岡山理科大学紀要 第44号 A pp. 81-90(2008)

Maximum Weight Trace問題に対する枝重みクリークを考慮した解構築法

西野 史芳・片山 謙吾*・南原 英生*・成久 洋之*
岡山理科大学大学院工学研究科情報工学専攻
*岡山理科大学工学部情報工学科

(2008年9月30日受付、2008年11月7日受理)

1. まえがき

本研究はMaximum Weight Trace問題を対象とし，枝重みクリークを考慮した解構築法を提案する．

Maximum Weight Trace問題とはマルチプルアラインメントに対する問題である．アラインメントとは，バイオインフォマティクスにおいて，複数のDNAの塩基配列やタンパク質のアミノ酸配列を並置し，相同性を求める手法である．相同性とは，複数の生物の共通祖先に由来する子孫間の類似性である．つまり，アラインメントによって共通祖先から分岐し，次第に進化していった複数の生物の共通点を見つけ出すことができる．バイオインフォマティクスが発達する以前，タンパク質の構造や機能の決定はタンパク質に対する直接的な実験により行われてきた．しかし，タンパク質に対する直接的な実験によって，タンパク質の構造や機能を決定するよりも，タンパク質に対応するDNAの配列に対し情報処理技術を用いて解析する方がはるかに易しい．そこで，情報処理技術を用いたアラインメントが重要な手法となっている[1]．特に，本研究で対象とするマルチプルアラインメントは3本以上の配列によるアラインメントである．マルチプルアラインメントに対する代表的な手法としては，厳密解法である動的計画法を複数の配列に対し，部分的に繰り返し適用させるClustalW[3]や，メタヒューリスティックアルゴリズムである遺伝的アルゴリズムを用いるSAGA[4]などが知られている．これらClustalWやSAGAなど既存の手法は，配列に対し，局所的な操作によって解を求めている．しかし局所的な操作では，配列が大規模になればなるほど，大域的に最適な解を求めにくくなる．そこで，本研究ではMaximum Weight Trace問題[5]を対象とする．Maximum Weight Trace問題では，アラインメントをアラインメントグラフとして表現する．アラインメントグラフを用いることで，配列の関係を広範囲に考慮することができる．また，このMaximum Weight Trace問題における最適解は，枝重みの合計が最大となる場合である．そこで本研究ではアラインメントグラフにおける枝重みクリークに着目した解構築法を提案する．

2. アラインメント

アミノ酸は20種類の文字からなる配列で表される．そして，複数の配列の相同性を求めるために整列させる操作がアラインメントである．特にペアワイズアラインメントについてはNeedleman-Wunsch法[7]によって最適なアラインメントを得ることができる．

2.1 アミノ酸配列

タンパク質はアミノ酸からなる化合物である．このアミノ酸は20種類あり，各アミノ酸は，以下に示すように1文字のアルファベットの20種類で表される．

A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y

したがって，あるタンパク質は1列のアミノ酸配列として表現される．このアミノ酸配列は生物が進化していく過程で，あるアミノ酸が他のアミノ酸と入れ替わる置換，新たにアミノ酸が加わる挿入，すでにあるアミノ酸が失われる欠失が行われる．しかし，進化していく過程の中で変化しないアミノ酸もある．そして，複数のタンパク質が同じ構造と機能を持っていれば相同性を保持しているという．

2.2 アラインメント

アラインメント (図 1) とは，複数の配列を縦に比較し，文字が一致するように，隙間を挿入し整列させる方法である．ここで，Σ を隙間である ' $-$ ' 以外の各アミノ酸を表す有限個の記号体系，' $-$ ' を含む有限個の記号体系を $\hat{\Sigma} = \Sigma \cup \{\text{`}-\text{'}\}$ とする．また各配列を $S_1, \ldots, S_n$，各配列の長さを $l_1, \ldots, l_n$ とし，S_n に含まれる l_n 個の各文字を $s_{n,1}, s_{n,2}, \ldots, s_{n,l_n}$ とする．このとき，$S_1, \ldots, S_n$ のアラインメント A は n 個の文字列 $\hat{S}_1, \ldots, \hat{S}_n \in \hat{\Sigma}$ からなる $n \times l$ 次元配列 $A = (a_{i,j})$ となる．この A は以下の特徴を持つ．

- $a_{i,j} \in \hat{\Sigma} \quad \forall 1 \leq i \leq n,\ 1 \leq j \leq n$
- $\hat{S}_i = S_i \setminus \{\text{`}-\text{'}\}$
- 縦列に隙間がない場合は $\max\{l_1, \ldots, l_n\} \leq n \leq \Sigma_{i=1}^{k} n_i$

A を得るための操作は $0 \leq a < n$，$0 \leq b < n$，$a \neq b$，$0 \leq \alpha < l_a$，$0 \leq \beta < l_b$ のとき以下のとおりである．

- $s_{a,l_\alpha}, s_{b,l_\beta} \in \Sigma$ かつ $s_{a,l_\alpha} \neq s_{b,l_\beta}$ の場合の置換
- $s_{a,l_\alpha} =$ ' $-$ ' かつ $s_{b,l_\beta} \in \Sigma$ の場合の挿入
- $s_{a,l_\alpha} \in \Sigma$ かつ $s_{b,l_\beta} =$ ' $-$ ' の場合の欠失

また，$n = 2$ の場合をペアワイズアラインメント，$n \geq 3$ の場合をマルチプルアラインメントと呼ぶ．

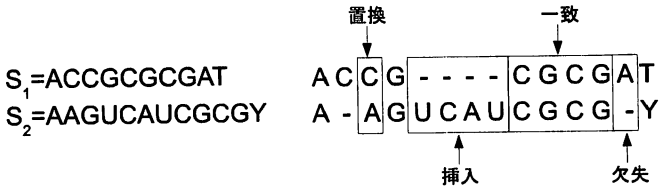

図 1　アラインメントの例

2.3 アラインメントスコア

最適なアラインメント A を得るためには，A を評価しなければならない．そのための得点をアライメントスコアと呼ぶ．アラインメントスコアを得るときは，スコア行列 (図 2)[6] を用いる．スコア行列は，置換されやすいものほど得点が高く，置換されにくいものほど得点が低く表されている．スコア行列により得られた得点に対し，隙間の長さに応じてペナルティを与えることで，アラインメントスコアを表す．スコア行列を m，アラインメントスコアを sc，隙間の数を g，ペナルティを p とすると，あるアラインメント A におけるアラインメントスコア sc は以下の式で得られる．

$$sc(A) = -(g \times p) + \Sigma_n m(\hat{S}_{1,l_n}, \hat{S}_{2,l_n})$$

2.4 Needleman-Wunsch 法

ペアワイズアラインメントは動的計画法を用いることで，最適なアラインメントを得ることができる．この動的計画法を用いることで最適なアラインメント A を得る方法を Needleman-Wunsch 法[7] と呼ぶ．2 本の配列 S_a，S_b について，最適な A を得るため，これらの配列をもとに $l_a \times l_b$ の 2 次元配列 $D_{i,j}$ を

$$\begin{aligned}
D_{0,0} &= 0 \\
D_{0,j} &= \Sigma_{n=1}^{j} m(-, s_{b,l_n}) \\
D_{i,0} &= \Sigma_{n=1}^{i} m(s_{a,l_n}, -) \\
D_{i,j} &= \max \left\{ \begin{array}{l} D_{i,j-1} + m(-, s_{b,l_j}) \\ D_{i-1,j-1} + m(s_{a,l_i}, s_{b,l_j}) \\ D_{i-1,j} + m(s_{a,l_i}, -) \end{array} \right\} \quad \forall i, j > 0
\end{aligned}$$

	G	A	S	T	P	L	I	M	V	D	N	E	Q	F	Y	W	K	R	H	C
G	5																			
A	1	2																		
S	1	1	2																	
T	0	1	1	3																
P	-1	1	1	0	6															
L	-4	-2	-3	-2	-3	6														
I	-3	-1	-1	0	-2	2	5													
M	-3	-1	-2	-1	-2	4	2	6												
V	-1	0	-1	0	-1	2	4	2	4											
D	1	0	0	0	-1	-4	-2	-3	-2	4										
N	0	0	1	0	-1	-3	-2	-2	-2	2	2									
E	0	0	0	0	-1	-3	-2	-2	-2	3	1	4								
Q	-1	0	-1	-1	0	-2	-2	-1	-2	2	1	2	4							
F	-5	-4	-3	-3	-5	2	1	0	-1	-6	-4	-5	-5	9						
Y	-5	-3	-3	-3	-5	-1	-1	-2	-2	-4	-2	-4	-4	7	10					
W	-7	-6	-2	-5	-6	-2	-5	-4	-6	-7	-4	-7	-5	0	0	17				
K	-2	-1	0	0	-1	-3	-2	0	-2	0	1	0	1	-5	-4	-3	5			
R	-3	-2	0	-1	0	-3	-2	0	-2	-1	0	-1	1	-4	-4	2	3	8		
H	-2	-1	-1	-1	0	-2	-2	-2	-2	1	2	1	3	-2	0	-3	0	2	6	
C	-3	-2	0	-2	-3	-6	-2	-5	-2	-5	-4	-5	-5	-4	0	-8	-5	-4	-3	12

図 2　スコア行列の例

と定義する．このとき $D_{i,j}$ =max となるのは，部分配列 s_{a,l_n}, s_{b,l_n} の一致または置換，挿入，欠失からなる各組合せのうち，最も得点の高い組合せ，すなわち最適なアラインメントスコアとなる組合せである．したがって，最適な A を得るには，最も高いアラインメントスコアとなる部分配列を通る経路を選択すればよい．

最適な A は，トレース行列を用いることで容易に手に入れられる．配列 S_a，S_b に対するトレース行列 $tr_{i,j}$ は，要素 $tr_{i,j} \subseteq \{\leftarrow, \uparrow, \nwarrow\}$ からなる行列と定義される．各要素は

$$
\begin{aligned}
tr_{0,0} &= \emptyset \\
tr_{0,j} &= \leftarrow \\
tr_{i,0} &= \uparrow \\
\nwarrow \in tr_{i,j} &\iff D_{i,j} = D_{i-1,j-1} + m(s_{a,l_i}, s_{b,l_j}) \quad \forall i,j > 0 \\
\uparrow \in tr_{i,j} &\iff D_{i,j} = D_{i-1,j} + m(s_{a,l_i}, -) \\
\leftarrow \in tr_{i,j} &\iff D_{i,j} = D_{i,j-1} + m(-, s_{b,l_j})
\end{aligned}
$$

となる．この $tr_{i,j}$ の右下角 tr_{l_a,l_b} から左上角 $tr_{0,0}$ にトレースバックする (図 3)．例えば，i 番目の $\nwarrow$ もしくは $\uparrow$ を s_{a,l_i} に一致または置換とし，すべての $\leftarrow$ を挿入とすることで，$\hat{S}_a$ を得る．同様に，j 番目の $\leftarrow$ もしくは $\nwarrow$ を s_{b,l_j} に一致または置換とし，すべての $\uparrow$ を欠失とすることで，$\hat{S}_b$ を得る．

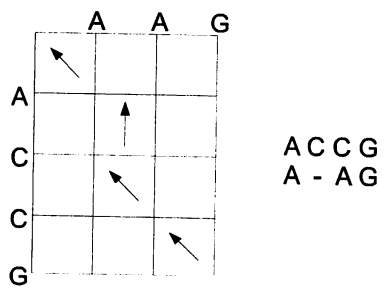

図 3　トレースバックにより得られるアラインメントの例

2.5 プログレッシブアラインメント

マルチプルアラインメントに対し最も一般的に用いられる方法がプログレッシブアラインメントである．この方法は，ペアワイズアラインメントを繰り返すことでアラインメントを計算する．具体的には，まず2本の配列に対しペアワイズアラインメントを行う．続いて，最初のアラインメントを1本の配列とみなし，3本目の配列とのアラインメントを計算する．この操作をすべての配列に対し行う．

2.6 案内木

プログレッシブアラインメントには，類似性の高い配列同士を最初にアラインメントするという方法がある．このことにより，アラインメント結果の信頼性を高めることができる．このときの道標となるものが案内木である．案内木では，木構造により配列の近縁関係を表現する．案内木の例を図4示す．各節点に子がある場合，その数は必ず2である．また，各枝長は配列間の距離を表現する．

この案内木を求める際に用いる方法として，近隣結合法[8]がある．近隣結合法は各配列の違いを数値で表し、最も距離の短いもの同士をまとめていき木構造を作る．

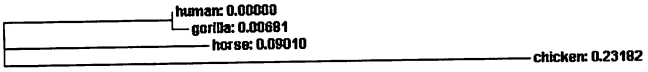

図4　案内木の例

2.7 ClustalW

プログレッシブアラインメントによるアラインメントはClustalW[3]と呼ばれるプログラムが広く利用されている．その流れを以下に示す．

1. 動的計画法によるペアワイズアラインメントを行い，各ペアに対する距離行列を求める．
2. 近隣結合法により案内木を求める．
3. 類似度が高い接点から低い接点へとアラインメントを順次計算する．

3. Maximum Weight Trace 問題

Maximum Weight Trace 問題はマルチプルアラインメントをグラフとして表す．しかし，このグラフから最適なアラインメントを得ることはNP困難である[5]．そのため，最適なアラインメントを得るために必要な情報のみからなるグラフを作る．

3.1 アラインメントグラフ

マルチプルアラインメントにおいて，n 個の配列の各文字を頂点とみなすと，頂点集合 V および各頂点間を結ぶ枝 e の集合 E からなる完全 n 部グラフ $G=(E,V)$ として表すことができる．そして，各枝には重み $w(e)$ がある．このとき V および E は

$$\begin{aligned} V &= s_{i,p},\quad \forall i=1,\ldots,n,\quad \forall p=1,\ldots,l_i \\ E &= \{(s_{i,p},s_{j,q})|1\le i<j\le n, 1\le p\le l_i, 1\le q\le l_j\} \end{aligned}$$

である．このように表されるグラフをアラインメントグラフ (図5) と呼ぶ．

3.2 Maximum Weight Trace 問題

ここであるアラインメント A を考える．この A によって表現される枝の全集合を A のトレース $T\subset E$ とする．そして，T の重みは $\Sigma_{e\in T}w(e)$ となる．このとき完全アラインメントグラフを元に $T\subset E$ から，最大の重みとなる T を選ぶことを Complete Maximum Weight Trace 問題と呼ぶ．同様に，アラインメントグラフを元に $T\subset E$ から，最大の重みとなる T を選ぶことを Maximum Weight Trace 問題と呼ぶ．

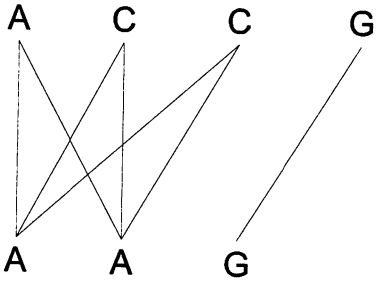

図5 アラインメントグラフの例

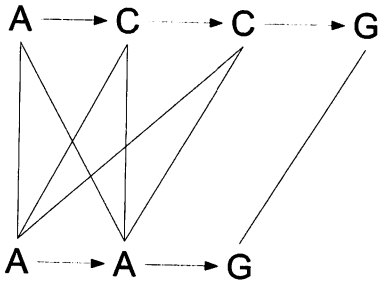

図6 拡張アラインメントグラフの例

3.3 アラインメントグラフの作成

Maximum Weight Trace 問題は NP 困難である．そこで，枝集合 E をアラインメントとなる可能性の高いものに限定することを考える．本研究では，ペアワイズアラインメントにおいて広く利用されている ClustalW を用いる．

3.4 Clustal W ' s similarity matrices with window filtering

アラインメントグラフにおける各枝重みの計算は，Clustal W's similarity matrices with window filtering (CSM) [9] を用いる．まず，対象とする配列に対し ClustalW を用いてアラインメントを得る．ここで，$\hat{S}$ を配列 S_i 及び S_j のペアワイズアラインメントの結果とする．このとき，$e_1, e_2, \ldots, e_m$ を $\hat{S}$ における m 個のアラインメントされたペア，$M(e_k)$ を e_k のスコアとする．また，任意の値 b からなる，e_k の前後 $(2 \cdot b + 1)$ の長さの範囲のスコアを反映させる．

$$\mathrm{CSM} = M(e_k) + \sum_{i=1}^{b}[M(e_k - i) \cdot (1 - \frac{i}{b+1}) \cdot g^-] + \sum_{i=1}^{b}[M(e_k + i) \cdot (1 - \frac{i}{b+1}) \cdot g^+]$$

$$g^- = \begin{cases} 1 & k-1 \text{から} k-i \text{の間にギャップがない場合} \\ 0 & \text{それ以外の場合} \end{cases}$$

$$g^+ = \begin{cases} 1 & k+1 \text{から} k+i \text{の間にギャップがない場合} \\ 0 & \text{それ以外の場合} \end{cases}$$

この計算により，各重みはその前後のアラインメントを反映した値になる．

3.5 拡張アラインメントグラフ

これまでに得られたアラインメントグラフを元にマルチプルアラインメントを得る際，トレースをグラフ理論的に特徴付ける必要がある[5]．そこで，拡張アラインメントグラフ (図6) を用いる．拡張アラインメントグラフでは新たに枝 $H = \{(s_{i,p}, s_{i,p+1}) | 1 \leq i \leq n, 1 \leq p \leq l_i - 1\}$ を定義し，矢印で表す．この枝 H の重みは 0 である．このとき拡張アラインメントグラフは $\overline{G} = (V, E, H)$ と表される．

このグラフ $\overline{G}$ では配列の頂点 $v_1, \ldots, v_n, n \geq 2$ について，$1 \leq i < n$ であり，かつ $(v_i, v_{i+1}) \in E$ または $(v_i, v_{i+1}) \in H$ とする．ここで，もし最初の頂点と最後の頂点が同じであるならば，このパスを cycle と呼ぶ．また，$\overline{G} = (V, E, H)$ が拡張アラインメントグラフであるならば，トレース T によって導かれた $T \subseteq E$ かつ $\overline{G}^T = (V, T, H)$ は拡張アラインメントグラフである．このとき，矢印を含まない $\overline{G}^T$ が全て cycle であったならば，T はトレース可能であると呼ぶ．

4. 解法

Maximum Weight Trace 問題は貪欲法によってアラインメントを得ることができる．しかし，貪欲法では最適なアラインメントが得られるとは限らない．そこで本研究では，貪欲法を用いる際にクリークの重みを考慮する．クリークの重みを考慮することで，よりよいアラインメントとなる枝が選ばれやすくなると考えられる．

4.1 貪欲法

Maximum Weight Trace 問題は貪欲法を用いることで，簡潔にアラインメントを得ることができる．この貪欲法ではアラインメントグラフに含まれる枝集合 E を重みについてソートし，重みの大きいトレース可能な枝 e から順にトレースに追加していく．そして，アラインメントグラフからトレース可能な枝が無くなったときに終了する．

しかし，この方法では大域的に最適なアラインメントになるとは限らない．例えば，ある e_a をトレースに追加して構成される部分配列よりも，e_a を追加することで，トレースに追加可能ではなくなった枝の部分集合 $E*\backslash e_a$ を追加した場合の部分配列の方がより良いアラインメントとなる場合がある．

4.2 クリークを考慮した解構築法

Maximum Weight Trace 問題において，最適解はトレースの枝重みの合計が最大となる場合である．これは，トレースに含まれる枝重みクリークにおける枝重みの合計が最大となる場合と等価である．また，このように枝重みクリークを考慮することで，貪欲法での問題点が解決することができると考えられる．そこで、本研究では枝重みクリークを考慮した解構築法を提案する．その手順を以下に示す．

1. 現在の探索空間からランダムな 1 頂点を初期解とする．そして，この頂点と他の頂点との間にトレース可能とならない枝を含む頂点を一時的に探索空間から除外する．
2. 初期解を含む最大枝重みクリークを求め，トレースに追加する．
3. トレースに追加した枝を含む頂点を探索空間から除き，一時的に除外した頂点を探索空間に戻す．もし，探索空間が無くなれば終了．さもなければ 1 へ戻る．

4.3 バイナリー 2 次計画問題

前節で述べた最大枝重みクリークを求める問題は最大枝重みクリーク問題 (Maximum Edge Weight Clique Problem, MEWCP) である．MEWCP はバイナリー 2 次計画問題に置き換えられることが知られている[14]．そこで，クリークを考慮した解構築法において，最大枝重みクリークを求める際に，バイナリー 2 次計画問題として解く．

このバイナリー 2 次計画問題 (BQP) とは $n \times n$ の対称行列 $Q = (q_{ij})$ が与えられたとき，次の目的関数を最大化する解を求める問題である．

$$f(x) = x^t Q x = \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} q_{ij} x_i x_j$$
$$x_i \in \{0,1\}, \qquad \forall\, i = 1, \ldots, n.$$

BQP は NP 困難であり，多数の応用例を有している．例えば，capital budgeting and financial analysis 問題，traffic message management 問題，マシンスケジューリング問題，分子構造問題などがある．さらに，BQP は様々な組合せ最適化問題と同等であることが知られている．例として，最大カット問題，最大クリーク問題，最大頂点パッキング問題，最小頂点カバー問題などがある[13]．

4.4 最大枝重みクリーク問題

最大枝重みクリーク問題 (maximum edge weighted clique problem, MEWCP) とは，ある完全グラフ $G = (V, E)$ が与えられたとき，枝重みの合計が最大となるクリークを求める問題である．これまでこの問題に対する解法の研究は線形モデルが中心であった．しかし，MEWCP は非線形モデルで形式化した場合の方がより自然に示すことが出来る[14]．完全グラフ G について，頂点を n，枝重みを c_{ij}，最大クリークサイズを b としたとき，MEWCP の 2 次形式は以下のとおりである．

$$\max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij} x_i x_j$$
$$\sum_{j=1}^{n} x_j \leq b$$
$$x_i \in 0, 1$$

このとき x_j が 1 ならば頂点 j はクリークに含まれる．さもなければ x_j は 0 である．これは BQP の目的関数でもある．

この目的関数をアラインメントグラフにおいて適用する場合は，以下のように拡張される．

$$\max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij} x_i x_j$$

$$\begin{aligned}
\sum_{j=1}^{n} x_j &\leq n \\
\sum_{i=0}^{S_1} x_i &\leq 1 \\
\sum_{i=S_1+1}^{S_2} x_i &\leq 1 \\
&\vdots \\
\sum_{i=S_{n-1}+1}^{S_n} x_i &\leq 1 \\
x_i &\in 0,1
\end{aligned}$$

このとき，与えられた配列が n 個であるとすると，$S_1, \ldots, S_2 - 1$ は一つ目の配列の各要素，$S_2, \ldots, S_3 - 1$ は二つ目の配列の各要素というように対応する．

5. 実験結果

クリークを考慮した解構築法を評価するために Maximum Weight Trace 問題に対する解法として貪欲法及び進化的アルゴリズム，Maximum Weight Trace 問題に対する解法ではない手法として ClustalW，SAGA と比較した．問題例はマルチプルアラインメントのベンチマークである BAliBASE[15] の Reference1 を使用した．また，評価値として用いているスコアは，BAliBASE におけるアラインメントスコア算出プログラム baliscore で求めた値である．この baliscore では最適なアラインメントのスコアを 1 とし，悪いアラインメントほど 0 に近づく．実験環境は CPU: Pentium 4 3.4GHz，RAM: 4GB，OS: Solaris10 である．なお，進化的アルゴリズムは文献[11]，SAGA は文献[16] より参照した．進化的アルゴリズムの時間については文献[11] に記載がないため省略している．表 1 に各手法の実験結果を示す．

表 1 より，提案法は貪欲法に比べ，スコアは多少低いが，高速な計算ができている．また，計算時間については SAGA よりも非常に高速になっている．貪欲法に比べスコアが低くなった理由としては，ある配列内の左側に寄っている文字と，その他の配列内の右側に寄っている文字のスコアが高ければ最大枝重みクリークとして選ばれるためであると考えられる．そのような場合は，ギャップが多く挿入されスコアも低くなる．そのため，ひとつの最大枝重みクリークを求めるのではなく，その前後の枝重みクリークとの合計が大きくなるような枝重みクリークを求める必要がある．一方，高速な計算が行われた理由としては，提案法は探索空間を狭めていくことで，探索する枝を限定しているためであると考えられる．ClustalW と比較すると，アラインメントグラフの枝重みは ClustalW を用いているにもかかわらず，スコアが改善している場合がある．これは，枝重みを求める際に，ある枝に対してその前後のスコアも考慮に入れることで、前後のアミノ酸の関係を考慮できたためであると考えられる．進化的アルゴリズムについては，多くの問題例において非常に高いスコアを示している．そして，進化的アルゴリズムにおいて使われている解構築法は貪欲法である．提案法は解構築法であるため、このような進化的アルゴリズムを用いることでより良いスコアを得ることができると考えられる．さらに，貪欲法に比べ高速な提案法を用いることで，より高速な解法を得ることができると考えられる．

表 1　実験結果

	提案法		貪欲法		進化的アルゴリズム		ClustalW		SAGA	
問題例	スコア	時間 (秒)	スコア	時間 (秒)	スコア	時間 (秒)	スコア	時間 (秒)	スコア	時間 (秒)
1aab	0.679	0.06	0.679	0.11	1.000	NA	0.679	0.01	0.839	4.00
1aboA	0.407	0.09	0.604	0.14	0.768	NA	0.625	0.01	0.521	19.00
1aho	0.632	0.11	0.642	0.20	1.000	NA	0.617	0.02	0.960	8.00
1csp	0.939	0.10	0.939	0.23	0.982	NA	0.939	0.02	0.955	5.00
1csy	0.642	0.23	0.684	1.16	0.977	NA	0.755	0.03	0.888	5.00
1dox	0.865	0.13	0.867	0.25	0.937	NA	0.857	0.02	0.864	10.00
1fjlA	0.938	0.15	0.923	0.50	1.000	NA	0.984	0.02	0.991	7.00
1fkj	0.798	0.21	0.786	0.74	1.000	NA	0.820	0.03	0.954	10.00
1fmb	0.864	0.12	0.881	0.30	0.983	NA	0.913	0.02	0.972	5.00
1hfh	0.613	0.32	0.597	1.35	0.966	NA	0.655	0.04	0.903	21.00
1hpi	0.783	0.06	0.771	0.13	0.989	NA	0.783	0.01	0.901	6.00
1idy	0.037	0.07	0.132	0.05	0.676	NA	0.377	0.01	0.348	6.00
1krn	0.924	0.12	0.924	0.35	1.000	NA	0.938	0.02	0.981	9.00
1pfc	0.602	0.29	0.590	1.00	0.986	NA	0.627	0.03	0.913	17.00
1plc	0.725	0.19	0.789	0.58	0.976	NA	0.779	0.03	0.951	10.00
1r69	0.194	0.06	0.175	0.08	0.483	NA	0.194	0.01	0.563	7.00
1tgxA	0.719	0.05	0.702	0.06	0.935	NA	0.724	0.01	0.760	5.00
1tvxA	0.059	0.05	0.132	0.06	0.583	NA	0.094	0.01	0.456	8.00
1wit	0.353	0.20	0.345	0.60	1.000	NA	0.378	0.02	0.810	12.00
1ycc	0.637	0.17	0.632	0.34	0.918	NA	0.683	0.02	0.779	10.00
2mhr	0.916	0.26	0.944	1.06	0.985	NA	0.963	0.04	0.961	11.00
2trx	0.395	0.09	0.386	0.21	0.737	NA	0.448	0.01	0.685	8.00
3cyr	0.568	0.15	0.565	0.34	0.898	NA	0.576	0.02	0.849	14.00
451c	0.307	0.15	0.385	0.36	0.820	NA	0.383	0.02	0.637	20.00
9rnt	0.891	0.20	0.859	0.74	0.995	NA	0.898	0.03	0.978	12.00
1ad2	0.707	0.65	0.707	3.17	0.963	NA	0.721	0.07	0.881	22.00
1amk	0.871	1.47	0.911	12.29	1.000	NA	0.944	0.14	0.978	32.00
1ar5A	0.865	0.63	0.870	2.62	0.985	NA	0.891	0.06	0.977	17.00
1aym3	0.756	0.81	0.825	4.31	0.950	NA	0.835	0.09	0.915	41.00
1bbt3	0.213	0.74	0.224	2.72	0.445	NA	0.344	0.06	0.522	52.00
1ezm	0.933	2.06	0.928	22.83	0.955	NA	0.933	0.20	0.926	119.00
1gdoA	0.722	0.89	0.726	4.77	0.927	NA	0.776	0.10	0.876	92.00
1havA	0.162	0.71	0.186	2.39	0.417	NA	0.311	0.06	0.311	64.00
1ldg	0.848	1.47	0.850	10.28	0.993	NA	0.846	0.15	0.938	31.00
1led	0.780	0.83	0.780	4.63	0.970	NA	0.793	0.09	0.834	28.00
1mrj	0.755	2.19	0.736	5.83	0.997	NA	0.773	0.10	0.923	63.00
1pgtA	0.865	0.63	0.859	3.04	1.000	NA	0.913	0.07	0.924	22.00
1pii	0.608	0.95	0.629	5.91	0.873	NA	0.622	0.10	0.848	83.00
1ppn	0.948	1.02	0.948	7.73	0.987	NA	0.948	0.10	0.985	50.00
1pysA	0.894	1.02	0.911	5.65	0.984	NA	0.913	0.10	0.922	38.00
1sbp	0.258	1.46	0.310	7.81	0.578	NA	0.304	0.13	0.440	85.00
1thm	0.773	1.10	0.752	6.82	0.972	NA	0.788	0.11	0.893	48.00
1tis	0.910	1.80	0.905	18.85	0.940	NA	0.901	0.18	0.947	103.00
1ton	0.567	1.17	0.502	9.33	0.955	NA	0.547	0.13	0.849	115.00
1uky	0.313	0.60	0.328	2.18	0.724	NA	0.289	0.06	0.443	62.00
1zin	0.909	0.66	0.900	3.31	0.934	NA	0.894	0.07	0.955	24.00
2cba	0.489	1.46	0.462	11.10	0.968	NA	0.464	0.12	0.831	63.00
2hsdA	0.248	0.79	0.222	4.34	0.696	NA	0.252	0.08	0.582	41.00
2pia	0.483	1.06	0.502	5.03	0.785	NA	0.571	0.11	0.623	94.00
3grs	0.093	0.75	0.115	2.63	0.469	NA	0.193	0.06	0.451	47.00
5ptp	0.850	3.10	0.842	9.62	0.961	NA	0.865	0.12	0.918	66.00
kinase	0.477	4.16	0.475	11.58	0.751	NA	0.471	0.14	0.644	68.00
1ac5	0.657	3.43	0.675	25.77	0.997	NA	0.677	0.29	0.798	326.00
1ad3	0.895	3.24	0.898	28.39	0.988	NA	0.896	0.28	0.739	84.00
1adj	0.755	2.81	0.829	23.99	1.000	NA	0.864	0.25	0.956	42.00
1ajsA	0.153	1.92	0.237	12.94	0.473	NA	0.251	0.17	0.503	79.00
1cpt	0.447	2.68	0.476	20.58	0.829	NA	0.466	0.21	0.774	121.00
1dlc	0.723	6.89	0.709	65.88	0.960	NA	0.722	0.50	0.826	185.00
1eft	0.675	2.16	0.674	18.09	0.935	NA	0.682	0.21	0.891	83.00
1fieA	0.868	20.46	0.873	104.21	0.984	NA	0.884	0.64	0.846	138.00
1gowA	0.680	3.67	0.731	33.05	0.922	NA	0.716	0.32	0.714	82.00
1gpb	0.892	56.09	0.906	439.97	0.994	NA	0.945	1.32	0.962	398.00
1gtr	0.798	5.42	0.811	67.72	0.991	NA	0.883	0.38	0.936	118.00
1lcf	0.814	60.29	0.818	492.09	0.997	NA	0.897	1.27	0.723	705.00
1lvl	0.819	57.98	0.818	504.59	0.557	NA	0.897	1.27	0.395	122.00
1pamA	0.270	14.49	0.264	67.77	0.600	NA	0.402	0.48	0.414	318.00
1ped	0.515	1.00	0.562	4.28	0.682	NA	0.600	0.11	0.812	35.00
1pkm	0.765	3.12	0.805	30.34	0.987	NA	0.808	0.28	0.895	71.00
1rthA	0.748	18.81	0.774	115.51	0.966	NA	0.793	0.59	0.908	351.00
1sesA	0.842	9.12	0.850	63.24	0.992	NA	0.857	0.40	0.907	257.00
2ack	0.113	10.22	0.493	75.56	0.917	NA	0.521	0.46	0.701	307.00
2myr	0.317	2.54	0.360	18.95	0.475	NA	0.400	0.24	0.185	200.00
3pmg	0.903	7.20	0.893	58.62	0.998	NA	0.879	0.44	0.937	224.00
4enl	0.532	1.04	0.599	5.50	0.626	NA	0.528	0.10	0.399	30.00
actin	0.836	4.89	0.834	44.76	0.991	NA	0.845	0.31	0.917	160.00
arp	0.674	5.73	0.669	48.65	0.945	NA	0.684	0.34	0.843	203.00
gal4	0.231	3.30	0.244	26.69	0.577	NA	0.350	0.27	0.403	168.00
glg	0.614	11.01	0.615	78.08	0.999	NA	0.633	0.47	0.856	285.00
平均	0.636	12.29	0.670	95.20	0.861	NA	0.695	0.45	0.740	195.85

6. むすび

本論文ではMaximum Weight Trace問題を対象とした．Maximum Weight Trace問題の最適解は複数の枝重みクリークにおける枝重みの合計が最大となる場合である．そこで，我々はMaximum Weight Trace問題に対するクリークを考慮した解構築法を提案した．提案法は枝を一つ一つ調べていくのではなく，クリークに着目している．そのため，従来の手法に比べ，効率的に解を構築することができると考えられる．そして，実験結果より，最大枝重みクリークを順次トレースに追加することで貪欲法に対し高速化することができた．今後の課題としては，複数の枝重みクリーク全体の枝重みの合計が最大となるクリークを探索すること，また，解改善法と組み合わせることが挙げられる．

参考文献

1) R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison, "バイオインフォマティクス–確率モデルによる遺伝子配列解析–," 医学出版, 2001.

2) P. Clote and R. Backofen, "統計物理化学から学ぶバイオインフォマティクス," 共立出版, 2004.

3) J. D. Thompson, D. G. Higgins and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice," Nucleic Acid Research, Vol. 22, pp.4673–4680, 1994.

4) C. Notredame and D.G. Higgins, "SAGA: Sequence Alignment by Genetic Algorithm," Nucleic Acid Research, Vol. 24, pp.1515–1524, 1996.

5) John D. Kececioglu, "The Maximum Weight Trace Problem in Multiple Sequence Alignment," In Proceedings of the 4th Symposium on Combinatorial Pattern Matching, LNCS(684), pp.106–119, Springer, 1993.

6) M. O. Dayhoff, "Atlas of protein sequence and structure," National Biomedical Reserch foundation, Georgetown University, Washington, D.C., Vol.5, pp.89–99, 1972.

7) S. B. Needleman and C. D. Wunsch, "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.

8) Saitou N, Nei M, "'The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees," Mol Biol Evol, 4(4), pp.406–425, 1987.

9) J. Fauster. "*Neue heuristische Lösungsansätze für das Multiple Sequence Alignment Problem*," Vienna University of Technology, Vienna, 2003.

10) G. Koller and G. R. Raidl, "An Evolutionary Algorithm for the Maximum Weight Trace Formulation of the Multiple Sequence Alignment Problem," Parallel Problem Solving from Nature, pp.302–311, 2004.

11) S. Leopold, "An Alignment Graph based Evolutionary Algorithm for the Multiple Sequence Alignment Problem," Master's thesis, Vienna University of Technology, Institute of Computer Graphics and Algorithms, February 2004.

12) C.Notredame, D. Higgins, and J. Heringa, "T-COFFEE: A novel method for fast and accurate multiple sequence alignment," Journal of Molecular Biology, 392, pp.205–217, 2000.

13) J.E. Beasley, "Heuristic algorithms for the unconstrained binary quadratic programming problem," Technical Report, Management School, Imperial College, UK, 1998.

14) Alidaee,B., Glover,F., Kochenberger,G., Wang,H., "Solving the Maximum Edge Weight Clique Problem via Unconstrained Quadratic Programming," European Journal of Operational Research, 181, pp.592–597,2007.

15) J. D. Thompson, F. Plewniak and O. Poch, "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs," Bioinformatics, Vol.15, pp.87–88, 1999.

16) 畝 哲広, "多重配列アライメントの Maximum Weight Trace 法に対する解法 ", 岡山理科大学, 岡山理科大学大学院工学研究科情報工学専攻修士学位論文 2005.

A Construction Method with Maximum Edge Weight Clique Finding for the Maximum Weight Trace Problem

Fumiyoshi NISHINO, Kengo KATAYAMA*,
Hideo MINAMIHARA* and Hiroyuki NARIHISA*

Graduate School of Engineering,
**Department of Information and Computer Engineering, Faculty of Engineering,*
Okayama University of Science.
1-1 Ridai-cho, Okayama, 700-0005, Japan.

(Received September 30, 2008; accepted November 7, 2008)

Most of the current techniques such as ClustalW and SAGA to the multiple alignment problem rely on local information to search near-optimal solutions. On the other hand, the maximum weight trace problem with alignment graph can express global information. The maximum weight trace problem is NP-hard. An optimal solution of the maximum weight trace problem is equal to one in which the total amount of edge weights in the solution graph is maximized. In this paper, we show a construction heuristic method that takes into account finding maximum edge weight cliques. We compare the method with the greedy one, evolutionary algorithm, ClustalW and SAGA on the benchmark set of the multiple sequence alignment problem. The outcomes show that some results of computation time are improved by our construction method with clique finding.

Keywords: multiple alignment; maximum weight trace problem; edge weight clique.