統合計算化学システム

MOE：高速分子構造検索ツール

CHEMICAL COMPUTING GROUP

MOEは、独自のmdbファイルフォーマットに多数の分子構造を効率的に格納し、構造検索を行うことができます。しかしながら、対象とするデータセットの中の分子構造とクエリとなる構造を順次比較していくため、数十万件程度の化合物ライブラリの検索にさえ、高性能なPCを使用しても1分程度を要します。そこで、大規模なデータセットをも短時間で検索できるようなアプリケーションの開発をめざし検討を行った結果、約1千万件のデータセットを数秒程度で検索することが可能となったので紹介します。

■検討の背景

MOEは、SMARTSによる分子構造比較を数千件／秒で行うことができますが、例えば数百～数千万に及ぶ市販化合物のリストを一括で検索したいというような場合は、SMARTSのみでは長時間を要し、専用の化合物データベース管理システムのような快適な検索速度は望めません。

通常、データベース管理システムでは、データ登録時に何らかのインデックスを作成しておくことで検索の高速化を図ります。MOE上で構造検索を行う場合も、部分構造をインデックス化し、SMARTS検索の前に候補構造を絞り込むことで高速化は可能となるはずです。そこで、一般的なデスクトップPCでもインデックスをメモリ上に展開して高速検索が可能となるよう、弊社では、独自にコンパクトで情報密度の高いフィンガープリント SSFPを考案しました。この稿では、SSFPの考え方と、それを用いた高速構造検索ツールSSQSについて紹介します。

■市販化合物の構造解析

部分構造を記述する代表的な手法としては、MOEにも搭載されているMACCSキーがあります。カルボニル基は154ビット目、芳香族環は162ビット目というように、官能基を166ビットのいずれかに当てはめておき、それぞれの存在を0か1で表すことで、分子の特徴を記述します。これは優れた手法ですが、MOEを含め通常のデスクトップPCで利用するソフトウェアが扱うことのできる整数は32ビットまでとなっていますので、一つの整数として166ビットのキーを保存することはできません。そのため、キーを6分割するか、例えばベンゼンであれば、[162,163,165]と1を立てるビット位置を並べて表現することになり、情報量の割には容量が大きくなってしまいます。

そこで視点を変え、部分構造としてどれくらいの種類を予め登録しておけば効率的な検索を行うことができるのか調べるべく、ナミキ商事様のご協力を得て、オンラインカタログChemCupid 2014年10月版に搭載されている約950万件データセットを解析しました。

MOEには、最小の環構造を検出するSmallestRingsという関数が搭載されており、例えばナフタレンを2つのベンゼン環として認識することができます。この関数を利用して、約950万件の分子構造から、その中に含まれるユニークな最小環構造を全て抽出したところ、わずか2,428種に過ぎないことがわかりました。二進法では12ビットで4,096の整数を表すことができますから、これらの構造に一つずつユニークな二進数をふることで、部分構造検索のための効率的なキーとすることができます。これをベースに、非環状構造の中の特徴的な重原子や代表的な縮環構造も加えてコア構造として定義し、それぞれにユニークな二進数を与え、約2,500のコア構造キーを定義しました。

なお、今後新規な最小環構造が合成される可能性もありますが、ここで定義したもの以外の環構造には、まとめて一つのキーを割り当てることとします。

■部分構造フィンガープリントSSFP

コア構造だけであれば12ビットでインデックスを定義できることがわかりましたが、一つの整数を表す32ビットをフルに使えば、より詳しい構造情報を記述することができるはずです。そこで、それぞれのコア構造に結合する原子の情報を残る19ビットで定義するルールを定め、独自の部分構造フィンガープリントSSFPを定義しました。

■部分構造検索ツールSSQS

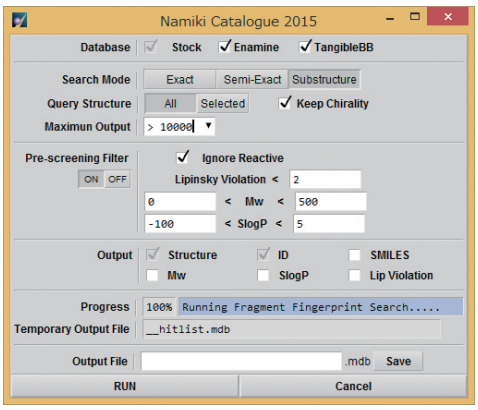

図1　SSQS構造検索インターフェース

SSFPに基づく検索を実行するためのGUIを(図1)に示します。このツールを使って、約5,500万件の分子構造を含むデータベースの部分構造検索を実行したところ、検索時間はクエリ構造の複雑さに依存するものの、ほとんどの場合、数秒～数十秒程度で完了しました。

弊社では、このツールのMOE/web化も検討中です。