多重アラインメントに基づく木データの分析

Multiple Alignments of Trees

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。

更に、申は、編集距離とパターン抽出問題の間に双対関係が成り立つことを示したが [4]、多重アラインメントとパターン抽出問題の関係は明らかではなかった。

本報告では、編集距離とパターン抽出問題の双対性

を利用して、多重アラインメントとパターン抽出問題との間の双対性を示す。この双対性により、申等によって拡張されたセンタースター法を用いてパターンを近似的に抽出することが可能となる。更に、この手法を現実の糖鎖のデータに適用したところ、白血病・結腸癌・嚢胞性線維症を特徴付ける糖鎖の構造パターンを明確に特定することができた。白血病の場合は特定された単一のパターンを含む場合に陽性と判定でき、結腸癌・嚢胞性線維症の場合は特定された複数のパターンのいずれかを含む場合に陰性と判定できる。

2 多重アラインメントとパターン抽出

以下では、[4] の内容を前提とする。[4] では、$(\mathscr{M},\Phi)$-パターン抽出問題を次のように定義した。

強 $(\mathscr{M},\Phi)$-パターン抽出 (PE) 問題:
$\{X_1,\ldots,X_n\} \subseteq \mathscr{W}$ とする。すべての $i \neq j \neq k \neq i$ に対して (1) $\mu_{ij} = \mu_{ji}^{-1}$ 及び (2) $\mu_{ij} = \mu_{ik} \circ \mu_{kj}$ が成り立つという条件のもとで、$\sum_{i=1}^{n}\sum_{j=i+1}^{n}\Phi(\mu_{ij})$ を最大にする $\{\mu_{ij} \in \mathscr{M}_{X_i,X_j} \mid i,j = 1,\ldots,n; i \neq j\}$ を求めよ。

この定義は MAST 問題の一般化であるが、申・久保山・宮原による多重アラインメントの一般化とはうまく対応しない。

弱 $(\mathscr{M},\Phi)$-パターン抽出 (PE) 問題:
$\{X_1,\ldots,X_n\} \subseteq \mathscr{W}$ とする。すべての $i \neq j \neq k \neq i$ に対して (1) $\mu_{ij} = \mu_{ji}^{-1}$ 及び (2) $\mu_{ij} \supseteq \mu_{ik} \circ \mu_{kj}$ が成り立つという条件のもとで、$\sum_{i=1}^{n}\sum_{j=i+1}^{n}\Phi(\mu_{ij})$ を最大にする $\{\mu_{ij} \in \mathscr{M}_{X_i,X_j} \mid i,j = 1,\ldots,n; i \neq j\}$ を求めよ。

上記のように条件 (2) を緩和することにより、多重アラインメントと双対関係が成立するようになる。即ち、距離の和を最小にする最適多重アラインメントを求めることと、弱 $(\mathscr{M},\Phi)$-パターン問題を求める音とは同値となる。また、申等によって拡張されたセンタースター法を双対的に弱 $(\mathscr{M},\Phi)$-パターン抽出 (PE) 問題に適用することにより、近似解を直接に求めることが可能となる。

3 糖鎖と疾病の関係の解明

疾病を特徴付ける糖鎖の構造パターンの特定を、弱 $(\mathscr{M},\Phi)$-パターン抽出問題を解くことにより行う。弱 $(\mathscr{M},\Phi)$-パターン抽出問題の解の計算は一般に NP-困難であるが、センタースター法により近似解を計算することができる。

本報告では、以下の手順により、白血病・結腸癌・嚢胞性線維症を特徴付ける糖鎖の部分構造を特定する。

1. センタースター法により弱 $(\mathscr{M},\Phi)$-パターン抽出問題の近似解を求める。
2. 得られたパターンを木構造パターンに分解する。
3. 得られた木構造パターンの予測性能を F 値で評価する。
4. 疾病を特徴付ける構造を抽出する。

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.

[4] 申吉浩, 射にもとづく類似性理論, **基本問題研究会（熊本）**, 2016