Vol.2012-BIO-28 No.5
2012/3/28

双対分解によるRNA構造アラインメント

佐藤健吾[†1,†5] 加藤有己[†2] 阿久津達也[†3]
浅井潔[†4,†5] 榊原康文[†1,†5]

本研究では，期待精度最大化原理に基づいて RNA 構造アラインメントを整数計画問題として定式化し，双対分解によってこれを高速に解くアルゴリズム DAFS を提案する．本手法では，双対分解によって RNA 構造アランメントを解きやすい部分問題，すなわち共通二次構造予測と（構造を考慮しない）配列アラインメントに分解して独立に効率よく解き，構造アランメントの制約を満たさない予測に対してペナルティを与える，という手順を繰り返すことによって最適解を得る．複数のデータセットにおける計算機実験から，とくに共通二次構造の精度において DAFS は既存の手法と比べて最も高精度，あるいはそれに匹敵する精度であり，かつ同程度の精度の手法と比べて明らかに高速であることが示された．

RNA structural alignments via dual decomposition

Kengo Sato,[†1,†5] Yuki Kato,[†2] Tatsuya Akutsu,[†3]
Kiyoshi Asai[†4,†5] and Yasubumi Sakakibara [†1,†5]

We develop DAFS, a novel algorithm that simultaneously aligns and folds RNA sequences based on maximizing expected accuracy of a predicted common secondary structure and its alignment. DAFS decomposes the pairwise structural alignment problem into two independent secondary structure prediction problems and one pairwise (non-structural) alignment problem by the dual decomposition technique, and maintains the consistency of a pairwise structural alignment by imposing penalties on inconsistent base pairs and alignment columns that are iteratively updated. The experiments on publicly available datasets showed that DAFS can produce reliable structural alignments from unaligned sequences in terms of accuracy of common secondary structure prediction.

1. はじめに

一本の配列からの RNA 二次構造予測の精度には限界があり，相同な配列群の信頼性が高いアラインメントが既に計算済みならば，配列群からの共通二次構造予測を用いた方が良い二次構造を得られることが知られている．このような共通二次構造予測の手法として RNAalifold[3], CentroidAlifold[8] などが開発されている．しかしながら，信頼性が高い RNA 配列のアラインメントを計算するためには RNA 二次構造の情報を用いる必要があるため，これはいわゆる「鶏と卵」問題となる．

このジレンマを克服するために，Sankoff は RNA 配列のアラインメントと二次構造予測を同時に行う動的計画法に基づくアルゴリズムを開発した[19]．しかしこのアルゴリズムは，配列長 L の N 本の配列に対して時間計算量 $O(L^{3N})$，空間計算量 $O(L^{2N})$ を必要とし実用的でないため，この計算量を削減するアルゴリズムが多数開発されてきた．しかしながら，以上のような state-of-the-art な実装を以ってしても Sankoff アルゴリズムを元にした手法は未だに多くの計算量を必要とする．

本論文では，双対分解による RNA 構造アラインメントアルゴリズム DAFS (Dual decomposition for Aligning and Folding RNA sequences Simultaneously) を提案する．本手法では，期待精度最大化（maximizing expected accuracy; MEA）原理に基づき，正しく予測される塩基対数とアラインメントカラム数の期待値の和を目的関数とする整数計画問題として RNA 構造アラインメントを定式化する．この整数計画問題を解く際，双対分解によって (1) RNA 構造アラインメントが満たすべき制約を 3 つの部分問題，すなわち 2 つの二次構造予測と 1 つの（構造を考慮しない）配列アラインメントに分解して，それぞれ独立に最適化する．それぞれの問題は Nussinov アルゴリズム[17] と Needleman–Wunsch アル

†1 慶應義塾大学 理工学部 生命情報学科
Department of Biosciences and Informatics, Keio University
†2 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†3 京都大学 化学研究所 バイオインフォマティクスセンター
Bioinformatics Center, Institute for Chemical Research, Kyoto University
†4 東京大学 大学院新領域創成科学研究科
Graduate School of Frontier Sciences, University of Tokyo
†5 産業技術総合研究所 生命情報工学研究センター
Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)

ゴリズム[16] によって高速に解くことができる．(2) 構造アラインメントが満たすべき制約と矛盾する塩基対とアラインメントカラムに対応するスコアにペナルティを与える．以上の手順を繰り返すことによって，RNA 構造アラインメントが満たすべき制約の下，目的関数すなわち期待精度を最大化する RNA 構造アラインメントを得る．本手法は，累進法によって容易にマルチプルアラインメントに拡張できる．

2. 手　　法

本節では，RNA 構造アラインメント問題を整数計画問題として定式化し，双対分解によってこの問題を高速に解くアルゴリズム DAFS を提案する．

2.1 Preliminaries

RNA 塩基の集合を $\Sigma = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ とし，すべての RNA 配列の集合を Σ^* とする．RNA 配列 $a = a_1 a_2 \cdots a_n \in \Sigma^*$ の長さを $|a|$ と書くこととする．2 本の RNA 配列 $a, b \in \Sigma^*$ が与えられた時，$\mathcal{A}(a,b)$ を a と b がとりうる（構造を考慮しない）すべての配列アラインメントの集合，$\mathcal{S}(a)$ を a がとりうるすべての二次構造の集合とする．アラインメント $z \in \mathcal{A}(a,b)$ は $|a| \times |b|$ 次元の二値行列 $z = (z_{ik})$ で表す．ここで $z_{ik} = 1$ は，塩基 a_i と b_k がアラインされることを表す．二次構造 $x \in \mathcal{S}(a)$ は $|a| \times |a|$ 次元の二値上三角行列 $x = (x_{ij})_{i<j}$ で表す．ここで $x_{ij} = 1$ は，塩基 a_i と a_j が塩基対を形成することを表す．$\mathcal{SA}(a,b)$ を a と b がとりうるすべての構造アラインメントの集合とし，構造アラインメント $\theta \in \mathcal{SA}(a,b)$ が配列アラインメント $z \in \mathcal{A}(a,b)$ と二次構造 $x \in \mathcal{S}(a)$，$y \in \mathcal{S}(b)$ から成る時，$\theta = (x, y, z)$ と書くこととする．

2.2 期待精度最大化に基づくスコア関数

RNA 配列が与えられた時，RNA 構造アラインメントの目的は，信頼性が高いアラインメントと共通二次構造を同時に得ることである．この目的のために，構造アラインメント $\hat{\theta} = (\hat{x}, \hat{y}, \hat{z})$ の正解構造アラインメント $\theta = (x, y, z)$ に対する利益関数を，二次構造に関する利益関数 $G_s(x, \hat{x})$，$G_s(y, \hat{y})$ と配列アラインメントに関する利益関数 $G_a(z, \hat{z})$ の重み付き線形和として次のように定義する：

$$G(\theta, \hat{\theta}) = \alpha \left\{ G_s(x, \hat{x}) + G_s(y, \hat{y}) \right\} + G_a(z, \hat{z}), \tag{1}$$

ここで $\alpha > 0$ は二次構造とアラインメントの間の重みを制御するパラメータである．二次構造 $\hat{x}$ に関する利益関数 $G_s(x, \hat{x})$ は次のように定義される：

$$G_s(x, \hat{x}) = (1 - \tau)\, TP_s(x, \hat{x}) + \tau\, TN_s(x, \hat{x}),$$

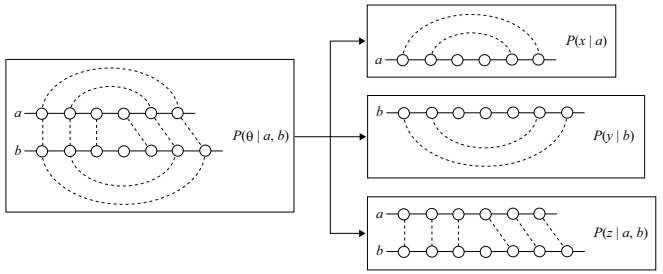

図 1 RNA 構造アラインメントの確率分布 $P(\theta \mid a, b)$ を分解し，二次構造の確率分布 $P(x \mid a)$，$P(y \mid b)$ と配列アラインメントの確率分布 $P(z \mid a, b)$ の積で近似する．
Fig. 1 An illustration of factorization of a probability distribution over a space of RNA structural alignments.

ここで $TP_s(x, \hat{x}) = \sum_{i<j} I(x_{ij} = 1) I(\hat{x}_{ij} = 1)$ は true positive の数，$TN_s(x, \hat{x}) = \sum_{i<j} I(x_{ij} = 0) I(\hat{x}_{ij} = 0)$ は true negative の数，$\tau \in [0, 1]$ は true positive と true negative のバランスを制御するパラメータである．また，$I(condition)$ は条件式 $condition$ が真なら 1，偽なら 0 を返す二値関数である．同様に，配列アラインメント $\hat{z}$ に関する利益関数 $G_a(z, \hat{z})$ は次のように定義される：

$$G_a(z, \hat{z}) = (1 - \sigma)\, TP_a(z, \hat{z}) + \sigma\, TN_a(z, \hat{z}),$$

ここで $TP_a(z, \hat{z}) = \sum_{i,k} I(z_{ik} = 1) I(\hat{z}_{ik} = 1)$ は true positive の数，$TN_a(z, \hat{z}) = \sum_{i,k} I(z_{ik} = 0) I(\hat{z}_{ik} = 0)$ は true negative の数，$\sigma \in [0, 1]$ は true positive と true negative のバランスを制御するパラメータである．

期待精度最大化原理に基づき，構造アライメントの空間 $\mathcal{SA}(a,b)$ 上に与えられた確率分布 $P(\theta \mid a, b)$ の下で，利益関数 (1) の期待値を最大化する構造アラインメント $\hat{\theta}$ を求める：

$$\mathbb{E}_{\theta|a,b}[G(\theta, \hat{\theta})] = \sum_{\theta \in \mathcal{SA}(a,b)} P(\theta \mid a, b) G(\theta, \hat{\theta}). \tag{2}$$

しかしながら，シュードノットを考慮しない二次構造を仮定したとしても，式 (2) の時間計算量は $O(|a|^3 |b|^3)$，空間計算量 $O(|a|^2 |b|^2)$ を必要とする．そこで，構造アラインメントの確率分布を次のように積近似する：

$$P(\theta \mid a, b) \approx P(x \mid a) P(y \mid b) P(z \mid a, b), \tag{3}$$

ここで $P(x \mid a)$ と $P(y \mid b)$ はそれぞれ RNA 二次構造の空間 $\mathcal{S}(a)$ と $\mathcal{S}(b)$ の上で定義された確率分布，$P(z \mid a, b)$ は 配列アラインメントの空間 $\mathcal{A}(a,b)$ 上の確率分布である（図 1）．

その結果，期待利益関数 (2) は次のように近似することができる：

$$\mathbb{E}_{\theta|a,b}[G(\theta,\hat{\theta})] \approx \sum_{i,k}\left[p_{ik}^{(a,b)}-\sigma\right]\hat{z}_{ik} + \alpha\left(\sum_{i<j}\left[p_{ij}^{(a)}-\tau\right]\hat{x}_{ij}+\sum_{k<l}\left[p_{kl}^{(b)}-\tau\right]\hat{y}_{kl}\right)+C, \tag{4}$$

ここで

$$p_{ij}^{(a)}=\sum_{x\in\mathcal{S}(a)}P(x\mid a)I(x_{ij}=1),\quad p_{kl}^{(b)}=\sum_{y\in\mathcal{S}(b)}P(y\mid b)I(y_{kl}=1)$$

は塩基対事後確率分布，

$$p_{ik}^{(a,b)}=\sum_{z\in\mathcal{A}(a,b)}P(z\mid a,b)I(z_{ik}=1)$$

はアライメント事後確率分布であり，C は $\hat{\theta}=(\hat{x},\hat{y},\hat{z})$ に依存しない定数である．配列 a，b がシュードノット構造を形成しないと仮定すると，塩基対確率行列 $p_{ij}^{(a)}$，$p_{kl}^{(b)}$ とアラインメント確率行列 $p_{ik}^{(a,b)}$ の時間計算量はそれぞれ $O(|a|^3)$，$O(|b|^3)$，$O(|a||b|)$ であり，空間計算量はそれぞれ $O(|a|^2)$，$O(|b|^2)$，$O(|a||b|)$ である．

2.3 整数計画法による定式化

本手法の目的は，RNA 構造アランメントが満たすべき制約の下，期待利益関数 (4) を最大化する RNA 構造アラインメントを計算することである．この最適化は次のような整数計画問題として定式化することができる：

$$\text{maximize:}\quad S(x,y,z;a,b)=\sum_{i,k}\left[p_{ik}^{(a,b)}-\sigma\right]z_{ik} + \alpha\left(\sum_{i<j}\left[p_{ij}^{(a)}-\tau\right]x_{ij}+\sum_{k<l}\left[p_{kl}^{(b)}-\tau\right]y_{kl}\right), \tag{5}$$

$$\text{subject to:}\quad \sum_{j<i}x_{ji}+\sum_{j>i}x_{ij}\le 1\quad(1\le\forall i\le|a|), \tag{6}$$

$$x_{ij}+x_{i'j'}\le 1\quad(1\le\forall i<\forall i'<\forall j<\forall j'\le|a|), \tag{7}$$

$$\sum_{l<k}y_{lk}+\sum_{l>k}y_{kl}\le 1\quad(1\le\forall k\le|b|), \tag{8}$$

$$y_{kl}+y_{k'l'}\le 1\quad(1\le\forall k<\forall k'<\forall l<\forall l'\le|b|), \tag{9}$$

$$\sum_{k}z_{ik}\le 1\quad(1\le\forall i\le|a|), \tag{10}$$

$$\sum_{i}z_{ik}\le 1\quad(1\le\forall k\le|b|), \tag{11}$$

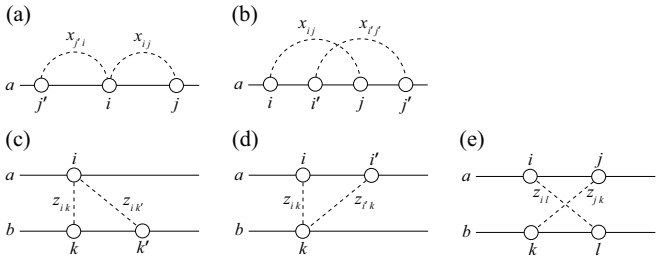

図 2 RNA 構造アラインメントが満たすべき制約．(a)〜(e) はそれぞれ，式 (6)，式 (7)，式 (10)，式 (11)，式 (12) に対応している．それぞれの図における破線は高々 1 つの変数だけが 1 となれることを示している．
Fig. 2 An illustration of the constraints of the IP formulation.

$$z_{il}+z_{jk}\le 1\quad(1\le\forall i<\forall j\le|a|;\,1\le\forall k<\forall l\le|b|), \tag{12}$$

$$x_{ij}=\sum_{k<l}w_{ijkl}\quad(1\le\forall i<\forall j\le|a|), \tag{13}$$

$$y_{kl}=\sum_{i<j}w_{ijkl}\quad(1\le\forall k<\forall l\le|b|), \tag{14}$$

$$z_{ik}\ge\sum_{j<i,l<k}w_{jilk}+\sum_{j>i,l>k}w_{ijkl}\quad(1\le\forall i\le|a|;\,1\le\forall k\le|b|), \tag{15}$$

ここで w_{ijkl} は，塩基対 (a_i,a_j) が塩基対 (b_k,b_l) とアラインされる時 $w_{ijkl}=1$，そうでない時 $w_{ijkl}=0$ となる二値変数である． 制約 (6) は，各々の塩基 a_i は高々 1 つの塩基とのみ塩基対を形成することができることを意味する（図 2a）．制約 (7) は，二次構造 $x\in\mathcal{S}(a)$ はシュードノットを許さないことを示している（図 2b）．同様に，制約 (8) と (9) は，二次構造 $y\in\mathcal{S}(b)$ に関する制約を表している．制約 (10) と (11) は，配列 a と b の各々の塩基は高々 1 つ塩基とのみアラインできることを意味する（図 2c，2d）．制約 (12) は，配列アラインメント $z\in\mathcal{A}(a,b)$ では図 2e のような交差するアラインメントを許さないことを示している．制約 (13)–(15) は，構造アラインメント $\theta=(x,y,z)\in\mathcal{SA}(a,b)$ が共通二次構造を持つために，それぞれの二次構造 x,y と配列アラインメント z において満たすべき制約である．もし，制約 (13)–(15) がなければ，構造アラインメントの要素 x，y，z は別々に効率よく解くことができることから，制約 (13)–(15) が RNA 構造アラインメント問題の複雑さの原因であると言える．

2.4 双 対 分 解

前節で述べたように，制約 (13)–(15) が，RNA 構造アラインメント問題を解きにくくしている．そこで，ラグランジュ緩和[14] によってこれらの制約を取り除いて，問題を解き易

くする．まず，制約 (13)–(15) を目的関数に移すことによってラグランジュ双対関数を定義する：

$$L(\lambda,\mu,\nu) = \max_{\substack{x\in\mathcal{S}(a),y\in\mathcal{S}(b),\\ z\in\mathcal{A}(a,b),w}} \Bigg\{ S(x,y,z;a,b) + \sum_{i<j}\lambda_{ij}\left(\sum_{k<l} w_{ijkl} - x_{ij}\right) + \sum_{k<l}\mu_{kl}\left(\sum_{i<j} w_{ijkl} - y_{kl}\right) + \sum_{i,k}\nu_{ik}\left(z_{ik} - \sum_{j<i,l<k} w_{jilk} - \sum_{j>i,l>k} w_{ijkl}\right)\Bigg\}, \tag{16}$$

ここで $\lambda = \{\lambda_{ij} \mid i<j\}$，$\mu = \{\mu_{kl} \mid k<l\}$，$\nu = \{\nu_{ik} \mid \nu_{ik} \geq 0\}$ はラグランジュ未定乗数である．式 (16) は次のように書き換えることができる：

$$\begin{aligned} L(\lambda,\mu,\nu) = {} & \max_{x\in\mathcal{S}(a)} \sum_{i<j}\left[\alpha(p_{ij}^{(a)} - \tau) - \lambda_{ij}\right] x_{ij} \\ & + \max_{y\in\mathcal{S}(b)} \sum_{k<l}\left[\alpha(p_{kl}^{(b)} - \tau) - \mu_{kl}\right] y_{kl} \\ & + \max_{z\in\mathcal{A}(a,b)} \sum_{i,k}\left[p_{ik}^{(a,b)} - \sigma + \nu_{ik}\right] z_{ik} \\ & + \max_{w} \sum_{i<j,k<l} \left[\lambda_{ij} + \mu_{kl} - \nu_{ik} - \nu_{jl}\right] w_{ijkl}, \end{aligned} \tag{17}$$

すなわち，式 (17) の各項は動的計画法によって独立に効率よく計算できることを意味している：第 1 項と第 2 項は Nussinov アルゴリズム[17]，第 3 項は Needleman–Wunsch アルゴリズム[16] で計算することができる．最後の項は，単純に係数が正となる w_{ijkl} だけを 1 にすればよい．このような解法を双対分解（dual decomposition）[13),22)] という．

双対目的関数 $L(\lambda,\mu,\nu)$ は主目的関数 (5) の上界となっているので，より良い上界を得るために未定乗数に関して式 (17) を最小化する．ラグランジュ関数 $L(\lambda,\mu,\nu)$ は凹関数であるが微分可能ではない[14] ため，劣勾配法を用いてラグランジュ未定乗数 λ_{ij}，μ_{kl}，ν_{ik} を反復的に最適化する．最終的に，図 3 で示すようなアルゴリズムを得る．ここで $\eta^{(t)} > 0$ はそれぞれの更新におけるステップ幅であり，$\lim_{t\to\infty}\eta^{(t)} = 0$ かつ $\sum_{t=1}^{\infty}\eta^{(t)} = \infty$ ならばラグランジュ関数 $L(\lambda,\mu,\nu)$ は最適値に収束することが知られている[14]．更新は，解が見つかるか，事前に与えられた上限回数 T に達するまで繰り返す．

ラグランジュ未定乗数は，二次構造と配列アラインメントの間の制約 (13)–(15) に関する矛盾に対するペナルティスコアとみなすことができる．

1: Calculate the posterior probabilities $p_{ij}^{(a)}$, $p_{kl}^{(b)}$ and $p_{ik}^{(a,b)}$.
2: Set $\lambda_{ij}^{(1)} = 0$, $\mu_{kl}^{(1)} = 0$ and $\nu_{ik}^{(1)} = 0$.
3: **for** $t = 1$ to T **do**
4: $\quad x^{(t)} \leftarrow \arg\max_{x\in\mathcal{S}(a)} \sum_{i<j}\left[\alpha(p_{ij}^{(a)} - \tau) - \lambda_{ij}^{(t)}\right] x_{ij}$
5: $\quad y^{(t)} \leftarrow \arg\max_{y\in\mathcal{S}(b)} \sum_{k<l}\left[\alpha(p_{kl}^{(b)} - \tau) - \mu_{kl}^{(t)}\right] y_{ij}$
6: $\quad z^{(t)} \leftarrow \arg\max_{z\in\mathcal{A}(a,b)} \sum_{i,k}\left[p_{ik}^{(a,b)} - \sigma + \nu_{ik}^{(t)}\right] z_{ik}$
7: $\quad w^{(t)} \leftarrow \arg\max_{w} \sum_{i<j,k<l}[\lambda_{ij}^{(t)} + \mu_{kl}^{(t)} - \nu_{ik}^{(t)} - \nu_{jl}^{(t)}] w_{ijkl}$
8: $\quad$ **if** $\theta^{(t)} = (x^{(t)}, y^{(t)}, z^{(t)})$ satisfies the constraints (13)–(15) **then**
9: $\quad\quad$ **return** $\theta^{(t)} = (x^{(t)}, y^{(t)}, z^{(t)})$
10: $\quad$ **end if**
11: $\quad \lambda_{ij}^{(t+1)} \leftarrow \lambda_{ij}^{(t)} - \eta^{(t)}\left(\sum_{k<l} w_{ijkl}^{(t)} - x_{ij}^{(t)}\right)$
12: $\quad \mu_{kl}^{(t+1)} \leftarrow \mu_{kl}^{(t)} - \eta^{(t)}\left(\sum_{i<j} w_{ijkl}^{(t)} - y_{kl}^{(t)}\right)$
13: $\quad \nu_{ik}^{(t+1)} \leftarrow \max\left\{0, \nu_{ik}^{(t)} - \eta^{(t)}\left(z_{ik}^{(t)} - \sum_{j<i,l<k} w_{jilk}^{(t)} - \sum_{j>i,l>k} w_{ijkl}^{(t)}\right)\right\}$
14: **end for**
15: **return** $\theta^{(T)} = (x^{(T)}, y^{(T)}, z^{(T)})$

図 3 双対分解による RNA 構造アラインメント．
Fig. 3 The algorithm for predicting RNA structural alignments using dual decomposition.

2.5 マルチプルアラインメントへの拡張

前節で述べた双対分解によるペアワイズアラインメントは，平均塩基対確率行列と平均アラインメント確率行列を用いることによって，累進法によるマルチプルアラインメントへ容易に拡張できる．

A と B を RNA 配列の 2 組のアラインメントとする．平均塩基対確率 $p_{ij}^{(A)}$ を A に含まれるすべての配列に関する $p_{ij}^{(a)}$ の平均と定義する．同様に，平均アラインメント確率 $p_{ik}^{(A,B)}$ を，アラインメント A と B に含まれる配列のすべての組に関する $p_{ik}^{(a,b)}$ の平均と定義する．そして，アラインメント A と B の構造アラインメントは，前節で述べた双対分解（図 3）において，塩基対確率 $p_{ij}^{(a)}$，$p_{kl}^{(b)}$ とアラインメント確率 $p_{ik}^{(a,b)}$ を平均塩基対確率 $p_{ij}^{(A)}$，$p_{kl}^{(B)}$ と平均アラインメント確率 $p_{ik}^{(A,B)}$ で置き換えて計算すれば得ることができる．

3. 結　　果

前節で述べたアルゴリズムに基づき `DAFS` を実装した．塩基対確率行列を計算するために，Vienna RNA package[10] で実装されている McCaskill アルゴリズム[15] を用いた．自由エネルギーパラメータは Andronescu らによって計算された Boltzmann likelihood パ

ラメータ[1] を用いた．アラインメント確率行列を計算するために，ProbCons[6] を用いた．DAFS のソースコードは http://www.ncrna.org/software/dafs/ から取得できる．

3.1 評価基準

アラインメントに対する 2 つの評価基準：(i) sum-of-pairs score (SPS)[21] (ii) structure conservation index (SCI)[23] と，共通二次構造に対する 3 つの評価基準：(iii) sensitivity (SEN) (iv) positive predictive value (PPV) (v) Matthews correlation coefficient (MCC) によって RNA 構造アラインメントの精度を評価した．SPS は計算されたアラインメントの中で正しく予測されたアラインメントカラムの割合であり，SCI は RNAalifold[3] で計算されたアラインメントの共通二次構造の最小自由エネルギー E_A とすべての配列に関する最小自由エネルギーの平均 $\bar{E}$ を用いて $\mathrm{SCI} = E_A/\bar{E}$ と定義される．これらの基準は，アラインメントのみから直接計算できる．SEN は正解二次構造に現れる塩基対を正しく予測できた割合 $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$，PPV は予測二次構造に現れる塩基対が正しい塩基対だった割合 $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$，MCC は次の式のように定義される：

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FN}\cdot\mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

ここで TP は正解構造，予測構造の両方に現れた塩基対の数，FP は予測構造に現れたが正解構造には現れない塩基対の数，TN は正解構造，予測構造の両方に現れない塩基対の数，FN は正解構造に現れたが予測構造には現れない塩基対の数である．SEN と PPV はトレードオフの関係にあり，MCC は両者のバランスを考えた評価基準である．

比較に用いた構造アラインメント法のいくつは共通二次構造を予測する機能を持っていないため，それぞれのアラインメント法でアラインメントを生成した後，CentroidAlifold[8] を用いて共通二次構造予測を行った．

計算時間 (TIME) は，Intel Xeon E5450 (3.0GHz) 上の Linux OS でデータセットすべての配列群の構造アラインメントの計算にかかる秒数である．

3.2 比較実験

DASF の性能を評価するために Murlet データセット[12] を用いた．Murlet データセットは Rfam[7] から取得した 17 ファミリー 85 組のアラインメントから構成されている．実験では以下のパラメータを用いた：$\alpha = 4.0$，$\tau = 0.2$，$\sigma = 0.01$，$T = 600$.

以下の構造アラインメント法と DAFS を比較した：(i) PicXAA-R version 1.0[18]; (ii) CentroidAlign version 1.00[9]; (iii) RAF version 1.0[4]; (iv) MAFFT version 6.861[11] with scarnapair (MAFFT-xinsi); (v) MXSCARNA version 2.1[20]; (vi) LARA version 1.3.2a[2]; (vii) LocARNA version 1.6.1[24]．さらにベースラインとして，次の構造を考慮しないアラインメント法との比較も行った：(viii) CONTRAlign version 2.01[5]; (ix) ProbConsRNA version 1.1[6].

表 1 Murlet データセットにおける結果.
Table 1 The results on the Murlet dataset.

	SPS	SCI	SEN	PPV	MCC	TIME
DAFS	0.75	0.46	0.67	0.77	**0.71**	416
PicXAA-R	0.78	0.48	0.65	0.78	0.70	167
CentroidAlign	0.78	0.48	0.62	0.80	0.69	169
RAF	0.75	0.46	**0.68**	0.75	**0.71**	4274
MAFFT-xinsi	**0.79**	0.53	0.64	0.80	**0.71**	401
MXSCARNA	0.75	0.44	0.65	0.78	0.70	152
LARA	0.75	0.50	0.62	0.78	0.68	5361
LocARNA	0.71	**0.61**	0.64	0.76	0.69	14540
CONTRAlign	0.77	0.41	0.57	0.83	0.66	169
ProbConsRNA	0.76	0.37	0.56	**0.84**	0.66	88

表 1 に Murlet データセットにおける実験結果を示す．これらの表から，DAFS は MCC において最も良い，あるいはそれに匹敵する精度を示しており，既存の手法の中で最も高精度である RAF に比べて明らかに高速であることがわかる．

4. おわりに

本論文では，双対分解による RNA 構造アラインメントアルゴリズム DAFS を提案した．複数のデータセットにおける計算機実験から，とくに共通二次構造の精度において DAFS は既存の手法と比べて最も高精度，あるいはそれに匹敵する精度であり，同程度の精度の手法と比べて明らかに高速であることが示された．

本手法の予測精度は，目的関数 (5) で用いている塩基対事後確率とアラインメント事後確率に大きく依存する．本手法では，高速に計算するために式 (3) のような積近似を導入しており，このことが精度の低下を招いている可能性がある．今後は，スコア関数を改善することにより予測精度の向上を目指す．

謝辞 本研究は一部，文部科学省科学研究費補助金若手研究（B）[#22700305 to K.S., #22700313 to Y.K] からの助成金を受けている．

参 考 文 献

1) Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H. and Murphy, K. P.: Computational approaches for RNA energy parameter estimation, *RNA*, Vol. 16, pp.2304–2318 (2010).
2) Bauer, M., Klau, G. W. and Reinert, K.: Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization, *BMC Bioinform.*, Vol.8, p.271 (2007).
3) Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R. and Stadler, P.F.: RNAalifold: improved consensus structure prediction for RNA alignments, *BMC Bioinform.*, Vol.9, p.474 (2008).
4) Do, C.B., Foo, C.S. and Batzoglou, S.: A max-margin model for efficient simultaneous alignment and folding of RNA sequences, *Bioinformatics*, Vol.24, pp.68–76 (2008).
5) Do, C.B., Gross, S. and Batzoglou, S.: CONTRAlign: Discriminative training for protein sequence alignment, *Proc. of the 10th Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, Vol.3909, pp.160–174 (2006).
6) Do, C.B., Mahabhashyam, M.S., Brudno, M. and Batzoglou, S.: ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res.*, Vol.15, pp. 330–340 (2005).
7) Gardner, P.P., Daub, J., Tate, J., Moore, B.L., Osuch, I.H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R. and Bateman, A.: Rfam: Wikipedia, clans and the "decimal" release, *Nucleic Acids Res.*, Vol.39, pp.D141–D145 (2011).
8) Hamada, M., Sato, K. and Asai, K.: Improving the accuracy of predicting secondary structure for aligned RNA sequences, *Nucleic Acids Res.*, Vol.39, pp.393–402 (2011).
9) Hamada, M., Sato, K., Kiryu, H., Mituyama, T. and Asai, K.: CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score, *Bioinformatics*, Vol.25, pp.3236–3243 (2009).
10) Hofacker, I.L.: Vienna RNA secondary structure server, *Nucleic Acids Res.*, Vol.31, pp.3429–3431 (2003).
11) Katoh, K. and Toh, H.: Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework, *BMC Bioinform.*, Vol.9, p.212 (2008).
12) Kiryu, H., Tabei, Y., Kin, T. and Asai, K.: Murlet: a practical multiple alignment tool for structural RNA sequences, *Bioinformatics*, Vol.23, pp.1588–1598 (2007).
13) Komodakis, N., Paragios, N. and Tziritas, G.: MRF Optimization via Dual Decomposition: Message-Passing Revisited, *Proc. of IEEE 11th International Conference on Computer Vision (ICCV 2007)*, pp.1–8 (2007).
14) Korte, B. and Vygen, J.: *Combinatorial Optimization: Theory and Algorithms*, Springer Verlag, Berlin, Germany (2008).
15) McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, Vol.29, pp.1105–1119 (1990).
16) Needleman, S.B. and Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, Vol.48, pp.443–453 (1970).
17) Nussinov, R., Pieczenick, G., Griggs, J. and Kleitman, D.: Algorithms for loop matching, *SIAM J. Appl. Math.*, Vol.35, pp.68–82 (1978).
18) Sahraeian, S.M. and Yoon, B.J.: PicXAA-R: efficient structural alignment of multiple RNA sequences using a greedy approach, *BMC Bioinform.*, Vol.12 Suppl 1, p.S38 (2011).
19) Sankoff, D.: Simultaneous solution of the RNA folding, alignment and protosequence problems, *SIAM J. Appli. Math.*, Vol.45, pp.810–825 (1985).
20) Tabei, Y., Kiryu, H., Kin, T. and Asai, K.: A fast structural multiple alignment method for long RNA sequences, *BMC Bioinform.*, Vol.9, p.33 (2008).
21) Thompson, J. D., Plewniak, F. and Poch, O.: A comprehensive comparison of multiple sequence alignment programs, *Nucleic Acids Res.*, Vol.27, pp.2682–2690 (1999).
22) Wainwright, M., Jaakkola, T. and Willsky, A.: MAP estimation via agreement on trees: message-passing and linear programming, *IEEE Trans. Inf. Theory*, Vol.51, pp.3697–3717 (2005).
23) Washietl, S., Hofacker, I.L. and Stadler, P.F.: Fast and reliable prediction of noncoding RNAs, *Proc. Natl. Acad. Sci. U.S.A.*, Vol.102, pp.2454–2459 (2005).
24) Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F. and Backofen, R.: Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering, *PLoS Comput. Biol.*, Vol.3, p.e65 (2007).