3ZC-08

多段階グラフ拡張による仮想化合物ライブラリ構築の研究

吉川 舜亮† 安尾 信明† 吉野 龍ノ介‡ 関嶋 政和†‡
†東京工業大学大学院情報理工学研究科計算工学専攻 ‡東京工業大学学術国際情報センター

1 はじめに

新薬の開発には，十数年もの期間が必要となっており，それに伴うコストも膨大となるため，これらの効率化が望まれている．しかし，創薬研究の対象となる化合物の理論的な総数は 10 の 60 乗にも昇るとされているのに対して，製薬会社が保有するデータ数が数百万化合物程度であり，探索範囲が充分であるとはいえないことが問題の一つであると考えられている．近年，情報技術を使用して仮想化合物ライブラリを構築し，その構築した仮想化合物ライブラリを用いることで，医薬品研究開発の初期段階における医薬品の候補となる化合物の絞込みの効率化が考えられており，既存の仮想化合物ライブラリ構築の研究では数億以上の化合物群とその合成経路を提供している [1]．しかし，データ数をいたずらに増加させてしまった場合，ユーザーが欲しいデータが含まれないライブラリを構築してしまう可能性がある点や，数億もの膨大なデータから欲しいデータを絞り込むことが困難である点が問題となっている．

そこで本研究では，化合物の持つ特性に偏りを持たせたライブラリを構築するシステムの開発を行なった．その際に，化合物を SMILES 記法 [2] と呼ばれる手法を用いてグラフとして扱い，グラフの拡張によって化合物の変換を表現する．このグラフの拡張を工夫することにより，化合物の特性を恣意的に変化させ，ユーザーが必要なデータのみを出力することができる．また，グラフの拡張を多段に行なうことで，特性に偏りを持ち，規模の大きいライブラリの構築を可能としている．

2 ライブラリ構築システムの構成

2.1 化合物の特性とグラフ拡張のルール

薬剤が体内で吸収されるためには，水に溶ける特性を保持する必要がある．そこで，化合物の特性の判断基準の例として，化合物に含まれる水素結合の供与体と受容体の個数が挙げられる．また，環構造の多い化合物は安定性が高く，薬剤として望ましいと考えられる．そこで，本システムではライブラリを構築する際に，

(1) 供与体の増加
(2) 受容体の増加
(3) 環構造の増加

の 3 種類のオプションから方向性を選択することで，よりユーザーにとって必要なデータが多く含まれたライブラリを提供することができる．

これらのオプションは複数を組み合わせて使用することもできる．本システムでは，選択されたオプションによって使用するグラフの拡張ルールを変更することで，よりユーザーが必要としているデータの出力を実現している．(1)，(2) が選択された場合，それぞれ供与体および受容体として扱われるグラフの部分構造の個数を増加させるようなグラフ拡張のルールが適用される．(3) が選択された場合，閉路を作り出すルールや閉路を付け加えるグラフ拡張のルールが適用される．本システムには合計で 302 種類のグラフの拡張ルールが導入されている．入力されたグラフが特定の構造を持っている場合，拡張後のグラフが出力されることで，化合物の反応を表現している．グラフの拡張にはケモインフォマティクス用ツールの RDKit[3] を利用した．

2.2 ライブラリ構築のフローチャート

本システムのフローチャートを図 1 に示す．ユーザーが入力した化合物データの集合をグラフの集合に変換し，ライブラリの構築を開始する．まず，集合の一番目のグラフに対して選択したすべてのグラフ拡張ルールが適用可能かを判断し，適用可能な場合は新しいグラフの集合に適用後のグラフを加える．集合内のすべてのグラフに対して同様の作業を行うことで，一段階のグラフの拡張が完了する．この拡張が完了したグラフの集合を再度システムの入力集合として適用することで，二段階目のグラフの拡張を行う．このように多段階に上記の作業を繰り返すことで，多数のデータを含んだライブラリの構築が可能となり，段階数の上限はユーザーが設定することで，取得するライブラリの規模を調節することができる．

Development of Virtual Compound Library based on Multiple Graphical Expanding
†Shunsuke YOSHIKAWA †Nobuaki YASUO ‡Ryunosuke YOSHINO †‡Masakazu SEKIJIMA
†Department of Computer Science, Graduate School of Science and Engineering, Tokyo Institute of Technology.
‡Global Scientific Information and Computing Center, Tokyo Institute of Technology.

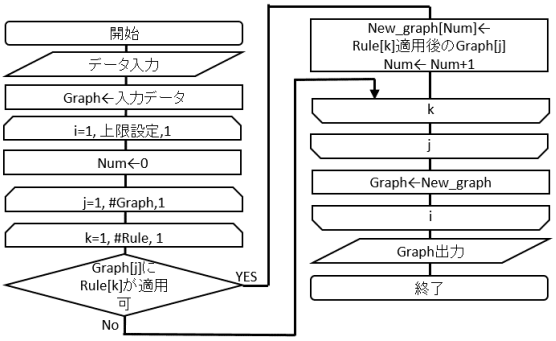

図 1: ライブラリ構築のフローチャート

2.3 開発環境

CPU　Intel core i7-3770 (3.4GHz)
メモリ　16GB
言語　Python 2.7.4
OS　Ubuntu 13.04

3 結果と考察

システムの評価を行なうために，ナミキ商事の building block 統合データベース (2013 年 6 月版) から 500 種類の化合物を抽出してシステムの入力化合物として用いた．グラフ拡張の段階数の上限は 3 とし，すべてのグラフ拡張ルールを適用してグラフの拡張を行ったところ，出力データ数は約 294 万となった．ここで構築されたライブラリを all_library とする．また，供与体を増やすグラフ拡張ルールを適用してグラフの拡張を行なったところ，出力データ数は 42758 となった．ここで構築されたライブラリを donor_library とする．

2 つの出力データの水素結合供与体の個数の平均と標準偏差を表 1 に示す．表 1 から，donor_library の方が供与体数の平均が 0.3 程度大きいことがわかる．また，2 つの出力データの供与体の個数のヒストグラムをそれぞれ図 2 に示す．図 2 から，all_library は供与体数が 2 のデータが最多となっているが，donor_library は供与体数が 3 のデータが最多となっており，供与体数の多いデータを得たいというユーザーの要求に答えたライブラリを構築できたことがわかる．供与体の個数はいずれも RDKit[3] によって計算されている．また，500 種類の入力データに対して all_library の構築を行なったところ，約 294 万種類のデータのライブラリの構築には約 38 時間を越える実行時間が必要となっていた一方で，donor_library の構築時間は 1 分程度で構築が終了している．このことから，反応の方向性を選択することで，all_library と比べて短時間で供与体数の多いデータを多数含んだライブラリを構築できたことがわかる．

表 1: 出力データの供与体数の平均と標準偏差

	平均	標準偏差
all_library	2.35	0.88
donor_library	2.68	0.70

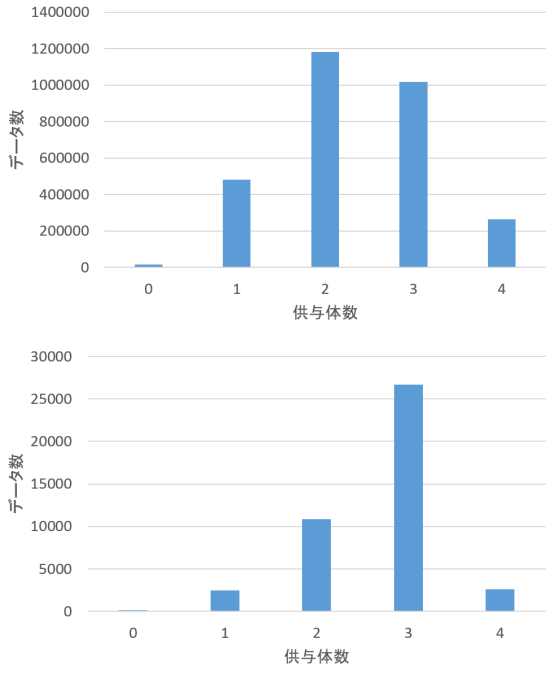

図 2: all_library(上) と donor_library(下) の供与体数の分布

4 まとめ

入力されたデータに対して方向性を選択してグラフを拡張させるルールを適用することで，特性に恣意的な偏りを持ったデータを，より短時間で提供することに成功した．オプションの種類や拡張ルールを増やすことで，よりユーザーの求める仮想ライブラリの構築ができると考えられる．また，規模の大きいライブラリを構築する際に，作成されたデータ数に応じて並列実行を可能とすることで，ライブラリ構築時間の短縮を目指すことができる．

参考文献

[1] 西村拓朗，船津公人，(2011)「大規模バーチャルライブラリ開発の試み」『日本化学会情報化学部会誌』, Vol.29, No.3, pp.49

[2] Weininger,D. (1988) SMILES, a Chemical Language and Information System. 1. Introduction and Encoding Rules. *J.Chem.Inf.Comput.Sci*, Vol.28, No.1, pp31

[3] http://www.rdkit.org/