人工知能学会研究会資料
SIG-FPAI-B501-02

木のマルチプルアラインメント

Multiple Alignments of Trees

森川優[1]　簗瀬大悟[1]　兵頭俊紀[1]　田中謙次[1]　申吉浩[1]　久保山哲二[2]

[1] 兵庫県立大学応用情報科学研究科
[1] Graduate School of Applied Informatics, University of Hyogo
[2] 学習院大学計算機センター
[2] Computer Center, Gakushuin University

Abstract: It was very recent that a formal definition of multiple alignments is given to general data structures that include not only strings but also trees and graphs. Also, it has been shown that the center star method, which is to compute approximately optimal multiple alignments for strings, is effective for the generalized multiple alignments. In this paper, we report the results of our experiments to prove effectiveness of the generalized multiple alignments and the extended center star method by taking trees as an example.

1　はじめに

配列（文字列）のペアに対するアラインメントは、Levenshtein 編集距離との関連で深く研究されている。例えば、

$$\alpha = \begin{pmatrix} - & \mathtt{C} & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{T} & - \\ \mathtt{T} & \mathtt{C} & - & - & \mathtt{G} & \mathtt{G} & \mathtt{A} \end{pmatrix}$$

は、二つの DNA 配列、CAAGT と TCGGA の間のアラインメントである。このアラインメントのコストは $\gamma(\alpha) = 5$ と計算され、与えられた配列のペアに対するアラインメントのコストの最小値が levenshtein 距離となる。

一方、3 つ以上の配列間のアラインメントは、マルチプルアラインメントと呼ばれ、例えば、

$$\alpha' = \begin{pmatrix} - & \mathtt{C} & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{T} & - \\ \mathtt{T} & \mathtt{C} & - & - & \mathtt{G} & \mathtt{G} & \mathtt{A} \\ \mathtt{T} & \mathtt{C} & \mathtt{A} & - & \mathtt{G} & \mathtt{G} & - \end{pmatrix}$$

は、TCAGG を加えた場合のマルチプルアラインメントである。ペアのアラインメントの場合と同様に、マルチプルアラインメントにもコスト関数を導入することが可能であり、任意の行のペアをペアのアラインメントと考えて、そのコストの総和を計算する、sum-of-pair 関数は広く使われている。マルチプルアラインメントのコストの最小値は、3 つ以上の配列の類似度を測る尺度と考えることができる。更に、コストの最小値を与えるアラインメントを最適アラインメントと呼ぶと、最適アラインメントはある特徴を共有する配列間のパターンを抽出する用途に利用できる。

n 個の配列に対して最適アラインメントを計算する計算量に関しては、$n = 2$ の時、即ち、配列 x と y の最適アラインメントは動的計画法により計算可能で、その時の計算量は $O(|x|\,|y|)$ である一方、$n > 2$ の場合は NP 困難であることが知られている。従って、$n > 2$ の場合にマルチアラインメントを利用するためには、最適アラインメントを求めるための近似アルゴリズムが必要であるが、近似誤差の上限が理論的に知られている良好な近似アルゴリズムとして、Center Star 法が知られている。

一方、配列以外のデータ構造に対するアラインメントとしては、木に対するアラインメント木が知られている [2]。しかしながら、最適アラインメント木のコストは、木の編集距離として最も一般的な Taï距離にはならず、三角不等式を満足しない別の編集距離を定義するにとどまる。一方、木の Taï距離に限らず、三角不等式を満足するグラフの任意の編集距離に対してアラインメントグラフを求める方法は、申 [3] によって提案されている。

配列以外のデータ構造に対するマルチプルアラインメントに関しては、木やグラフも含めて、殆ど何も知られていなかったが、最近になって、申・久保山・宮原は、有限個の要素からなるデータ構造に対して、任意の三角不等式を満足する編集距離が与えられた場合、最適アラインメントのコストが距離と一致するようなマルチプルアラインメントを定義する、極めて一般的な方法を提案した。更に、$n > 2$ の時、最適マルチアラインメントを求める問題は一般に NP 困難であるが、Center Star 法が同じ近似誤差で成立することを示した。

本報告では、申・久保山・宮原のアルゴリズムを実

装し、更に、実際の木のデータセットを用いて、マルチプルアラインメントが有効であるかを検証する予備的実験を行ったのでその結果を報告する。

2 実験結果

2.1 プログラム

実験に用いるプログラムは、木の間の Tai編集距離を計算し、その結果に対して、申・久保山・宮原が開発した Center Star 法のアルゴリズムを適用し、マルチプルアラインメントを出力するものである。

与えられた木の集合に対して、最適アラインメントはひとつとは限らないが、このプログラムはその中の一つを無作為に出力するものである。

プログラムは Scala を用いて作成した。

2.2 データセット

実験に用いたデータセットは、赤血球に付着する糖鎖と白血病との関係を示すもので、クラスラベルは白血病の発病の有無である。糖鎖は、赤血球に固着しているタンパク質を根と考え、それに連なる糖鎖の構造を順序木として表現したものである。

データセットは、糖鎖データのレポジトリである [1] から取得した。

2.3 実験の目的と方法

実験の目的は、正例のみからなるデータ、負例のみからなるデータ、正例と負例が混合したデータから生成される最適マルチプルアラインメントをそのサイズで比較することにある。当初の予想としては、正例データから作成されるマルチプルアラインメントは、発病の理由となる共通構造を含んでいるため、データの数が増えてもマルチプルアラインメントのサイズは急速には増大せず、一方、負例データの場合は、木の構造に多様性があるため、マルチプルアラインメントのサイズは正例データの場合に比べて、マルチプルアラインメントのサイズはより急速に増大するであろうというものである。正例と負例が混合したデータの場合は、この傾向は更に強くなるものと予測した。

実験は、前述のデータセットから、正例、負例、正例と負例の混合したデータ（木）を、個数を 3 から 177 の間で変えながら、ランダムに選び、選んだデータに対して作成したプログラムによって最適アラインメントを計算することで実施した。

2.4 実験結果と考察

下図に示すように、当初の予想を裏付ける実験結果が得られた。

正例のみからなるデータ場合（グラフ a)、データの個数が 177 の場合に理由不明のはずれ値が存在するが、それ以外は低い値にまとまっている。赤の+印で示した平均値は、概ね、$y = 7.3x^{0.57}$ の曲線に沿って分布している。

一方、負例のみからなるデータの場合（グラフ b)、最適アラインメントのサイズは広い範囲でバラついており、平均値は、概ね、$y = 5.8x^{0.74}$ の曲線のまわりに分布する。

正例と負例が混ざったデータでは（グラフ c)、バラツキの幅はフレイのみの場合と大きな差はないが、平均値は、概ね、$y = 6.2x^{0.76}$ の曲線の周りに分布する。

近似曲線の次数を見ると、$0.57 < 0.74 < 0.76$ と、正例のみ、負例のみ、正例・負例混合の順で増加しており、特に、正例のみの場合の次数は他に比較して際立って小さい。

これは、正例のデータである木構造の間では、白血病発症の原因となる共通構造を含んでおり、その共通構造はマルチプルアラインメント中では木の多くのノードを吸引するため、マルチプルアラインメントのサイズが小さくなることが理由だと考えられる。

2.5 今後の課題

今回は、申・久保山・宮原によって考案されたマルチプルアラインメントの定義と、拡張された Center Star 法によるマルチプルアラインメントの利用の有効性を確かめる予備的実験である。今後、マルチプルアラインメントからの共通構造の抽出、共通構造の有効性の検証、より詳細なマルチプルアラインメントの統計的性質の検証を行い、配列以外の構造、特に木構造における、マルチプルアラインメントの利用の有効性を検証していく。

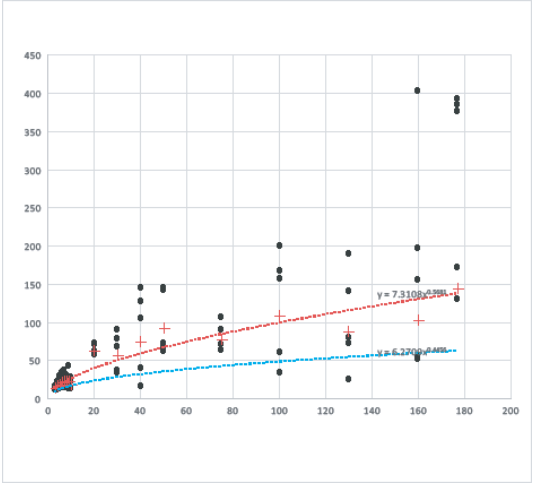
参考文献

[1] K. Hashimoto, S. Goto, S. Kawano, K. F. Aoki-Kinoshita, and N. Ueda. Kegg as a glycome informatics resources. *Glygobiology*, 16:63R – 70R, 2006.

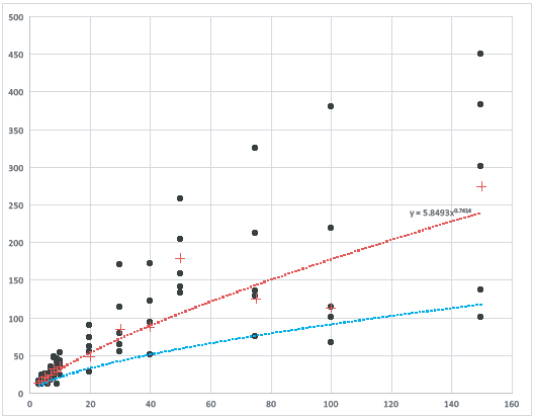
[2] T. Jiang, L. Wang, and K. Zhang. Alignment of trees — an alternative to tree edit. *Theoretical Computer Science*, 143:137–148, 1995.

[3] K. Shin. Alignment kernels based on a generalization of alignments. *IEICE Trans. on Information and Systems*, E97. D(1):1–10, 2014.

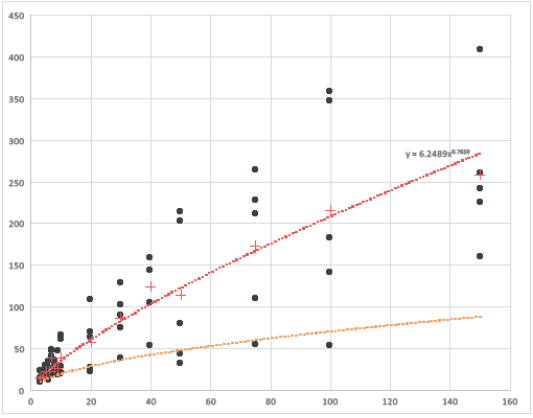
図 1: 実験結果



(a) 正例の場合の最適アラインメントのサイズ



(b) 負例の場合の最適アラインメントのサイズ



(c) 正例・負例が混じった場合の最適アラインメントのサイズ