Vol.2015-BIO-43 No.2
2015/9/12

# 双対分解によるマルチプルアラインメント

穴水 拓郎[1,a)]　榊原 康文[1,b)]　佐藤 健吾[1,c)]

**概要**：バイオインフォマティクスにおいて塩基配列やアミノ酸配列を解析する手法として，配列アラインメントが広く用いられている．ペアワイズアラインメントは動的計画法で効率的に解けることが知られている．一方，マルチプルアラインメントは，その計算量の問題から累進法と呼ばれる近似手法で解くことが多いが，最適解が保証されない．本研究では，マルチプルアラインメントを整数計画問題として定式化し，双対分解によって総当たりのペアワイズアラインメントに相当する部分問題に分割することによって最適解を効率よく求める手法を提案する．

Takuro Anamizu[1,a)]　Yasubumi Sakakibara[1,b)]　Kengo Sato[1,c)]

## 1. はじめに

配列アラインメントとは，二本以上の配列を入力として類似する塩基どうしが縦に揃うように並べる操作である．特に，配列本数が二本のときをペアワイズアラインメント，三本以上のときをマルチプルアラインメントと呼ぶ．ペアワイズアラインメントは動的計画法で効率的に解ける一方，マルチプルアラインメントは累進法と呼ばれる近似手法で解くことが多い．しかし累進法では最適解が保証されず，繰り返し最適化法による精度改善は，計算時間とトレードオフになってしまう．本研究では，マルチプル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 [4] によって総当たりのペアワイズアラインメントに相当する部分問題に分割することによって最適解を効率よく求める手法 DMSA (Dual decomposition for Multiple Sequence Alignment) を提案する．

## 2. 手法

配列 $a$ の長さを $|a|$ と書くこととする．二本の配列 $a, b$ が与えられたとき，$\mathcal{A}(a,b)$ を $a$ と $b$ がとりうる全ての配列アラインメントの集合とする．アラインメント $z \in \mathcal{A}(a,b)$ は $|a|\times|b|$ 次元の二値行列 $z = (z_{ij})$ で表わす．ここで $z_{ij} = 1$ は，塩基 $a_i$ と $b_j$ がアラインされていることを表す．$n$ 本の配列 $s_1, \ldots, s_n$ のマルチプルアラインメント $\theta \in \mathcal{A}(s_1, \ldots, s_n)$ が総当たりのペアワイズアラインメントからなるとき，$\theta = (z^{(1,2)}, \ldots, z^{(n-1,n)})$ と書くこととする．

マルチプルアラインメントの利益関数を，総当たりのペアワイズアラインメントの利益関数の総和として定義する：

$$G(\theta, \hat{\theta}) = G(z^{(1,2)}, \hat{z}^{(1,2)}) + \cdots + G(z^{(n-1,n)}, \hat{z}^{(n-1,n)}) \tag{1}$$

ペアワイズアラインメント $z$ に関する利益関数 $G(z, \hat{z})$ は次のように定義される：

$$G(z, \hat{z}) = (1-\sigma)TP(z, \hat{z}) + \sigma TN(z, \hat{z}) \tag{2}$$

ここで $TP(z, \hat{z}) = \sum_{i,j} I(z_{ij} = 1)I(\hat{z}_{ij} = 1)$ は true positive の数，$TN(z, \hat{z}) = \sum_{i,j} I(z_{ij} = 0)I(\hat{z}_{ij} = 0)$ は true negative の数，$\sigma$ 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 [2] に基づき，アラインメント空間 $\mathcal{A}(s_1, \ldots, s_n)$ 上に与えられる確率分布 $P(\theta \mid s_1, \ldots, s_n)$ の下で，利益関数の期待値を最大化するマルチプルアラインメント $\hat{\theta}$ を求める：

$$\mathbb{E}[G(\theta, \hat{\theta})] = \sum_{\theta \in \mathcal{A}(s_1, \ldots, s_n)} P(\theta \mid s_1, \ldots, s_n) G(\theta, \hat{\theta}) \tag{3}$$

ここで，マルチプルアラインメントの確率分布を次のように積近似する：

$$P(\theta \mid s_1, \ldots, s_n) \approx \prod_{1 \le k < l \le n} P(z^{(k,l)} \mid s_k, s_l) \tag{4}$$

その結果，期待利益関数は次のように近似することができる：

[1] 慶應義塾大学理工学部生命情報学科
Department of Biosciences and Informatics, Keio University
[a)] anamizu@dna.bio.keio.ac.jp
[b)] yasu@bio.keio.ac.jp
[c)] satoken@bio.keio.ac.jp

$$\mathbb{E}[G(\theta,\hat{\theta})] \approx \sum_{1\le k<l\le n}\sum_{i=1}^{|s_k|}\sum_{j=1}^{|s_l|}\left[p_{ij}^{(k,l)}-\sigma\right]z_{ij}^{(k,l)}+C \tag{5}$$

ここで

$$p_{ij}^{(a,b)} = \sum_{z\in\mathcal{A}(a,b)} P(z\mid a,b)I(z_{ij}=1) \tag{6}$$

はアラインメント事後確率であり，$C$ は $\theta$ に依存しない定数である．

本手法の目的は，マルチプルアラインメントが満たすべき制約の下，期待利益関数を最大化するマルチプル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：

maximize:

$$S(z;s) = \sum_{1\le k<l\le n}\sum_{i=1}^{|s_k|}\sum_{j=1}^{|s_l|}\left[p_{ij}^{(k,l)}-\sigma\right]z_{ij}^{(k,l)} \tag{7}$$

subject to:

$$\sum_{j=1}^{|s_l|} z_{ij}^{(k,l)} \le 1 \quad (1\le\forall k\le n, 1\le\forall l\le n, 1\le\forall i\le|s_k|) \tag{8}$$

$$z_{iq}^{(k,l)} + z_{jp}^{(k,l)} \le 1 \tag{9}$$
$$(1\le\forall k<\forall l\le n, 1\le\forall i<\forall j\le|s_k|, 1\le\forall p<\forall q\le|s_l|)$$

$$z_{ij}^{(k,l)} - z_{jh}^{(l,m)} + z_{hi}^{(m,k)} \le 1 \tag{10}$$

$$z_{ij}^{(k,l)} + z_{jh}^{(l,m)} - z_{hi}^{(m,k)} \le 1 \tag{11}$$

$$-z_{ij}^{(k,l)} + z_{jh}^{(l,m)} + z_{hi}^{(m,k)} \le 1 \tag{12}$$
$$(1\le\forall k<\forall l<\forall m\le n, 1\le\forall i\le|s_k|, 1\le\forall j\le|s_l|, 1\le\forall h\le|s_m|)$$

$$z_{i_1i_2}^{(\rho_1,\rho_2)} + \cdots + z_{i_{k-1}i_k}^{(\rho_{k-1},\rho_k)} + z_{i_ki_1}^{(\rho_k,\rho_1)} \le k-1 \tag{13}$$
$$(\forall\rho\in P_k, 1\le\forall i_p\le|s_{\rho_p}| \text{ for } 1\le\forall p\le k)$$

ただし，$P_k$ は $\{1,\ldots,n\}$ の大きさ $k$ の巡回順列の集合である．制約 (8) は，二本の配列の各々の塩基は高々 1 つの塩基とのみアラインされることを意味する．制約 (9) は交差するアラインメントを許さないことを表す．制約 (10)-(12) は，consistency transformation [3] に相当する．制約 (13) は，アラインメント列の順番に矛盾がないことを表す．

制約 (10)-(12) および制約制約 (13) 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．

$$\begin{aligned}L(\lambda,\mu,\nu,\xi) = {} & \max S(z;s) \\ & + \sum_{1\le k<l<m\le n}\sum_{i=1}^{|s_k|}\sum_{j=1}^{|s_l|}\sum_{h=1}^{|s_m|}\lambda_{ijh}^{(k,l,m)}(1-z_{ij}^{(k,l)}-z_{jh}^{(l,m)}+z_{hi}^{(m,k)}) \\ & + \sum_{1\le k<l<m\le n}\sum_{i=1}^{|s_k|}\sum_{j=1}^{|s_l|}\sum_{h=1}^{|s_m|}\mu_{ijh}^{(k,l,m)}(1-z_{ij}^{(k,l)}+z_{jh}^{(l,m)}-z_{hi}^{(m,k)}) \\ & + \sum_{1\le k<l<m\le n}\sum_{i=1}^{|s_k|}\sum_{j=1}^{|s_l|}\sum_{h=1}^{|s_m|}\nu_{ijh}^{(k,l,m)}(1+z_{ij}^{(k,l)}-z_{jh}^{(l,m)}-z_{hi}^{(m,k)}) \\ & + \sum_{3\le k\le n}\sum_{\rho\in P_k}\sum_{i_1=1}^{|s_{\rho_1}|}\cdots\sum_{i_k=1}^{|s_{\rho_k}|}\xi_{i_1\cdots i_k}^{(\rho)}(k-1-(z_{i_1i_2}^{(\rho_1,\rho_2)}+\cdots+z_{i_{k-1}i_k}^{(\rho_{k-1},\rho_k)}+z_{i_ki_1}^{(\rho_k,\rho_1)}))\end{aligned} \tag{14}$$

ここで，$\lambda,\mu,\nu,\xi$ はラグランジュ未定乗数である．上の式は次のように書きかえることができる．

$$L(\lambda,\mu,\nu,\xi) = \sum_{1\le k<l\le n}\sum_{i=1}^{|s_k|}\sum_{j=1}^{|s_l|}\left[p_{ij}^{(k,l)}-\sigma+\phi_{ij}^{(k,l)}\right]z_{ij}^{(k,l)} \tag{15}$$

ただし，

$$\begin{aligned}\phi_{ij}^{(k,l)} = {} & \sum_{m>l}\sum_{h=1}^{|s_m|}(-\lambda_{ijh}^{(k,l,m)}-\mu_{ijh}^{(k,l,m)}+\nu_{ijh}^{(k,l,m)}) \\ & + \sum_{k<m<l}\sum_{h=1}^{|s_m|}(\lambda_{ihj}^{(k,m,l)}-\mu_{ihj}^{(k,m,l)}-\nu_{ihj}^{(k,m,l)}) \\ & + \sum_{m<k}\sum_{h=1}^{|s_m|}(-\lambda_{hij}^{(m,k,l)}+\mu_{hij}^{(m,k,l)}-\nu_{hij}^{(m,k,l)}) \\ & - \sum_{3\le m\le n}\sum_{\substack{\rho\in P_m \text{ s.t.}\\ \rho_l=k,\rho_{l+1}=l}}\sum_{i_1=1}^{|s_{\rho_1}|}\cdots\sum_{i_{l-1}=1}^{|s_{\rho_{l-1}}|}\sum_{i_{l+1}=1}^{|s_{\rho_{l+1}}|}\cdots\sum_{i_k=1}^{|s_{\rho_k}|}\xi_{i_1\cdots i_k}^{(\rho)}\end{aligned}$$

劣勾配法によって式 (15) を $\lambda,\mu,\nu,\xi$ に関して最小化することによって元々の問題 (7) の解を得ることができる．そのアルゴリズムは以下のように動作する：(1) 入力配列の総当たりのペアワイズアラインメントに分解して，それぞれ独立に最適化する．各ペアワイズアラインメントは Needleman-Wunsch アルゴリズムによって高速に解くことができる．(2) マルチプルアラインメントが満たすべき制約 (10)-(13) と矛盾する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，マルチプル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マルチプルアラインメントを得る．

## 3. 結果

前節で述べたアルゴリズムに基づき DMSA を実装した．アラインメント確率行列 (6) を計算するために，ProbCons [1] を用いた．実験結果は発表時に示す．

### 参考文献

[1] Do, C. B., Mahabhashyam, M. S., Brudno, M. and Batzoglou, S.: ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res.*, Vol. 15, No. 2, pp. 330–340 (2005).

[2] Hamada, M. and Asai, K.: A classification of bioinformatics algorithms from the viewpoint of maximizing expected accuracy (MEA), *J. Comput. Biol.*, Vol. 19, No. 5, pp. 532–549 (2012).

[3] Notredame, C., Higgins, D. G. and Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J. Mol. Biol.*, Vol. 302, No. 1, pp. 205–217 (2000).

[4] Sato, K., Kato, Y., Akutsu, T., Asai, K. and Sakakibara, Y.: DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition, *Bioinformatics*, Vol. 28, No. 24, pp. 3218–3224 (2012).