バイオインフォマティクスのための共通パターン抽出アルゴリズムの研究

(課題番号　１３６８０３９４)

平成１３年度～１５年度科学研究費補助金
基盤研究(C)(2) 研究成果報告書

平成１７年５月

京都大学図書
1050712757
阿久津達也氏寄贈
附属図書館

研究代表者　阿久津　達也
(京都大学化学研究所教授)

バイオインフォマティクスのための共通パターン抽出アルゴリズムの研究

(課題番号　１３６８０３９４)

平成１３年度～１５年度科学研究費補助金
基盤研究(C)(2) 研究成果報告書

平成１７年５月

研究代表者　阿久津　達也
(京都大学化学研究所教授)

研 究 組 織

研究代表者: 阿久津　達也　（京都大学化学研究所　教授）
研究分担者: 宮野　悟　（東京大学医科学研究所　教授）
研究分担者: 上田　展久　（京都大学化学研究所　助手）

交付決定額

（金額単位：千円）

	直接経費	間接経費	合　計
平成13年度	900	0	900
平成14年度	700	0	700
平成15年度	1,000	0	1,000
総　計	2,600	0	2,600

研究発表

学会誌等

1. T. Akutsu, A local search algorithm for local multiple alignment: special case analysis and application to cancer classification, *Proc. 2001 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2001)*, 1284-1290, 2001.
2. T. Akutsu and K. Horimoto: Local multiple alignment of numerical sequences: detection of subtle motifs from protein sequences and structures, *Genome Informatics*, **12**, 83-92, 2001.
3. T. Akutsu, H. Bannai, S. Miyano and S. Ott: On the complexity of deriving position specific score matrices from examples, *Proc. 13th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, **2373**, 168-177, 2002.
4. T. Akutsu and S. Ott: Inferring a union of halfspaces from examples, *Proc. 8th International Conference on Computing and Combinatorics, Lecture Notes in Computer Science*, **2387**, 117-126, 2002.
5. K.C.D. Bahadur, T. Akutsu, E. Tomita, T. Seki and A. Fujiyama: Point matching under non-uniform distortions and protein side chain packing based on an efficient maximum clique algorithm, *Genome Informatics*, **13**, 143-152, 2002.
6. T. Akutsu, S. Miyano and S. Kuhara: A simple greedy algorithm for finding functional relations: efficient implementation and average case analysis, *Theoretical Computer Science*, **292**, 481-495, 2003.
7. T. Akutsu, K. Kanaya, A. Ohyama and A. Fujiyama: Point matching under non-uniform distortions, *Discrete Applied Mathematics*, **127**, 5-21, 2003.
8. T. Akutsu, S. Kuhara, O. Maruyama and S. Miyano: Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model, *Theoretical Computer Science*, **298**, 235-251, 2003.
9. D. Shinozaki, T. Akutsu and O. Maruyama: Finding optimal degenerated patterns in DNA sequences, *Bioinformatics*, **19**, ii206-ii214, 2003.
10. M. Hayashida, N. Ueda and T. Akutsu: Inferring strengths of protein-protein interactions from experimental data using linear programming, *Bioinformatics*, **19**, ii58-ii65, 2003.
11. H. Saigo, J-P. Vert, N. Ueda and T. Akutsu: Protein homology detection using string alignment kernels, *Bioinformatics*, **20**, 1682-1689, 2004.
12. T. Akutsu: Algorithms for point set matching with k-differences, *Proc. 10th Int. Computing and Combinatorics Conference, Lecture Notes in Computer Science*, **3106**, 249-258, 2004.

口頭発表

1. T. Akutsu and S. Miyano: Selecting informative genes for cancer classification using gene expression data, *2001 IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 2001年6月4日.
2. 阿久津達也, 坂内英夫, 宮野悟, Sascha Ott: 位置依存スコア行列の例からの学習の計算複雑度について, 情報処理学会 第*82*回アルゴリズム研究会, 2002年1月24日.
3. 阿久津達也, Sascha Ott: 半空間の和集合の例からの推定, 情報処理学会 第*84*回アルゴリズム研究会, 2002年5月23日.
4. 阿久津達也: バイオインフォマティクスにおけるアルゴリズム論的諸問題, 情報処理学会 第*86*回アルゴリズム研究会, 2002年9月19日.
5. T. Akutsu: A Gibbs sampling algorithm for numerical sequences: Detection of subtle motifs from protein sequences and structures, *2002 Korean Society for Bioinformatics Annual Meeting*, 2002年11月15日.
6. T. Akutsu: Optimization problems and metaheuristics in Bioinformatics, *5th Metaheuristics International Conference*, 2003年8月28日.
7. T. Akutsu: Computational and statistical methods in Bioinformatics, *14th Int. Symp. on Methodologies for Intelligent Systems*, 2003年10月28日.

出版物

1. T. Akutsu and S. Miyano: Selecting informative genes for cancer classification using gene expression data, In: W. Zhang & I. Shmulevich (eds.): *Computational and Statistical Approaches to Genomics*, Kluwer Academic Publishers, 2002.
2. 阿久津達也: 遺伝子発現情報解析のための数理モデルとアルゴリズム, (財) 国際高等研究所, 2003年2月28日.

研究成果による工業所有権の出願・取得状況

無し

目 次

1 はしがき **1**

2 局所アライメントによる共通配列パターン抽出 **4**

2.1 A local search algorithm for local multiple alignment: special case analysis and application to cancer classification 5

2.2 Local multiple alignment of numerical sequences: detection of subtle motifs from protein sequences and structures 12

3 正負の例からの配列パターン抽出 **22**

3.1 On the complexity of deriving position specific score matrices from examples 23

3.2 Inferring a union of halfspaces from examples 33

3.3 Finding optimal degenerated patterns in DNA sequences 43

4 電気泳動画像およびタンパク質立体構造のパターンマッチング **52**

4.1 Point matching under non-uniform distortions 53

4.2 Point matching under non-uniform distortions and protein side chain packing based on an efficient maximum clique algorithm 70

4.3 Algorithms for point set matching with k-differences 80

5 カーネル法に基づく配列分類 **90**

5.1 Protein homology detection using string alignment kernels 91

6 ブーリアンネットワークに基づく遺伝子ネットワーク推定 **99**

6.1 A simple greedy algorithm for finding functional relations: efficient implementation and average case analysis 100

6.2 Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model 115

7 線形計画法に基づくタンパク質相互作用推定 **132**

7.1 Inferring strengths of protein-protein interactions from experimental data using linear programming 133

1 はしがき

ヒトゲノム計画を始めとする各種ゲノム計画の急速な進展により、DNA配列データ（文字列データ）のみならず、遺伝子発現量の時系列データ（実数値配列データ）、タンパク質二次元電気泳動画像やDNAマイクロアレイなどの画像データ（２次元データ）、タンパク質立体構造データ（３次元データ）などの多種多様なデータが大量に生成されつつある。これらの大量データの中から有用な情報を引き出すことが強く求められているが、そのための主要な解析法の一つとして、共通の生物学的性質をもつ遺伝子群やタンパク質群が与えられた時に、対応するDNA配列群や立体構造群から共通のパターンを検出するという方法があげられる。パターンと機能とは深く関連する場合が多いことが広く知られているので、これまでにも数多くの研究が行なわれてきた。特に大きく分類すると以下の２種類の定式化のもとでの研究が行なわれてきた。

(i) 最適化問題としての定式化
　パターンに対してスコアづけを行い、スコアが最適となるパターンを抽出する

(ii) 学習問題としての定式化
　機能既知のデータから隠れマルコフモデルなどの数理モデルを学習させ、機能未知のデータが入力された時にそのモデルを用いて推定を行なう

しかしながら、いずれの定式化においても計算論的な困難に直面する。(i) では入力配列の個数に制限がない場合にはNP困難という計算論的に困難なクラスに属してしまうことがほとんどである。(ii) では最適な数理モデルの学習は困難であることが多く、EMアルゴリズムなどの局所探索アルゴリズムが幅広く利用されている。また、これまでの研究では配列パターン（配列モチーフ）については多くの研究がなされたが、その他のデータからの共通パターン抽出についてはあまり研究されていないという問題もあった。そこで、本研究では、(i)(ii) の両者のアプローチについて更なる理論的研究を行うとともに、より多様なデータに対応可能なアルゴリズムの開発を試みた。具体的には、以下の研究を行った。

局所アライメントによる共通配列パターン抽出

局所アラインメントは配列データからのモチーフ抽出において広く利用されており、特にEMアルゴリズムに基づく方法や、Gibbsサンプリングに基づく方法が良く用いられている。そこで、EM型のアルゴリズムの特殊な場合についての収束性に関する理論的研究を行うとともに、その遺伝子発現データ解析への応用の研究を行った。また、Gibbsサンプリング型のアルゴリズムを数値データに適用できるような修正を行い、タンパク質立体構造データからのモチーフ抽出に適用した。

正負の例からの配列パターン抽出

正負の例から配列パターンを抽出するために、正規表現などの文法規則を用いた研究がこれまで数多く用いられてきた。しかしながら、位置依存スコア行列（PSSM）と呼ばれる確率的なモデルがバイオインフォマティクスにおいては幅広

く利用されているが、このモデルを正負の例から抽出することに関して理論面からはほとんど研究されていなかった。そこで、この問題について研究を行い、パターン位置が既知、かつ、正負の例を誤り無く分類できる場合は多項式時間でPSSMを学習できることを示し、それ以外のほとんどの場合にはNP困難となることを示した。さらに、近似可能性に関する困難性も示した。また、これらの問題は半空間の和集合を学習することと深い関連があることも示した。一方、正負の配列から最適なdegenerateパターンを求める実用的な手法についても研究を行った。

電気泳動画像およびタンパク質立体構造のパターンマッチング

DNA二次元電気泳動画像のスポットマッチング問題を幾何的なマッチング問題として定義し、そのNP困難性を証明した。一方、この問題と、ある定式化のもとで既にNP困難であることが知られているタンパク質立体構造の構造アライメント問題に対し、これらの問題を最大クリーク問題に帰着し、最大クリークに対する既存のアルゴリズムを適用することにより最適解を求める手法を開発した。実データを用いた計算機実験の結果、サイズ（点もしくはアミノ酸の個数）が100〜200程度であれば最適解が計算できることが判明した。また、上記問題と関連のある最大共通部分点集合を求める問題について理論的研究を行い、類似性の高い点集合が入力された場合に、計算効率に関する従来の結果を改善することができた。

カーネル法に基づく配列分類

タンパク質配列から共通部分を抽出するためには、タンパク質配列をあらかじめ分類しておくことが有用である。最近、配列間の類似性を測るカーネル関数を定義し、それに基づきサポートベクタマシンを適用することにより配列を分類するという研究が行われるようになった。そこで本研究においては、配列アライメントにおいて広く利用されているローカルアライメントに基づくカーネル関数を開発した。ただし、ローカルアライメントをそのまま用いたのではカーネル関数としての数学的性質を満たさないため、ペアHMMという確率モデルを用いてローカルアライメントと似た計算を行うカーネル関数を開発した。ベンチマークデータを用いて比較を行った結果、既存のカーネル関数より良好な分類精度を得ることができた。

ブーリアンネットワークに基づく遺伝子ネットワーク推定

遺伝子ネットワーク推定は、多くの遺伝子発現データからその背後にあるネットワーク構造を推定するという意味で、パターン抽出問題として解釈することができる。本研究においては、この問題に関して我々が以前に行った研究の詳細にわたる見直しを図り、Journal論文という形式にまとめた。

線形計画法に基づくタンパク質相互作用推定

既知のタンパク質相互作用データから未知のタンパク質相互作用を推定することはバイオインフォマティクスにおいて重要な問題であり、多くの研究が行われている。そのための一つのアプローチとして、個々のタンパク質をドメインと呼ばれる部品の集まりとみなして、既知の相互作用データからドメイン間の相互作用を抽出

することにより、新たな相互作用を予測する研究が行われている。この問題も、ドメイン間の相互作用パターンを抽出する問題と考えることができる。そこで、この問題を線形計画問題として定式化することによりドメイン間の相互作用パターンを抽出する手法を開発した。計算機実験を行った結果、相互作用頻度データをもとに推定を行った場合には既存手法を大きく上回る予測精度を得ることができた。

上記のように本研究は順調に進展し、バイオインフォマティクスにおける多くのパターン抽出問題に対して、理論的に有用な結果、もしくは、実データに対して有効に動作するアルゴリズムを開発することができた。

なお、当初の計画では平成16年度まで研究を行うことになっていたが、本研究で得られた成果を発展させ、より実用的なアルゴリズムを研究開発することが必要になり、また、バイオインフォマティクス分野の発展に伴い、配列データや立体構造データ以外にも、化学構造、糖鎖構造、RNA二次構造などのグラフ構造を持ったデータからの共通パターン抽出について研究することも必要になった。そこで、より大きな規模で研究を展開することを目的として申請を行ったところ、平成16年度より科学研究費補助金基盤研究(B)(2)「構造を持つ生物情報データからの共通パターン抽出法」を開始することが認められたため、本研究は平成15年度にて終了した。

2 局所アライメントによる共通配列パターン抽出

2.1 A local search algorithm for local multiple alignment: special case analysis and application to cancer classification

2.2 Local multiple alignment of numerical sequences: detection of subtle motifs from protein sequences and structures