Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag

Tairan Liu,† Misagh Naderi,‡ Misagh Naderi,‡ Chris Alvin,§ Supratik Mukhopadhyay,‖ and Michal Brylinski†‡§∥

© 2017 American Chemical Society

ABSTRACT: Constructing high-quality libraries of molecular building blocks is essential for successful fragment-based drug discovery. In this communication, we describe eMolFrag, a new open-source software to decompose organic compounds into nonredundant fragments retaining molecular connectivity information. Given a collection of molecules, eMolFrag generates a set of unique fragments comprising larger moieties, bricks, and smaller linkers connecting bricks. These building blocks can subsequently be used to construct virtual screening libraries for targeted drug discovery. The robustness and computational performance of eMolFrag is assessed against the Directory of Useful Decoys, Enhanced database conducted in serial and parallel modes with up to 16 computing cores. Further, the application of eMolFrag in de novo drug design is illustrated using the adenosine receptor. eMolFrag is implemented in Python, and it is available as stand-alone software and a web server at www.brylinski.org/emolfrag and https://github.com/liutairan/eMolFrag.

INTRODUCTION

Hit identification, lead generation, and lead optimization are the key steps at the outset of a drug discovery process. Briefly, compounds showing promising activity identified by high-throughput screening as initial hits are filtered and modified to generate lead compounds, which satisfy basic drug-likeliness properties.¹ These lead compounds are further optimized to enhance the potency toward the target protein as well as to reduce their nonspecificity and toxicity.² Conventional hit identification is not only limited to already synthesized compounds often leading to low discovery rates, but it is also expensive and requires time-consuming screening experiments.³ Consequently, virtual screening that can rapidly evaluate millions of compounds has become an integral part of lead identification protocols.⁴

In order to enhance the chemical diversity of virtual screening libraries, large collections of drug-like compounds can be generated through combinatorial chemistry.⁵ Since constructing and screening the entire chemical space are not feasible even with the most advanced computers, building extensive yet targeted libraries is critical for the success of virtual screening. A number of fragment- and atom-based techniques have been developed to generate novel chemical compounds for virtual screening, including binding-site point connection methods (LUDI⁶), fragment connection methods (LEA3D),⁷ LigBuilder,⁸ and eSynth⁹), sequential build-up algorithms (LEGEND¹⁰ and SPROUT¹¹), and random connection techniques (CoG¹² and Flux¹³). These de novo methods require an initial set of building blocks or molecular fragments, which ultimately control the properties of the resulting screening compounds and their affinity toward the target protein. Consequently, there is a great interest in efficient fragmentation techniques to generate sets of chemically feasible building blocks for the subsequent molecular synthesis. Retrosynthetic combinatorial analysis procedure (RECAP¹⁴) and breaking retrosynthetically interesting chemical substructures (BRICS¹⁵) are examples of systematic fragmentation methods. In RECAP, compounds are dissected based on a set of 11 bond types, following simple rules such as leaving cyclic bonds and alkyl groups smaller than five carbons intact. These rules ensure that major structural features of organic compounds, such as ring motifs, are preserved. BRICS expands the bond type criteria used by RECAP from 11 to 16 taking into account the chemical environment of each bond type and the surrounding substructures. Additional filters are also applied in order to prevent generating small and unwanted fragments. Other methods extract and classify chemical scaffolds by pruning side chains and removing peripheral ring moieties.¹⁶

In general, the performance of fragment-based chemical synthesis tools such as eSynth, CONFIRM,¹⁷ and AutoGrow¹⁸ could significantly be improved by employing building blocks annotated with empirical connectivity patterns. Although this information could help explore pharmacologically relevant regions of the diverse chemical space,⁹ many existing fragmentation tools, e.g., Fragmenter and molBLOCKS,¹⁹

Received: October 1, 2016
Published: March 27, 2017
do not consider the chemical context of the fragments. In other words, the connectivity information on a fragment is not stored while extracting building blocks. To address this issue, we developed cMolFrag, a new open-source molecular fragmentation software. cMolFrag decomposes either a single ligand or a library of compounds into two types of chemical building blocks, bricks and the connecting linkers. The resulting complete and nonredundant sets of building blocks are annotated with the comprehensive connectivity information in order to facilitate the construction of novel compounds with combinatorial synthesis software. cMolFrag has been parallelized to decrease the computing time required to analyze large collections of molecules.

**METHODS**

cMolFrag employs a graph-based notation, where molecules are sets of nodes representing atoms connected by edges corresponding to chemical bonds. A fragment is a substructure, which has either all or only some atoms and bonds of a given molecule; fragments are categorized as either bricks or linkers. Given a collection of molecules, the complete set of unique fragments is constructed in two steps shown in Figure 1. The first step, labeled as Part I, involves creating an initial set of fragments, whereas the second step, labeled as Part II, guarantees the uniqueness of the resulting set of fragments.

**Part I: Fragmentation.** In cMolFrag, a set of molecules are first decomposed into constituent fragments with the BRICS algorithm, implemented in RDKit. Chemical compounds are broken down into larger moieties called bricks connected by linkers based on 16 chemical environments defined by the BRICS model; a pseudocode for the fragmentation process is given in the Supporting Information (Algorithm S1). A brick fragment is a molecular construct having at least four non-hydrogen atoms. Subsequently, bricks are removed from a molecule and the remaining fragments are classified as linkers (see Algorithm S2 in the Supporting Information). Broken bonds are replaced by dummy atoms, which are placeholders for those atoms removed from a particular bond. The complete information, including the type of atoms involved in those bonds that were broken, is stored for each brick in order to provide empirical connectivity patterns. Linkers have different auxiliary connectivity information, i.e. these fragments are annotated only with the maximum number of bonds at various positions. Examples of bricks and linkers are provided in the Supporting Information (Examples S1 and S2, respectively). We found that this approach allows to efficiently construct series of new molecules, whose chemistry is similar to that of parent compounds.

**Part II: Mitigation of Fragment Redundancy.** Since one of the objectives of an effective fragmentation procedure is to employ the resulting fragments in a synthesis procedure, the cardinality of the final set of fragments is critical. On that account, cMolFrag attempts to minimize the size of sets of bricks and linkers by removing redundancy with a partitioning and sieve-based removal scheme presented in the Supporting Information (Algorithm S3). Two fragments are equivalent if the Tanimoto coefficient (TC) calculated for topologically constrained maximum common substructures by the kcombu program is equal to 1.0. Information on equivalent atoms provided by kcombu as well as their connectivity information is then used to consolidate identical fragments into a single, unique construct.

**RESULTS AND DISCUSSION**

**Benchmarks against the DUD-E Database.** We validate the cMolFrag algorithm by conducting a self-reconstruction test as described previously. Briefly, given an input molecule \( m \), a set of fragments extracted from \( m \) by cMolFrag are passed to a fragment-based construction procedure with eSynth employing its chemical rules. A molecule with the highest chemical similarity to \( m \) measured by the TC calculated for Daylight fingerprints is selected from a series of compounds constructed by eSynth. Here, we employ a fingerprint-based assessment of chemical similarity with OpenBabel because this technique is computationally much faster than kcombu. A TC of \( \geq 0.8 \) indicates that a molecule highly similar to \( m \) was generated, whereas a TC of 1.0 indicates that compound \( m \) has been reconstructed. As a testing set, we use 20 408 active compounds for 102 protein targets from the Directory of Useful Decoys, Enhanced database covering a diverse chemical space of pharmacologically relevant molecules.

The performance of cMolFrag is compared to molBLOCKS, another fragmentation software employing the RECAP algorithm. Figure 2 shows a two-way box plot of the number of atoms per fragment and the number of fragments per molecule for these two programs. Fragments generated by MolFrag are passed to a fingerprint-based similarity search against the DUD-E database using Tanimoto coefficients calculated for Daylight fingerprints.
mOLBLOCKS typically contain 6–10 (the default protocol) and 6–11 (an extensive mode) atoms, whereas most fragments extracted by eMolFrag consist of 2–7 atoms. The median numbers of fragments per molecule are 3, 8, and 6 for mOLBLOCKS (default), mOLBLOCKS (extensive), and eMolFrag, respectively. Finally, molecular synthesis with eSynth was conducted employing fragments generated by eMolFrag for active compounds in the DUD-E database. Encouragingly, 82.8% of active compounds were reconstructed with a TC of 0.8. An inspection of the failed cases revealed that the major reason for not generating a relatively small fraction of testing compounds is the fact that the synthesis software does not allow to directly connect two bricks. Overall, the self-reconstruction benchmarking results demonstrate that eMolFrag properly extracts molecular fragments providing sufficient connectivity information to rebuild the majority of parent molecules.

Computational Performance. Decomposing large compound libraries can be time-consuming depending on the number of input molecules; therefore, we parallelized the eMolFrag code. The serial and parallel performance of eMolFrag is assessed by fragmenting subsets of DUD-E actives with sizes varying from 100 to 12800 molecules. All tests were performed on a machine equipped with two 2.6 GHZ 8-core Sandy Bridge Xeon 64-bit processors, 32GB 1666 MHz RAM and 500GB HD, running Red Hat Enterprise Linux 6. Figure 3 shows the wall time for the complete fragmentation procedure plotted over a fixed input data set of 3200 molecules and the number of computing cores varying from 1 to 16 roughly corresponds to a hypothetical code consisting of 47–60% parallel calculations. Note that eMolFrag does not conform exactly to Amdahl’s law because the workload related to removing redundancy (Part II in Figure 1) is unevenly distributed across computing cores. Although the total execution time of eMolFrag diverges from Amdahl’s law, the parallel processing is faster than the serial execution. The average processing speed for the parallel code running on 16 computing cores ranges from 24 molecules per second for the smallest data set to 11.8 molecules per second for the largest data set (see Table S1 in the Supporting Information). This shorter processing time for parallel eMolFrag becomes particularly beneficial for larger data sets. For instance, decomposing 20408 active compounds from the DUD-E data set for the self-benchmarking test takes 1 h and 18 min on a single core compared to only half an hour on 16 computing cores.

Application to Antagonists of the Adenosine Receptor. To illustrate the application of eMolFrag in de novo drug discovery, we show that bioactive compounds can successfully be constructed from molecular fragments extracted from chemically dissimilar binders of the same target protein. Here, we selected the human adenosine A2a receptor (AA2AR), a member of the G protein-coupled receptor (GPCR) superfamily containing targets for about 27% of all FDA-approved drugs. Figure 4 presents individual steps of the cross-validation procedure, in which CHEMBL144979, a known bioactive ligand for AA2AR, is the target molecule. Four other AA2AR antagonists, called donors, are shown in Figure 4A. Since the chemical similarity of donors to the target, measured by the TC reported by kcombi, is lower than 0.5, CHEMBL144979 can be considered novel with respect to the donor molecules.

Without removing redundancy, the average processing speed of serial eMolFrag is 9.8 molecules/s for the smallest data set and 23.2 molecules/s for the largest data set. For comparison, a serial version of mOLBLOCKS, which does not remove redundancy, is capable of processing 6.6 and 12.5 molecules/s for the smallest and the largest data sets, respectively. Thus, eMolFrag is 1.2–1.9x faster than mOLBLOCKS. Further, algorithms implemented in eMolFrag are polynomial in complexity; the best-fit curves in Figure 3 are $y = 0.022x^{1.238}$ ($R^2 = 0.99989$) for serial and $y = 0.013x^{2.201}$ ($R^2 = 0.99989$) for parallel execution. This near-linear scaling gives empirical evidence of the efficient implementation of eMolFrag.

The impact of the number of computing cores on parallel processing is assessed by comparing the performance of parallel eMolFrag to the theoretical speedup estimated with Amdahl’s law.26 The inset in Figure 3 shows that executing eMolFrag in parallel for a fixed input data set of 3200 molecules and the number of computing cores varying from 1 to 16 results in a speedup of 2.4 compared to only half an hour on 16 computing cores.

Figure 3. Serial and parallel performance of eMolFrag. The main graph shows the wall time for the complete fragmentation procedure plotted over a fixed size input data set of 3200 molecules and the number of computing cores varying from 1 to 16 roughly corresponds to a hypothetical code consisting of 47–60% parallel calculations. Note that eMolFrag does not conform exactly to Amdahl’s law because the workload related to removing redundancy (Part II in Figure 1) is unevenly distributed across computing cores. Although the total execution time of eMolFrag diverges from Amdahl’s law, the parallel processing is faster than the serial execution. The average processing speed for the parallel code running on 16 computing cores ranges from 24 molecules per second for the smallest data set to 11.8 molecules per second for the largest data set (see Table S1 in the Supporting Information). This shorter processing time for parallel eMolFrag becomes particularly beneficial for larger data sets. For instance, decomposing 20408 active compounds from the DUD-E data set for the self-benchmarking test takes 1 h and 18 min on a single core compared to only half an hour on 16 computing cores.

Unique sets of 10 bricks and 7 linkers extracted by eMolFrag from 4 donors are shown in Figures 4B and C, respectively. For instance, the triazolo-quinoxaline fragment highlighted in pink carrying the chlorine moiety was obtained from CHEMBL95229. This compound is a member of a series of pyrazolo-triazolo-pyrimidines with subnanomolar affinity against ARs created via N5-phenylcarbamoyl substitutions.29 Bricks contain information on atom types that can be attached to various positions (small boxes in Figure 4B), whereas linkers are annotated with the maximum number of allowed bonds (small circles in Figure 4C). The sets of bricks and linkers are complete and nonredundant, i.e. each unique fragment carries
the connectivity information extracted from multiple donor compounds. For example, the connectivity information for a benzene ring, which is present in all donors, is consolidated by eMolFrag into a single fragment shown in cyan in Figure 4B.

Subsequently, molecular fragments extracted by eMolFrag were passed to eSynth in order to generate a series of compounds. A serial version of eSynth produced 4,492,609 virtual compounds in 12 h. Encouragingly, the first compound in Figure 4D (shown in a box) is CHEMBL144979; therefore, the target molecule has been successfully constructed. Further, the set of virtual molecules comprises 845 compounds, whose TC to CHEMBL144979 is ≥ 0.7 and as many as 239,656 molecules with a TC of ≥ 0.5. Three randomly selected virtual molecules are presented in Figure 4D to demonstrate the chemical diversity of compounds generated by eSynth. It is important to note that these retrospective cross-validation benchmarks are designed to mimic real applications by attempting to construct target molecules using building blocks extracted from chemically dissimilar compounds. This case study demonstrates that high-quality fragment sets generated by eMolFrag can be used in fragment-based drug discovery to create targeted screening libraries likely containing novel bioactives.

**CONCLUSIONS**

eMolFrag is a fast and robust tool to extract molecular fragments, classified as bricks and linkers, from small molecule data sets. Subsequently, these fragments can be used to construct targeted libraries for virtual screening. A unique feature of eMolFrag is that it stores the connectivity information for the extracted building blocks to help generate new series of chemically feasible compounds. Although eMolFrag was optimized to work with eSynth, a recently developed molecular synthesis algorithm, it can also be integrated into other cheminformatics toolkits utilizing chemical fragments. eMolFrag is freely available as stand-alone software and a Web server at www.brylinski.org/emolfrag and https://github.com/liutairan/eMolFrag.

**ASSOCIATED CONTENT**

5 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00596.

eMolFrag algorithms, computing speeds, and file formats (PDF)
**REFERENCES**

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. | Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 2001, 46, 3–26.

(2) Palermo, G.; De Vivo, M. | Computational Chemistry for Drug Discovery. In *Encyclopedia of Nanotechnology*; Bhushan, B., Ed.; Springer: Netherlands, Dordrecht, 2014; pp 1–15.

(3) Leelananda, S. P.; Lindert, S. | Computational methods in drug discovery. In *Encyclopedia of Nanotechnology*; Bhushan, B., Ed.; Springer: Netherlands, Dordrecht, 2014; pp 1–15.

(4) Mestres, J. | Virtual screening: a real screening complement to high-throughput screening. *Biochem. Soc. Trans.* 2002, 30, 797–799.

(5) Lipinski, C.; Hopkins, A. | Navigating chemical space for biology and medicine. *Nature* 2004, 432, 855–861.

(6) Bohm, H. J. | The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* 1992, 6, 61–78.

(7) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Poche, S. | LEASD: a computer-aided ligand design for structure-based drug design. *J. Med. Chem.* 2005, 48, 2457–2468.

(8) Yuan, Y.; Pei, J.; Lai, L. | LigBuilder 2: a practical de novo drug design approach. *J. Chem. Inf. Model.* 2011, 51, 1083–1091.

(9) Naderi, M.; Alvin, C.; Ding, Y.; Mukhopadhyay, S.; Brylinski, M. | A graph-based approach to construct target-focused libraries for virtual screening. *J. Cheminf.* 2016, 8, 14.

(10) Nishihata, Y.; Itai, A. | Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* 1991, 47, 8985–8990.

(11) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. | SPROUT: a program for structure generation. *J. Comput.-Aided Mol. Des.* 1993, 7, 127–153.

(12) Brown, N.; McKay, B.; Gilar, G.; Gasteiger, J. | A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* 2004, 44, 1079–1087.

(13) Fechner, U.; Schneider, G. | Flux (2): comparison of molecular mutation and crossover operators for ligand-based de novo design. *J. Chem. Inf. Model.* 2007, 47, 656–667.