Vol.2015-MPS-103 No.25
Vol.2015-BIO-42 No.25
2015/6/23

根付き順序木の圧縮における分割型文法とオイラー文字列との比較

劉 立偉[1]　森 智弥[2]　趙 楊[3]　林田 守広[2]　阿久津 達也[2]

概要：辺にラベルの付いた根付き順序木を圧縮する手法として文法に基づく手法を検討する．木文法としては文脈自由文法 (CFG) を拡張した Simple Elementary Ordered Tree Grammar (SEOTG) と呼ばれる分割型文法を使用する．一方で根付き順序木はオイラー文字列として文字列に変換できるので，オイラー文字列を CFG に基づいて圧縮する．本研究では与えられた根付き順序木のみを生成する最小の SEOTG のサイズと，変換後のオイラー文字列のみを生成する最小の CFG のサイズとを比較し，ある特定の木に対しては，76 頂点までであるが，頂点数の増加とともにサイズの差も大きくなることを報告する．

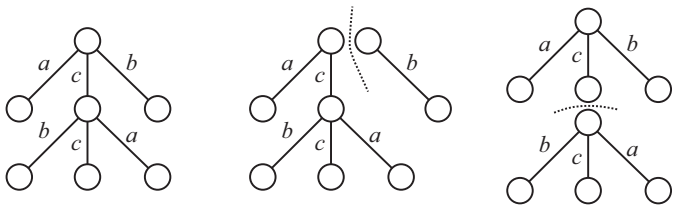

図 1　根付き順序木の分割の例．左）もとの木．中）根での横分割．右）内部頂点での縦分割．

1. はじめに

根付き順序木を圧縮する手法として，分割型文法である SEOTG [1] に基づく手法とオイラー文字列を用いた手法を紹介する．

1.1 分割型文法 SEOTG に基づく圧縮

SEOTG は木の分割方法として，根における横分割と，根と葉以外の内部頂点での縦分割をもつ (図 1)．分割された各部分木に SEOTG の一つの非終端記号が対応し，$T \to T_1 + T_2$ (横分割), $T \to T_1 \oplus T_2$ (縦分割), $T \to a$ (終端記号への置換) の形の規則からなるとする．

与えられた根付き順序木 T を生成する最小の SEOTG を見つけるために，非終端記号の数を最小化する整数計画問題 MinSEOTGMul [2] が以下のように提案されている．ここで U は T のすべての部分木の集合であり，p_u は部分木 u に対応する非終端記号が最小文法に現れるとき 1 を取り，そうでなければ 0 となる．s_u のみ実数値をとる．

最小化 $\sum_{u \in U} p_u$

制約条件

$$x_{i,\epsilon,j,j} = 1 \quad \text{for all } i,\ j \in ch(i)\ (|ch(j)| = 0),$$

$$x_{i,j,j,j} = 1 \quad \text{for all } i,\ j \in ch(i)\ (|ch(j)| > 0),$$

$$x_{1,\epsilon,lch(1),rch(1)} = 1,$$

$$x_{i,\epsilon,h,k} \le \sum_{l=h}^{k-1} y_{i,\epsilon,h,l,k} + \sum_{t \in I(T_{i,\epsilon,h,k})} z_{i,\epsilon,h,k,t} \quad \text{for all } i,\ h \le k \in ch(i),$$

$$y_{i,\epsilon,h,l,k} \le \frac{1}{2}(x_{i,\epsilon,h,l} + x_{i,\epsilon,l+1,k}) \quad \text{for all } i,\ h \le l < k \in ch(i),$$

$$z_{i,\epsilon,h,k,t} \le \frac{1}{2}(x_{i,t,h,k} + x_{t,\epsilon,lch(t),rch(t)}) \quad \text{for all } i,\ h \le k \in ch(i), t \in I(T_{i,\epsilon,h,k}),$$

$$x_{i,j,h,k} \le \sum_{l=h}^{k-1} y_{i,j,h,l,k} + \sum_{t \in anc(j)} z_{i,j,h,k,t} \quad \text{for all } i,\ h \le k \in ch(i),\ j \in I(T_{i,\epsilon,h,k}),$$

$$y_{i,j,h,l,k} \le \frac{1}{2}(x_{i,\epsilon,h,l} + x_{i,j,l+1,k}) \quad \text{for all } i,\ h \le l < k \in ch(i),\ j \in I(T_{i,\epsilon,l+1,k}),$$

$$y_{i,j,h,l,k} \le \frac{1}{2}(x_{i,j,h,l} + x_{i,\epsilon,l+1,k}) \quad \text{for all } i,\ h \le l < k \in ch(i),\ j \in I(T_{i,\epsilon,h,l}),$$

$$z_{i,j,h,k,t} \le \frac{1}{2}(x_{i,t,h,k} + x_{t,j,lch(t),rch(t)}) \quad \text{for all } i,\ h \le k \in ch(i),\ j \in I(T_{i,\epsilon,h,k}),\ t \in anc(j),$$

$$s_u \le p_u < 1 + s_u \quad \text{for all } u \in U,$$

$$s_u = \frac{\sum_{\{(i,j,h,k) | es(T_{i,j,h,k}) = u\}} x_{i,j,h,k}}{|\{(i,j,h,k) | es(T_{i,j,h,k}) = u\}|}.$$

[1] 大連交通大学理学部応用数学科，116028 中国遼寧省大連市沙河口区黄河路 794 号
[2] 京都大学化学研究所バイオインフォマティクスセンター, 611-0011 京都府宇治市五ヶ庄
[3] 産業技術総合研究所創薬分子プロファイリング研究センター，135-0064 東京都江東区青海 2-4-7

Vol.2015-MPS-103 No.25
Vol.2015-BIO-42 No.25
2015/6/23

1.2 オイラー文字列を用いた圧縮

オイラー文字列 $es(T)$ は木 T を根から深さ優先探索順に辺を辿ったときの辺のラベルを並べた文字列であり，行きが a のとき戻りは $\bar{a}$ とする．図 1 の左図の木のオイラー文字列は $a\bar{a}cb\bar{b}c\bar{c}a\bar{a}c\bar{b}\bar{b}$ となる．辺にラベルの付いた 2 つの根付き順序木 T_1, T_2 について，$es(T_1) = es(T_2)$ であればそのときに限って T_1 と T_2 とは同型であることが知られている．

SEOTG の場合と同様に，根付き順序木 T が与えられたとき $es(T)$ を生成する CFG において非終端記号の数を最小化する問題は以下のように定式化できる．ただし規則の形としては A, B, C を非終端記号，a を終端記号として $A \to BC$, $A \to a$ のみとする．

最小化 $\sum_{u \in U} p_u$

制約条件

$$
\begin{aligned}
& x_{i,i} = 1 \quad \text{for all } i, \\
& x_{i,j} \le \sum_{k=i}^{j-1} y_{i,k,j} \quad \text{for all } i, j, \\
& y_{i,k,j} \le \frac{1}{2}(x_{i,k} + x_{k+1,j}) \quad \text{for all } i, j, k, \\
& s_u \le p_u < 1 + s_u \quad \text{for all } u \in U, \\
& s_u = \frac{1}{n} \sum_{\{(i,j) | es(T)[i,j] = u\}} x_{i,j}.
\end{aligned}
$$

2. 結果

n を自然数として，$3n + 1$ 要素の頂点集合 $V = \{v_0, \ldots, v_n, u_1, \ldots, u_n, w_1 \ldots, w_n\}$ と辺集合 $E = \{(v_i, u_{i+1}), (v_i, v_{i+1}), (v_i, w_{i+1}) | i = 0, \ldots, n-1\}$ からなる，v_0 を根とする木 $T(V, E)$ を考える．v_i の子は $u_{i+1}, v_{i+1}, w_{i+1}$ の順であり，L を定数として以下のように辺にラベルを付加する．

(1) $\lfloor \frac{i}{L} \rfloor \bmod 2 = 0$ なら (v_i, u_{i+1}) のラベルは a，そうでなければ b.

(2) (v_i, v_{i+1}) のラベルは c.

(3) $\lfloor \frac{i}{L-1} \rfloor \bmod 2 = 0$ なら (v_i, w_{i+1}) のラベルは a，そうでなければ b.

$n = L^2 = 4, 9, 16, 25$ について SEOTG，CFG それぞれに対して最小の非終端記号を整数計画問題を解くことで求めた結果を図 2 に示す．n が大きくなるごとにサイズの差が大きくなっていることがわかる．さらに $\sqrt{n}$, $\log(n)$ を当てはめた結果，SEOTG は $4.21\sqrt{n - 1.07}$，オイラー文字列を用いた圧縮は $4.70 \log(n + 1.56)$ で平均二乗誤差の平方根はそれぞれ 0.69, 0.48 であった．逆に $\log(n)$, $\sqrt{n}$ を当てはめたとき，SEOTG は $6.20 \log(n - 1.03)$，オイラー文字列は $2.89\sqrt{n + 4.54}$ であり，平均二乗誤差の平方根はそれぞれ 0.69, 0.91 であった．

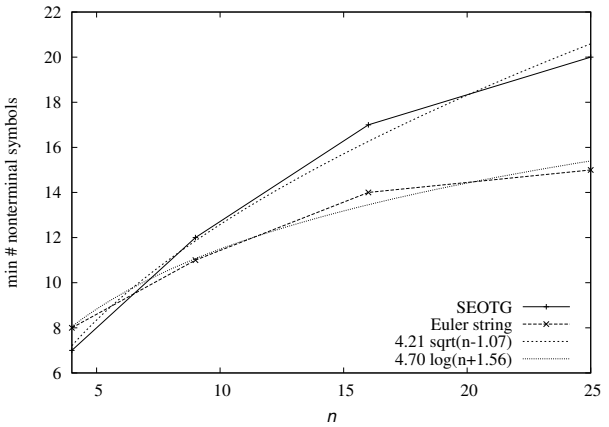

図 2 文法圧縮による SEOTG と CFG の最小の非終端記号数の結果.

$L = n + 1$ の場合は b にラベル付けされる辺は存在せず，SEOTG とオイラー文字列の両方で，二等分するような分割が最小文法として選択され，n の増大とともに差が増大するとは考えられない．

3. おわりに

根付き順序木の圧縮について，分割型文法 SEOTG とオイラー文字列を用いる手法を比較した．本発表で定義した木について，圧縮後のサイズが頂点数の増加とともに SEOTG とオイラー文字列とで差が増大することを $n = 25$ までで確認した．n がさらに増大する場合に差が増大することの証明は今後の課題である．

オイラー文字列の方が圧縮後のサイズは小さくなるが，木の構造に基づいているとは限らない．n が十分大きければ，最小の SEOTG では $(v_i, u_{i+1}), (v_i, v_{i+1}), (v_i, w_{i+1})$ の三辺からなる部分木を単位としてもとの木 T が構成される．一方オイラー文字列では，$es(T) = a\bar{a}ca\bar{a}c \cdots$ となるため $a\bar{a}c$ などを単位として構成される．このことからも木の構造を考慮した文法圧縮のために SEOTG は有用であることが示唆される．

参考文献

[1] Akutsu, T.: A Bisection Algorithm for Grammar-based Compression of Ordered Trees, *Information Processing Letters*, pp. 815–820 (2010).

[2] Zhao, Y., Hayashida, M., Cao, Y., Hwang, J. and Akutsu, T.: Grammar-based Compression Approach to Extraction of Common Rules Among Multiple Trees of Glycans and RNAs, *BMC Bioinformatics*, Vol. 16, p. 128 (2015).