Vol.2020-MPS-127 No.10
2020/3/2

一般化 Series-Parallel グラフの文法圧縮と時間計算量の実験的解析

小村亘平[1]　林田守広[1]　小谷野仁[2]　阿久津達也[3]

概要：一般化 Series-Parallel (GSP)グラフを生成する文法を定義し，この文法のもとで入力グラフのみを生成する最小文法を求める問題を整数線形計画問題として定式化する．GSP グラフとは，ソースとシンクというラベルの付いた 2 つの頂点を持ち，その 2 点で直列構成と並列構成が可能なグラフであり，木グラフ，外平面グラフを含む．本研究では頂点数と辺数が同じ場合について GSP グラフを文法圧縮したときの頂点数の増加に伴う時間計算量を計算機実験の結果から推定する．

1. はじめに

データ圧縮とは，データの実質的な内容を損なわず，総容量を小さくする技術のことである．文法圧縮はデータの共通部分を 1 つの生成規則にまとめ，集約することでデータの圧縮を実現する．圧縮データを用いて情報の抽出を行い，データをコンパクトで扱いやすい表現へ変換することが注目されている．中でも高分子化合物や細胞内の生物学的ネットワークの生成規則を求めることは，その構造及び機能を調べるうえで非常に有用な手段である．

Zhao ら[1]は化合物であるグリカンや RNA の分子構造を SEO(U)TG という文法[2]に基づいて圧縮した．この文法は木構造のみを対象としており，閉路を含む分子構造や生体分子ネットワークなどのグラフには適用できない．そこで本研究では SEO(U)TG を拡張し，閉路を含む一般化 Series-Parallel (GSP)グラフに対する GSPGG という文法と，それに基づく圧縮手法を提案する．GSP グラフとはソース，シンクというラベルの付いた 2 つの頂点を持ち，その 2 点で直列構成と並列構成が可能なグラフであり，木グラフ，外平面グラフを含む．本研究では GSPGG に基づく文法圧縮手法の正当性を検証し，時間計算量を計算機実験の結果から推定する．

2. 辺ラベル付き無向 GSP グラフに対する文法

圧縮の対象となる入力 GSP グラフはラベル付けされた無向の辺を持つとする．GSP グラフに対する文法 GSPGG は 4 タプル(Σ,Γ,S,Δ)で定義され，Σ は終端記号の集合，Γ は非終端記号の集合，S は Γ の開始の非終端記号，Δ は生成規則の集合を表す．終端記号はラベル付けされた無向の辺である．図 1 に GSPGG の生成規則のテンプレートを示す．図 1 において各矢印の先頭と末尾は，その辺の 2 つの端点を区別するために示されている．白と黒の四角は左側の白（黒）長方形の頂点が右側の白（黒）長方形の頂点に対応することを意味する．また，非終端記号から生成された終端記号のみからなるグラフが対称の場合，ソースとシンクの頂点は相互に交換することができる．図 1 の(j)から(m)の生成規則では，ソースとシンクの間の辺が 2 つの辺に置き換えられ閉路を生成する．これらの生成規則から，GSP グラフ，外平面グラフ，木グラフが生成できる．

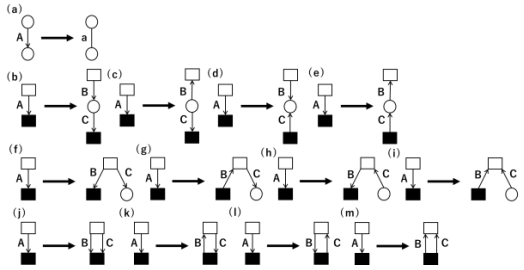


図 1　GSPGG の生成規則のテンプレート

3. 最小 GSPGG を求める整数線形計画問題

入力の GSP グラフ（G とする）の最小 GSPGG を求めることは最小の非終端記号の数になる GSPGG を求めることに等しい．つまり G を 1 つの辺のみが残るまで部分グラフに分割していき，G を構成するすべての生成規則の組み合わせを求め，その中から最小の非終端記号の数になる GSPGG を決定することで最小 GSPGG は求められる．G の部分グラフ $G_{i,S,j,T}$ を，切断点 i,j とそれぞれに隣接する頂点の集合 S,T によって表す．また，G の頂点集合を V，辺集合を E で表す．以下に特定の G を生成する最小 GSPGG を求める整数線形計画問題を示す．

$$\text{Minimize} \quad \sum_{(i,S,j,T)\in\mathcal{S}(G)} x_{i,S,j,T} \tag{1}$$

Subject to

$$x_{\epsilon,\emptyset,\epsilon,\emptyset} = 1, \tag{2}$$

$$\text{for all } (i,S,j,T) \in \mathcal{S}(G) \text{ s.t. } \left|E_{i,S,j,T}\right| = 1$$

$$x_{i,S,j,T} = 1, \tag{3}$$

$$\text{for all } (i,S,j,T) \in \mathcal{S}(G) \text{ s.t. } \left|E_{i,S,j,T}\right| \geq 2$$

$$x_{i,S,j,T} \leq \sum_{(i',S',j',T',i'',S'',j'',T'')\in\mathcal{C}(G_{i,S,j,T})} y_{i',S',j',T',i'',S'',j'',T''} , \tag{4}$$

1 松江工業高等専門学校
National Institute of Technology, Matsue College
2 農業・食品産業技術総合研究機構
National Agriculture and Food Research Organization
3 京都大学
Kyoto University

$$\text{for all}(i',S',j',T',i'',S'',j'',T'') \in \mathcal{C}(G_{i,S,j,T})$$

$$y_{i',S',j',T',i'',S'',j'',T''} \leq \frac{1}{2}\left(x_{i',S',j',T'} + x_{i'',S'',j'',T''}\right), \quad (5)$$

$$\text{for all } (i,S,j,T) \in U(G)$$

$$x_{i,S,j,T} = 0, \quad (6)$$

$$x_{i,S,j,T}, y_{i',S',j',T',i'',S'',j'',T''} \in \{0,1\} \quad (7)$$

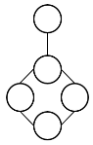
ここで$\mathcal{S}(G)$は G を分割して得られる，互いに同型でない部分グラフの添字(i,S,j,T)の集合を表し，$\mathcal{C}(G_{i,S,j,T})$は $G_{i,S,j,T}$ を構成する 2 つの部分グラフ $G_{i',S',j',T'}$と $G_{i'',S'',j'',T''}$の添字の集合を表し，$U(G)$は$\mathcal{S}(G)$の部分集合であり，それ以上分割できない，辺を 2 つ以上もつ部分グラフの添字の集合を表す．式(1)は目的関数である非終端記号の数を最小にすることを意味している．式(7)は各変数が 0 か 1 の値をとることを示す．$x_{i,S,j,T}$=1 になるのは対応する部分グラフ $G_{i,S,j,T}$がこの文法により生成されるとき，$y_{i',S',j',T',i'',S'',j'',T''}$=1 になるのは対応する部分グラフが 2 つの部分グラフ $G_{i',S',j',T'}$，$G_{i'',S'',j'',T''}$により生成されるときである．式(2)は入力グラフ G がこの文法によって構成されなければならないので，$x_{\varepsilon,\varepsilon,\varepsilon,\varepsilon}$=1 の制約を課す．式(3)は $G_{i,S,j,T}$ が 1 つの辺のみであるとき，$x_{i,S,j,T}$=1 の制約を課す．式(4)，(5)は $G_{i,S,j,T}$ を構成する生成規則がこの文法に属することを示す．式(6)では，$U(G)$に含まる部分グラフを構成する生成規則は存在しないため，$x_{i,S,j,T}$=0 の制約を課す．

4. 実験方法

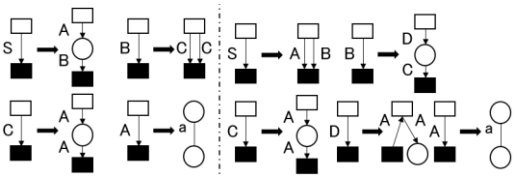
入力となる GSP グラフ G の頂点数を $n=|V|$，辺の数を $m=|E|$とする．n は $2<n<23$ の整数値で $m=n$ とした G を 100 個ランダムに生成し，文法圧縮にかかる平均時間（t [ms]）を，グラフ分割による整数線形計画問題の生成に要する時間（t_1 [ms]）と，この問題を解くのに要する時間（t_2 [ms]）の 2 つに分けて計測した．G は $n=2$, $m=1$ を初期状態として，指定した n と m を満たすように図 1 の生成規則をランダムに適用し生成した．整数線形計画問題の求解には，数理計画法ソルバーである IBM 社の CPLEX(Ver.12.9)を用いた．また計算機環境は OS に Linux，CPU に Xeon E5-2687W v4 を用いて行った．さらに，Excel 2016 のデータ分析機能の「回帰分析」を使用し，10 を底とする対数 $\log(t)$と頂点数 n から，最小二乗法に基づく近似式を求めた．

5. 実験結果

図 2 に $n=5$, $m=5$ の入力グラフ G の例とこのグラフに対する本手法の出力結果から得られる最小文法を示す．(a)の入力グラフを生成する最小文法として提案手法により(b)が得られた．(a)を生成する文法として(c)などが考えられるが，(b)と比べると(c)は非終端記号が一つ多くなっていることがわかる．



(a) $n=5,m=5$ の入力グラフ G



(b) 最小 GSPGG (c) その他の GSPGG

図 2 最小文法の例

図 3 に各 G の分割による整数線形計画の平均生成時間(t_1)と最小文法を求める計算にかかる平均時間(t_2)を示す．またそれぞれについて指数関数による近似の結果は式(8)，(9)となった．t_2 に比べ t_1 の指数は小さい値となることがわかった．

$$t_1 = 10^{0.13496n-1.5314} = O(1.364457^n) \quad (8)$$

$$t_2 = 10^{0.39544n-2.4505} = O(2.48565^n) \quad (9)$$

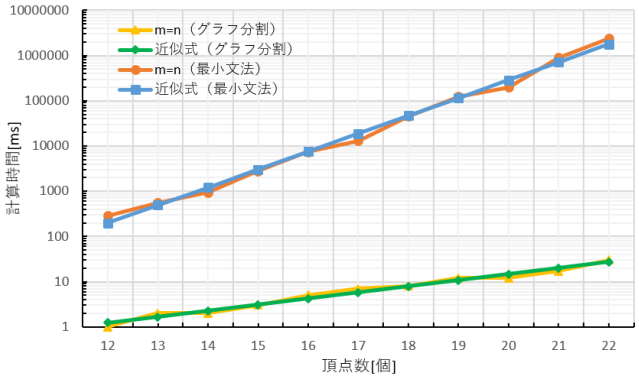


図 3 グラフ分割による整数線形計画問題の平均生成時間と最小文法の平均求解時間[ms]

6. おわりに

最小 GSPGG を求める整数線形計画問題への定式化を提案した．またグラフ分割による定式化に要する時間と最小文法を求める時間を，頂点数と辺数が等しい場合に実験的に解析した．今後の研究として，次数を制限した場合の整数線形計画問題への定式化に要する時間を推定したい．

参考文献

[1] 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*, 16, 128, 1-13, 2015.

[2] Akutsu, T. A Bisection Algorithm for Grammar-based Compression of Ordered Trees, *Information Processing Letters*, 110, 815-820, 2010.