Supplementary Material for Sharing Chemical Information Without Revealing Structures

Matthew Matlock and S Joshua Swamidass

July 8, 2013

1 An Algorithm for Estimating the Number of Scaffolds in an Equivalence Class

The calculation given in the discussion suggests that the number of distinct scaffolds consistent with a equivalence class is an astronomically high. This simple calculation is not exact, because it does not take into account the structural constraints implied by the scaffold network topology. Some of these, such as symmetry and degeneracy with respect to a scaffold's molecular topology are discussed in the paper. In this section, we describe the algorithm which we used to extract the implied constraints from a given scaffold network topology. This algorithm constructs constraints using an example scaffold belonging to the scaffold equivalence class having this topology. We then use these constraints to give an estimated lower bound on the number of possible scaffolds that could exist within this equivalence class, by building other scaffolds belonging to this equivalence class using rings and linker chains observed in the PubChem data.

Figure 1 provides a visual description of the 5 step process for generating lower bounds on the potential size of a given scaffold network equivalence class. The first three steps of this process involve decomposing a sample scaffold into a set of constraints. These constraints can then be utilized (step 4) to identify scaffold components (specific rings and linker chains) which could be used to build other scaffolds within the class. Finally (step 5), we utilize these identified rings and linkers to count the number of possible scaffolds fitting this equivalence class using a simple combinatoric calculation. We discuss each step in detail below.

In step 1, we search the PubChem database for an example scaffold which belongs to the scaffold equivalence class using SNG.

In step 2, we identify the constituent components of the scaffold (Rings and Linker Chains). We compute the smallest set of smallest rings (SSSR) using the openbabel software, and construct canonical SMILES representations of each of these rings. Next, each atom not belonging to a ring, is identified as a âĂIJlinkerâĂİ atom. We additionally identify the ring atom connected to an adjacent linker as a part of the linker. For fused ring systems, the linkers are defined as those atoms which are shared between the rings. We find the connected subsets among these atoms (the Linker Chains) and compute canonical SMILES representations of each of these linker chains. Next, we count the distinct rings and linker chains present in the molecule by comparing canonical SMILES representations.

In step 3, we identify molecular symmetries which impact the scaffold network topology. For each ring r in the example scaffold, we find all neighboring rings (that is, rings connected by at most 1 linker chain), and then, for each possible pair of neighbors n_i, n_j with linkers l_i, l_j linking n_i, n_j to r, we construct a sub-scaffold consisting of n_i, n_j, r, l_i, l_j. We then evaluate this sub-scaffold for any axes of symmetry. Then, we assign ring r a symmetry value s, corresponding to the number of identified axes of symmetry. We also compute required binding positions along each axis of symmetry, distinguishing between equatorial and axial binding positions around the axis of symmetry.

In step 4, we use the symmetry constraints found in step 3 to select potential replacement rings from the PubChem database which have exactly this number of symmetries, as well as the required number of axial and equatorial binding positions for each axis of symmetry. We identify satisfying ring sets for each distinct ring identified in the original example scaffold. If multiple rings of the same variety were observed with different symmetry constraints, the strictest set of symmetry constraints are used to select potential replacement rings from the database.

Finally, in step 5 we use these sets of rings and linkers to calculate a lower bound on the size of the given scaffold network equivalence class.

Finding the exact size of a scaffold network equivalence class using combinatoric methods can involve complex calculation. In particular, the overlaps among the sets of potential rings R and among the sets of potential linkers L are frequently complex and vary significantly between scaffold equivalence classes. We must count the number of possible

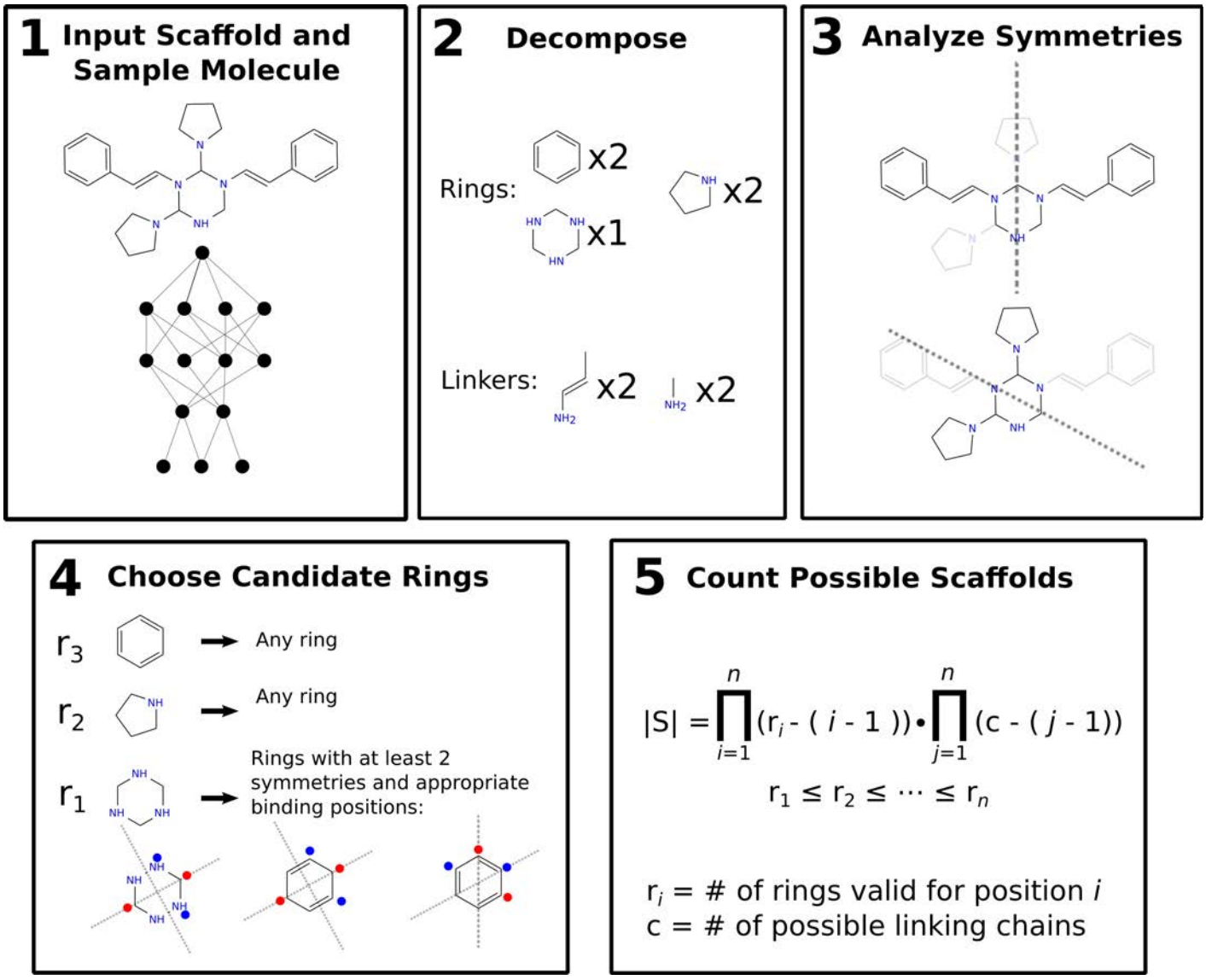

Figure 1: Algorithm for estimating the number of scaffolds in an equivalence class. Each scaffold defines constraints that are sufficient for identifying scaffolds in the same equivalence class. The scaffold is first decomposed into rings and linkers. The number of unique rings and linkers are counted. Next, each ring is compared for axes of symmetry with respect to each pair of adjacent rings. These symmetry constraints are then used to select rings from the PubChem database that fit each position. Finally, we count the number of potential scaffolds belonging to this equivalence class.

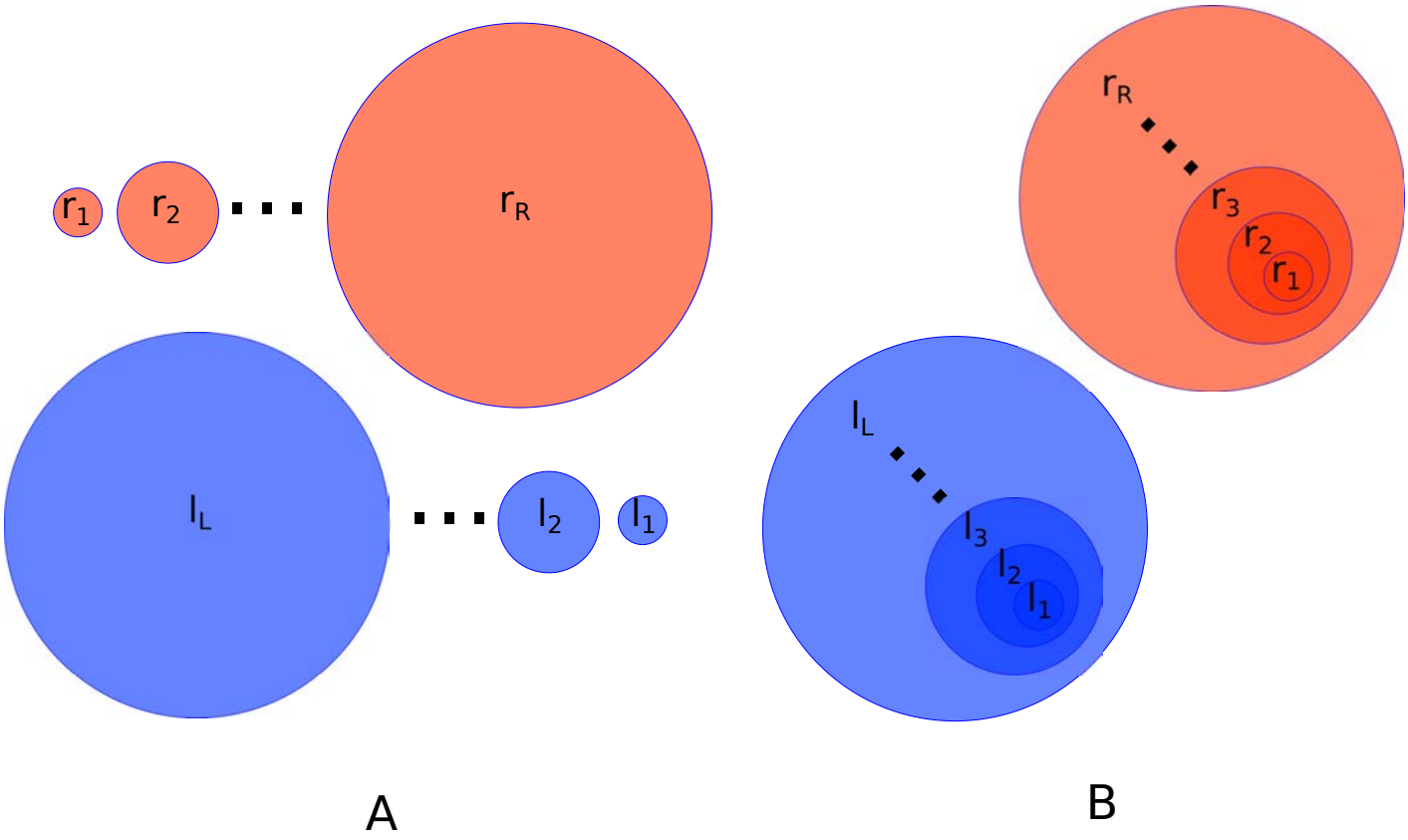

Figure 2: Examples of potential sets of Rings and Linkers. A. The number of possible choices of $R + L$ distinct items is maximized when these sets are perfectly disjoint. B. When the sets are perfectly overlapping, the number of possible choices of $R + L$ distinct items is minimized.

choices of $R + L$ distinct items from $R + L$ possibly overlapping sets. In the simplest case, each of the $R + L$ sets are distinct and non-overlapping, as shown in Figure 2A. In this case, we can easily compute the exact combinatoric value as:

$$\prod_{i=1}^{R} r_i \cdot \prod_{j=1}^{L} l_j$$

However, as mentioned, this is not the case in many of our scaffold constraint systems. Instead, we chose to underestimate the size of an equivalence class by assuming that each of the R ring sets and L linker sets are perfectly overlapping, as shown in Figure 2B. The computation in this case is equally simple:

$$\prod_{i=1}^{R} [r_i - (i-1)] \cdot \prod_{j=1}^{L} [l_j - (j-1)]$$

Since we calculate the worst-case overlap, and find constraints on the possible rings and linkers that are at least as stringent as those implied by the scaffold network topology, we are guaranteed to compute lower bound estimates for the size of a given scaffold network equivalence class. As noted in the paper, this can result in significant undercounting in some cases, but still reveals that the average size of the scaffold equivalence classes found in the PubChem data studied in this paper is astronomically large (average lower bound of 10^{52}).