JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

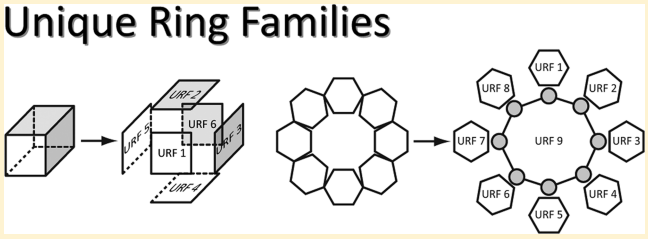
Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies

Adrian Kolodzik,[†,‡] Sascha Urbaczek,[†] and Matthias Rarey*[,†]

[†]Center for Bioinformatics (ZBH), University of Hamburg, Bundesstr. 43, 20146 Hamburg, Germany

Supporting Information

ABSTRACT: The perception of a set of rings forms the basis for a number of chemoinformatics applications, e.g. the systematic naming of compounds, the calculation of molecular descriptors, the matching of SMARTS expressions, and the generation of atomic coordinates. We introduce the concept of unique ring families (URFs) as an extension of the concept of relevant cycles (RCs).[1,2] URFs are consistent for different atom orders and represent an intuitive description of the rings of a molecular graph. Furthermore, in contrast to RCs, URFs are polynomial in number. We provide an algorithm to efficiently calculate URFs in polynomial time and demonstrate their suitability for real-time applications by providing computing time benchmarks for the PubChem Database.[3] URFs combine three important properties of chemical ring descriptions, for the first time, namely being unique, chemically meaningful, and efficient to compute. Therefore, URFs are a valuable alternative to the commonly used concept of the smallest set of smallest rings (SSSR) and would be suited to become the standard measure for ring topologies of small molecules.



INTRODUCTION

Ring perception is a crucial step in many chemoinformatics applications, including the calculation of molecular descriptors, the matching of SMARTS expressions, and the generation of two- and three-dimensional atomic coordinates. In order to obtain consistent results, a set of rings has to be unique in the sense that it depends only on the molecule's topology. Efficient algorithms and ring perception concepts that lead to a limited number of cycles provide the means for interactive applications. Chemically meaningful rings allow for an easy analysis and interpretation of the resulting set of rings. Due to their high relevance in chemistry, several computational methods for automatic ring perception have been developed over the past 35 years.[4] Each of these methods has deficiencies in being either not unique or not polynomial in number or not chemically meaningful. The paper of Berger et al.[5] impressively demonstrates this for a number of ring perception concepts including the widely used SSSR.[4]

A molecule can be interpreted as a simple, connected, unweighted and undirected graph $G = (V,E)$ where the atoms are interpreted as a set of vertices V and bonds are considered a set of edges E. A cycle is a subgraph of G such that any vertex degree is exactly two. A connected cycle is called elementary. Since elementary cycles meet our expectation of rings in a molecular graphs we will use the terms elementary cycle and ring synonymously. $E(v_1,v_2)$ is the edge connecting the vertices v_1 and v_2. For the set of vertices or edges of a cycle (or a general subgraph) C, we will write $V(C)$ and $E(C)$, respectively. A cycle C containing the edges $E(C)$ has a length of $|C|$ which is equal to its number of edges $|E(C)|$. It can be described by the incidence vector of its edges. A cycle with n edges is called n-cycle. A connected n-cycle is called n-ring. A chord is an edge e connecting two vertices of a ring C with $e \notin E(C)$. A ring is chord-less if it has no chord. Cycles can be combined to larger ones by forming the symmetric difference of their edges; this operation is considered the "addition" of cycles. In order to describe the addition of cycles, we utilize the xor operator $\oplus$ in agreement with the nomenclature used by Berger et al.[5] Thus, the addition of two cycles C_A and C_B that forms the cycle C_C will be written as $C_A \oplus C_B = C_C$. All cycles of G form the cycle space $S(G)$. A cycle base $B(G)$ is a subset of $S(G)$ that allows to construct all cycles of $S(G)$ by the addition operation. The length of $B(G)$ is equal to the sum of the lengths of its cycles. All cycles of a cycle base are elementary.

In the following, we will discuss common concepts of ring perception in order to motivate our new approach. The **set of all rings**[6] (Ω) includes all elementary rings of a molecular graph and efficient algorithms for its calculation have been developed. The number of rings and the computational runtimes, however, grow dramatically for complex ringsystems. Additionally, not all resulting rings are meaningful in a chemical context, and Ω is, thus, a unique description that is neither chemically meaningful nor polynomial in size.

The most frequently applied strategy of ring perception is the calculation of the **smallest set of smallest rings**[4] (SSSR) which is a subset of Ω. An SSSR represents a minimum cycle base (MCB). It contains a polynomial number of rings and can be calculated in polynomial time.[7] If a molecular graph contains only a single MCB, the SSSR is unique and intuitive. If this is

Received: December 31, 2011
Published: July 10, 2012

dx.doi.org/10.1021/ci200629w | J. Chem. Inf. Model. 2012, 52, 2013−2021

not the case, the resulting SSSR is arbitrary and depends on the specific algorithm used for its construction. Furthermore, the selected SSSR often depends on the input atom order.[8]

The problems arising from nonunique ring descriptions can be exemplified with SMARTS pattern matching. According to page 20 of the Daylight Theory Manual,[9] the SMARTS pattern [R3] describes an atom which is part of three SSSR rings. The matching of this SMARTS pattern on the highly symmetric molecule cubane (SMILES = C12C3C4C1C5C2C3C45) using the Daylight web service[10] illustrates the problems arising from the SSSR's lack of uniqueness (see Figure 1). Any combination

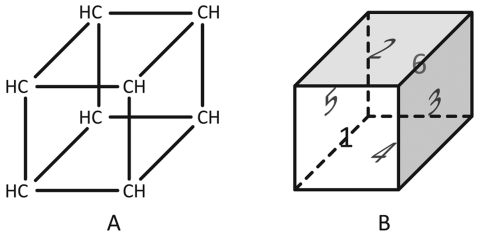


Figure 1. Cubane contains six alternative MCBs. Each combination of five of the 4-rings forms an SSSR.

of five of the shown 4-rings forms a valid SSSR. The sixth ring can be constructed by adding the rings of the SSSR.[11] Consequently, the SMARTS pattern [R3] only matches four of the eight equivalent carbon atoms depending on the selected SSSR.

The **essential set of essential rings** (ESER)[12] and the approaches published by Corey[13] and Wipke[14] try to perceive chemically meaningful rings by calculating an MCB and adding rings up to a certain size or rings including certain elements. Due to their heuristic nature, these approaches lack a mathematical foundation and are not suitable for all kinds of molecular graphs.[5]

In addition, there is a number of graph theoretical ring perception concepts which are limited to planar graphs. The **minimum planar cycle base** and the **extended set of smallest rings**[15] are examples of such concepts. Since molecular graphs are not necessarily planar, these ring perception concepts are of limited use for general applications in chemoinformatics.

The **set of β-rings**[16] is defined on a plane embedding of a molecular graph. The chord-less faces of the embedding are processed by increasing size. The set of β-rings includes all faces representing 3-rings or 4-rings. Additionally, it contains all faces which are linearly independent of three or less shorter faces already contained in the set. Berger et al.[5] suggested to use the **β*-rings** instead. These rings are calculated on all chord-less rings of a graph instead of the chord-less faces of a specific plane embedding. In contrast to the set of β-rings, the set of β*-rings is unique but contains an exponential number of rings.

An additional set of rings which is defined for general graphs is the **set of smallest cycles at edges** (SSCE).[17] The SSCE is calculated on the basis of Ω by recursively deleting all edges included in more than one ring. The SSCE does, however, not necessarily contain a cycle base. Consequently, it does not provide a complete description of the rings of a molecular graph.

Relevant cycles[1,2] (RCs) are defined as the union of all MCBs. They comprise a unique set of rings and an intuitive description of most molecular graphs. Some molecules, however, contain an exponential number of RCs. Examples are cyclophane-like structures which will be discussed in more detail in the following sections.

To tackle the exponential number of rings, Gleiss et al.[11] suggested to classify RCs into **interchangeability classes** (ICs). ICs are calculated by dividing RCs into essential and interchangeable rings. An essential ring is included in all MCBs. Rings which are not essential are called interchangeable. An IC contains either a single essential ring or all interchangeable rings which can be constructed from a subset of the IC and shorter cycles. While treating the rings of an interchangeability class as a union can be suitable for the prediction of RNA secondary structures, this concept is not generally applicable in chemoinformatics. For example, the description of the six RCs of cubane or the 6-rings of fulleren as single ICs is too coarse for most applications and, especially in the case of fullerene, it is not chemically meaningful.

Relevant cycle families (RCFs)[1] are conceptually similar to ICs. An RCF contains all RCs generated on the basis of a single relevant cycle prototype (RCP). RCPs are not unique and their number depends on the order of the molecule's atoms. Since each RCP results in an RCF, the RCFs are also not unique and their number can vary for a molecule.

None of the mentioned concepts of ring perception efficiently calculate a complete and polynomial set of unique and chemically intuitive rings for molecular graphs. We introduce the concept of **unique ring families** (URFs), which meets all of these requirements.

UNIQUE RING FAMILIES

Generation of Relevant Cycles. Since unique ring families (URFs) are defined on the basis of RCs, we provide a short outline of Vismara's RC detection algorithm.[1] The perception of RCs involves five consecutive steps which are explained below (see Figure 2):

1. Calculate all 2-connected components of the molecular graph G.
2. For each 2-connected component, calculate the shortest paths from each vertex r to each other vertex, only passing through vertices which follow r in an arbitrary but fixed order π.
3. Calculate RCPs by combining pairs of shortest paths of identical size starting from the same vertex r.
4. Eliminate RCPs which linearly depend on strictly smaller cycles with respect to cycle addition.
5. Calculate RCs by a backtracking procedure on the basis of the RCPs.

2-Connected components of the molecular graph can be calculated using the algorithm published by Tarjan.[18] The 2-connected components will be called ringsystems in the following sections. An order π of the vertices is established by sorting them according to their degree in descending order. Vertices of identical degree are ordered arbitrarily. This ordering guarantees polynomial runtime complexity for the calculation of RCPs. In the second step, a breadth-first-search is used to calculate a single shortest path $P(r,t)$ from each vertex r to each other vertex through vertices following r in the ordering π. Thus, only paths through vertices of equal or lower degree are considered. If two shortest paths $P(r,p)$ and $P(r,q)$ of identical size solely share the vertex r, and if furthermore p and q are directly connected by an edge, an uneven ring is identified. If p and q are both directly connected to a vertex z which is neither a member of $P(r,p)$ nor a member of $P(r,q)$, an

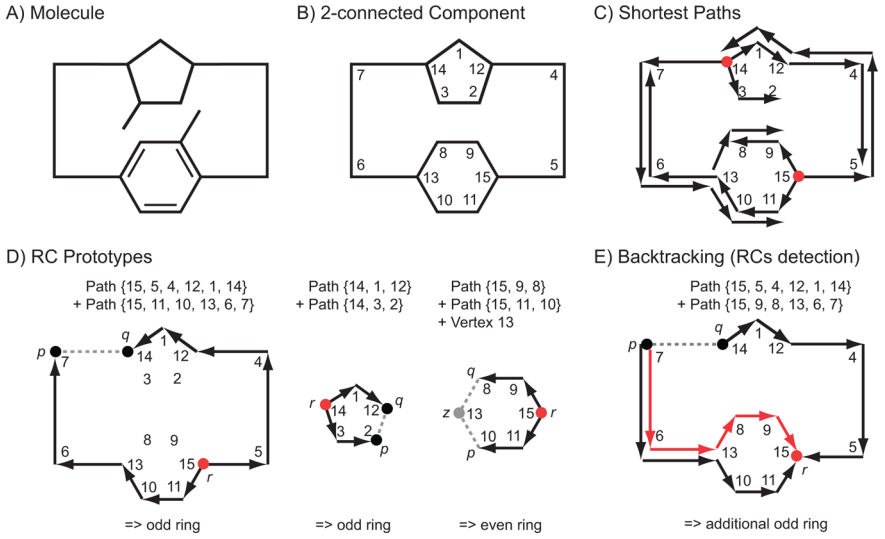


Figure 2. Process to identify the RCs of a molecular graph. (A) At first, 2-connected components are calculated (B) and vertices are ordered according to their degree. Vertices of higher degree are labeled with higher numbers than vertices of lower degree. (C) Shortest paths only passing through vertices following r in the order π are calculated from each vertex r to each other vertex of the graph (shown for vertices 14 and 15). (D) The polynomial number of RCPs are calculated on the basis of the identified shortest paths. Two shortest paths of equal lengths which only share the vertex r form an uneven RCP if their end points (p, q) are adjacent. If they share an adjacent vertex z, they form an even RCP. (E) RCs are enumerated on the basis of RCPs by combining alternative shortest paths (red arrows) connecting p or q to r.

even RCP is identified. The length of the shortest paths used to identify an RCP of size n is therefore given by the following equation:

$$|E(P(r,p))| = |E(P(r,q))| = \begin{cases} \frac{n-1}{2} & \text{if } n \text{ is odd} \\ \frac{n}{2} - 1 & \text{if } n \text{ is even} \end{cases} \quad (1)$$

As described above, only a single shortest path is considered for each pair of vertices. Multiple shortest paths connecting two vertices may exist, however. Thus, the polynomial number of RCPs represent only a subset of the exponential number of RCs. To identify all RCs on the basis of the RCPs, Vismara's algorithm uses a backtracking procedure. The set of RCs calculated during backtracking on the basis of a single RCP is defined as an RCF. This backtracking procedure includes the following steps:

First, for each RCP the set S_p of all shortest paths from p to r and the set S_q of all shortest paths from q to r are calculated. If an RCP is uneven, each combination of $P(r,p) \in S_p$ and $P(r,q) \in S_q$ forms an uneven RC with the edge $E(p,q)$ (see, for example, the 11-ring in Figure 2E). If an RCP is even, each combination of $P(r,p) \in S_p$ and $P(r,q) \in S_q$ forms an even RC with the edges $E(p,z)$ and $E(q,z)$.

Note that all RCs of an RCF have the same size. If their size is uneven, they share at least the vertices r, p, and q and the edge $E(p,q)$. Otherwise, they share at least the vertices r, p, q, and z and the edges $E(p,z)$ and $E(q,z)$. All RCFs of a molecular graph are disjoint with respect to their rings and their union forms the set of all RCs of a graph. In the following, the RCF of a ring C_x will be called RCF_x. Furthermore, we will write $E(RCF_x)$ and $V(RCF_x)$ to denote the union of the edges or vertices of all rings of an RCF_x, respectively.

Introduction of Unique Ring Families. On the basis of the RCs of a molecular graph, we define the terms URF-pair-related and URF-related as follows:

Definition 1. Let C_1 and C_2 be two RCs of a graph G, then C_1 and C_2 are *URF-pair-related* if and only if all of the following conditions hold:

1. $|C_1|=|C_2|$
2. $E(C_1) \cap E(C_2) \neq \emptyset$
3. It exists a set S of strictly smaller rings in G such that $C_1 \oplus (\oplus_{c \in S} c) = C_2$

Definition 2 . The URF-relation is defined as the transitive closure of the *URF-pair-relation*. A URF is defined as the set of URF-related RCs and hence represents an equivalence class. The length |URF| is defined as the length of each of its RCs. The number of URFs of a graph is called *URF-number*.

For an efficient calculation of molecular ring topologies in case of complex ringsystems, a description of rings should be at most polynomial in number with respect to the size of the graph. In the following, we estimate the URF-number of a molecular graph by comparing it to the polynomial number of RCFs.

Theorem 1. Any two rings of an RCF are URF-related.

Due to the construction of RCFs as described above, any two RCs of an RCF have identical lengths and share at least either an edge $E(p,q)$ or the edges $E(p,z)$ and $E(q,z)$. Thus, all rings of an RCF meet conditions 1 and 2 of definition 1. Furthermore, the RCs of an RCF differ only by alternative shortest paths replacing $P(r,p)$ or $P(r,q)$. As a consequence of eq 1, the following equation describes the length of two shortest paths used to construct an RCP of size n:

$$|P(r,p) \in S_p| = |P(r,q) \in S_q| < \frac{n}{2} \quad (2)$$

Since $P(r,p)$ contains less than half of the edges of the RCP, the symmetric difference of any two alternative shortest paths of S_p forms a set of rings which are smaller than n. Since the same is true for any two alternative paths of S_q, each two rings of an RCF can be constructed by cycle addition of each other and a set of smaller rings. Hence, all rings of an RCF meet condition 3 of definition 1. Consequently, any two RCs of an

dx.doi.org/10.1021/ci200629w | J. Chem. Inf. Model. 2012, 52, 2013−2021

RCF are URF-related and the URF-number is less or equal to the number of RCFs. Since the number of RCPs and RCFs is polynomial according to Theorem 4 of Vismara's paper,[1] the URF-number is at most polynomial, too.

Calculation of URFs. In the following, we provide an algorithm to calculate the polynomial number of URFs in polynomial time on the basis of the RCPs. The algorithm uses the described properties of RCPs as well as their linear dependency with respect to cycle addition in order to describe URFs by their edges sets.

Lemma 1 . Let C_A and C_B be two URF-related RCs, then C_A and C_B linearly depend on each other and a set of smaller rings with respect to cycle addition.

According to condition 3 of definition 1, two URF-pair-related RCs linearly depend on each other and a set of smaller rings with respect to cycle addition. Since a URF consists of the transitive closure of the URF-pair-relation, any two URF-related RCs linearly depend on each other and a set of smaller rings. Thus, URFs can be calculated in three steps.

1. Calculate RCPs according to Vismara's algorithm.
2. Let $B_<(G)$ be a subset of a minimum cycle basis B with $B_<(G) = \{C \in B | |C| < |C_A| = |C_B|\}$. Identify all 2-pairs of RCPs (C_A, C_B) with

$$C_A \oplus \left(\bigoplus_{c \in B_<(G)} c \right) = C_B \tag{3}$$

   Note that this operation is already performed during the calculation of RCPs. In Vismara's ring construction algorithm, a Gaussian elimination is used to eliminate rings which depend on smaller rings. Any ring C_A which depends on smaller and equal sized rings is marked as relevant. If the set of equal sized rings on which C_A depends on, only consists of a single ring C_B, C_A and C_B are marked as potentially URF-related. Furthermore, please note that any two rings of $RCF_A \cup RCF_B$ meet conditions 1 and 3 for being URF-pair-related.
3. If any two rings of RCF_A and RCF_B share an edge, these two rings are URF-pair-related. Since the URF-relation is an equivalence relation, C_A and C_B are URF-related if

$$E(RCF_A) \cap E(RCF_B) \neq \emptyset \tag{4}$$

In order to calculate RCPs according to Vismara's algorithm, rings which linearly depend on strictly smaller rings are eliminated. If a ring depends linearly on rings of the same size and strictly smaller rings, it is marked as relevant. All RCs of identical size which are identified in this step to be linearly dependent on each other and a set of smaller rings form pairs of possibly URF-related RCPs. For each RCP, all edges and vertices belonging to the same RCF can be identified using a simple breadth first search starting from r followed by a backtracking procedure involving the following steps:

1. Starting from r each vertex v is labeled according to its distance d_v to r using a breadth-first-search.
2. E_{cur} and V_{cur} represent the vertices and edges currently identified as belonging to E_{RCF} and V_{RCF}, respectively. V_{cur} is initialized with $V_{cur} \leftarrow \{p,q,z\}$ if C_A has even size and $V_{cur} \leftarrow \{p,q\}$ if C_A has uneven size. E_{cur} is initialized with $E_{cur} \leftarrow \{E(p,z),E(q,z)\}$ if C_A has even size and $E_{cur} \leftarrow \{E(p,q)\}$ if C_A has uneven size. A list Q of vertices is initialized with $Q \leftarrow \{p,q\}$.
3. For a vertex $v_{cur} \in Q$ identify each directly connected vertex v_{adj}. If $d_{vcur} - 1 = d_{v_{adj}}$, then
   - $E_{cur} \leftarrow E_{cur} \cup E(v_{cur},v_{adj})$
   - $V_{cur} \leftarrow V_{cur} \cup (v_{adj})$
   - $Q \leftarrow Q \cup v_{adj}$ if $v_{adj} \notin V_{curr}$
4. $Q \leftarrow Q \backslash v_{cur}$.
5. If $Q = \emptyset$, then $E_{cur} = E(RCF_A)$ and $V_{cur} = V(RCF_A)$. Otherwise, continue with step 3.

For a connected graph containing $|E|$ edges and $|V|$ vertices, RCPs can be calculated in $O(Z|E|^3)$ with $Z = |E| - |V| + 1$ being the cyclomatic number of G.[1] A Gaussian elimination to identify RCPs of identical size, which depend on each other and strictly smaller rings, can be performed in $O(|E|R^2)$ operations with R being the number of RCPs. The sets of edges belonging to each RCF are calculated in $O(|E|R)$. Finally, the edge set intersections of all 2-pairs of RCFs can be calculated in $O(|E|R^2)$. According to Visamara[1] the number of RCPs (R) is limited by the following relation:

$$R \leq 2|E|^2 + Z|V| \Rightarrow R \leq 2|E|^2 + |E||V| \tag{5}$$

Consequently, the Gaussian elimination and the calculation of the edge intersection of 2-pairs of RCPs are the speed-limiting steps and URFs can be perceived in $O(|E|^5+|V|^2)$. Thus, URFs represent a polynomial description of the ring topologies of a molecular graph and can be calculated in polynomial time.

Interpretation of URFs. From a chemical perspective, URFs can be best understood by calculating the union of the edges of all URF-related rings. Since a URF can contain smaller URFs, it can be illustrated by merging these smaller URFs to single nodes. This illustration represents a quotient graph of the partition of smaller URFs. Examples of molecular graphs and their corresponding RCs, RCPs and URFs are shown in Figures 3 and 4.

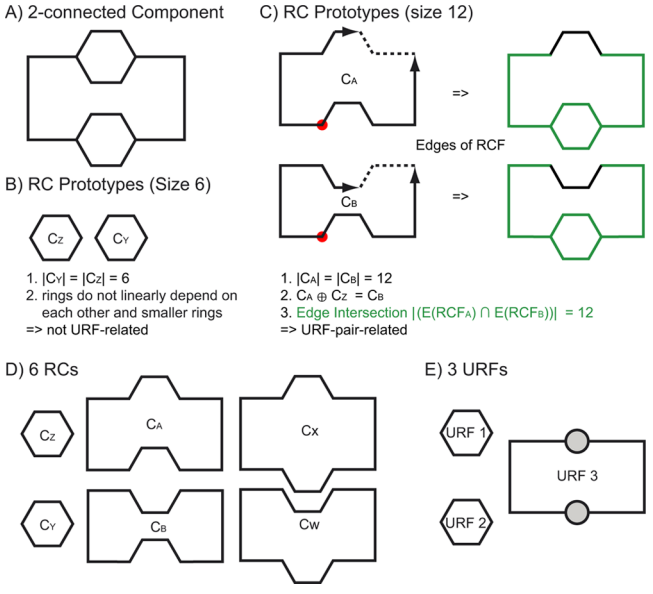


Figure 3. Ring system (A) containing 2 RCPs of size 6 (B) and 2 RCPs of size 12 (C). The two small rings form individual URFs (E). The two 12-rings belong to the same URF since they have the same size, share edges, and are linearly dependent on each other and one of the 6-rings. The molecular graph contains a total of six RCs (D) and three URFs (E). The URFs are illustrated as a quotient graph with the smaller URFs merged to individual nodes.

Compared to common strategies of ring perception, URFs have the major advantages that they are unique, intuitive, polynomial in number and provide a complete description of

dx.doi.org/10.1021/ci200629w | J. Chem. Inf. Model. 2012, 52, 2013−2021

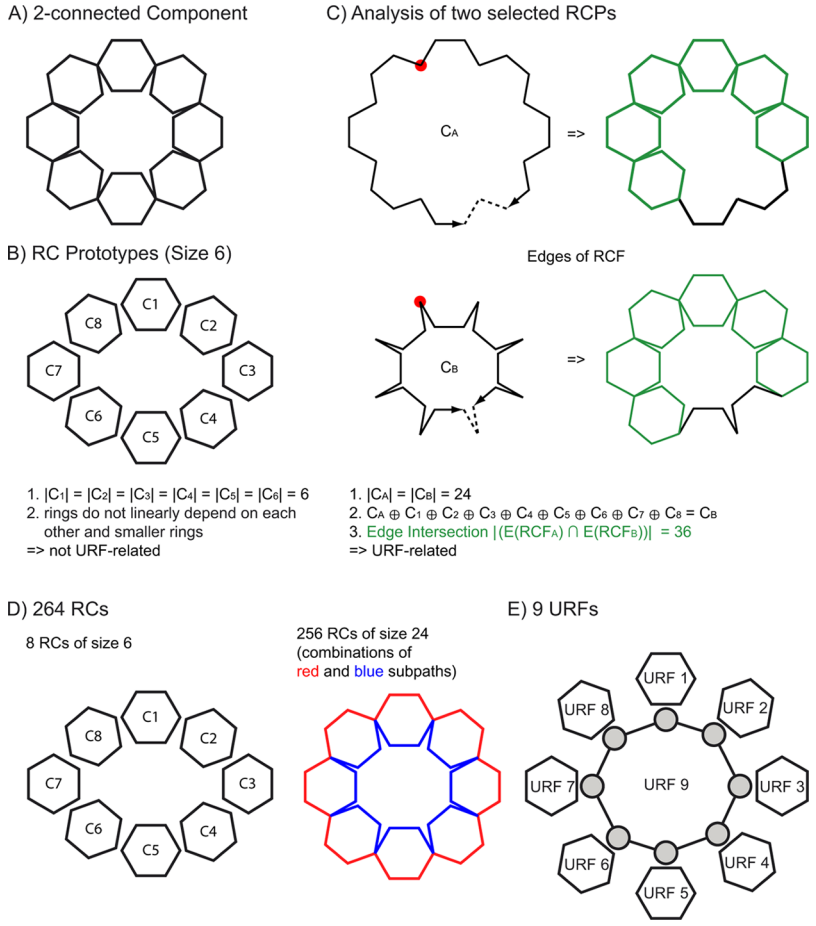
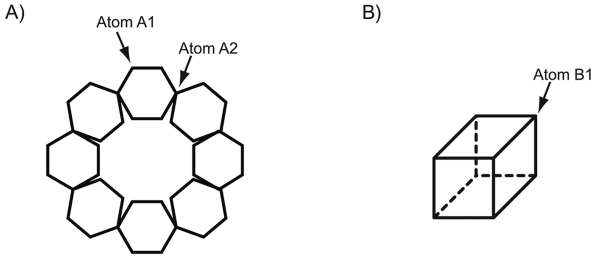


Figure 4. Ring system (A) consisting of 8 RCPs of size 6 (B) and 4 RCPs of size 24 of which 2 are illustrated (C). While the large RCPs have the same size and are linearly dependent according to condition 3 of definition 1, they do not share any edge. Their RCFs, however, share 36 edges. Note that this demonstrates, that two URF-related rings are not necessarily URF-pair-related. (D) The molecular graph contains 8 RCs of size 6 and 256 RCs of size 24. The set of all 264 RCs can be represented by 9 URFs. Eight URFs each contain a single 6-ring. One URF represents a macrocycle including the small URFs. This URF is illustrated as a quotient graph of the partition of smaller URFs. Note that the number of RCs increases exponentially with the number of para-bridged 6-rings, while the number of URFs increases linearly and stays intuitive.

the ring topology of a molecular graph. Macrocycles with para-substituated rings are a well-known problem (see Figures 3 and 4). The molecular structure shown in Figure 4 contains 264 RCs and 256 different possible SSSR cycle bases. The 256 large RCs belong to the same URF, resulting in 9 URFs. Thereby, URFs model the intuitive description of the molecule as a macrocycle containing eight smaller rings.

A frequently found specification in chemical patterns is the number of rings an atom is involved in. In the pattern language SMARTS, this is modeled with the R-feature. As discussed in the introduction, the R-feature is based on an SSSR which causes problems due to nonuniqueness. So far, no alternative approach resulting in a unique and polynomial number of ring representatives was available. Describing atoms by the number of URFs they are involved in represents an easy to implement solution to this problem.

Figure 5A shows the number of rings that contain the atoms A1 and A2. Using SSSRs, the result depends on the selected cycle base. In contrast, the number of RCs is large and chemically nonintuitive. Similar problems occur for symmetric cyclic structures like cubane (see Figure 5B). The calculation of URFs results in a consistent and chemically meaningful value for each atom. Furthermore, if an application requires the construction of an MCB, this can be easily achieved by selecting a



Ring Concept	Molecule A	Atom A1	Atom A2
URFs	9	2	3
SSSR	9	1 or 2	3
RCs	264	129	258

Ring Concept	Molecule B	Atom B1
URFs	6	3
SSSR	5	2 or 3
RCs	6	3

Figure 5. Two complex ring systems with their number of SSSR-rings, relevant cycles, and unique ring families. Additionally, ring memberships for the marked atoms are listed.

single arbitrary RCP of each URF followed by a Gaussian elimination of the resulting set of rings. Since the number of URFs is greater than or equal to the number of cycles of an MCB and smaller than or equal to the number of RCPs, the URF-number can be estimated by the following equation:

$$(E - V + 1) \leq \text{URF} - \text{number} \leq (2E^2 + EV) \qquad (6)$$

dx.doi.org/10.1021/ci200629w | J. Chem. Inf. Model. 2012, 52, 2013−2021

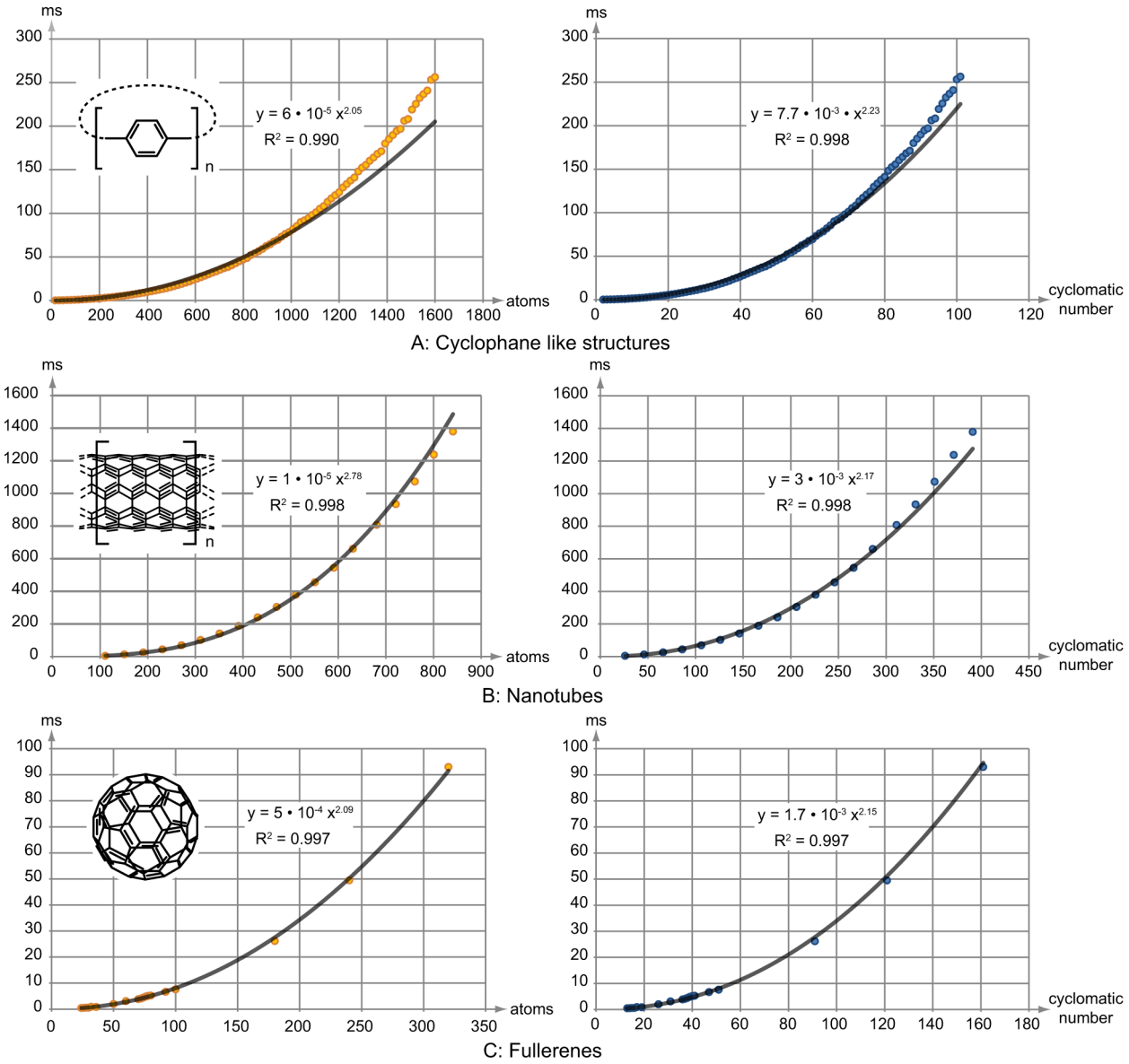


Figure 6. Required runtimes for the calculation of URFs depending on the number of atoms (left) and the cyclomatic number (right) for cyclophane-like structures (A), nanotubes (B), and fullerenes (C).

■ COMPUTING TIME BENCHMARKS

Ring perception is an important step in almost all chemoinformatics tasks. Applications which process large data sets thus require a fast method to identify the rings of molecular graphs. To check the large-scale applicability of the described method to calculate URFs, we measured the runtimes for the perception of URFs for a number of test sets. Time measurements were performed in a single thread on a PC with an Intel Core2 Quad Q9550 CPU (2.83 GHz) and 4 GB of main memory. For each molecule of the data set, the runtime for 100 iterations of ring perception was measured and on the basis of this measurement, the average runtime for a single ring perception was calculated. For file-IO we used the NAOMI framework.[19] Measured runtimes shown in Figure 7 do not include file I/O and molecular preprocessing. The data structures of the NAOMI framework are not specifically optimized for the detection of URFs but focus on the correct chemical modeling of small molecules. The listed runtimes thus provide an estimate of URF detection in the context of a common cheminformatic application.

To investigate the maximum runtime for the perception of URFs, we generated a number of molecules containing highly complex ring systems. First, we generated cyclophane-like structures that contain a large macrocycle with n para-bridged 6-rings. The generated molecules have a cyclomatic number Z_n of $Z_n = n + 1$, contain $n^2 + n$ RCs and $n + 1$ URFs. The runtime for the calculation of the URFs of these molecules is shown in Figure 6A. The required runtime for molecules containing $|V|$ atoms and a cyclomatic number of Z increases approximately with $|V|^2$ and Z^2.

As a second type of molecules that contain complex rings, single walled nanotubes were generated using ConTub.[20] While the parameters i and k were set to 5 nm, the length of the nanotube was increased in steps of 5 nm starting with a length of 10 nm up to a maximum of 100 nm. Both V and Z increase linearly with the length of the nanotube. As shown in Figure 6B,

dx.doi.org/10.1021/ci200629w | J. Chem. Inf. Model. 2012, 52, 2013−2021

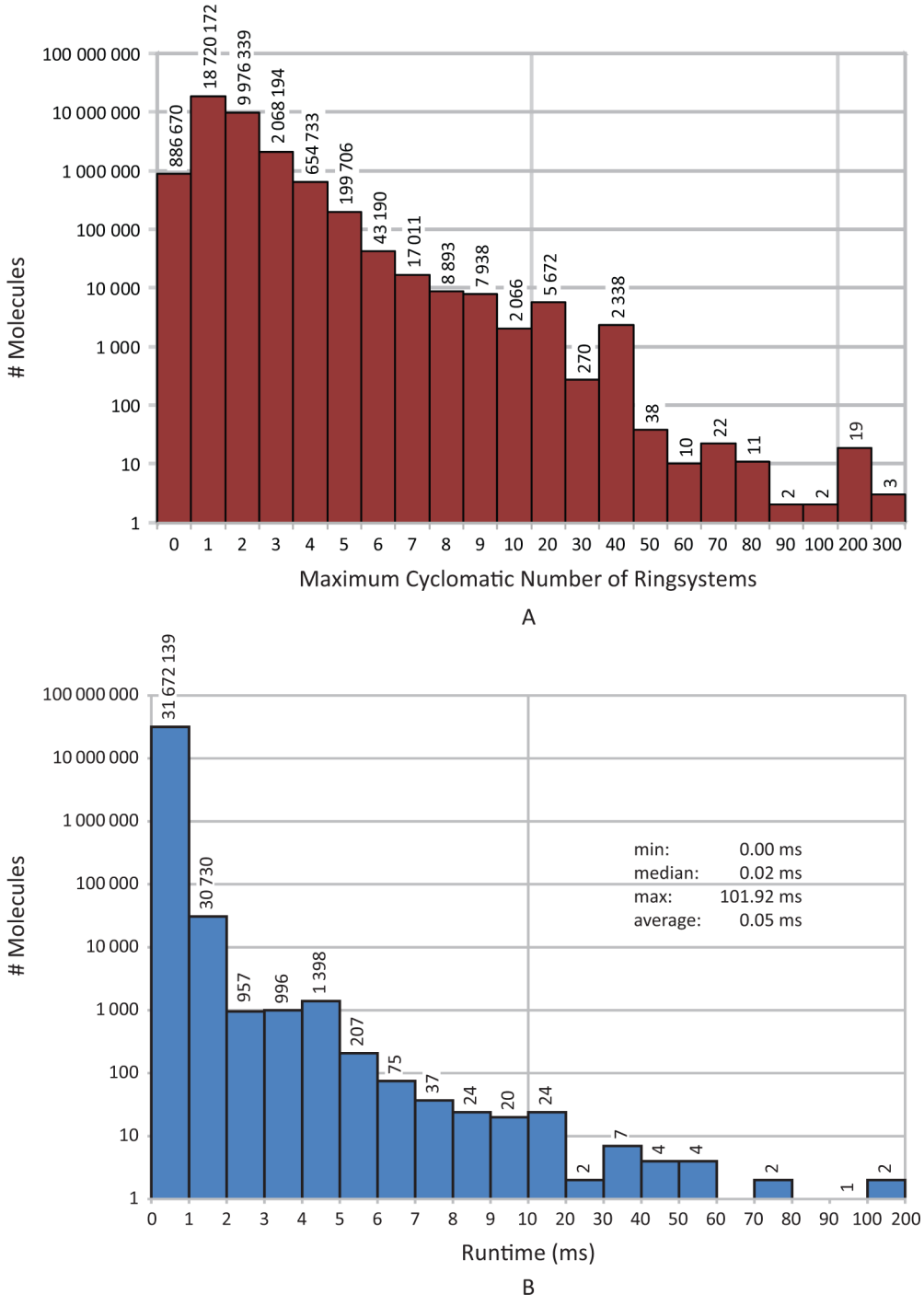


Figure 7. (A) Maximum cyclomatic number of the ringsystems of the molecules of the Pubchem-2D data set. (B) Benchmarks for URF perception for those molecules of the PubChem-2D data set having a cyclomatic number of at least one.

the runtime for the calculation of URFs increases slower than V^3 or Z^3.

As a third set of complex molecules, a number of fullerenes ranging from C24 to C320 were generated. Coordinates of these molecules were taken from a Fortran program specialized in the generation of fullerenes.[21] The runtime requirement again increased approximately with V^2 as well as with Z^2 (see Figure 6C).

Finally, to investigate the runtime which is required to perceive rings of commonly used molecules, we perceived URFs for the PubChem Compound 2D data set.[3] The data set was downloaded on March 27th, 2011 from ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/ and contains 32 593 299 molecular structures. These include a number of molecules of high complexity not present in the respective 3D data set. Figure 7A illustrates the complexity of the data set by showing the maximum cyclomatic number for the ringsystems of each molecule.

Shown runtimes represent the required runtime for 100 iterations of ring perception. Nevertheless, these runtimes are close to zero for most common molecules. The median for the percerption of URFs for a molecule of the Pubchem Data set is

dx.doi.org/10.1021/ci200629w | J. Chem. Inf. Model. 2012, 52, 2013−2021

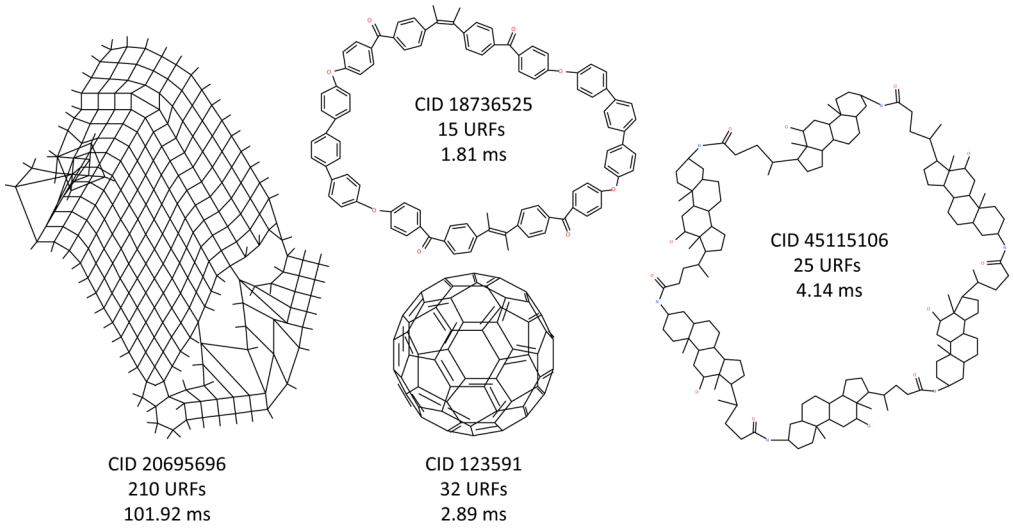


Figure 8. Runtimes and compound IDs for a selection of molecules of the Pubchem-2D data set.

0.02 ms, the average runtime is 0.05 ms, and the maximum runtime is 102 ms. This demonstrates that URFs can be calculated on the fly even for interactive applications and large databases. Only 34 490 molecules (0.11% of the database) show runtimes of more than 1 ms for the calculation of URFs. A list of those 100 molecules which require the highest runtimes for the calculation of URFs is added to this paper as Supporting Information. Some representative examples are shown in Figure 8.

A common molecular file format conversion, tested with Open Babel for the ZINC-everything data set, requires approximately 2 ms.[19] Due to the low runtime for calculating URFs of about 0.02 ms for commonly used molecules, the perception of URFs is suitable for high throughput chemoinformatics applications. Even for an artificially complex cylophane-like structure containing $100 + 2^{100}$ RCs, the URFs can be calculated in less than 2 s.

CONCLUSION

We have introduced the concept of unique ring families (URFs). In contrast to common ring perception approaches, URFs are polynomial in number, unique, and provide a complete description of the rings of a molecular graph. Furthermore, we have described an efficient method to calculate URFs in polynomial time. We demonstrated its applicability on large scale by showing computing time benchmarks for the Pubchem 2D data set. For these reasons, URFs represent a valuable alternative to common ring perception concepts and are worthwhile to be considered as a standard description for ring topologies in molecular graphs.

ASSOCIATED CONTENT

Supporting Information

100 molecular structures of the PubChem Database which require the highest runtimes for the perception of URFs. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Present Address

‡Evotec AG, Essener Bogen 7, 22419 Hamburg. Phone: 0049 40 56081 230. Email: Adrian.Kolodzik@evotec.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Christian Ehrlich and J. Robert Fischer for helpful comments and proofreading of the manuscript. Furthermore, the authors thank J. Robert Fischer and Tobias Lippert for their work on the NAOMI framework, which was used for reading the molecules of the Pubchem data set.

REFERENCES

(1) Vismara, P. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.* **1997**, *4*, 1−15.
(2) Plotkin, M. Mathematical Basis of Ring-Finding Algorithms in CIDS. *J. Chem. Doc.* **1971**, *11*, 60−63.
(3) Wang, Y.; Xiao, J.; Suzek, T.; Zhang, J.; Wang, J.; Bryant, S. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, 623−33.
(4) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40−43.
(5) Berger, F.; Flamm, C.; Gleiss, P.; Leydold, J.; Stadler, P. Counterexamples in Chemical Ring Perception. *J. Chem. Inf. Model.* **2004**, *44*, 323−331.
(6) Hanser, T.; Jauffret, P.; Kaufmann, G. A New Algorithm for Exhaustive Ring Perception in a Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1146−1152.
(7) Balducci, R.; Pearlman, R. S. Efficient exact solution of the ring perception problem. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 822−831.
(8) Carta, G.; Onnis, V.; Knox, A.; Fayne, D.; Lloyd, D. Permuting input for more effective sampling of 3D conformer space. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 179−190.
(9) *Daylight Theory Manual 4.9*. http://www.daylight.com/dayhtml/doc/theory/index.pdf (accessed June 9th, 2012).
(10) *Daylight Depictmatch*. http://www.daylight.com/daycgi_tutorials/depictmatch.cgi (accessed June 9th, 2012).
(11) Petra M. Gleiss, J. L.; Stadler, P. F. Interchangeability of Relevant Cycles in Graphs. *Electron. J. Comb.* **2000**, 1−16.
(12) Fujita, S. A new algorithm for selection of synthetically important rings. The essential set of essential rings for organic structures. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 78−82.

(13) Corey, E.; Perersson, G. Algorithm for machine perception of synthetically significant rings in complex cyclic organic structures. *J. Am. Chem. Soc.* **1972**, *94*, 460−465.
(14) Wipke, W.; Dyott, T. Use of Ring Assemblies in Ring Perception Algorithm. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140−147.
(15) Downs, G.; Gillet, V.; Holliday, J.; Lynch, M. Theoretical aspects of ring perception and development of the extended set of smallest rings concept. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187−206.
(16) Nickelsen, H. Ringbegriffe in der Chemie-Dokumentation. *Nachr. Dok.* **1971**, *3*, 121−123.
(17) Dury, L.; Latour, T.; Leherte, L.; Barberis, F.; Vercauteren, D. A new graph descriptor for molecules containing cycles. Application as screening criterion for searchingmolecular structures within large databases of organic compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1437−1445.
(18) Tarjan, R.; Vishkin, U. An Efficient Parallel Biconnectivity Algorithm. *SIAM J. Comput.* **1985**, *14*, 862−874.
(19) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.
(20) Melchor, S.; Martin-Martinez, F. J.; Dobado, J. A. CoNTub v2.0 - Algorithms for Constructing C3-Symmetric Models of Three-Nanotube Junctions. *J. Chem. Inf. Model.* **2011**, *51*, 1492−1505.
(21) Schwerdtfeger, P. *Topological Analysis of Fullerenes - A Fortran Program*. http://ctcp.massey.ac.nz/index.php?group=page=fullerenes menu=fulleren (accessed March 11th, 2012).