NeuroFold: RNA二次構造予測システム

NeuroFold: an RNA secondary structure prediction system using a Hopfield neural network

秋山 泰, 金久 實
Yutaka Akiyama, Minoru Kanehisa
京都大学化学研究所
Institute for Chemical Research, Kyoto University

E-mail: akiyama@kuicr.kyoto-u.ac.jp

We have developed a quick algorithm for RNA secondary structure prediction which employs a combinatorial optimization technique based on the Hopfield neural network theory[1]. The detail of our algorithm was presented[2] at the last Genome Informatics Workshop in 1991. Briefly, the algorithm consists of four independent stages. At the first stage all possible stacking region candidates are listed via a string matching process along a given RNA sequence. (Several different search strategies can be designated by options.) Then the second module executes selective deletions of redundant candidates. And the third, a Hopfield neural network module performs the central task of constructing the most feasible combination set from proposed candidates. The technique of dynamic modification of neural I/O function is used in order to escape from local minima problem. The last stage is to build an output display from the selected set of stack region candidates.

In this workstation session, we would like to demonstrate performance of the "NeuroFold" system which is an X11 window-based implementation of the proposed algorithm. The algorithm itself has also been enhanced in many aspects: for example, now pseudoknots can be effectively predicted in the NeuroFold system.

[概要]

我々はHopfield型ニューラルネットにおけるエネルギー極小化の性質[1]を利用した高速なRNA二次構造予測法を提案した。そのアルゴリズムの詳細については、昨年度の公開ワークショップにおいて発表[2]している。本ワークステーションセッションにおいては、昨年度に提案した予測法にもとづいて現在作成中であるX11ウィンドウベースのRNA二次構造予測システム "NeuroFold"について、デモンストレーションを通じて紹介する。昨年度の発表の時点と比べてアルゴリズム自身も種々の点で改良されており、例えばpseudoknot の予測やG-U間水素結合を適切に扱えるようになった。

[NeuroFoldシステムの構成と各部の動作]

Fig. 1にNeuroFoldシステムの構成を示す。NeuroFoldシステムにおける二次構造予測の処理は大きく４つの実行ステージに分けられる。

はじめに、入力された一次構造をもとにMatchモジュールがスタック領域（互いに相補的な塩基配列が存在するとき、対応する塩基間が水素結合で結ばれて生ずる構造）の候補を抽出する。このときオプションの指定により、どれほど短いスタック領域候補まで抽出するかを指定できる。（配列が長大な場合には探索を粗くして、例えば長さ５以上の候補に絞ることにより以降の処理が高速化できる）またG-U間水素結合を許すか否か、配列上である距離以上離れたスタック領域を禁止するか、等もオプションとして指定可能としている。これらはスタック領域候補数を減らして計算時間を短縮したい場合に有効である。

Matchモジュールから出力される候補には冗長性がある。たとえばある場所に長さ６のスタック領域候補が見つかった場合には、その部分構造である長さ５のスタック２個、長さ４のスタック３個...、等も同時に出力されている。そこで次にRedundモジュールによって、候補のうち冗長性の高いものはその一方を

消去する。具体的には、２つの候補が存在する時、一方が他方に位置的に包含され、しかも他者との矛盾関係に差異がない場合は短い方の候補を消去している（包含されていても、他者との矛盾関係が軽減されるような場合は短い候補を消去してはいけない）。

このようにして選択されたスタック領域の候補はNeuroSelectモジュールに受け渡される。NeuroSelectモジュールはまず、全ての候補対の間の位置関係を調べて衝突行列C_{ij}を作成する。２つの候補の間の関係はFig. 2に掲げる４種類に分類される。１つの塩基が複数の水素結合に使われるような場合は矛盾となり$C_{ij}=1$、それ以外は$C_{ij}=0$とする。受け渡されたN個の候補から、できる限り安定度が高くなるように候補の組み合わせを選出する過程はHopfield型ニューラルネットのダイナミクスを用いて実行される。受け渡された各候補ごとに１つのニューロンユニットを割り当て、計算終了時における各ユニットの活性／不活性により各候補の選択／不選択を表現する。このときニューラルネットの動作を規定するエネルギー関数[1]として次式を設計した[2]。この式を最小化する二値変数x_iの組が最適な候補選択を表現している。

$$E=\sum_{i=1}^{n} e_i x_i + \lambda \cdot \frac{\max(|e_i|)}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_i x_j \qquad e_i \ = \ (\text{候補 } i \text{ の表すスタック領域のエネルギー})$$

Hopfield型ニューラルネットはニューロン集団内での並列的な競合ダイナミクスにより自然にこのエネルギー関数の最小値（準最小値）を与えるような状態へと収束する性質があり[1]、収束後の状態を読み出せば無矛盾かつエネルギー的に安定となるようなスタック領域候補の組を得ることが可能となる。

各ニューロンの入出力関数の勾配は計算中に次第に急峻化し、各ニューロンが平均して１００回ずつ状態遷移をした時点(100 iterations)で、各ニューロンの出力値が二値をとるようにした。計算の進行に伴う各ニューロンユニットの活性値の変化の例をFig. 3に示す。序盤から中盤にかけては目まぐるしくパターンは変化するが、やがていくつかの有力候補間の争いとなり最終的には矛盾のない答に収束している。

Fig. 4に、NeuroFoldシステムで抽出されたpseudoknotsの例を示す。(a)はTYMV (Turnip Yellow Mosaic Virus)の3'端付近の二次構造である。矢印で示された相補鎖がpseudoknotを形成することにより全体がL字型にねじれtRNA様の立体構造をとると考えられている。(b)はHIV2 (Human Immunodeficiency Virus 2) のgag-pol領域を入力して得た二次構造予測である。この領域付近では翻訳時のフレームシフトが起きる事が知られているが、ヘアピンの頭にフレームシフトの原因と目されるpseudoknotが形成される事がわかった。

[議論]

NeuroFoldシステムではヘアピン以外のループ類の安定度評価を行っていない。このため例えばFig. 5に示すように本来の正解(a)と同程度の比率でループ部の安定度がやや劣る解(b)が出てきてしまう事がある。ループ類の正しい評価のためには大別して２つの方法がある。第一の方法は、Match処理においてバルジループや内部ループが途中に挟まっても一連の長いスタック領域として候補を生成することである。この場合には部分候補の数をどう抑えるかが課題となる。第二の方法は、NeuroSelect処理において、ループが形成される場合を動的に察知して評価値を修正することである。いずれの場合でも、Redund処理において現在のような大胆な削除戦略をとれなくなるため、処理時間が長くなるという問題がある。NeuroFoldにおける計算時間はスタック候補数に依存する[3]ので候補削減のための別の工夫をさらに進める必要がある。

なお本手法とZuker法[4]の計算量および時間性能比較については文献3)を参照されたい。

本研究は文部省科学研究費重点領域研究「ゲノム情報」の助成を受けています。また計算機実験の実行に際して、京都大学化学研究所スーパーコンピュータラボラトリより計算機のCPU timeの提供を受けました。

[参考文献]

1) Hopfield and Tank: *Biol. Cybern.*, 52, 131-152 (1985).

2) Akiyama and Furuya: *Proc. of the 2nd Genome Informatics Workshop*, 124-127, (1991). [in Japanese]

3) Akiyama and Furuya: *Proc. of 1992 Informatics Symp.*, Info. Proc. Soc. Japan, 125-134, (1992). [in Japanese]

4) Zuker and Stiegler: *Nucl. Acids. Res.*, 9, 1, 133-148 (1981).

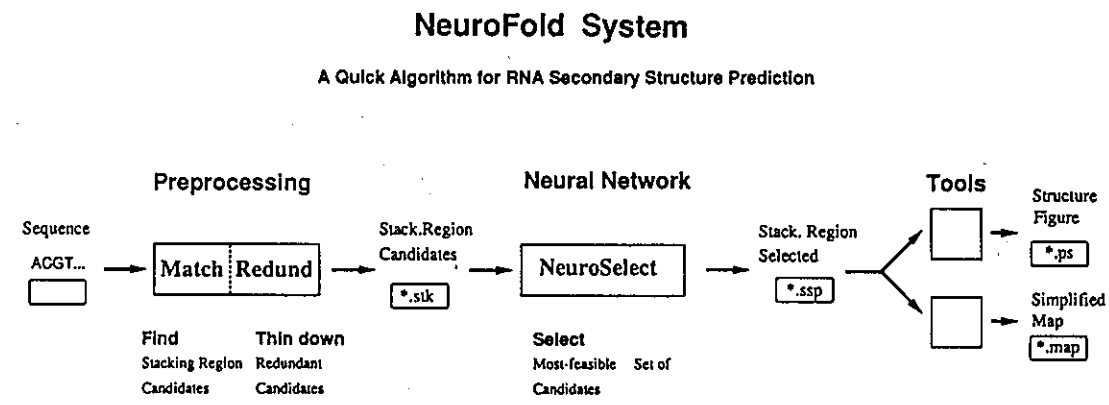


Fig.1 Schematic diagram of the NeuroFold system

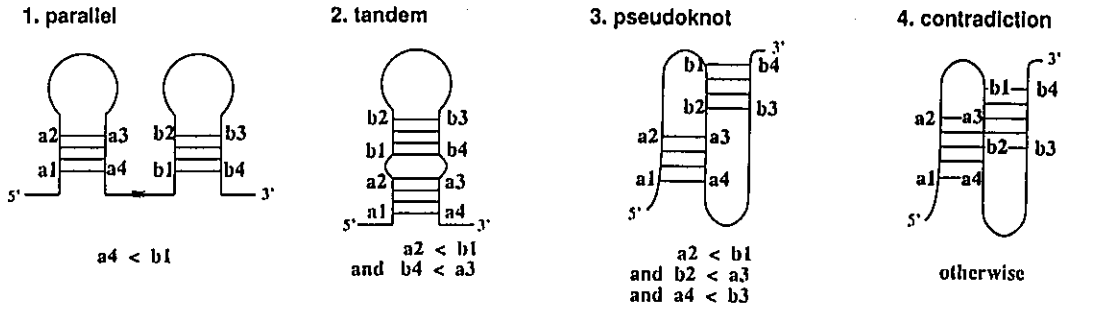


Fig.2 Relations between two stack region candidates

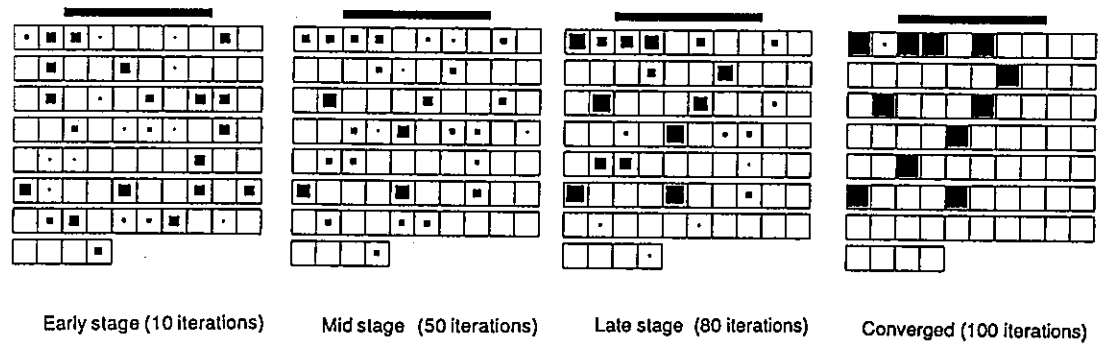
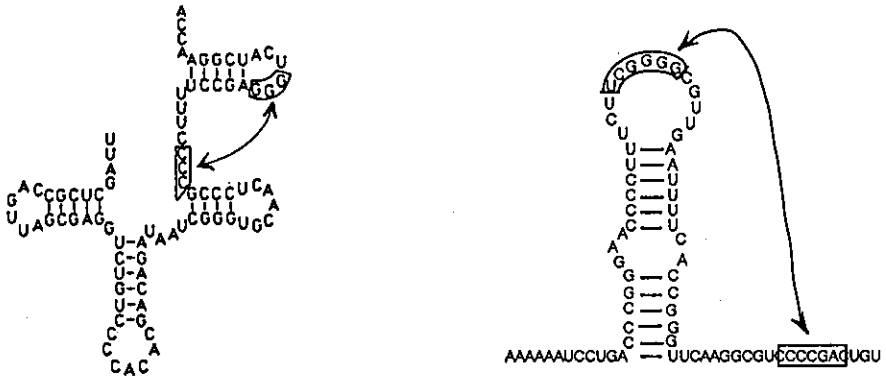
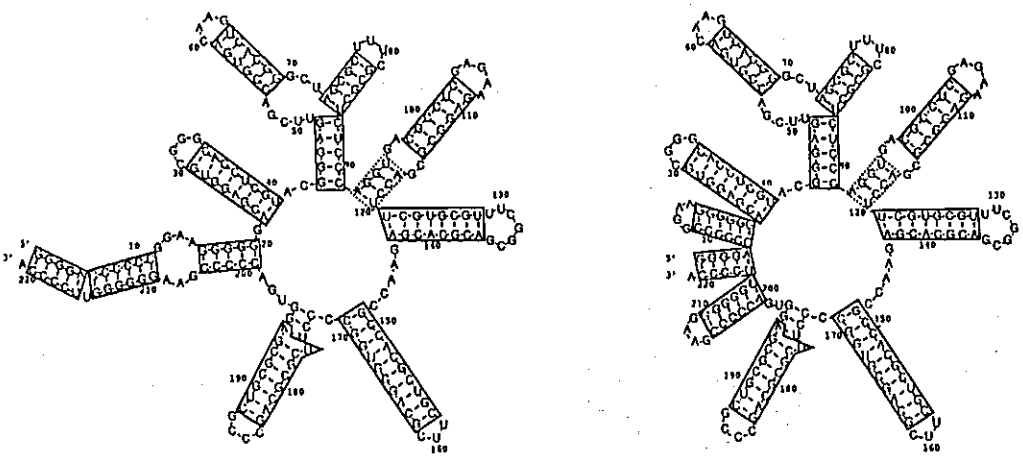


Fig.3 An example state transitions of the neural network
(Item= MDV-1(+) 221bp, 74 stack region candidates)



(a) TYMV 3'-end, 84bp (b) HIV2 gag-pol region, 109bp

Fig.4 Example pseudoknots found by the NeuroFold system



(a) a desirable prediction (b) a solution with less stable loops

Fig.5 A false prediction case caused by inaccurate evaluation of loop stabilities

Item= MDV-1(+) 221bp