use in the selection process.

(a) The similarity probe method calculates a numerical measure of similarity between two molecular structures by resolving each into a set of descriptors. The number of descriptors in common between the two molecules is counted and a similarity index is calculated as:

$$S = 2C/T$$

where, S is the calculated similarity value, C is the number of descriptors in common, and T is the total number of descriptors in the two molecules. This similarity index varies between 0 and 1, being 0 for molecules that share no descriptors and 1 for identical molecules. In the present work we used two descriptors developed in our laboratory,[2,3] viz., the atom pair and the topological torsion.

(b) Trend vector analysis. Unlike the similarity probe, which uses only chemical information, trend vector analysis uses biological activity information. The statistical relationship between biological activity and descriptor content of a "training set" of real compounds is used to calculate a trend vector, which can then be used to predict the biological activity of untested compounds. Details have been published elsewhere.[2,3]

* To whom enquiries should be addressed.

0095-2338/91/1631-0527$02.50/0

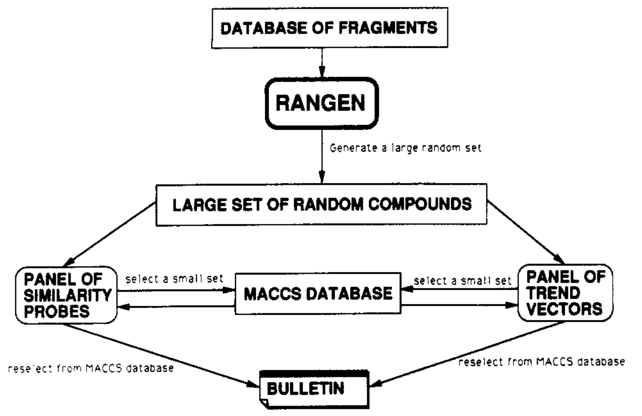
DATABASE OF FRAGMENTS
RANGEN
Generate a large random set
LARGE SET OF RANDOM COMPOUNDS
PANEL OF SIMILARITY PROBES
select a small set
MACCS DATABASE
select a small set
PANEL OF TREND VECTORS
reselect from MACCS database
BULLETIN
reselect from MACCS database

Figure 1. Schematic illustration of the random structure generation, selection, and archival system.

We use the trend vector model to evaluate the randomly generated structures and to select the top ranking few for further consideration. In selecting a structure we require that at least 90% of the descriptors contained in it were also seen in the compounds that went into the building of the trend vector. This step eliminates most unrealistic structures.

We would like to reiterate the fact that the above procedures result in the rejection of 99.9% or more of the generated structures.

3. Archival. The few surviving compounds are registered into a MACCS[4] database, assigning them new permanent ID numbers. From time to time, we run the two selection procedures (similarity probe and trend vector analysis) again on the MACCS database. Of course we select the same compounds again, but now we can identify them by their new permanent ID numbers. This reselection is a convenience, eliminating complex bookkeeping and taking only a few seconds, as now we deal with hundreds of structures rather than tens of thousands. The selected structures are printed out as a bulletin to consider for synthesis. The method is outlined schematically in Figure 1. The remaining structures in the original randomly generated set are usually discarded.

RESULTS AND DISCUSSION

Medicinal chemists have explored a variety of methods to generate novel chemical candidates for synthesis.[5-9] Automatic generation of chemical structures consistent with a given molecular formula has been accomplished by earlier workers.[10-12] Work has also been done on discovering new chemical reactions and structures based on chemical principles.[13] Computer generation of chemical structures from fragments has been accomplished by others[14] as part of an expert system for analysis of C-13 NMR spectra. Here our purpose is merely to generate a profusion of different structures without any constraints and then to use statistical techniques to select a small fraction of these.

To give an idea of the kind of structures produced by the generation program we show a random sample of these in Figure 2. Since these structures have not been selected in any way, some of them are rather unusual.

Results from Similarity Probe. For the similarity probe selection step, we chose five drugs, viz., atenolol, captopril, ranitidine, cimetidine, and naproxen. They were chosen because of their simple structure and because they are widely used and very effective. Some of the structures chosen because of their similarity to atenolol are shown in Figure 3. It can be seen that most of these compounds look very reasonable from a chemist's viewpoint. Each is an interesting variation of the probe compound.

Results from Trend Vector Selection. For the present work, we built trend vectors using biological activity data from five different biological assays being carried out in our laboratory. In each case, we used both the atom pair and the topological torsion descriptors.[2,3] Each random structure was then rated against each of these trend vectors. If a structure rated high on any one of the trend vectors, it would be selected.

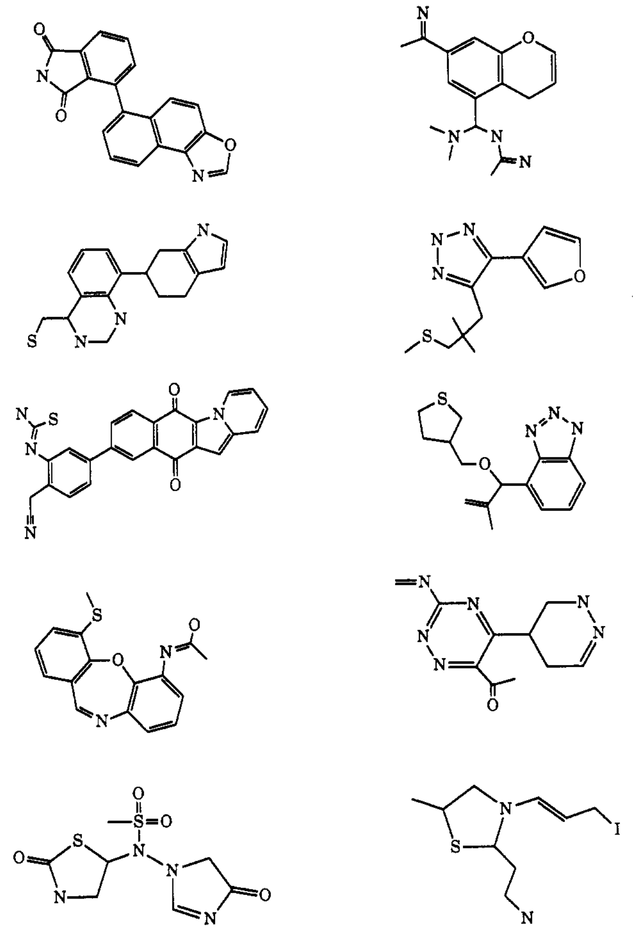

Figure 2. Unselected sample of the structures generated by the random structure generator.

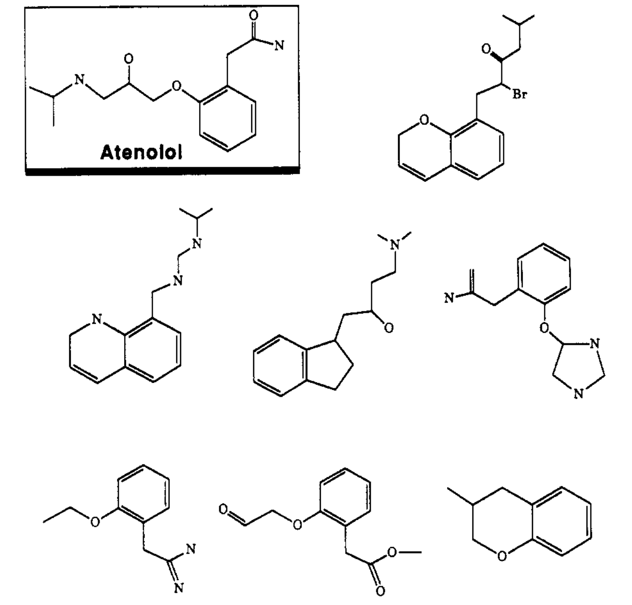

Figure 3. Some random structures similar to atenolol.

We discuss the results obtained with one of these assays, viz., the antiarrhythmic assay. This is an in vivo test that evaluates compounds for their ability to prevent induced arrhythmia in mice. Figure 4 shows some randomly generated structures selected by the topological torsion trend vector.

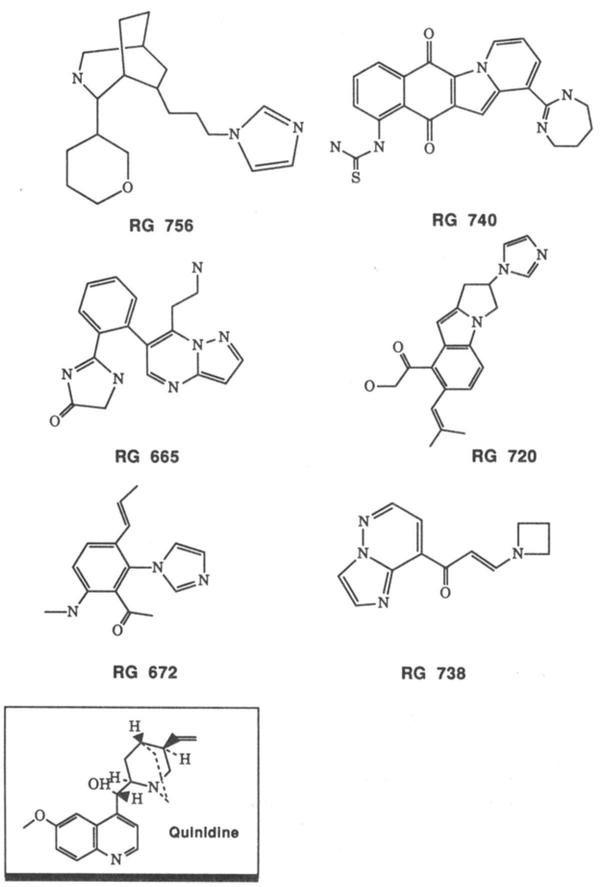

Figure 4. Random structures selected by the antiarrhythmic trend vector. The topological torsion descriptor was used in this case. Note the resemblance between quinidine, a well-known antiarrhythmic (inset) and RG 756.

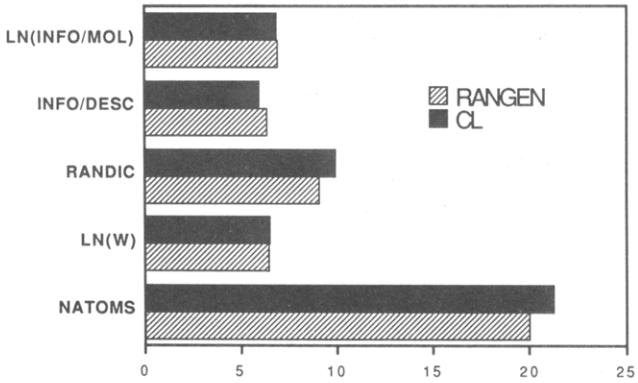

Figure 5. Histogram showing average values of the Wiener index [LN(W)], Randic index, information content per molecule [LN(INFO/MOL)], and per descriptor (INFO/DESC) for random structures (RANGEN) and for CL-file compounds (CL).

There is some resemblance between quinidine (a well-known antiarrhythmic) and RG 756. The important point, however, is that this is only intended as an idea generator; unreasonable compounds can be ignored by the chemist.

Studies on Topological Indexes. In order to characterize these machine-generated structures, we calculated their molecular weights, number of non-hydrogen atoms, and a series of topological indexes. We calculated the Wiener index,[15] Randic index,[16] and information content (see Appendix) per molecule and per atom. As can be seen from Figure 5, these general measures for the randomly generated structures closely mimicked those for compounds in our database (indeed, the weighting factors and molecular weight cutoffs had been chosen with this in mind).

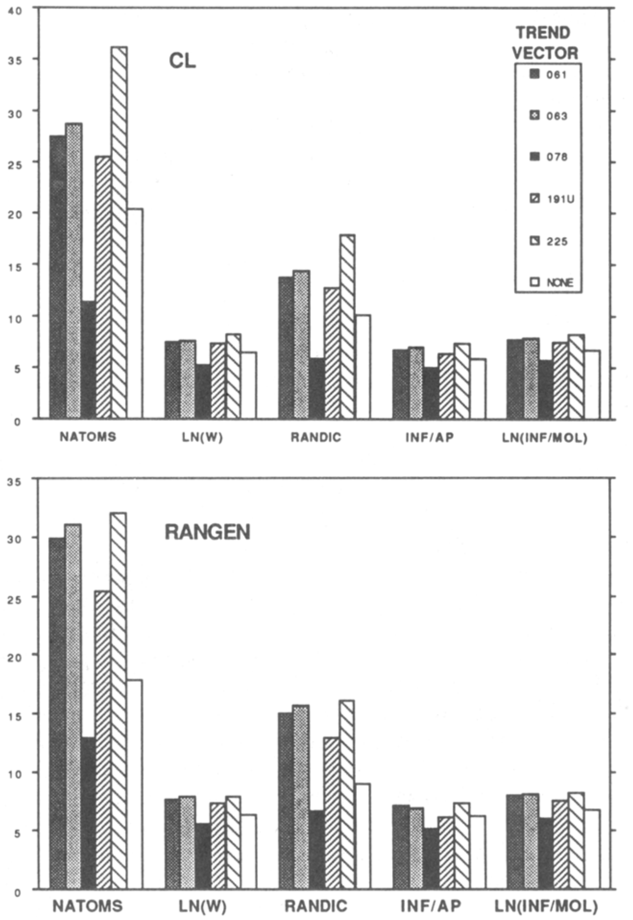

Figure 6. Histogram showing the distribution of topological indexes for random compounds selected by various trend vectors (upper panel) and CL-file compounds selected by the same trend vectors (lower panel).

We carried out similar calculations on the random structures chosen by each of the five trend vectors as described above and compared these to compounds chosen from our corporate database by the same trend vectors. Here we note (Figure 6) that the variations among the trend vectors cause similar variation among the real compounds and the randomly generated compounds. This shows that the compounds chosen by a trend vector are indeed dependent on the information in the trend vector, rather than on the method of generation of the compound.

Since we have shown on other occasion[1,3] that choice via trend vector enhances the hit rate in real compounds and the generated structures resemble real compounds, there is every reason to believe that it would do the same for random compounds, irrespective of the exact algorithm for compound generation.

Overlaps with the Parent Set. It is quite possible that by joining together fragments derived from a parent database one regenerates compounds actually present in the database. An idea of the frequency of such an occurrence can be obtained by generating hashcodes[17] for the random structures and comparing them with the hashcodes for the parent database. Experience with several runs of 10 000 randomly generated structures shows that this is a relatively rare occurrence, with an incidence of at most 0.3%. Thus, the generated structures are over 99% novel.

Estimation of the Size of the Virtual Random Set. Since nothing was done in the course of random generation to prevent the same compound from being generated twice, measurement of the frequency of this accident gives an estimate of the size of the virtual list of random compounds.

Thus, in 20000 compounds generated, we observed 30 duplicates. By Poisson statistics, the ratio of the probability of two hits on the same compound to one hit is just $m/2$, where m is the mean number of hits per compound. Therefore we can estimate m as $2 \times 30/20000 = 0.003$. Thus the total universe of compounds is of the order of 20000/0.003, or 6.7 million. (Of course, this is a virtual figure; the conceivable number of compounds is greater, but the probability weighting scheme reduces the effective number.) Since 20–30000 compounds per overnight run is well within the power of the method, the total space can be explored in about 200 runs.

It should be understood however, that this estimate applies only to our particular set of fragments and statistical weights. One is free to add to the fragment database or manipulate the weights; these variations might expand the universe tremendously.

The Power of the Method. The power of this method lies mainly in the rapidity with which compounds can be produced. For example, on our VAX 8650 we can produce 1000 randomly generated compounds in less than 1 min of CPU time. Thus, in an overnight run it is very easy to produce say, 20000 compounds, evaluate each compound against a panel of 10 different trend vectors, and conduct 10 different similarity probe calculations. Thus we have the capability to generate and to screen a very large number of hypothetical compounds. We believe that by repeatedly going through the process of generating and screening we can come up with interesting structures which will by themselves or with some modification show the desired biological activity. This method expands the universe of selection beyond the realm of known compounds and hence has the potential to bring new ideas to the chemist.

This method could conceivably also be used in a more restricted sense by a chemist to produce new recombinations of fragments extracted from a known series of active compounds. Some of the random structures built by using this small set of fragments could contain novel rearrangements of functional groups not considered before.

Also, we have used a variant of our program to build many random structures around a fixed 'core' considered essential for a particular biological activity. Stringent selection, as described above, may then yield potential structures for lead optimization.

Planned Enhancements. The very simple method of joining fragments employed here was adequate to test the idea, but still led to compounds that the organic chemists deemed unsynthesizable. However, any or all of the following methods of linkage could be used, choosing the link at random. Fragments could be linked by insertion of -CONH-, -NHCO-, -COO-, OCO, -O-, -S-, -NH-, -SS-, etc.

Furthermore, variation of the rings could be obtained by occasional substitution of -CH_2CH_2- for -CH_2-, or -CH=CHCH= for -CH=, or a hetero atom for carbon, etc. Note that these changes can be introduced at random without any perception of ring structure; they might often lead to nonsense, but they would occasionally lead to significant variants.

CONCLUSIONS

Preliminary experiments with a random structure generator indicate that it might be a useful tool in the hands of medicinal chemists. It could be used as an idea generator that creates novel combinations of a large diverse set of fragments. By restricting the fragments to a small chosen set, one could also conceivably use it to create new variants of known drugs.

ACKNOWLEDGMENT

We thank Dr. Raymond Carhart, who originally conceived of the notion of automated random assembly of structures followed by selection.

APPENDIX

Information per Descriptor. This topological index is calculated from the representation of a compound as a collection of descriptors; it is defined as:

$$\sum_i - p_i \log (p_i)$$

where

$$p_i = \frac{\text{number of occurrences of descriptor `}i\text{'}}{\text{total number of descriptors}}$$

It is a measure of the variety among descriptors. Total information per compound is (information per descriptor) × (total number of descriptors); in general, it is larger with larger compounds.

REFERENCES AND NOTES

(1) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. Methods in Computer-Assisted Drug Design and Discovery. In *Molecular Aspects of Chemotherapy*; Borowski, E., Shugan, J., Eds.; Pergamon: Elmsford, NY, 1989; Chapter 1, p 1–10.

(2) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(3) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.

(4) MACCS is an acronym for Molecular Access System, a chemical database management system supplied by Molecular Design Limited, San Leandro, CA.

(5) Martin, Y. C. Computer Design of Potentially Bioactive Molecules by Geometric Searching with ALADDIN. *Tetrahedron Comput. Methodol.* **1990**, *3* (1), 15–25.

(6) Lewis, R. A.; Dean, P. M. Automated Site-Directed Drug Design: The Concept of Spacer Skeletons for Primary Structure Generation. *Proc. R. Soc. London, B* **1989**, *236*, 125–140.

(7) Lewis, R. A.; Dean, P. M. Automated Site-Directed Drug Design: The Concept of Spacer Skeletons for Primary Structure Generation. *Proc. R. Soc. London, B* **1989**, *236*, 141–162.

(8) Desjarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor-Binding Site of Known 3-Dimensional Structure. *J. Med. Chem.* **1989**, *31* (4), 722–729.

(9) Desjarlais, R. L.; Seibel, G. L.; Kuntz, I. D.; Furth, P. S.; Alvarez, J. C.; Demontellano, P. R. O.; Decamp, D. L.; Babe, L. M.; Craik, C. S. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus-1 protease. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87* (17), 6644–6648.

(10) Carhart, R.; Smith, D.; Brown, H.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference XVII—Approach to Computer-Assisted Elucidation of Molecular Structure. *J. Am. Chem. Soc.* **1975**, *97* (20), 5755–5762.

(11) Bangov, I. P.; Kanev, K. D. Computer-assisted structure generation from a gross formula II multiple bond unsaturated and cyclic compounds. Employment of fragments. *J. Math. Chem.* **1988**, *2*, 31–48.

(12) Novak, B.; Szotyory, L. (1978) Structural formula generating computer algorithm. *Textes Conf. Cadre Congr. Int. Contrib. Calc. Electron. Dev. Genie Chim. Chim. Ind.* **1978**, *Vol. A*, 78–82.

(13) Bauer, J.; Ugi, I. Chemical reactions and structures without precedent generated by computer program. *J. Chem. Res. Synop.* **1982**, *No. 11*, 298.

(14) Razinger, M.; Zupan, J.; Novic, M. Computer generation of chemical structures from known fragments. *Mikrochim. Acta* **1986**, *II*, 411–421.

(15) Weiner, H. Relation of physical properties of the isomeric alkanes to molecular structure. *J. Phys. Chem.* **1948**, *52*, 1082–1089.

(16) Randic, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1976**, *97*, 6609–6615.

(17) Nilakantan, R.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. A Ring Based Chemical Structural Query System: Use of a Novel Ring-Complexity Heuristic. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 65–68.