様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　９　日現在

機関番号：　３４５０４
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２６５

研究課題名（和文）有機化合物の新規骨格創製アルゴリズムの開発

研究課題名（英文）Development of Algorithm for Enumerating Organic Molecules

研究代表者
猪口　明博（Inokuchi, Akihiro）

関西学院大学・理工学部・准教授

研究者番号：７０４５２４５６

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：グラフは構造をもつデータを表現するために最も自然はデータ構造の1つである．近年，構造をもつデータからの知識発見手法が様々な実データに適用できるようになってきたことで，グラフマイニング技術に関する研究は大変注目を浴びている．本研究では，化合物をラベル付きグラフで表現し，以下の3つの手法を提案した．(1)グラフを効率良く列挙するアルゴリズム．提案手法は，13マイクロ秒で各グラフを列挙できる．(2) グラフの集合Gとクエリqが与えられたとき，qに含まれるGの要素を効率良く検索するアルゴリズム．(3) グラフとその特性との関係を学習し，特性が未知のグラフの特性を予測するアルゴリズム．

研究成果の概要（英文）：A graph is one of the most natural data structures for representing structured data. For instance, a chemical compound is represented as a graph, where each vertex corresponds to an atom, each edge corresponds to a bond between two atoms therein, and the label of the vertex corresponds to an atom type. Recently, the topic of graph mining has become of great interest because knowledge discovery from structured data can be applied to various real-world datasets.
In this work, we proposed the following three algorithms by representing chemical molecules as labeled graphs: (1) an algorithm for enumerating a complete set of labeled graphs efficiently each of which output in 13 microseconds in average, (2) an algorithm for searching labeled graphs of which q is a supergraph, when given a set of labeled graphs G and a query graph q, and (3) an algorithm for learning the relationship between labeled graphs and their properties and for forecasting a property of a new labeled graph.

研究分野：　データマイニング

キーワード：　データマイニング

１．研究開始当初の背景

新規医薬品の開発には数十億から数百億の莫大な研究開発費が必要であり，プロジェクトを立ち上げてから市場に出るまでの創薬期間は最低でも 10 年はかかると言われている．創薬における最近の問題は，市場に出る新規医薬品の数が鈍化していることである．効率化された実験装置の導入により，研究開発段階で合成される新規化合物が飛躍的に増加する一方で，その実用化は進んでいない．実社会で実際に合成された化合物は 10^8 程度であるが，理論的に存在し得る有機化合物は 10^{60} 程度あると言われており，我々人類は存在し得る化合物のうち，まだごく一部しか手に入れていない．

２．研究の目的

10^{60} 個の化合物を実際に合成するのは容易ではないが，計算機上で存在可能性の高い化合物をグラフ理論に基づいて列挙し，その薬理効果を予測することは可能である．また，薬理効果の高いものから優先的に合成していけば，新規化合物の開発に有用であると考えられる．本研究の目的は，グラフ理論と列挙アルゴリズムに基づいて，存在可能な化合物を高速に列挙するアルゴリズムを考案し，世界最大の化合物ライブラリを作成することである．さらに産業界と協力し作成した化合物ライブラリを網羅性の観点から評価し，研究成果を産業界に還元することである．

３．研究の方法

本研究では，以下に挙げる 3 つの技術に関して研究を行った．

(1) 化合物列挙

化合物は，ラベル付きグラフで表すことができる．ここで，化合物の原子，原子間の化学結合，原子の種類，結合の種類は，それぞれグラフの頂点，辺，頂点ラベル，辺ラベルに対応する．我々は，グラフで表現された化合物の集合が与えられたときに，それらに頻繁に出現する部分グラフを列挙するアルゴリズムを構築してきた．本研究では，そこで培った頻出部分グラフ列挙アルゴリズムを活用し，化合物を列挙するアルゴリズムを検討する．

(2) グラフ検索

(1)で列挙されたグラフに対応する化合物のすべてが自然界で存在するわけではない．幾つかの特定の部分構造を含んでいると存在しないことが分かっている．その問題は，列挙されたグラフが部分構造を部分グラフとして含んでいるかどうかを判定する部分グラフ同型問題で解くことができる．しかし，部分グラフ同型問題は NP 完全であるので，列挙された大量のグラフと多数の部分グラフの各ペアに対して部分グラフ同型問題を解くことは計算時間の観点から現実的ではない．そこで，2 つのグラフの集合に対して，この問題を効率的に解く方法を検討し，グラフデータ ベースを構築した．

(3) グラフ分類

グラフ分類問題とは，グラフとそれが属するクラスの対からなる集合のデータが与えられたときに，グラフとクラスの関係性を計算機に学習させる，また，クラスが未知のグラフが与えられたときに，このグラフが属するクラスを学習結果に基づいて予測するという問題である．化合物はグラフで表現でき，それが薬理活性や副作用を有するかをクラスとするときに，新規に合成された化合物の薬理活性や副作用を計算機で予測することができるので，この問題は創薬において非常に重要な問題である．部分構造が類似する化合物は同じ薬理活性や副作用をもつことが多いことからも分かるように，グラフとクラスの関係性を学習するには，2 つのグラフの類似度を正確に測る指標やそのアルゴリズムが重要となるため，グラフの類似度を測るための指標とアルゴリズムを検討した．

４．研究成果

(1) 化合物列挙

既存研究として，スイスの研究グループが 360 個の CPU で並列計算し，10 万 CUP 時間で，1660 億化合物の列挙に関する報告を行っている．この方法では，①まず， 約 1140 億個のラベルなしグラフを列挙する．②次に，そのうち，化合物の表現として適切でないグラフを削除し，約 500 万個のグラフを残す．③その後，頂点に炭素，窒素，酸素などのラベルを付けたり，辺に結合の種類のラベルを付けたり，幾つかのフィルタを通したりして，最終的に 1660 億個の化合物を得る．この手法では，①から②への段階で，列挙されたグラフのうち 99.995%のグラフが削除されるので計算時間の無駄である．

そこで，我々は頻出部分グラフ列挙問題で培った技術を用いることで，1 グラフ当たり 13 マイクロ秒程度でラベル付きグラフを列挙できるソフトウェアを作成した．グラフは AcGM コードで表現されるが，グラフが列挙されるたびに，それが正準コードであるかをチェックするため，同じグラフが 2 度列挙されることはない．このため，計算時間の無駄が少ない．また，正準コードの接頭部分は正準であるように，AcGM コードが定義されているので，すべてのラベル付きグラフを漏れなく列挙することができる．

(2) グラフ検索

ここでは，グラフの集合 G とグラフ $q \in Q$ が与えられたとき，q に含まれる G の要素を出力する問題を対象とする．提案手法では，グラフを AcGM コードや DFS コードで表し，そのコードによるグラフ検索のための索引を構築する．検索の際には，その索引を深さ優先に探索し，部分グラフ同型問題を解く．こ

の時に，索引の浅い部分は複数のコードが共通の枝で表現されているので，索引を用いることで複数のグラフ同型問題を平行して解くことができる．従来手法の 1 つである LW-Index は，以下のような課題を抱えていたが，提案手法は LW-index よりも 1 桁以上高速であることを実証した．

- G の頻出部分グラフパターンを F とする．F から LW-index で索引に用いられる頻出部分グラフパターン P を選択するのに，LW-index では O(k|F||Q||G|) を要する．ここで，k は F から選択されるパターンの数である．O(k|F||G|) でその近似解を求める改善手法も提案されているが，計算時間は F や Q に依存する．また，頻出部分グラフを列挙するアルゴリズムは非常に多くの計算時間を要する．
- あるクエリ q が与えられたとき，頻出部分グラフパターン p は q に含まれないグラフを filtering するのに適していると考えられているが，p が G で頻出ならば，p は Q でも頻出である可能性が高いので，必ずしも頻出部分グラフパターンが filtering に適しているとは限らない．むしろ，G にあまり出現しない部分グラフのほうがを索引に適している可能性がある．
- 既存手法の多くは，事前にクエリの集合 Q の分布が既知であることを前提として，G と Q から索引に適した頻出部分グラフパターンを選択する手法をとっている．このため，Q が事前に既知ではない場合に適用できない，また Q の分布が変わった際に，索引を再構築する必要がある．

提案手法の特徴は以下の通りである．

- Q や F を必要としない．また，索引を構築するのに必要な計算時間は O(|G|) である．
- 提案手法の索引は木構造であるが，その浅い位置に G において非頻出な部分グラフを配置する．
- 索引構築時に Q を必要としないため，Q の分布が変わっても，索引を再構築する必要がない．また，G にグラフが追加されたり，G からグラフが削除されたり，G の要素が更新されても，簡単に索引を変更することが可能である．
- 検索が高速であり，現在最速の LW-index よりも 1 桁以上高速である．

(3) グラフ分類

提案手法の 1 つ目は，グラフの各頂点をアダマール符号でラベル付けし，2 つのグラフの各頂点のラベルを比較することで 2 つのグラフの類似度を測る．ただし，頂点同士の比較では，部分構造の類似度を測れないので，ある頂点のラベルとその頂点に隣接する頂点のラベルの和を繰り返しとることにより，その頂点の周辺の部分構造の表現を可能にした．アダマール符号を用いることで，ラベルの和が平均 0 の 2 項分布に従うので，繰り返し和をとったとしても省メモリで動作することを可能にした．

提案手法の 2 つ目は，1 つ目の改良手法である．まず頂点同士の比較において，一致/不一致で判定していたものを類似の程度に応じて数値化した．また，すべての頂点の組み合わせをとるのではなく，最小 2 部グラフマッチング問題を取り入れることで，最適な頂点の対応関係をとるようにした．これにより 2 つのグラフの類似度をより正確に測ることを可能にした．

ベンチマークデータを用いて提案手法の性能を評価したところ，この分野の代表的な従来手法に比べ分類精度が 10%も大幅に向上したケースもあった．また，提案されたアルゴリズムは，グラフの頂点数に対して，多項式時間で動作するので，大規模なデータにも適用できる．

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

1. 長村 佳歩，奥井 颯平，猪口 明博．滑らかな変化を検出するためのグラフ系列クラスタリング．情報処理学会論文誌，58(1), pp. 278-287, 2017. （査読あり）
   http://id.nii.ac.jp/1001/00176897

2. 片岡 哲也，猪口 明博．アダマール符号を用いたグラフカーネルによるグラフクラス分類．情報処理学会論文誌 57(9), pp. 2122-2130, 2016. （査読あり）
   http://id.nii.ac.jp/1001/00174645

〔学会発表〕（計 7 件）

1. Tetsuya Kataoka, Eimi Shiotsuki, and Akihiro Inokuchi. Mapping Distance Graph Kernels using Bipartite Matching. Proc. of the 6th International Conference on Pattern Recognition Applications and Methods, pp. 61-70, 2017.2.24-26, Porto, (Portugal).

2. Sohei Okui, Kaho Osamura, and Akihiro Inokuchi. Detecting Smooth Cluster Changes in Evolving Graphs. Proc. of International Conference on Machine Learning and Applications (ICMLA), pp.369-374, 2016.12.18-20. Anaheim, (USA).

3. Tetsuya Kataoka and Akihiro Inokuchi. Hadamard Code Graph Kernels for Classifying Graphs. Proc. of the fifth International Conference on Pattern Recognition Applications and Methods, pp.24-32, 2016.2.24-26, Rome, (Italy).

4. **猪口明博, 磯村哲. グラフコーディングを用いたスーパーグラフ検索の効率化, 第 105 回人工知能学会知識ベースシステム研究会**, pp.26-33, **2015 年 8 月 7 日, 関西学院大学 梅田キャンパス（大阪府大阪市）**

〔図書〕（計 1 件）

1. Tetsuya Kataoka and Akihiro Inokuchi. Experimental Evaluation of Graph Classification with Hadamard Code Graph Kernels. ICPRAM (Revised Selected Papers), 260 pages, (**分筆** pp. 1-19), 2016.

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
該当なし

６．研究組織
(1)研究代表者
猪口 明博 （INOKUCHI, Akihiro）
関西学院大学・理工学部・准教授
研究者番号：70452456

(2)研究分担者
岡田 孝（OKADA, Takashi）
関西学院大学・理工学部・理工学部研究員
研究者番号： 00103135