Capturing Alternative Secondary Structures of RNA by Decomposition of Base Pairing Probabilities(塩基対確率行列の分離による RNA の複数安定二次構造予測)

メディカル情報生命専攻　ゲノム情報解析分野 浅井研究室
指導教員：浅井　潔　教授
修了年月：2017 年 3 月 学籍番号：47-156437 氏名：萩尾　太一
キーワード: RNA 二次構造予測

1.序論

近年の研究の進展により、RNA の二次構造が RNA の機能を解明するための重要な手掛かりとなる例が数多く示されてきた。しかし、実験による RNA の二次構造の決定を、転写されうる全ての RNA について網羅的に実施する事は困難である。このため、計算機を用いた RNA 二次構造予測が重要である。既存の RNA 二次構造予測ソフトウェアである CentroidFold や mFold を用いれば、それぞれのソフトウェアが最適とする二次構造を予測することができる。しかし、RNA にはリボスイッチや RNA サーモメーターのように複数の安定構造を持つ RNA が存在する。既存のソフトウェアでは、ただ一つの構造しか予測することができず、RNA の構造変化を捉えることができないという問題が存在する。本研究では、この問題点を解決する新規手法を提案する。

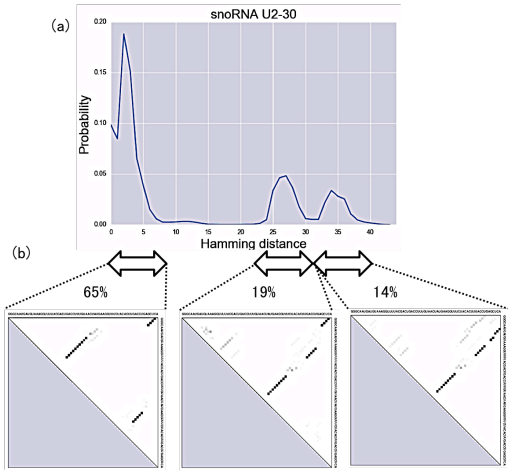

Figure 1.塩基対確率行列の分離

2.方法

森らや E. Freyhult らは、任意の複数の参照二次構造からのハミング距離に対応する二次構造の存在確率分布を計算する手法を示している。この手法を用いると、存在確率の高い複数の構造集合を確認できる(Figure 1(a))。本研究では、確認できた構造集合は、クラスタ(安定構造の類似構造集合)であるという仮説に基づいて、ハミング距離で定義された構造集合の塩基対確率行列とγ −centroid 構造を予測できるアルゴリズムを構築した(Figure 1(b))。森らや E. Freyhult らは、多項式拡張することによって、RNA 配列$r = r_1..r_k..r_l..r_N$について、参照構造S_Rからハミング距離h離れた構造集合の分配関数がx^hの係数になるような多項式${Z^b}_{kl}[x]$、$Z_{1N}[x]$、${Z^m}_{kl}[x]$を動的計画法で計算するアルゴリズムを提案した。同様に、ハミング距離h離れた構造の塩基kとlが結合する確率がx^hの係数になるような多項式${P^b}_{kl}[x]$は、塩基対確率の定義より、${Z^b}_{kl}[x]$、$Z_{1N}[x]$、$W^b_{kl}[x]$を用いて計算することができる。本研究では、新たに$W^b_{kl}[x]$を計算する再帰式を導出した。以下にその再帰式を示す。

Initialization: $Z_{1,0}[x] = 1.0, Z_{N+1,N}[x] = 1.0$

Recursion($1 \le k < l \le N$):

if ($(r_k, r_l) \in \mathcal{B}$): ${W^b}_{kl}[x] = Z_{1,k-1}[x] Z_{l+1,N}[x] \cdot x^{g_1(i,k,j,l,S_R)}$

else: ${W^b}_{kl}[x] = 0.0$

Recursion($1 \le i < k < l < j \le N$): ($d = j - i$としたとき、$d = N - 1$から降順に計算)

if ($(r_k, r_l) \in \mathcal{B}$):

$$
\begin{aligned}
{W^b}_{kl}[x] \mathrel{+}= & \sum_{\substack{i,j \\ i<k<l<j}} {W^b}_{ij}[x] \cdot e^{-[F_1(i,k,j,l)/KT)]} \cdot x^{g_2(i,k,j,l,S_R)} \\
& + \sum_{\substack{i,j \\ i<k<l<j}} {W^b}_{ij}[x] (e^{-[F_2(i,k)/KT)]} \cdot {Z^m}_{l+1,j-1}[x] \cdot x^{g_3(i,k,j,l,S_R)} + {Z^m}_{i+1,k-1}[x] \cdot e^{-[F_2(j,l)/KT)]} \cdot x^{g_4(i,k,j,l,S_R)} \\
& + {Z^m}_{i-1,k-1}[x] \cdot {Z^m}_{l+1,j-1}[x] \cdot x^{g_5(i,k,j,l,S_R)})
\end{aligned}
$$

$\mathcal{B} = \{(A,U),(U,A),(G,C),(C,G),(G,U),(U,G)\}$、$g_k(\cdot)(k = 1,2,3,4,5)$はハミング距離の増分である。この再帰式を動的計画法で計算する。さらに、多項式乗算を離散フーリエ変換で高速化することで、時間計算量$O(N^4 H_{max})$で${P^b}_{kl}[x]$を計算できる。H_{max}は参照構造S_Rと配列rが取りうる構造とのハミング距離の最大値を示す。

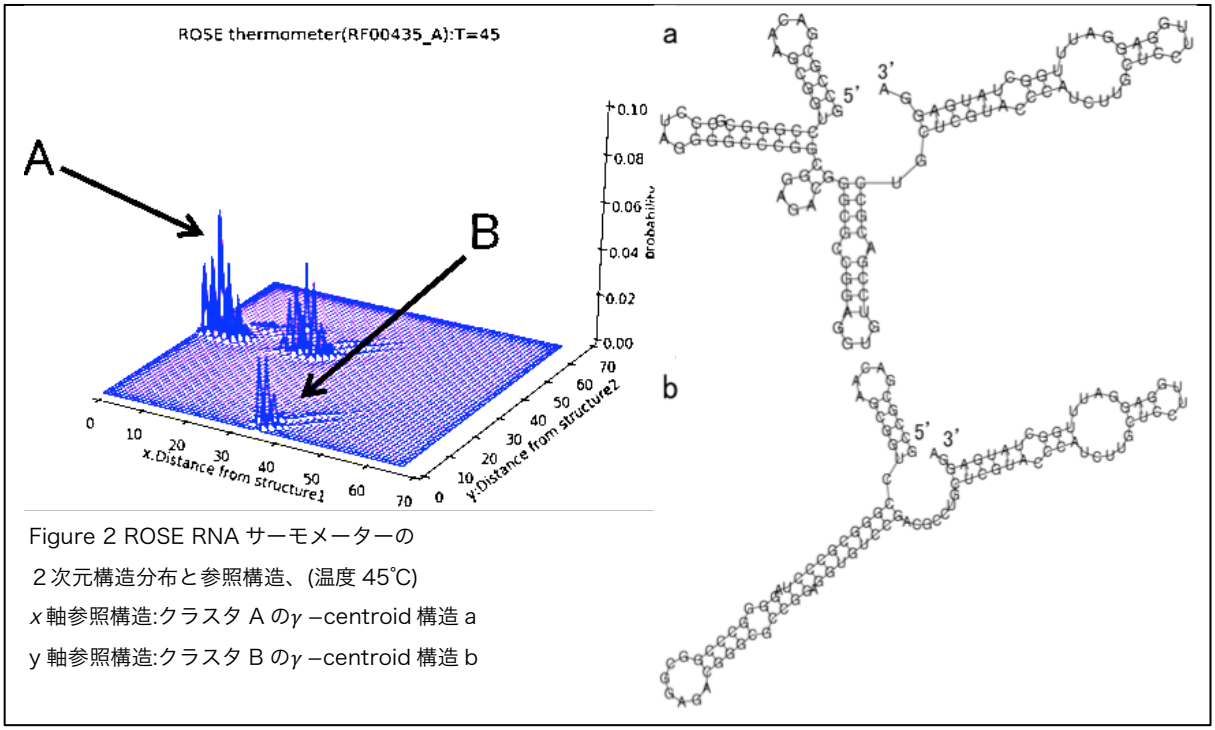

Figure 2 ROSE RNA サーモメーターの
２次元構造分布と参照構造、(温度 45℃)
x 軸参照構造:クラスタ A のγ –centroid 構造 a
y 軸参照構造:クラスタ B のγ –centroid 構造 b

3.結果と考察

本研究で提案した手法を用いて、ROSE RNA サーモメーターの複数の安定構造を予測できた。予測した構造が生物学的な知見と矛盾しないことを考察する。ROSE RNA サーモメーターは、mRNA の 5'UTR に位置し、温度変化によって構造を変化することで、下流の遺伝子の翻訳を制御している。また、約 30℃で翻訳抑制、約 45℃で翻訳促進することが知られている。37℃の 1 次元構造分布の結果から、3 つのクラスタ A,B,C を取り出すことができた。ここで、温度 30℃,45℃で、クラスタ A,B のγ –centroid 構造を参照構造とし、ハミング距離を座標とした 2 次元構造分布を計算した。Figure2、Figure 3 より温度を 30℃から 45℃に上昇させると、クラスタ A の確率が減少していることがわかる。対照的にクラスタ B の確率が増加していることがわかる。このことからクラスタ A が翻訳抑制構造、クラスタ B は翻訳促進構造と考えると、生物学的な知見と矛盾しない。構造的に、ROSE RNA サーモメーターは、5'末端に最も近いヘアピンループは、翻訳抑制構造、翻訳促進構造に関わらず、安定であることが知られている。さらに、3'末端以外のヘアピンループは温度が上昇すると、Zipper-like に分解されることが知られている。提案手法により構造 a から構造 b に変化するためには、切断しなければいけない塩基対が存在することがわかった(Figure2,a,b)。このことからも、生物学的知見と予測した構造に矛盾がないことがわかる。

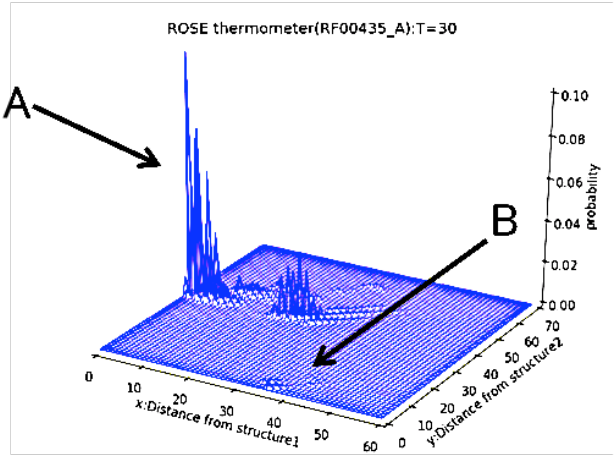

Figure 3 ROSE RNA サーモメーターの２次元構造分布(温度 30℃)
x 軸参照構造:クラスタ A のγ –centroid 構造(Figure2.a)
y 軸参照構造:クラスタ B のγ –centroid 構造(Figure2.b)

4.結論

本研究で提案する手法は、RNA 配列だけから RNA の構造変化を捉えることができることを示した。このことから、本研究を用いて網羅的に RNA 配列を解析すれば、今まで構造変化をすることが知られていなかった RNA を発見できる可能性がある。合成生物学は DNA や RNA、タンパク質などの生体高分子を設計することで、望みの特性をもつ生体システムを作ることを目的としている。本研究の提案手法を用いて、RNA の構造変化を予測することで、機能性 RNA の配列設計に貢献することが期待できる。