公募研究：2003～2004年度

生体高分子と結合する低分子化合物の効率的な比較、探索、発見アルゴリズムの開発

●馬見塚　拓　◆山口　敦子

京都大学化学研究所

〈研究の目的と進め方〉

タンパク質等の生体高分子に結合する比較的分子量の小さな化合物の比較、探索、発見のための効率的なアルゴリズムを開発・実装することが本研究の目的である。本研究は、比較、探索のためのアルゴリズムの開発と、発見のためのアルゴリズムの開発の二つに分けられる。

前者の研究は、理論計算機科学を背景とし、低分子化合物の化学構造を分子グラフとみなす。さらに、生体内の分子グラフに対して仮定可能な合理的な制約を有効に利用することによって、理論計算機科学の観点からも新規性のある効率的な比較・探索アルゴリズムを開発していく。

後者の研究は、機械学習・データマイニングを背景とし、低分子化合物データに即した表現系やモデルを設定し、それらの学習手法を開発する。特に、近年多様な生体内のデータが蓄積されつつあり、多種のデータと化合物データを組み合わせることにより、より精度の高い効率的な学習・マイニングが可能になると予想される。従って、異種データの融合が可能な学習・発見手法を開発していく。

〈研究開始時の研究計画〉

上記、目的と進め方に記載したように、本研究は主に２つに分けられ、それぞれ異なるアプローチにより進めることを計画した。

前者は、比較・探索のためのアルゴリズムであり、理論計算機科学の面から評価可能な効率的なアルゴリズムの開発を計画した。具体的には、最大共通部分グラフの探索である。加えて、低分子化合物の中でも比較的強い制約を持つ特殊な化合物を対象としたアルゴリズムの実装も計画した。具体的には、糖鎖に焦点をあてたマッチングアルゴリズムの実装である。

後者は、発見のためのアルゴリズムであり、機械学習・データマイニングといった文脈において評価可能なアルゴリズムの構築を計画した。具体的には、化合物間の相互作用および生体低分子化合物のネットワークである代謝パスウェイを対象とし、それらの解析、さらに未知の相互作用ないしは代謝経路の予測・発見が可能な手法の構築を目論んだ。

〈研究期間の成果〉

比較・探索アルゴリズムの開発においては、主に二つの成果を挙げた。第一の成果について説明する。ここでは、二つのグラフ（分子）を入力とし、最大共通部分グラフを出力する問題をまず考慮した。特に、化学構造に対して合理的な制約をおき効率的なアルゴリズムを構築した。合理的な制約とは、具体的には以下の２つである。まず各原子（ノード）の結合数（次数）には上限があること、および木幅（tree-width）と呼ばれるグラフの複雑性の尺度の一つに上限があることである。これらの制約をもったグラフと、より弱い制約（具体的には全域木の数が多項式）をもったグラフの二つを入力とした時に、多項式時間でこの問題を解決するアルゴリズムを構築した。さらに、実際の化合物における木幅の値や全域木の数を測定し、木幅は非常に小さな値で抑えられること、また、全域木の数も比較的小さなオーダーの多項式で抑えられることを確認し、想定した制約が合理的だったことがわかる。さらに、ここで使用した制約、特に木幅を尺度とした化合物の特性についての解析を開始し、様々な知識発見としての成果を挙げた。特に、このアルゴリズムの入力グラフは、最大共通部分グラフを発見する問題において、最も広いクラスに属するグラフであり、この成果は計算機科学の観点からも新しく、理論計算機科学の雑誌及び国際会議予稿集に受理・採録された（成果リスト：５、１０）。第二の成果は、第三の生体分子とも呼ばれる糖鎖をlabeled unordered treeと見なして、効率的にペアワイズアライメントする手法を実装した。この成果は、国際会議の他、雑誌論文としてもまとめられた（成果リスト：４、７、８、１１）。

発見アルゴリズムに関しては、主に２つの成果を挙げた。第一は、相互作用データより一般的には共起データを解析するための機械学習手法の構築を行った。より具体的には、共起データのみならずデータの背景知識を導入しそれらの観点から共起データをクラスタリングする確率モデルである。タンパク質相互作用データ等を用いた実証実験から既存手法に対する優位性を検証した。この成果は、バイオインフォマティクスあるいは機械学習分野の有力な雑誌もしくは国際会議予稿集に受理・採録された（成果リスト：１、２、３、１２、１３、１４、１５、１６）。第二は代謝系における化合物生成反応の中で、遺伝子発現により起きていると考えられる反応集合のみを自動的に抽出し、それらを反応鎖としてクラスタリングする枠組みを提案した。より具体的には、化合物化学反応を一次マルコフモデルとして捉え既存の代謝パスウェイをマルコフモデルとみなす、反応を活性化する酵素の遺伝子発現情報から反応鎖（系列）を生成し、マルコフモデルミクスチャの確率パラメータを反応鎖データから学習する。見方を変えると、この手法は、代謝パスウェイの既存の知識と遺伝子発現データを組み合わせ活性度の高いパスウェイを発見する手法と言える。この手法により、遺伝子発現の観点から活性化される反応鎖を自動的に抽出出来、代謝パスウェイ上の化学反応の知識発見が可能となることを実験により示した。この成果は、バイオインフォマティクスおよびデータマイニングの国際会議予稿集・雑誌に投稿し、受理・採録された（成果リスト：６，９）。

〈国内外での成果の位置づけ〉

薬物等の低分子化合物に関する問題を扱うケモインフォマティクスと呼ばれる分野は長い歴史を持つ。近年、情報処理技術の生命科学応用への注目とともに、再び脚光を浴びつつある。この分野には、本研究課題の関連研究が数多くある。しかし、比較・探索に関して、例えば上記最大共通部分グラフ探索問題に対しては、上記分野

を含めて関連手法の多くは、ad hocなヒューリスティクスを使用することが多く、最適解の保証や計算量の議論が無いことがままある。本研究の最大共通部分グラフ探索手法は、生体内低分子化合物の特徴を有効に利用し最適解を求める新しいアルゴリズムであり、計算量の上限を保証するという利点を持つ。さらに、計算機科学の観点からも入力グラフは最も広いクラスに属すると考えられる。また、糖鎖のペアワイズアライメント手法も、計算量が保証されている他、実装により糖鎖インフォマティクスのツールの世界的な先駆けとなり、糖鎖研究でも指折りのツールとなっている。発見に関しては、上記分野を中心として多くの手法があるが、いずれも既存の学習・マイニング手法の適用の域を出ない。本研究の発見手法は、化合物データをはじめとする多様なデータの各々に合致した表現系の構築と異種のデータを融合可能な効率的な学習手法の構築を目標とし、相応しい成果を収めた。すなわち、新規手法の構築のみならず、予測精度の大幅な向上かつ既存手法では不可能な科学的発見を可能とした。

〈達成できなかったこと、予想外の困難、その理由〉

比較・探索アルゴリズムの構築に関しては、上記成果の次のステップとして、入力グラフの最大共通部分のみならず、グラフ間の相同性（距離）を出力するアルゴリズムを研究課題とした。この目的のために、入力分子グラフ間の最短反応（編集）ステップを距離と考え、この最短編集距離問題を効率的に解くアルゴリズムの構築・実装を目標としていた。しかしながら、この問題の解決には十分な時間が無く、効率的なアルゴリズムを提案するには至らなかった。

〈今後の課題〉

比較・探索アルゴリズムの構築に関しては、上記未達成の問題の解決が今後の課題である。具体的な方策としては、入力グラフ比較での編集距離の計算に際し、どのような編集を対象とすべきかをまず考慮する。そのために、最大共通部分グラフ問題の解法で得た低分子化合物に関する知見を踏まえて、まず、必要な編集セットと分子グラフの性質を反映した制約を考察する。さらに、これら前提条件の下で、この問題への効率的なアルゴリズムを構築することが課題となる。

発見アルゴリズムに関しては、低分子化合物に関連する他の様々な問題に取り組んでいく。この際、対象とする問題に合わせて、発見効率の改善に有効なデータの調査をより進め、必要と思われるより多種のデータを選択し、それらの融合のための表現系の構築、学習アルゴリズムの設計・実装を行い、構築アルゴリズムの性能を試していくことが今後の課題となる。

〈研究期間の全成果公表リスト〉

1）Mining New Protein-Protein Interactions - Using a Hierarchical Latent-variable Model to Determine the Function of a Functionally Unknown Protein. Mamitsuka, H., IEEE Engineering in Medicine and Biology Magazine, 24 (3), 103-108, 2005.

2）Essential Latent Knowledge for Protein-Protein Interactions: Analysis by an Unsupervised Learning Approach. Mamitsuka, H., IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2 (2), 119-130, 2005.

3）Efficient Unsupervised Mining from Noisy Co-occurrence Data. Mamitsuka, H., New Mathematics and Natural Computation, 1 (1), 173-193, 2005.

4）A Score Matrix to Reveal the Hidden Links in Glycans. Aoki, K. F., Mamitsuka, H., Akutsu, T. and Kanehisa, M., Bioinformatics, 21 (8), 1457-1463, 2005.

5）Finding the Maximum Common Subgraph of a Partial k-Tree and a Graph with a Polynomially Bounded Number of Spanning Trees. Yamaguchi, A., Aoki, K. F. and Mamitsuka, H., Information Processing Letters, 92 (2), 57-63, 2004.

6）A Hierarchical Mixture of Markov Models for Finding Biologically Active Metabolic Paths using Gene Expression and Protein Classes. Mamitsuka, H. and Okuno, Y.,Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB 2004), 341-352, Stanford, CA, August, 2004, IEEE Computer Society Press.

7）KCaM (KEGG Carbohydrate Matcher): A Software Tool for Analyzing the Structures of Carbohydrate Sugar Chains. Aoki, K. F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S. and Kanehisa, M., Nucleic Acids Research, 32, W267-W272, 2004.

8）Managing and Analyzing Carbohydrate Data. Aoki, K. F., Ueda, N., Yamaguchi, A., Akutsu, T., Kanehisa, M. and Mamitsuka, H., ACM SIGMOD Record, 33 (2), 33-38, 2004.

9）Mining Biologically Active Patterns in Metabolic Pathways using Microarray Expression Profiles. Mamitsuka, H., Okuno, Y. and Yamaguchi, A., ACM SIGKDD Explorations, 5 (2):113-121, 2003.

10）Finding the Maximum Common Subgraph of a Partial k-Tree and a Graph with a Polynomially Bounded Number of Spanning Trees. Yamaguchi, A. and Mamitsuka, H., Proceedings of the Fourteenth International Symposium on Algorithm and Computation (ISAAC 2003, Lecture Notes in Computer Science, 2906), 58-67, Kyoto, Japan, December, 2003, Springer-Verlag.

11）Efficient Tree-Matching Methods for Accurate Carbohydrate Database Queries. Aoki, K. F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M. and Mamitsuka, H., Proceedings of the Fourteenth International Conference on Genome Informatics (GIW 2003, Genome Informatics, 14), 134-143, Yokohama, Japan, December, 2003, Universal Academy Press.

12）Efficient Mining from Heterogeneous Data Sets for Predicting Protein-Protein Interactions. Mamitsuka, H., Proceedings of the Fourteenth International Workshop on Database and Expert Systems, 32-36, Prague, Czech Republic, September, 2003, IEEE Computer Society Press.

13）Selective Sampling with a Hierarchical Latent Variable Model. Mamitsuka, H., Proceedings of the Fifth International Symposium on Intelligent Data Analysis (IDA 2003, Lecture Notes in Computer Science, 2810), 352-363, Berlin, Germany, August, 2003, Springer-Verlag.

14）Hierarchical Latent Knowledge Analysis for Co-occurrence Data. Mamitsuka, H., Proceedings of the Twentieth International Conference on Machine

Learning (ICML 2003), 504-511, Washington DC, August, 2003, AAAI Press.
15) Efficient Unsupervised Mining from Noisy Data Sets: Application to Clustering Co-occurrence Data. Mamitsuka, H., Proceedings of the Third SIAM International Conference on Data Mining (SDM 2003), 239-243, San Francisco, CA, May, 2003, SIAM.
16) Detecting Experimental Noise in Protein-Protein Interactions with Iterative Sampling and Model-based Clustering. Mamitsuka, H., Proceedings of the Third IEEE International Symposium on Bioinformatics and Bioengineering (BIBE 2003), 385-392, Bethesda, MD, March, 2003, IEEE Computer Society Press.