Mining a Large Database for Peptidomimetic Ring Structures Using a Topological Index

Alan H. Lipkus

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210-0012

Received October 3, 1998

A topological index that characterizes the way in which a substructure is imbedded in the rings of a structure was recently introduced as a tool for analyzing large substructure-search answer sets. This paper describes a technique that uses the ring-imbedding index to mine a chemical database for novel compounds. The strategy is to run a very general substructure search and then extract from the answer set those structures with the least common index values. An experimental test was performed on the CAS Registry File. Using a peptide-like substructure query that retrieved over 33 000 answers, this technique was able to detect a number of unusual peptidomimetic ring structures in the database. This technique is potentially useful when the problem of incorporating rings into a particular type of structure is of interest, as is often the case in medicinal chemistry.

INTRODUCTION

The substructure searching of a large database often retrieves an answer set containing many more structures than can be manually inspected. Simple tools for the analysis of substructure-search answer sets thus have the potential to be very helpful. There are a number of structural features that could serve as a basis for such tools; one of the most interesting of these is ring topology. Recently, it was proposed that answer sets can be analyzed based on how the substructure query is imbedded in the rings of each file structure, and a topological index that characterizes the extent and nature of this ring imbedding was introduced.[1] It was shown that a large answer set could be analyzed by assigning this index to each answer-set structure and grouping together structures having the same index value. This made it possible to sample more fully the diversity of ring imbeddings represented in the answer set. This approach to answer-set analysis was motivated partly by the fact that medicinal chemists often look for ways to imbed portions of a lead molecule in rings to create analogs with less conformational flexibility.[2]

It was also proposed that the ring-imbedding index might be used to look for unusual structures in an answer set.[1] The basis for this proposal is that those structures having the least common index values display the rarest types of ring imbeddings and should therefore exhibit a significant degree of structural novelty. This suggests that the ring-imbedding index could be used to mine a chemical database for novel compounds by extracting from a very large answer set those structures with the least common index values. The present paper reports an experiment designed to test this idea. The goal of the experiment was to find unusual peptidomimetic ring structures in the CAS Registry File. The database was searched using a very general peptide-like query that retrieved an extremely large number of answers, and the answers were then analyzed with the ring-imbedding index. One reason for choosing peptidomimetic structures as the target is that rings play an especially important role in these compounds. Rings are routinely used to try to impose on peptidomimetics a conformation similar to that of the natural receptor-bound ligand so as to enhance potency or selectivity.[3−6]

THE RING-IMBEDDING INDEX

A topological index called RIQI (for Ring-Imbedding-of-Query Index) has already been described in detail[1] and is only summarized here. Unlike most topological indexes, the RIQI depends on a pair of chemical graphs: a substructure query and a matching file structure. As given by the formula

$$\mathrm{RIQI} = RS \times 10^5 + IC$$

the RIQI is a combination of two quantities describing different aspects of the query imbedding in the file structure. *RS* is the number of separate file-structure rings, or ring systems, in which the query is imbedded. *IC* is an integer that quantifies the "complexity" of this ring imbedding; its magnitude reflects the number of query bonds in rings as well as the number of query atoms acting as bridgeheads or ring fusion sites. The RIQI is easily calculated from the query and file-structure connection tables and the atom-to-atom mapping between them. If more than one mapping is possible, a structure may have more than one valid RIQI.

The conceptual basis for the RIQI is the *ring-imbedding graph*, an abstract graph (i.e., no element or bond type information) that represents how the substructure query is imbedded in the rings of the file structure. The quantities *RS* and *IC* can be interpreted as specific properties of the ring-imbedding graph. *RS* is the number of connected components in the graph. *IC* is the topological complexity of the graph as characterized by a simple graph invariant: the sum of the squares of the degrees of its nodes; this invariant was originally proposed as a branching index.[7] The ring-imbedding graph represents only the basic topology of the imbedding and does not take into account structural features lying outside the immediate environment of the query. The graph thus contains a relatively small amount of structural information. As a result, the information content of the RIQI is also low.

10.1021/ci980151+ CCC: $18.00
Published on Web 04/14/1999

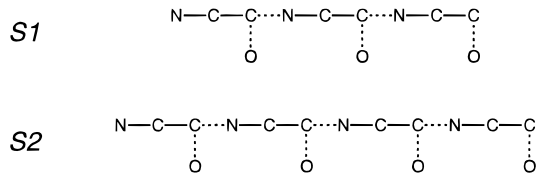

Figure 1. Peptide-like substructures used to form the search query. Solid lines are "single exact" bonds. Dotted lines are "unspecified" bonds.

When RIQI values are generated for each of the structures in a substructure-search answer set, those structures with ring imbeddings that are sufficiently similar will have the same value. Groups of structures having the same index value can be collected together for analysis; each group is called a *RIQI class*. It should be emphasized that the ability of the RIQI to group structures together is dependent on its low information content. An index containing much more information would be too discriminating to group structures effectively.

The set of all RIQI classes is useful in sampling the ring-imbedding diversity of an answer set. It is only the smallest classes, however, that are of interest with regard to database mining for structural novelty. The smallest classes contain the structures that, by definition, have the least common RIQI values (as previously noted, these structures are presumed to be the most novel). The size criterion for defining what constitutes the "smallest" classes is necessarily relative and will depend partly on the total number of structures one wishes to examine.

THE SEARCH QUERY

What distinguishes the present study as a database-mining experiment rather than an ordinary answer-set analysis is the intentional use of a search query that retrieves an exceptionally large answer set. The search query is based on the tripeptide-like substructure (*S1*) shown in Figure 1. The "single exact" bonds are allowed to match only single bonds in the file structure. The "unspecified" bonds are allowed to match any bond type, i.e., single, double, triple, or normalized; these bonds thus match the standard peptide linkage, which is normalized as a tautomer on the CAS Registry File,[8] as well as other bond patterns. This use of "unspecified" bonds is compatible with the RIQI because no information about specific bond types, or atom types, is required for the index calculation. Hence, so-called variable atoms, which are allowed to match more than one atom type (e.g., any halogen), may also be used. Variable atoms and bonds are one way to make the query more general so that it retrieves enough structures to permit a meaningful exploration of the database.

Since the goal of the search is to retrieve a wide variety of ring forms, all of the bonds in substructure *S1*, except for the C−O bonds, are allowed to match either chain or ring bonds in the file structure. The C−O bonds are allowed to match only chain bonds. Each oxygen is constrained to have only one non-hydrogen attachment, thus prohibiting ether linkages at these atoms.

As a search query, *S1* by itself is too general; it would retrieve over 300 000 answers from the Registry File. While a large answer set is desirable, the analysis of a set this large is inappropriate for a primarily illustrative study. For that reason, substructure *S2* was introduced (see Figure 1). This substructure is the same as *S1* but extended by one "monomer". If *S1* NOT *S2* were used as the search query, it would exclude all peptides or peptide derivatives having four or more consecutive amino acid residues, thereby greatly reducing the answer set.

There is another group of structures that was also excluded from the answer set: those in which *S1* maps onto an acyclic portion of the structure. These structures can be discarded because the substructure of interest is not ring imbedded. Such structures could be readily identified and eliminated at a later stage since each will have a RIQI equal to zero (as is always the case when there is no ring imbedding of the query). However, since these structures are numerous, it is more efficient to reject them at the search stage. This was done by using *S1* NOT [*S1(all-chain)* OR *S2*] as the search query, where *S1(all-chain)* is the same as *S1* but with all its bonds allowed to match only chain bonds in the file structure.[9]

Table 1. RIQI Classes in the Answer Set

RIQI	answers	RIQI	answers	RIQI	answers
100010	12286	100042	5	200026	14
100014	2139	100044	31	200028	30
100016	2	100046	89	200030	2
100018	629	100048	4	200032	52
100020	1	100050	162	200034	192
100022	469	100052	2	200036	14
100024	13227	100054	11	200040	8
100026	33	100056	7	200042	209
100028	6	100058	470	200044	8
100030	384	100062	26	200046	9
100032	35	100064	5	200048	1
100034	156	100072	3	200052	3
100036	11	100074	15	200054	4
100038	335	200020	2026	200056	6
100040	199	200024	177	300030	167

RESULTS AND DISCUSSION

A search of the CAS Registry File using the query just described retrieved 33 664 Registry Numbers.[10] The RIQI for each of these answers was calculated based on the mapping between substructure *S1* and the file structure. Most of these structures contain only one occurrence of *S1*, and for them only a single mapping is possible.[11] In a small fraction of the answers, multiple mappings are possible, e.g., 375 answers have two or more nonoverlapping occurrences of *S1*. However, only one RIQI was calculated for each structure because the algorithm as currently implemented looks for only one mapping between query and file structure.

Answers having the same index value were grouped together, yielding 45 RIQI classes. These classes are listed in Table 1 along with the number of answers each contains. As expected, there is no class of structures with a RIQI of zero; if such structures had not been excluded by the search query, more than 40 000 of them would have been retrieved. The distribution of class sizes is seen to be extremely uneven; the five largest classes comprise 90% of the answers. The large size of some classes appears to result from a few structural types that occur very frequently in the database. Class 100024, the largest, consists predominantly of structures based on the penicillin and cephalosporin frameworks. Class 100010, the second largest, is dominated by derivatives of linear peptides containing one proline residue, while class

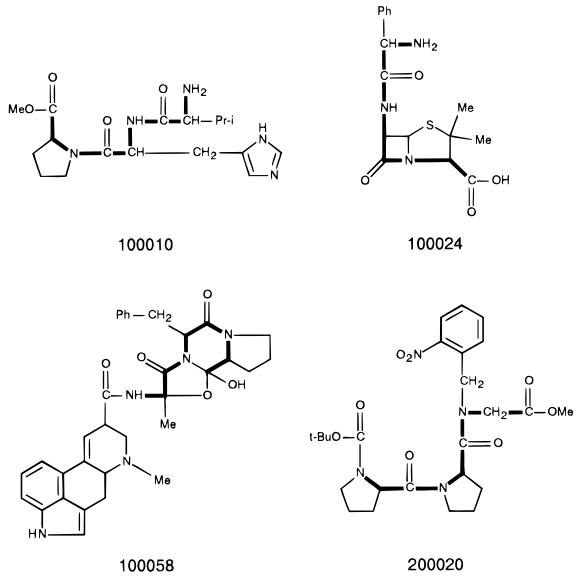

Figure 2. Typical structures from some of the largest RIQI classes. In this and subsequent figures, the carbon−nitrogen "backbone" of substructure *S1* (see Figure 1) is indicated by boldface bonds.

Table 2. RIQI Classes with Less Than 10 Answers and Imbedding Complexity (*IC*) ≥ 40

class	RIQI	answers	class	RIQI	answers
a	100042	5	*h*	200044	8
b	100048	4	*i*	200046	9
c	100052	2	*j*	200048	1
d	100056	7	*k*	200052	3
e	100064	5	*l*	200054	4
f	100072	3	*m*	200056	6
g	200040	8			

200020 is composed mostly of derivatives of linear peptides containing two proline residues. Class 100058 consists almost entirely of derivatives of the ergot peptide alkaloids. A typical structure from each of the above four classes is shown in Figure 2. It is clear that a random sample of this answer set would consist largely or entirely of compounds representing such common structural types.

To explore the kinds of structures found in the smallest RIQI classes, a manageable number of these classes had to be selected. It was decided that classes should be selected not only on the basis of size but also based on the imbedding complexity, as measured by the quantity *IC*. Every class containing fewer than 10 answers and having *IC* ≥ 40 was selected. These classes are listed in Table 2, divided into two groups according to whether *RS* equals one or two. The use of *IC* as a selection criterion was suggested by the observation that the structures with low *IC* are somewhat uninteresting with regard to conformational restrictions on the imbedded query because relatively few of the query bonds are in rings.

The first group of RIQI classes listed in Table 2 (*a*−*f*) contains a total of 26 answers. Because so many of these answers are related derivatives, most of them can be adequately represented by the set of structures in Figure 3. There is one structure from each class except for structures *a1*−*a3*, which are all from class *a*. In every structure shown, most of the query bonds are in rings. This is the result of selecting only classes with a high *IC*; a larger value for this quantity tends to indicate a greater number of ring-imbedded query bonds. The incorporation of query atoms into ring fusions also contributes to *IC*. This is well illustrated by structure *f*, in which the spiro ring fusion adds to the value of *IC* and thereby differentiates this structure, with a RIQI of 100072, from the large number of very similar structures having a RIQI of 100058 (see Figure 2).

In structures *a1*−*a3*, the query is imbedded in very different ring systems. Nevertheless, these structures all have the same RIQI. The underlying reason these structures have the same index value is that their ring-imbedding graphs are isomorphic; this happens because these graphs take into account only a small amount of structural information. The ring-imbedding graph is intentionally focused on the query and its immediate environment to ensure that structures are not grouped based on features distant from the query. An unavoidable result of this, however, is an intrinsic limit on the ability of the RIQI to discriminate between structures, as shown by this example.

The second group of RIQI classes in Table 2 (*g*−*m*) contains 39 answers. As with the previous group, these answers include many related derivatives and can be adequately represented by a smaller set of structures, shown in Figure 4. Several of the classes (*g*, *h*, *k*, and *m*) are represented by more than one structure. In each of the structures in Figure 4, the imbedded query spans two separate rings or ring systems. This is in contrast to the structures of Figure 3, in which the query is imbedded in a single ring system.

The RIQI classes listed in Table 2 were selected as those whose ring imbeddings are the rarest and most complex. It was expected that this strategy would tend to extract compounds showing a high degree of structural novelty. It is interesting that almost half of the representative structures shown in Figures 3 and 4 can be identified as novel peptidomimetics based on their literature references. Structures *a2*,[12] *a3*,[13] *c*,[14] *h1*,[15] *h2*,[16] *h3*,[16] *k2*,[17] *l*,[18] and *m1*[19] were all designed and synthesized as new types of peptidomimetics. Finding such a high concentration of unusual peptidomimetics among the selected structures is encouraging. This is not, however, an example of true discovery via database mining since these compounds were already identified as peptidomimetics in the literature. The best way to validate this approach would be to use it to select known compounds that act as peptidomimetics but which have *not* been identified as such in the literature. Unfortunately, that sort of test is beyond the scope of the present study.

One factor that limits the RIQI as a database-mining tool is that the imbedding complexity *IC* is degenerate. In other words, the same value of *IC* may be assigned to nonisomorphic ring-imbedding graphs. Because *IC* is an extremely simple graph invariant (the sum of the squares of the degrees of the nodes), it is expected to exhibit degeneracy, especially since it does not depend on the actual connectivity of the nodes, only their degrees. However, this degeneracy can result in a small group of interesting structures being missed because it falls into the same RIQI class as a much larger group of structures.

For example, consider graphs *X* and *Y* in Figure 5. Each is a disconnected graph with two connected components. There are five structures in the answer set that have *X* as their ring-imbedding graph. These include structure *1* in Figure 5, which is a known peptidomimetic,[20] and four

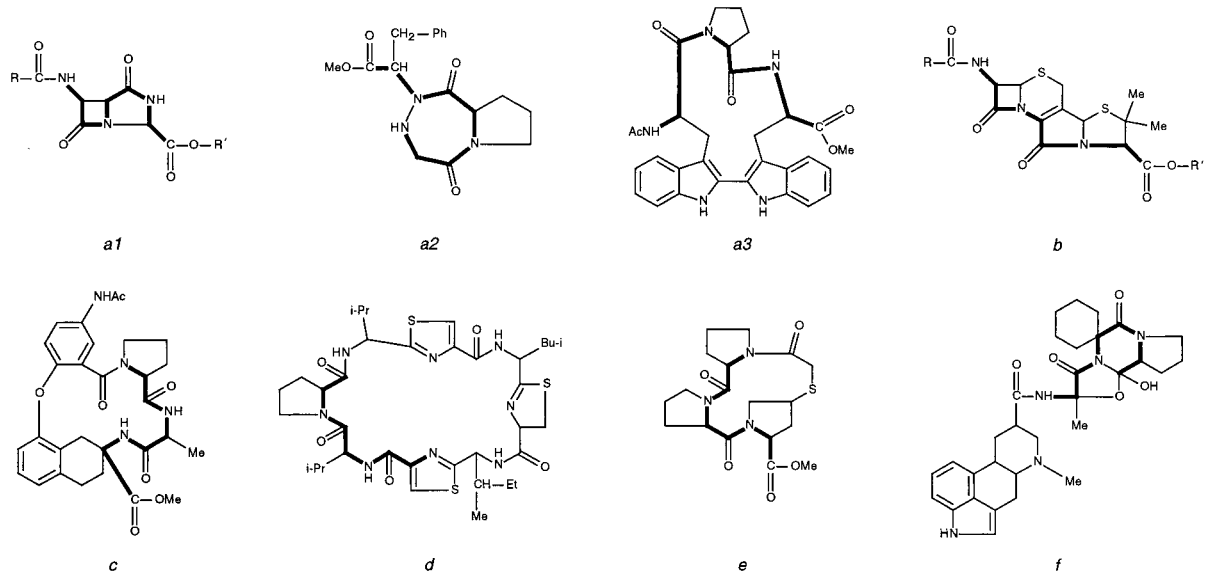

Figure 3. Representative structures for RIQI classes *a*−*f* (see Table 2). The letter below each structure is that of the corresponding class, with a number appended where necessary to distinguish structures from the same class. R-groups have been used to compress some diagrams.

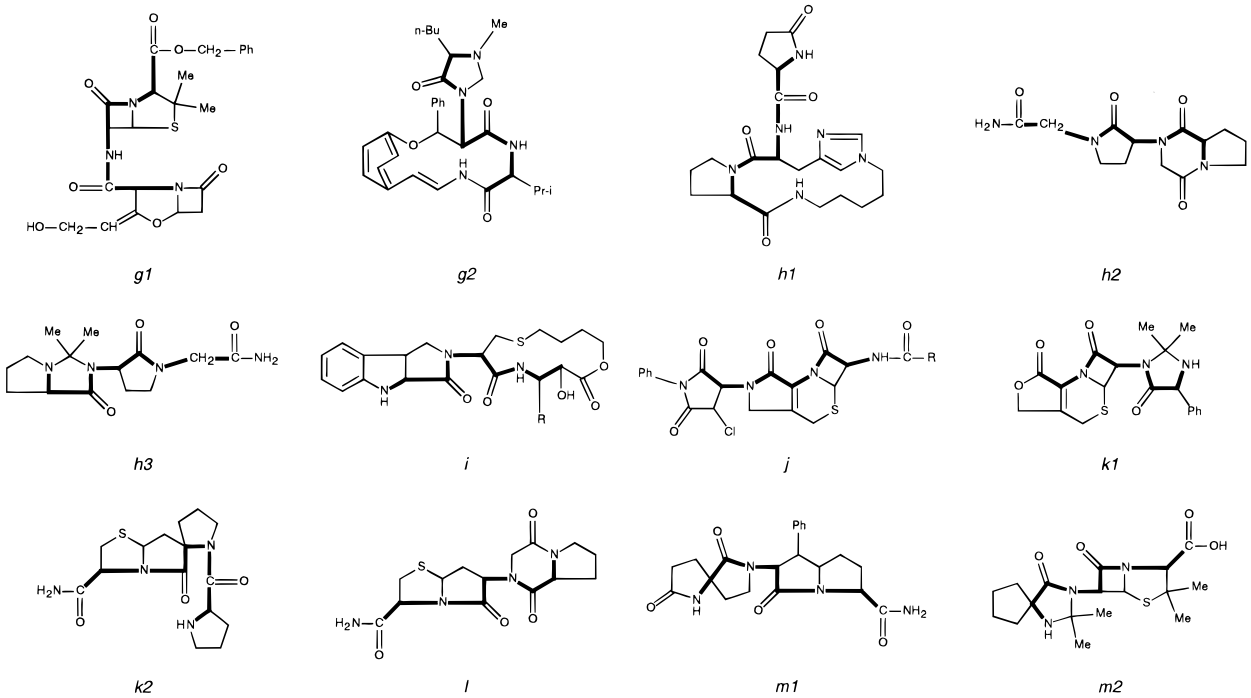

Figure 4. Same as Figure 3 but for RIQI classes *g*−*m*.

related spirocycles. There are 204 structures that have *Y* as their ring-imbedding graph. Each of these consists of a penicillin or cephalosporin framework containing a substituted imidazolidinone ring as exemplified by structure *2* in Figure 5. Although graphs *X* and *Y* are nonisomorphic, *IC* has the same value (42) for both. As a result of this degeneracy, the five spirocycles do not constitute a distinct RIQI class but fall into the same class (200042) as the penicillin/cephalosporin derivatives. These spirocycles were therefore missed by the analysis that selected the small classes listed in Table 2.

A way to improve the effectiveness of the RIQI as a database-mining tool would be to replace *IC* with a less degenerate measure of imbedding complexity. A number of different indexes to quantify the topological complexity of a graph have been proposed,[21] and some of these, applied to the ring-imbedding graph, would undoubtedly yield plausible measures of imbedding complexity having less degeneracy than *IC*. Such a change has a potential disadvantage, however. The current algorithm for calculating the RIQI is very simple and fast, due in part to the simplicity of *IC*. The use of a different measure of imbedding complexity is likely to make the algorithm more complicated and time-consuming.

CONCLUSION

The present study demonstrates the potential value of a technique that uses the ring-imbedding index to mine a chemical database for novel structures. This technique is appropriate when the problem of incorporating rings into a particular type of structure is of interest; such a problem often

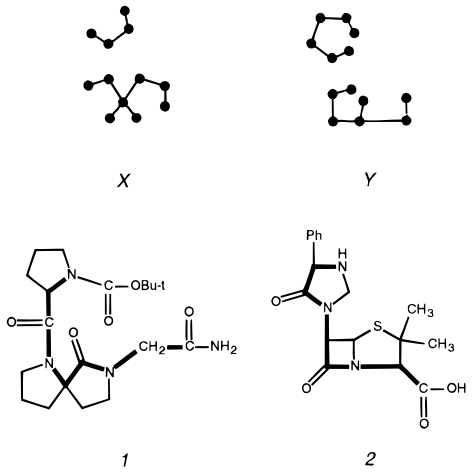

Figure 5. Two ring-imbedding graphs, *X* and *Y*. Structures *1* and *2* are examples of answer-set structures that have *X* and *Y*, respectively, as their ring-imbedding graph. The graphs have been drawn so as to suggest their relationship to the corresponding structure.

arises in medicinal chemistry when designing analogs of a lead molecule. In the example presented here, this technique was able to detect a number of unusual peptidomimetic ring structures within an exceptionally large answer set. The strategy used for finding novel compounds, i.e., classifying substructure-search answers according to their RIQI and then examining the smallest RIQI classes, implicitly assumes that the size distribution of the classes will be uneven enough to result in many small classes. This is certainly true for the results shown in Table 1, where 40% of the classes together contain less than 0.26% of the answers.

It is worth comparing this technique with another method that could be used to select novel compounds. If cluster analysis were applied to the answer set, one would expect to find the most unusual compounds in the smallest clusters. The compounds selected in this manner would be novel in terms of their overall structural characteristics (since conventional similarity measures take into account the entire structure). In contrast, the technique presented here finds compounds that are novel only with respect to the imbedded query. This difference emphasizes the fact that the RIQI disregards global structural characteristics and focuses on a localized structural feature. This distinguishes it from other topological indexes, which generally characterize the entire structure.

The global structural information which the RIQI disregards can be introduced by using additional sources of information. This would help compensate for the inherently limited ability of the RIQI to discriminate between structures. For instance, one could use additional ring-analysis tools. These might include other ring-based topological indexes[22] or algorithms for classifying structures according to their ring systems.[23,24] This would allow more complete information about the rings in a structure to be taken into account.

REFERENCES AND NOTES

(1) Lipkus, A. H. A Ring-Imbedding Index and Its Use in Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 92−97.

(2) Wermuth, C. G. Ring Transformations. In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: London, 1996; pp 239−260.

(3) Giannis, A.; Kolter, T. Peptidomimetics for Receptor Ligands−Discovery, Development, and Medical Perspectives. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 1244−1267.

(4) Wiley, R. A.; Rich, D. H. Peptidomimetics Derived from Natural Products. *Med. Res. Rev.* **1993**, *13*, 327−384.

(5) Hruby, V. J.; Li, G.; Haskell-Luevano, C.; Shenderovich, M. Design of Peptides, Proteins, and Peptidomimetics in Chi Space. *Biopolymers* **1997**, *43*, 219−266.

(6) Giannis, A.; Rübsam, F. Peptidomimetics in Drug Design. In *Advances in Drug Research*; Testa, B., Meyer, U. A., Eds.; Academic Press: London, 1997; Vol. 29, pp 1−78.

(7) Gutman, I.; Ruscić, B.; Trinajstić, N.; Wilcox, C. F., Jr. Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399−3405.

(8) Mockus, J.; Stobaugh, R. E. The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18−22.

(9) Even when all the query bonds map onto chain bonds, a query atom could still map onto a ring atom (as long as the query atom can accommodate the two additional substituents needed to form a ring). This kind of ring imbedding, which does not involve any query bonds, is ignored in the RIQI calculation (see ref 1 for further discussion) and is thus of no interest here. To ensure that structures showing only this kind of ring imbedding were rejected at the search stage, atoms in *S1(all-chain)* were allowed to match ring, as well as chain, atoms in the file structure.

(10) It was found necessary to break this search into several smaller searches in order not to exceed certain limits imposed by the online system. This was done primarily by using screens to search separately over ranges of carbon numbers and, to a lesser extent, ranges of Registry Numbers. Screens were also used to exclude certain unwanted substance categories: structures containing metal atoms, structures representing the monomer of a homopolymer, and substances consisting of two or more components, e.g., mixtures.

(11) The exception is when *S1* closes to form a nine-membered cycle, in which case there are at least three possible mappings. Each of the mappings, however, leads to the same value for the RIQI.

(12) Lenman, M. M.; Ingham, S. L.; Gani, D. Synthesis and Structure of Cis-Peptidyl Prolinamide Mimetics Based Upon 1,2,5-Triazepine-3,6-diones. *Chem. Commun.* **1996**, 85−87.

(13) Stachel, S. J.; Habeeb, R. L.; Van Vranken, D. L. Formation of Constrained, Fluorescent Peptides via Tryptophan Dimerization and Oxidation. *J. Am. Chem. Soc.* **1996**, *118*, 1225−1226.

(14) Abrecht, C.; Mueller, K.; Obrecht, D.; Trzeciak, A. Preparation of Peptides Cyclized to Tetrahydronaphthalene Moieties as Research Tools (Models of Protein α-Helicity) and as Potential Drugs. Eur. Pat. Appl. 640618, 1995.

(15) Jones, J. H.; Wyatt, P. B. Cyclic TRH Analogs. In *Peptides 1988*; Jung, G., Bayer, E., Eds.; de Gruyter: Berlin, 1989; pp 289−291.

(16) Baures, P. W.; Ojala, W. H.; Gleason, W. B.; Mishra, R. K.; Johnson, R. L. Design, Synthesis, X-Ray Analysis, and Dopamine Receptor-Modulating Activity of Mimics of the "C5" Hydrogen-Bonded Conformation in the Peptidomimetic 2-Oxo-3(*R*)-[(2(*S*)-pyrrolidinylcarbonyl)amino]-1-pyrrolidineacetamide. *J. Med. Chem.* **1994**, *37*, 3677−3683.

(17) Genin, M. J.; Mishra, R. K.; Johnson, R. L. Dopamine Receptor Modulation by a Highly Rigid Spiro Bicyclic Peptidomimetic of Pro-Leu-Gly-NH_2. *J. Med. Chem.* **1993**, *36*, 3481−3483.

(18) Baures, P. W.; Ojala, W. H.; Costain, W. J.; Ott, M. C.; Gleason, W. B.; Mishra, R. K.; Johnson, R. L. Design, Synthesis, and Dopamine Receptor-Modulating Activity of Diketopiperazine Peptidomimetics of L-Prolyl-L-leucylglycinamide. *J. Med. Chem.* **1997**, *40*, 3594−3600.

(19) Rutledge, L. D.; Perlman, J. H.; Gershengorn, M. C.; Marshall, G. R.; Moeller, K. D. Conformationally Restricted TRH Analogs: A Probe for the Pyroglutamate Region. *J. Med. Chem.* **1996**, *39*, 1571−1574.

(20) Genin, M. J.; Ojala, W. H.; Gleason, W. B.; Johnson, R. L. Synthesis and Crystal Structure of a Peptidomimetic Containing the (*R*)-4.4-Spiro Lactam Type-II β-Turn Mimic. *J. Org. Chem.* **1993**, *58*, 2334−2337.

(21) Bonchev, D. The Problems of Computing Molecular Complexity. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 33−63.

(22) Randić, M. On Characterization of Cyclic Structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1063−1071.

(23) Balaban, A. T.; Filip, P.; Balaban, T. S. Computer Program for Finding All Possible Cycles in Graphs. *J. Comput. Chem.* **1985**, *6*, 316−329.

(24) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database Diversity Assessment: New Ideas, Concepts, and Tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.

CI980151+