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

令和　元 年　 ５ 月 ２６ 日現在

機関番号：１４３０１
研究種目：基盤研究(A)（一般）
研究期間：2014～2018
課題番号：２６２４００３４

研究課題名（和文）離散的手法と統計的手法の融合による構造設計法

研究課題名（英文）An Approach to Novel Structure Design by Combining Discrete Methods and Statistical 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

交付決定額（研究期間全体）：（直接経費）　　28,300,000円

研究成果の概要（和文）：化学構造の列挙に関して多くの進展を得た。具体的には、ベンゼン環およびナフタレン環を含む木状化合物列挙、木に２個および３個の辺を加えた部分構造クラスの化合物列挙、任意の環構造を頂点として与えた場合の木状化合物列挙などのアルゴリズムを開発し、また、以前に開発した分枝限定法に基づくアルゴリズムにResource Cutという新たな限定操作を導入し計算速度を大きく改善するという成果も得た。化学構造以外にも、タンパク質相互作用予測、タンパク質複合体予測、タンパク質切断部位予測、RNA配列解析などについて新規計算手法を開発し、その一部は生物学実験により有効性を確認した。

研究成果の学術的意義や社会的意義
本研究により、従来手法と比較し、より効率的に列挙可能な化学構造のクラスを大きく拡大することができた。化学構造の列挙は新規薬剤の設計などに応用できる可能性があるため、重要な結果であると考えられる。ただし、実際に応用するには様々な制約を取り入れることが必要であり、さらなる改良、拡張が必要である。タンパク質相互作用予測、複合体予測、切断部位予測、RNA配列解析についても新規手法を開発することができた。これらはタンパク質やRNAの機能をその配列データから推定するために有用であり、その結果、生体生体高分子の機能や役割の解明に貢献する可能性があり、さらには、医療などに役立つ可能性がある。

研究成果の概要（英文）：We developed enumeration algorithms for various subclasses of chemical graphs, which include tree structured graphs with benzene and naphthalene rings as nodes, subclasses of 2-edge and 3-edge augmented tree structured graphs, and tree structured graphs with arbitrary ring structures as nodes. Furthermore, we improved our previously developed branch-and-bound algorithm for enumeration of chemical trees by introducing a new cut operation named Resource Cut. In addition, we developed computational methods for protein complex prediction, protein-protein interaction prediction, protein cleavage site prediction and RNA sequence analysis, the effectiveness of some of which was verified via biological experiments.

研究分野：数理生物情報学

キーワード：ケモインフォマティクス　構造列挙　グラフアルゴリズム　カーネル法　生物情報ネットワーク　化学構造

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

1990 年代から 2000 年代にかけて大きく発展をとげたカーネル法は、現在でも統計学や機械学習において精力的に研究されていると同時に、画像理解、自然言語処理、バイオインフォマティクスなど様々な分野に幅広く応用されている。カーネル法の主要手法の一つであるサポートベクターマシン（SVM）を用いる予測手法では、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上の点（特徴ベクトル）に写像し、写像された空間における超平面との位置関係により分類や予測を行う。

本研究では、以前に行っていた基盤研究(A)に引き続き、従来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。つまり、望ましい性質を持つと考えられる特徴ベクトルを統計的手法により選択した後で、「特徴ベクトルから、もとの構造を推定する」ことにより新規構造を導出する方法について、理論および実用アルゴリズムの両面から研究する。実際、代表者らは本研究の端緒となる論文を 2005 年に発表したが、それらを参考にして 2009 年に Waterloo 大学の研究者らにより、カーネル法と構造推定の組合せによる新規構造推定法が開発され、コンピュータシミュレーションによりその有用性が示された。しかしながら、彼らの手法では化合物の構造が線型に近いものに限定されており、実用性・発展性がないという問題がある。我々のこれまでの研究により、真に有用な設計法の開発のためには、できるだけ多くの化学構造に対応できることが必須であることを認識するに至った。さらに、構造推定においては与えられた制約のもとでの構造列挙が主要な役割を果たすため、構造列挙の高速化が不可欠であることも認識された。

２．研究の目的

本研究の第一の目的は化学構造列挙のための世界最高性能のアルゴリズムを開発することである。これまでの２度の基盤研究(A)により木状構造を持つ化学構造についてはその目的を達成し、ケモインフォマティクスの中心的学術誌などに論文発表するとともに、web ツール Enumol を構築し公開してきた。さらに、構造異性体列挙という課題についても新規な技法を開発し本質的な進歩をもたらした。しかしながら、木状化合物の割合は全体の 10%程度であるため、実用性を高めるためには、より複雑な構造に対応できるようにする必要がある。

化学構造列挙は 1870 年代における数学者 Cayley の研究より始まり、数学者 Polya による数え上げ理論などを通じて発展してきた長い歴史を持つ研究課題である。さらに、薬剤設計以外にも質量分析計からの構造推定などにおいても主要な役割を果たす重要研究課題であり、現時点ではドイツの研究グループにより 20 年以上にわたり開発されてきた Molgen システムが最も有名、かつ、最高性能を持つと考えられている（http://www.molgen.de/）。木状化合物の列挙に関しては阿久津と永持らが開発してきたアルゴリズムが既に Molgen を上回っているが、Molgen は一般的な化学構造を扱えるという利点がある。すべてについて Molgen を大きく凌駕するシステムを今後 5 年以内に開発するのは困難であるが、現実の化合物の数十%程度をカバーし Molgen より高速に列挙するアルゴリズムを開発できる可能性はあると考えている。現在、Molgen の開発は滞っていると考えられるため、本研究により化学構造列挙において国際的に第一線の成果を得ることが期待できる。なお、構造推定と統計手法の組み合わせは、機械学習およびケモインフォマティクスの両分野において研究されてきたが、ヒューリスティクスな研究が多く一般性が不十分である。化学構造列挙はデータマイニングにおいても研究されているが、それらは既存データを用いて大幅な枝刈りを行うものであり、新規構造設計に直接は適用できない。よって、本研究により化学構造列挙技術が大きく進展することが期待できる。

３．研究の方法

本研究で発展させる「特徴ベクトルからの構造推定」による新規化学構造設計の基盤技術を確立するために、化学構造の列挙についての理論的研究およびアルゴリズム開発、回帰に基づく特徴ベクトル選択法の開発を行い、これらを統合することにより新規化学構造設計の枠組みを構築し、計算機シミュレーションにより有用性を評価する。さらに、方法論の一般性を確立するために配列設計､代謝ネットワーク設計にも取り組み､列挙に基づく設計手法を開発する。この研究を実施するにあたり、バイオインフォマティクス、グラフ理論・アルゴリズム、神経生物学・分子生物学それぞれの専門家による共同研究を行う。

４．研究成果

本研究により得られた主要な成果は以下のとおりである。

（１）　化学構造列挙

以前の基盤研究(A)において開発した､木構造に１個の辺を加えた場合に対する列挙アルゴリズムを完成させた。そして、木構造に 2 個の辺を追加した場合に対する列挙アルゴリズムをほぼ完成させた。さらに、木構造に３個の辺を追加し、閉路が２頂点でのみ合流し、かつ、２連結成分が１個であるようなグラフ構造のクラスを定義し、そのグラフクラス内において与えれた制約を満たす化学構造を効率的に列挙する分枝限定法アルゴリズムを開発した。

これまで分枝限定法に基づく木状化合物の列挙手法を開発してきたが、Resource Cut という新たな限定手法を開発した。これは途中まで生成した時点で、残りの原子組成を用いては結合数の制約などを満たすことができないと判断した場合にそれ以上の探索を打ち切り、前の時点

に戻るというものである。その結果、生成可能な木状化合物のサイズを４０原子から５０原子に拡大することができた。
　一方、ベンゼン環やナフタレン環などの環構造を頂点として許すように拡張した木構造に対する２種類の列挙アルゴリズムを開発した。一つは木構造の列挙と動的計画法を組み合わせたアルゴリズムであり、以前に立体異性体の列挙に用いた動的計画法による手法に基づきつつも対称性などを効率的に扱うための新規のアイデアを導入することにより高速に列挙する方法を開発した。さらに、この手法は列挙以外にも順番を指定してその構造を高速に抽出することが可能であり、構造のサンプリングなどに応用できる可能性がある。もう一つは、効率は多少落ちるものの幅優先探索に基づくより単純な手法であり、ユーザが指定した任意の環構造を扱えるという特徴がある。

（２） タンパク質情報解析
　病原体タンパク質と宿主タンパク質間の相互作用予測問題に取り組み、両者のタンパク質配列データから長さ３のアミノ酸断片の出現頻度に基づく特徴量と宿主タンパク質についての既知のタンパク質相互作用ネットワークから抽出した特徴を組み込んだ特徴ベクトルを開発し、これに確率的勾配降下法などの最適化手法を組み合わせた手法を開発した。また、２種類のタンパク質が複合体を成すかどうかをタンパク質配列データなどから予測する新規手法を開発した。具体的には、タンパク質相互作用ネットワークデータ、配列データをもとに得たドメイン組成データ、系統プロファイルデータ、細胞内局在性データなどをカーネル関数を用いて統合し、それとサポートベクターマシンを組み合わせることにより学習および予測を行う手法を開発した。また、以前より研究を行っていたタンパク質切断部位予測について、既存手法および以前に開発した手法をベンチマークデータにより比較し、さらに、開発手法の有用性を生物学的実験によって示した。

（３）RNA 情報解析
　RNA 結合予測に関して以前より整数計画法を用いる方法を開発していたが、生物学実験を通じた評価・改良を行い、その成果を論文としてまとめた。また、非コード RNA とタンパク質の相互作用のなすネットワークの情報解析を行い、ネットワーク構造が二つに大きく分断されることを見出すとともに、ネットワーク制御において重要な役割を果たすと考えられる相当する非コード RNA には疾患と関連するものが多いことを見出した。

（４）代謝ネットワーク解析
　代謝ネットワークの改変について研究し、生成不可能にすべき化合物と新たに生成可能にすべき化合物を指定した際に、それらの制約を満たすようなネットワーク改変のうち、最小限の手間で済むものを見出すアルゴリズムを整数計画法に基づき開発した。

（５）当初計画との変更点
　上記のようにいくつもの研究成果を得ることができたが、カーネル回帰などの回帰手法との組み合わせについては、近年の深層学習技術の急速な進展により、ニューラルネットワークを用いた回帰を適用する方がより有効性が高いと判断した。そこで、ニューラルネットワークを用いた回帰と列挙との組み合わせを含む基盤研究(A)を前年度申請し採択されたため、新たな枠組みで研究に取り組むことにした。

５．主な発表論文

〔雑誌論文〕（計 34 件）

① Y. Nishiyama, A. Shurbevski, H. Nagamochi, T. Akutsu: Resource cut, a new bounding procedure to algorithms for enumerating tree-like chemical graphs. IEEE/ACM Trans. Comput. Biology Bioinform., 査読有, 16:77-90, 2019.
https://doi.org/10.1109/TCBB.2018.2832061
② J. Li, H. Nagamochi, T. Akutsu: Enumerating substituted benzene isomers of tree-like chemical graphs. IEEE/ACM Trans. Comput. Biology Bioinform., 査読有, 15:633-646, 2018.
https://doi.org/10.1109/TCBB.2016.2628888
③ T. Tamura, W. Lu, J. Song, T. Akutsu: Computing minimum reaction modifications in a Boolean metabolic network. IEEE/ACM Trans. Comput. Biology Bioinform., 査読有, 15: 1853-1862, 2018.
https://doi.org/10.1109/TCBB.2017.2777456
④ L. Liu, T. Mori, Y. Zhao, M. Hayashida, T. Akutsu: Euler string-based compression of tree-structured data and its application to analysis of RNAs. Current Bioinformatics, 査読有, 13:25-33, 2018.
https://doi.org/10.2174/1574893611666160608102 23
⑤ P. Ruan, M. Hayashida, T. Akutsu, J-P. Vert: Improving prediction of heterodimeric protein complexes using combination with pairwise kernel. BMC Bioinformatics, 査読有,

19-S(1):73-84, 2018.
https://doi.org/10.1186/s12859-018-2017-5
⑥ M. Ishitsuka, T. Akutsu, J. C. Nacher: Critical controllability analysis of directed biological networks using efficient graph reduction. Scientific Reports, 査読有, 7:14361 (10 pages), 2017.
https://doi.org/10.1038/s41598-017-14334-8
⑦ Y. Kato, T. Mori, K. Sato, S. Maegawa, H. Hosokawa, T. Akutsu: An accessibility-incorporated method for accurate prediction of RNA-RNA interactions from sequence data. Bioinformatics, 査読有, 33:202-209, 2017.
https://doi.org/10.1093/bioinformatics/btw603
⑧ J. Jindalertudomdee, M. Hayashida, Y. Zhao, T. Akutsu: Enumeration method for tree-like chemical compounds with benzene rings and naphthalene rings by breadth-first search order. BMC Bioinformatics, 査読有, 17:113 (16 pages), 2016.
https://doi.org/10.1186/s12859-016-0962-4
⑨ J. Jindalertudomdee, M. Hayashida, T. Akutsu: Enumeration method for structural isomers containing user-defined structures based on breadth-first search approach. Journal of Computational Biology, 査読有, 23: 625-640, 2016.
https://doi.org/10.1089/cmb.2016.0056
⑩ Y. Bao, M. Hayashida, T. Akutsu: LBSizeCleav: improved support vector machine (SVM)-based prediction of Dicer cleavage sites using loop/bulge length. BMC Bioinformatics, 査読有, 17:487 (11 pages), 2016.
https://dor.org/10.1186/s12859-016-1353-6
⑪ T. Tamura, W. Lu, T. Akutsu: Computational methods for modification of metabolic networks. Computational and Structural Biotechnology Journal, 査読有, 13:376-381, 2015.
http://dx.doi.org/10.1016/j.csbj.2015.05.004
⑫ H. Kagami, T. Akutsu, S. Maegawa, H. Hosokawa, J. C. Nacher: Determining associations between human diseases and non-coding RNAs with critical roles in network control. Scientific Reports, 査読有, 5:14577 (11 pages), 2015.
https://doi.org/10.1038/srep14577

〔学会発表〕(計 15 件)
① Y. Tamura, A. Shurbevski, H. Nagamochi, T. Akutsu: Enumerating chemical mono-block 3-augmented trees with two junctions. The 8th International Conference on Bioscience, Biochemistry and Bioinformatics, 2018.
② Y. Nishiyama, A. Shurbevski, H. Nagamochi, T. Akutsu: Resource cut, a new bounding procedure to algorithms for Enumerating tree-like chemical graphs. Sixteenth Asia Pacific Bioinformatics Conference, 2018.
③ T. Mori, H. Ngouv, M. Hayashida, T. Akutsu, J. Nacher: ncRNA-disease association prediction based on sequence information and tripartite network. Sixteenth Asia Pacific Bioinformatics Conference, 2018.
④ J. Jindalertudomdee, M. Hayashida, J. Song, T. Akutsu: Host-pathogen protein interaction prediction based on local topology structures of a protein interaction network. IEEE 16th International Conference on Bioinformatics and Bioengineering, 7-12, 2016.
⑤ T. Tamura, C-Y. Lin, J-M. Yang, T. Akutsu: Finding influential genes using gene expression data and Boolean models of metabolic networks. IEEE 16th International Conference on Bioinformatics and Bioengineering, 2016.
⑥ M. Hayashida, J. Jindalertudomdee, T. Akutsu: Parallelization of enumerating tree-like chemical compounds by breadth-first search order. 8th International Conference on Systems Biology, 2014.

〔図書〕(計 1 件)
① T. Akutsu: Algorithms for Analysis, Inference, and Control of Boolean Networks. World Scientific, 2018.

〔産業財産権〕
○出願状況(計 0 件)
○取得状況(計 0 件)

〔その他〕
なし

６．研究組織

(1)研究分担者
研究分担者氏名：永持　仁
ローマ字氏名：(NAGAMOCHI, hitoshi)
所属研究機関名：京都大学
部局名：大学院情報学研究科
職名：教授
研究者番号（8 桁）：70202231

研究分担者氏名：細川　浩
ローマ字氏名：(HOSOKAWA, hiroshi)
所属研究機関名：京都大学
部局名：大学院情報学研究科
職名：講師
研究者番号（8 桁）：90359779

研究分担者氏名：林田　守広
ローマ字氏名：(HAYASHIDA, morihiro)
所属研究機関名：松江工業高等専門学校
部局名：電気情報工学科
職名：准教授
研究者番号（8 桁）：40402929

(2)研究協力者
研究協力者氏名：前川　真吾
ローマ字氏名：(MAEGAWA, shingo)
所属研究機関名：京都大学
部局名：大学院情報学研究科
職名：助教
研究者番号（8 桁）：30467401