最大重みクリーク抽出アルゴリズムのRNAの二次構造予測への適用

3U−9

若月 光夫　田中 健夫　富田 悦次

電気通信大学 電子情報学科

1　まえがき

RNAの二次構造予測問題は，その機能解明等のためゲノム情報解析の一つとして重要であり，現在までに様々な研究報告がなされてきている．この予測法の一つとして，Zuker and Stiegler[1]を代表とする，塩基対単位による動的計画法を用いた多項式時間アルゴリズムが広く利用されている．しかし，この手法では，例えばHIV-2のgag-pol領域(図2)中に見られるようなシュードノット構造を扱うことはできず，予測結果と実構造とが必ずしも一致しないことがあり得る．これに対し，Pipas and McMahon [2]は，スタック領域と呼ばれる，連続する塩基対結合によって形成される梯状構造を単位として候補の抽出を行い，この可能な組合せの中でエネルギーが最小となるものを全探索で求める手法を提案した．この方法で最適解を得るにはスタック領域候補数の指数オーダの時間を要するため，秋山と古谷[3]は，ニューラルネットワークを用いた多項式時間の近似解法を提案した．

本研究では，文献[2],[3]の考え方に基づき，RNAの二次構造予測問題を，グラフ理論における組合せ最適化問題の一つである最大重みクリーク問題として形式化し，この分枝限定アルゴリズムを適用した結果等を示す．

2　RNAの二次構造と最大重みクリーク

RNAの塩基は主として，アデニン(A)，シトシン(C)，グアニン(G)，ウラシル(U)であり，A-U間およびC-G間(まれに，G-U間)に水素結合ができることで，一本鎖RNAは折り畳まれ二次構造が形成される．

与えられたRNA配列から重み付き無向グラフを生成する手順は，次の通りである．まず，RNA配列に対し，スタック領域の候補を全て抽出し，これをグラフの節点にそれぞれ対応させる．また，スタック領域候補のエネルギー(負値)を符号反転した数値を，対応する各節点の重みとして与える．最後に，2つのスタック領域候補が両立する場合にのみ，それに対応する節点間に枝を張る．

こうして得られたグラフのある部分グラフにおいて，節点間全てに枝が張られているとき，これをクリークと呼ぶ．また，このようなクリークのうち，それに含まれる節点の重みの総和が最大のものを最大重みクリークと呼ぶ．エネルギー最小となるスタック領域候補の組合せは，上記手順で生成されたグラフにおける最大重みクリークに対応している．本研究では，文献[4]で提案した，分枝限定アルゴリズムMWCLiqueによって最大重みクリークの抽出を行い，この厳密解を所与のRNAに対する二次構造予測の結果とする．

3　重み付き無向グラフの生成

ここで，RNA配列から生成されるグラフの節点数および枝存在率について，理論的解析を行なう．簡単のため，RNA配列中の各塩基が，その位置に関係なく塩基の種類に特定のある確率によってランダムに生起するものと仮定する．塩基A, C, G, Uの生起確率をそれぞれ，p_A, p_C, p_G, p_Uで表す．A-U間およびC-G間のみ結合があるものと仮定すると，一対の塩基が水素結合を形成する確率p_Bは$p_B = 2(p_A p_U + p_C p_G)$で与えられる．RNAの配列長を$N$，ヘアピンループ，およびシュードノット構造中で制約となる一本鎖部分の最短の長さをそれぞれm, nとして，長さK以上のスタック領域候補を全て抽出するものとする．このとき，生成されるグラフの節点数$|V|$の期待値$E(|V|)$は，

$$E(|V|) = \frac{1}{2} \sum_{k=K}^{\lfloor (N-m)/2 \rfloor} p_B^k \prod_{l=1}^{2} (N - m - 2k + l)$$

で与えられる．また，このときのグラフの枝数$|E|$の期待値$E(|E|)$は，

$$E(|E|) = \frac{1}{24} \sum_{k_1=K}^{\lfloor N/2 \rfloor - K - m} p_B^{k_1} \sum_{k_2=K}^{\lfloor N/2 \rfloor - m - k_1} p_B^{k_2} \prod_{l=1}^{4} (N - 2m - 2k_1 - 2k_2 + l) + \frac{1}{24} \sum_{k_1=K}^{\lfloor (N-m)/2 \rfloor - K} p_B^{k_1} \sum_{k_2=K}^{\lfloor (N-m)/2 \rfloor - k_1} p_B^{k_2}$$

An Application of an Algorithm for Finding a Maximum Weight Clique to RNA Secondary Structure Prediction

Mitsuo Wakatsuki　Takeo Tanaka　Etsuji Tomita

The University of Electro-Communications

1-5-1 Chofugaoka, Chofu, Tokyo 182, Japan

$$\prod_{l=1}^{4}(N-m-2k_1-2k_2+l)$$
$$+\ \frac{1}{24}\sum_{k_1=K}^{\lfloor N/2\rfloor-K-n} p_B^{k_1}\sum_{k_2=K}^{\lfloor N/2\rfloor-n-k_1} p_B^{k_2}$$
$$\prod_{l=1}^{4}(N-2n-2k_1-2k_2+l)$$

となる．従って，枝存在率 ξ の期待値 $E(\xi)$ は，

$$E(\xi) = E(|E|)/\left\{\frac{1}{2}E(|V|)^2-\frac{1}{4}\sum_{k=K}^{\lfloor (N-m)/2\rfloor} p_B^{2k}\prod_{l=1}^{2}(N-m-2k+l)\right\}$$

で与えられる．

4 最大重みクリーク抽出アルゴリズム

本研究で使用の最大重みクリーク抽出アルゴリズム MWCLique の性能評価を，計算機実験により行った (アルゴリズムの詳細については，文献 [4] 参照)．節点数 100 ～ 1000，枝存在確率 (p) 0.1, 0.3, 0.5, 0.7 の各組合せに対して 20 個のランダムグラフを生成し，節点重みは 1 ～ 10 の整数値を一様乱数により割当て，実験対象のグラフを作成した．これらのグラフについて，基本アルゴリズム MaxWeightClique およびアルゴリズム MWCLique を実働化し，平均実行時間 (単位：ミリ秒) を測定した結果を図 1 に示す．なお，使用計算機は SUN SPARCstation 10(model 40) である．図 1 から，アルゴリズム MWCLique は基本アルゴリズムに比べ，枝存在率の高いグラフに対して優位性を発揮することが確認できる．

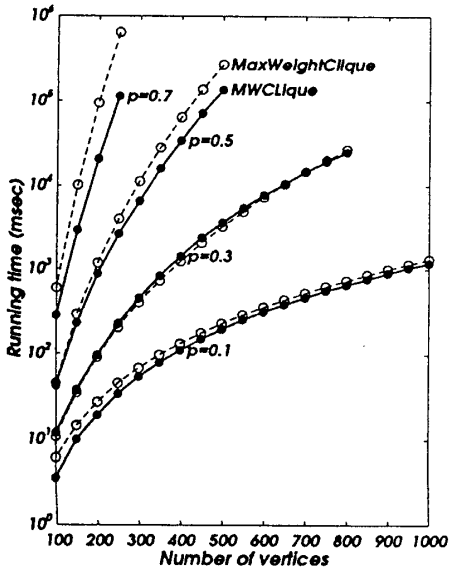

図 1: 最大重みクリーク抽出の平均実行時間

5 実行例

実際の RNA 配列として，HIV-2 の gag-pol 領域 (GenBank の HIV2ROD 中，配列番号 1829-1910 の部分配列) を与え，$K = 5$ として本手法を適用したところ，節点数 5，枝数 5 のグラフが生成され，図 2 のような結果を得た．ここで，図中の点線の矩形で囲まれた部分が抽出されたスタック領域である．

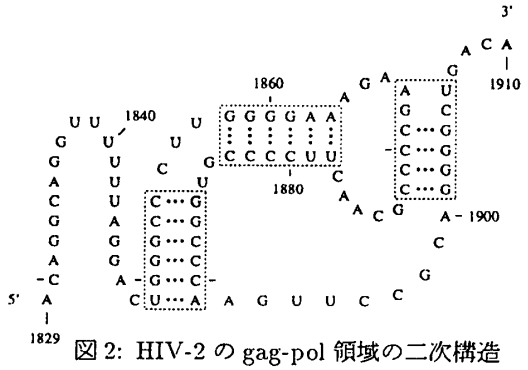

図 2: HIV-2 の gag-pol 領域の二次構造

6 むすび

RNA の二次構造予測問題を最大重みクリーク抽出問題として形式化し，この分枝限定アルゴリズムを適用した手法について検討した．

RNA 配列が長い場合やスタック領域候補の長さの設定を短くした場合には，グラフの節点数が大きくかつ枝存在率が 1 に漸近するため，最大重みクリークの厳密解の抽出は困難となる．そこで，今後は，比較的短時間で妥当な精度の解が得られるような最大重みクリークの近似アルゴリズムを開発し，RNA の二次構造予測に適用する予定である．

謝辞 最大重みクリーク抽出アルゴリズムの実働化に御協力頂いた本学卒業生の今松 憲一氏 (現，エスコム) に感謝致します．

参考文献

[1] Zuker, M. and Stiegler, P. : "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information", *Nucl. Acids Res.*, 9, pp.133-148 (1981).

[2] Pipas, J. M. and McMahon, J. E. : "Method for predicting RNA secondary structure", *Proc. Natl. Acad. Sci. USA*, 72, pp.2017-2021 (1975).

[3] 秋山 泰, 古谷 立美 : "対称相互結合型ニューラルネットワークを 用いた大規模な RNA の二次構造予測", 信学技報 NC90-62, pp.57-64 (1991).

[4] 今松 憲一, 富田 悦次, 若月 光夫 : "近似彩色を用いた単純な最大重みクリーク抽出アルゴリズム", 電通大紀要, 8, 1, pp.17-21 (1995).