タンパク質ドメイン構成に基づくプロテオーム圧縮

Proteome compression via protein domain compositions

化学研究所バイオインフォマティクスセンター数理生物情報研究領域　林田守広

背景と目的

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．生物一個体の持つ情報量はどのくらいだろうか？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．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．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．

検討内容

生体の持つタンパク質の集合をP，ドメインの集合をDとする．各タンパク質P_i $(\in P)$に含まれるドメインD_m $(\in D)$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Pを圧縮し，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3つの規則を考える．

1. タンパク質P_iがドメインのみから構成される場合．
   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，$|P_i|\lceil\log|D|\rceil$のコストがかかるとする．
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場合．
   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に$\lceil\log|P|\rceil + \left|P_j\right| + |P_i - P_j|\lceil\log|D|\rceil$のコストがかかるとする．
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場合．
   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$2\lceil\log|P|\rceil + |P_i - P_j - P_k|\lceil\log|D|\rceil$のコストがかかるとする．

最小コストをも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．ハイパーエッジの持つ頂点の数が 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 以上では NP 困難になることが知られている．本研究では，最適解を見つけることが困難であるので，上の 3 つ目の規則を除いて最適解を見つけた後，除いた規則のうちある基準を満たすものだけを再び加えて最適解を求める，ヒューリスティックな手法を提案する（詳細は文献[1]）．

結果

タンパク質データベース UniProt 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．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．

考察

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ヒューリスティックな手法を提案し，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．

参考論文

1. M. Hayashida, P. Ruan, and T. Akutsu, Proteome compression via protein domain compositions, *Proc. IEEE International Conference on Systems Biology (ISB2013)*, 2013.