様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５月２０日現在

研究種目：基盤研究（A）
研究期間：2007～2009
課題番号：19200022
研究課題名（和文）グラフ理論とカーネル法の融合による化学構造設計法

研究課題名（英文）New Methods for Designing Chemical Structures Using Graph Theory and Kernel Methods

研究代表者
阿久津　達也（AKUTSU TATSUYA）
京都大学・化学研究所・教授
研究者番号：90261859

研究成果の概要（和文）：本研究ではグラフ理論と機械学習手法の一つであるカーネル法を組み合わせることにより新規化学構造を設計するための方法について研究を行い、ラベルつきパスの頻度からなる特徴ベクトルから木状化合物を列挙するアルゴリズムおよびそれを実装した Web サーバー、木状および外平面グラフ構造を持つ化学構造の立体異性体列挙アルゴリズム、立体異性体を区別可能な化学構造に対するカーネル関数の開発などの成果を得た。

研究成果の概要（英文）：In this project, we studied a new approach for design of novel chemical compounds by combining graph theory and kernel methods. We developed branch-and-bound algorithms for enumerating tree-like chemical structures from a given feature vector consisting of frequency of labeled paths, dynamic programming-based algorithms for enumerating stereoisomers from a given tree-like or outer-planar chemical structure, and kernel functions that can discriminate stereoisomers. Some of these algorithms were implemented in a newly developed web server.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	7,400,000	2,220,000	9,620,000
2008 年度	10,000,000	3,000,000	13,000,000
2009 年度	9,300,000	2,790,000	12,090,000
年度			
年度			
総　計	26,700,000	8,010,000	34,710,000

研究分野：数理生物情報学
科研費の分科・細目：情報学・生体生命情報学
キーワード：カーネル法、グラフアルゴリズム、特徴ベクトル、サポートベクターマシン、列挙アルゴリズム、動的計画法、ケモインフォマティクス

１．研究開始当初の背景

新薬の開発が国際的にも行き詰まりつつある中、新たな化合物の設計法の開発が求められている。一方、機械学習の研究から 1990 年代に発明されたサポートベクターマシンという学習手法はパターン認識やバイオインフォマティクスなど様々な分野に応用され、近年は化合物の活性の予測にも応用されている。研究代表者は以前に行っていた基盤研究(B)などにおいてサポートベクターマシ

ンの化合物への応用について研究を行っていた。その過程で「特徴ベクトルからの逆写像を計算すれば、化合物の設計ができるのではないか」との着想に至り、博士課程学生と一緒に研究を開始した。その結果、逆写像の計算可能性に関する基本的な理論的結果を得ることができ、また、小さな化学構造に対して厳密解を実際に計算できるアルゴリズムを開発した。この学生の博士論文の審査委員の一人であった研究分担者の永持もこの問題に興味を持ち、特殊な場合について代表者らの理論的結果を大幅に改善する結果を得た。その後、阿久津と永持は、実用的な手法の開発をめざし、共同研究を開始した。一方、川端は薬剤設計について長年にわたり実験的研究をしてきた。阿久津は実用的手法開発のためには実際に薬剤設計の実験的研究に携わっている研究者の協力が不可欠と考え、同じ研究所に所属している川端に共同研究を打診し、川端も快諾したため、３名による共同研究を開始することとなった。なお、研究体制をより強化するために阿久津研究室の助教の林田も共同研究に加わることになった。

２．研究の目的

新規化合物の設計は薬学、医学、農学などにおいて重要であり、ポストゲノム時代における生体生命情報学の主要目標の一つである。そのために様々な情報解析手法が研究されてきたが、その一つに化学構造とその活性の関係を解析・推定する「構造活性相関」がある。この構造活性相関について、近年、様々な機械学習が応用されるようになってきた。特に、サポートベクターマシンに代表されるカーネル法の有効性を示唆する論文がいくつも出版されてきた。サポートベクターマシンを用いる方法では、通常、対象となるデータを、有限次元のユークリッド空間、もしくは、無限次元のヒルベルト空間の点（特徴ベクトルとよばれる）に写像し（ただし、実際の計算にあたっては陽には写像しない場合もある）、写像された空間における超平面を用いることにより、分類や予測を行う。活性予測のためには、化学構造を特徴ベクトルに変換する必要があるが、通常、化学構造はグラフ構造として入力され、そのグラフが特徴ベクトルに変換される。変換法により予測精度が異なることになるため、グラフから特徴ベクトルへの様々な変換法が研究されている。

本研究は、このような従来手法についての改良を目指すのではなく、従来手法の逆を行うことにより新規化学構造を計算機により導き出す計算手法について研究する。具体的には、「特徴ベクトルから、もとの化学構造を推定する」ことにより新規な化学構造を導き出す方法について、理論基盤構築および実用的アルゴリズムの両面から研究する。この方法が開発できると様々な応用の可能性がある。例えば、化合物 A と化合物 B の中間の性質を持つ化合物を設計したいとする。この場合、A の特徴ベクトルと B の特徴ベクトルの中点を計算し、それを逆写像することにより中間の性質を持つと期待される化合物の構造を得ることができる。もちろん、これらのシナリオは楽観的過ぎるかもしれないが、様々な改良を行うことにより、役立つようになる可能性がある。申請者らは、既存研究の延長線上で細かな改良を行うのではなく、困難かもしれないが新たな考え方に基づき、理論的基盤を構築しつつ、実用的手法を開発していきたいと考えている。

３．研究の方法

本研究では特徴ベクトルからの化学構造の推定について理論および実用的アルゴリズム開発の両面から研究を行う。特に、異なる専門の研究者が共同して研究を行うことにより、研究が健全に進展し、かつ、新規性のある結果を得ることが期待できる。具体的には、バイオインフォマティクスの専門家（阿久津、林田）、グラフ理論・アルゴリズムの専門家（永持）、実験的研究を行っている薬剤設計の専門家（川端）が協力して研究を行う。なお、実用的アルゴリズムの開発、WEB サーバーの構築、計算機実験には多大な労力が必要となるため、ポスドク研究員を１名雇用する。さらに、アルゴリズム開発・実装、および、計算機実験のために大学院生の協力も得る。

具体的には以下の研究を行う。

(1) グラフの 推定問題の計算論的側面について現在の研究をより深化させ、どのような種類のグラフ、どのような種類の特徴ベクトルであれば、厳密解、もしくは、近似解が理論的に効率よく（多項式時間で）計算や数え上げができるのかを明らかにする。
(2) 特徴ベクトルから化学構造に対応するグラフ構造を実用的な時間で計算するためのアルゴリズムを開発する。なお、既に一般的な場合における計算困難性（NP 困難性）がわかっているため、できるだけ大きなサイズの化合物に対して実用的な時間で解を計算できるようにする。
(3) 開発したアルゴリズムを実際に利用可能とする WEB サーバーを作成・公開し、商業目的以外にはフリーで利用できるようにする。
(4) 既存のカーネル関数以外に、化学構造の特徴をより適確にとらえる新たなカーネル関数を開発する。
(5) 化学構造以外にも関連する物質、具体的

には RNA やタンパク質に対する新たな解析手法を開発する。

４．研究成果

上記で述べた計画や方法に従い研究を行い以下の成果を得た。

(1) 木状化合物の数え上げアルゴリズム

ラベルつきパスの個数からなる特徴ベクトルが与えられた時に、その制約を満たす木構造を持つ化学構造を列挙するアルゴリズムを開発した。具体的には従来から提案されていた無順序木の数え上げ手法に、特徴ベクトルに基づくカットと、次数制約に基づくカットを組み込んだ分枝限定法アルゴリズムを開発した。そして既存の Alkane 族の数え上げアルゴリズムとの比較を行い、はるかに少ないメモリー量で、同程度以上に高速に動作することを理論および計算機実験の両面から示した。さらに永持が以前に開発した detachment 手法を組み込み高速化を図った。

(2) 立体異性体の数え上げアルゴリズム

木構造を持つ化合物に対し、光学異性体を高速に列挙するアルゴリズムを開発した。このアルゴリズムは動的計画法に基づいており、いったん個数を数え上げた後に異性体を順次出力する。そして、個数の計算自体は原子数に対する線形時間で計算でき、数え上げについては異性体１個あたり原子数の線形時間で実行できることを示した。また、このアルゴリズムの考え方を拡張し外平面的グラフを持つ化学構造に対する立体異性体列挙アルゴリズムのプロトタイプを開発した。

(3) 外平面的グラフの数え上げアルゴリズム

既知の化合物の多くは外平面的グラフの構造をもつが、このクラスの根付きグラフを１個当たり定数時間で列挙するアルゴリズムが存在することを証明した。外平面的グラフを重複なく列挙するには、平面構造の軸対称性と子の並びに対する順列の二種類の対称性を同時に扱う困難が生じるが、軸対称性を子の順列として疑似的に扱う工夫により，複雑ではあるが根付き外平面的グラフを１個当たり定数時間遅れで生成するアルゴリズムを設計することができた。

(4) 立体異性体を区別可能なカーネル関数

立体異性体を区別できるカーネル関数を開発した。この手法は、既存の木パターングラフカーネルとよばれる、部分木の出現頻度に基づくカーネル関数を利用している。そして、部分木の出現頻度を計算する際に立体異性に関する情報をチェックすることにより立体異性体を区別する。光学異性体を含む構造活性相関データを用いて計算機実験を行い、その有効性を確認した。

(5) RNA およびタンパク質解析手法

新規化合物設計のためには、化合物データのみならずタンパク質データや RNA データの解析も有用である。そこで上記研究と関連して、整数計画法を用いた RNA 二次構造予測法、動的計画法を用いたタンパク質のβシート領域の予測法、動的計画法を用いた相互作用する RNA 二次構造の予測法、サポートベクターマシンを用いたタンパク質残基露出度の推定法などを開発した。

(6) WEB サーバーEnuMol の開発

木状化学構造列挙 WEB サーバーの第一版を完成させ EnuMol を名づけ公開した。このサーバーは与えられた特徴ベクトルを満たす木状の化学構造を列挙するものであり、結果はグラフィクス形式で見ることができ、また、ファイル形式でダウンロードすることもできる。

これらの研究は国際的な評価が高まりつつあり、ケモインフォマティクス・アルゴリズムに関するハンドブックの分担執筆やケモインフォマティクス・アルゴリズムに関する国際ワークショップへの招待講演などにつながっている。特に化合物の数え上げの実用的アルゴリズムについては、ケモインフォマティクスの中心的な研究者の一人から、当該グループは国際的に活躍する数グループの一つであるとの高い評価を受けている。

なお、本研究をさらに発展させ、より実用的な数え上げシステム、構造設計支援システムを開発するために、新たな基盤研究 A「離散的手法とカーネル法の融合による構造設計法」を前年度申請したところ採択されたため、最終年度を待たずに本研究は終了し、その基盤研究 A に引き継がれた。

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 14 件、すべて査読有）

① Y. Ishida, Y. Kato, L. Zhao, H. Nagamochi, T. Akutsu: Branch-and-bound algorithms for enumerating treelike chemical graphs with given path frequency using detachment-cut, Journal of Chemical Information and Modeling, 50, 934-946 (2010).

② T. Imada, S. Ota, H. Nagamochi and T. Akutsu: Enumerating stereoisomers of tree structured molecules using dynamic programming, Lecture Notes in Computer Science, 5878, 14-23 (2009).

③ Y. Kato, T. Akutsu and H. Seki: Dynamic programming algorithms and grammatical modeling for protein beta-sheet prediction, Journal of Computational Biology, 16, 945-957

(2009).

④ Y. Kato, T. Akutsu and H. Seki: A grammatical approach to RNA-RNA interaction prediction, Pattern Recognition, 42, 531-538 (2009).

⑤ T. Kawabata, C. Jiang, K. Hayashi, K.Tsubaki, T. Yoshimura, S. Majumdar, T. Sasamori and N. Tokitoh: Axially chiral binaphthyl surrogates with an inner N-H-N hydrogen bond, Journal of the American Chemical Society, 131, 54-55 (2009).

⑥ H. Fujiwara, J. Wang, L. Zhao, H. Nagamochi and T. Akutsu: Enumerating tree-like chemical graphs with given path frequency, Journal of Chemical Information and Modeling, 48, 1345-1357 (2008).

⑦ M. Hayashida, T. Akutsu and H. Nagamochi: A clustering method for analysis of sequence similarity networks of proteins using maximal components of graphs, IPSJ Transactions on Bioinformatics, 49-Sig 5, 15-24, (2008).

⑧ T. Kawabata, W. Muramatsu, T. Nishio, T. Shibata and H. Schedel: A Catalytic One-Step Process for the Chemo- and Regioselective Acylation of Carbohydrates. Journal of the American Chemical Society, 129, 12890-12895 (2007).

〔学会発表〕(計 15 件)

① T. Imada, S. Ota, H. Nagamochi and T. Akutsu: Enumerating stereoisomers of tree structured molecules using dynamic programming, 20th International Symposium on Algorithms and Computation, Hawaii, USA, 2009/12/16.

② 阿久津達也: 木構造および化学構造に対する特徴ベクトル：埋め込み、検索、構造推定, 第 12 回情報論的学習理論ワークショップ, 九州大学（福岡）, 2009/10/20.

③ J. Wang and H. Nagamochi: Enumerating colored and rooted outerplanar graphs, 情報処理学会, 第 129 回アルゴリズム研究会, 東芝科学館（東京）, 2009/3/5.

④ Y. Ishida, L. Zhao, H. Nagamochi and T. Akutsu: Improved algorithms for enumerating tree-like chemical graphs with given path frequency, The 19th Int. Conference on Genome Informatics, Gold Coast, Australia, 2008/12/1.

⑤ 阿久津達也: 列挙型アルゴリズムのバイオインフォマティクスへの応用, 第 7 回情報科学技術フォーラム, 慶応義塾大学（神奈川）, 2008/9/2.

⑥ 浦田隆史, J.B. Brown, 田村武幸, 川端猛夫, 阿久津達也: 不斉炭素原子を考慮した化合物に対するグラフカーネル法, 情報処理学会バイオ情報学研究会, 九州大学（福岡）, 2008/3/4.

⑦ T. Urata, J.B. Brown, T. Tamura, T. Kawabata and T. Akutsu: A graph kernel method incorporating chirality, 日本バイオインフォマティクス学会年会, 日本未来館（東京）, 2007/12/19.

〔図書〕(計 0 件)
〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕
木状化合物列挙 WEB サーバー：
http://sunflower.kuicr.kyoto-u.ac.jp/tools/enumol/index-j.html

６．研究組織
(1)研究代表者
阿久津　達也（AKUTSU　TATSUYA）
京都大学・化学研究所・教授
研究者番号：90261859

(2)研究分担者
川端　猛夫（KAWABATA　TAKEO）
京都大学・化学研究所・教授
研究者番号：50214680

林田　守広（HAYASHIDA　MORIHIRO）
京都大学・化学研究所・助教
研究者番号：40402929

永持　仁（NAGAMOCHI　HIROSHI）
京都大学・情報学研究科・教授
研究者番号：70202231