Vol.2014-MPS-100 No.11
2014/9/25

タンパク質ドメイン構成に基づくプロテオーム圧縮

林田 守広[1,a)]　阮 佩穎[1,b)]　阿久津 達也[1,c)]

概要：生物は進化の過程において，突然変異や組み換えなどによって DNA の塩基配列情報を変化させながらも自らの生命を維持させてきた．生物の持つ情報を DNA の塩基配列とすると, この配列を圧縮することによって大体の情報量を知ることができる．本研究では塩基配列の代わりにタンパク質ドメイン構成に基づき，個体の持つすべてのタンパク質について圧縮する．遺伝子重複や遺伝子融合などの進化現象により同じドメイン構成を持つタンパク質が複数生成されるとすると，複製元のタンパク質を参照することでデータ量を減らすことができる．このような参照によるネットワークは有向ハイパーグラフとなり，多数の参照候補を持つグラフから最小大域木を見つけることで圧縮する．しかし現実的な時間での，ハイパーグラフからの最小大域木の抽出は困難であるので，前処理としてハイパーエッジを削減する発見的な手法を提案する．本手法を数種の生物種に適用した結果，タンパク質進化における遺伝子融合の重要性が示唆された．

1. はじめに

生体は自らの生命を維持していることからも一つの非平衡開放系とみなすことができる．孤立系においては不可逆過程であればエントロピーは増大する．生物は進化の過程において，突然変異や組み換えなどによって DNA の塩基配列情報を変化させながらも自らの生命を維持させてきた．生物の持つ情報を DNA の塩基配列とすると，この配列を圧縮することによって生物一個体の持つ大体の情報量を知ることができる．圧縮率が高ければ DNA 配列に繰り返しや冗長な部分が多く，配列長に比べて情報量は少ないと考えられ，逆に圧縮率が低ければ情報量は多いと考えられる．現在までに DNA 塩基配列やタンパク質アミノ酸配列を圧縮するための様々な手法が開発されてきた [1], [2]．多くは部分配列の繰り返しや頻度に基づいている．一方で，タンパク質はドメインと呼ばれる部分構造を持ち，同種のドメインが異なる種類のタンパク質に含まれている例も存在する．本研究ではこのタンパク質のドメイン構成を個体の持つすべてのタンパク質について圧縮することによって，生体の持つ情報量について考察する．

2. 提案手法

生体の持つタンパク質の集合を $\mathscr{P}$，ドメインの集合を $\mathscr{D}$ とする．各タンパク質 $P_i(\in \mathscr{P})$ に含まれるドメイン $D_m(\in \mathscr{D})$ の集合も P_i で表す．ここで同種のドメインが複数含まれていれば P_i は多重集合となる．本研究ではプロテオーム $\mathscr{P}$ を圧縮し，$\mathscr{P}$ を構成するための最小の文法を見つける．文法としては，以下の三種類の規則を考える [3].

[規則 1] タンパク質 P_i がドメインのみから構成される．
この規則に対するコストを以下のように定める．ドメインの番号が情報として必要であるので，P_i に含まれるドメインの数を $|P_i|$ として，$|P_i| \cdot \lceil \log |\mathscr{D}| \rceil$ のコストがかかるとする．

[規則 2] タンパク質 P_i がタンパク質 P_j からドメインの削除と新たなドメインの挿入によって構成される．
遺伝子重複と呼ばれる現象に対応し，進化的に P_j が複製されて P_i が形成されたと考える．P_j の指定と，P_j に含まれるドメインの取捨選択，また $|P_i - P_j|$ 個の新たなドメインの指定に $\lceil \log |\mathscr{P}| \rceil + |P_j| + |P_i - P_j| \cdot \lceil \log |\mathscr{D}| \rceil$ のコストがかかるとする．

[規則 3] タンパク質 P_i が二つのタンパク質 P_j, P_k から新たなドメインの挿入によって構成される．
遺伝子融合と呼ばれる現象に対応し，進化的に P_j と P_k が融合し複製されて P_i が形成されたと考える．ドメインの削除は可能な組み合わせの数が膨大になるため考慮しない．この場合に $2 \cdot \lceil \log |\mathscr{P}| \rceil + |P_i - P_j - P_k| \cdot \lceil \log |\mathscr{D}| \rceil$ のコストがかかるとする．

最小コストを持つ上のような文法を見つける問題は，辺に重みの付いた有向ハイパーグラフに対する最小大域木を見つける問題に変換できる．ここで，ハイパーエッジの持つ

[1] 京都大学
Kyoto University, Uji, Kyoto 611–0011, Japan
[a)] morihiro@kuicr.kyoto-u.ac.jp
[b)] ruan@kuicr.kyoto-u.ac.jp
[c)] takutsu@kuicr.kyoto-u.ac.jp

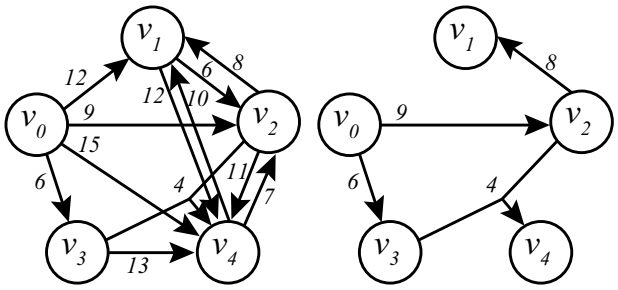

図 1 $\mathscr{P} = \{P_1, P_2, P_3, P_4\}$, $P_1 = \{D_1, D_1, D_2, D_3\}$, $P_2 = \{D_1, D_1, D_2\}$, $P_3 = \{D_4, D_5\}$, $P_4 = \{D_1, D_1, D_2, D_4, D_5\}$ に対するハイパーグラフ (左図) とその最小大域木 (右図)．頂点 v_i はタンパク質 P_i に対応する．

頂点の数が 2，つまり普通の辺だけの場合は多項式時間で最小大域木を見つけることができるが，ハイパーエッジの頂点の数が 3 以上の場合は NP 困難になることが知られている．例として，$\mathscr{P} = \{P_1, P_2, P_3, P_4\}$, $\mathscr{D} = \{D_1, D_2, D_3, D_4, D_5\}$, $P_1 = \{D_1, D_1, D_2, D_3\}$, $P_2 = \{D_1, D_1, D_2\}$, $P_3 = \{D_4, D_5\}$, $P_4 = \{D_1, D_1, D_2, D_4, D_5\}$ に対しては，図 1 左が，可能な規則を有向辺とし，その辺の重みをその規則のコストとして，プロテオーム $\mathscr{P}$ をハイパーグラフに変換したものである．このハイパーグラフから得られる最小大域木が右図となる．v_0 はドメインをもたない仮想タンパク質を意味し，P_2, P_3 はそれぞれ規則 1 によりドメインから構成され，P_1 は P_2 から規則 2 の遺伝子重複により生成され，P_4 は P_2 と P_3 から規則 3 の遺伝子融合により生成される．

本研究では現実的な時間で最適解を見つけることが困難な本問題に対して，規則 3 のハイパーエッジを除いて最適解を見つけた後，同じ頂点に入る規則 3 のハイパーエッジのうち，重みが最適解の辺の重みよりも小さいものだけを再び加えて最適解を求める，発見的な手法を提案する [3]．

3. 結果

UniProt データベース [4] から 14 の生物種，*D. discoideum*, *E. coli*, *S. cerevisiae*, *S. pombe*, *C. elegans*, *D. melanogaster*, *A. thaliana*, *O. sativa*, *D. rerio*, *X. laevis*, *G. gallus*, *M. musculus*, *P. troglodytes*, *H. sapiens* について，タンパク質ドメイン構成の情報を取得し，提案手法を適用した．図 2 は Pfam ドメイン [5] を使った場合の各生物種の圧縮率を示す．規則 1, 2 のみを使った場合の圧縮サイズは常に，圧縮前のサイズよりも小さく，すべての規則を使った場合の圧縮サイズよりも僅かではあるが大きかった．また元のサイズからの圧縮率の比較からは，*M. musculus* と *H. sapiens* が他の生物種に比べて圧縮率が高く，同じドメインが高等な生物種ほど頻繁に活用されていることが示唆された．さらに抽出された規則 3 の文法からは，一度他の二つのタンパク質から遺伝子融合によって形成されたタンパク質が，他のタンパク質の遺伝子融合の材料になっている例がいくつか発見された．

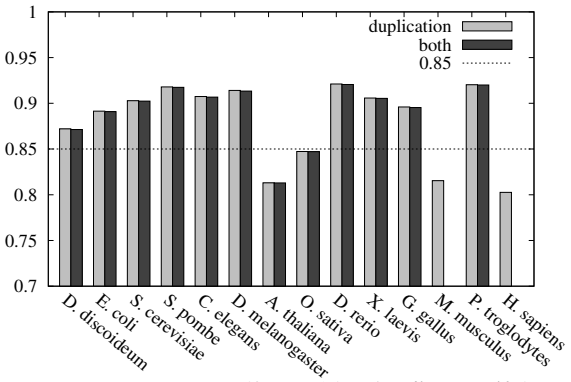

図 2 Pfam ドメインを使った場合の各生物種の圧縮率．

4. おわりに

本研究では，タンパク質ドメイン構成に基づくプロテオーム圧縮のための発見的な手法を提案し，実際に 14 の生物種に対して適用した．これまでに DNA 塩基配列やタンパク質アミノ酸配列に対する圧縮は研究されてきたが，ドメイン構成に基づく圧縮では最初の研究である．また生物の進化過程でみられる，遺伝子重複，遺伝子融合という現象に基づいて文法を構成した．圧縮率では，*M. musculus* と *H. sapiens* が他の生物種に比べて圧縮率が高く，高等な生物種ほど同じドメインが頻繁に活用されていることが示唆された．しかしながら，生物種間の比較のためには，より自由度の高い文法に対する最適化アルゴリズムの開発が求められる．さらに現実的な時間で解を得るために効率的なアルゴリズムの開発が求められる．

参考文献

[1] Grumbach, S. and Tahi, F.: A New Challenge for Compression Algorithms: Genetic Sequences, *Information Processing & Management*, pp. 875–886 (1994).
[2] Cao, M., Dix, T., Allison, L. and Mears, C.: A simple statistical algorithm for biological sequence compression, *Proc. Data Compression Conference (DCC '07)*, pp. 43–52 (2007).
[3] Hayashida, M., Ruan, P. and Akutsu, T.: Proteome compression via protein domain compositions, *Methods*, Vol. 67, pp. 380–385 (2014).
[4] The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, Vol. 40, pp. D71–D75 (2012).
[5] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A. and Finn, R. D.: The Pfam protein families database, *Nucleic Acids Research*, Vol. 40, pp. D290–D301 (2012).