情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-BIO-22 No.3
2010/7/28

ペア間推移量を用いた配列アライメント法

原　　利英[†1]　佐藤圭子[†1]　大矢雅則[†1]

アミノ酸配列に対するアライメントにおいて，アミノ酸ペア間推移量を用いるようアルゴリズムを改良することで，生成されるアライメントの精度が向上することを確認した．本発表では，この新たに開発したアライメント手法である MTRAP 法と，既存の手法とのアライメント精度面での比較および検証を行う．

A sequence alignment algorithm using the transition quantity

Toshihide Hara,[†1] Keiko Sato[†1]
and Masanori Ohya[†1]

We have been developed a sequence alignment algorithm using the transition quantity. The transition quantity is a new measure based on transition probability between two consecutive pairs of residues. In this paper, we compare the performance of our new algorithm called MTRAP and that of the other usual algorithm.

1. はじめに

現在，ヒトゲノム計画の進展に代表されるように多種多様な生物のゲノム配列が決定され，扱われるデータ量は指数的に増え続けている．特に，2007 年ごろに次世代ゲノムシーケンサとよばれる装置が登場してからは，ゲノム配列の読み取りが革新的に高速化し，その後に続く配列解析の精度および速度の重要性が今までになく高まっている時代であるといえる．

現在までに，様々な配列アライメント法が開発されてきた．FASTA[18] や BLAST[2] に代表されるデータベース検索を目的としたアライメント法では，その速度に重点が置かれる一方，マルチプルアライメントを目的としたアライメント法ではその正確さに重点が置かれ開発されている．現在，こうしたマルチプルアライメントを行うための手法として，ClustalW[21], DIALIGN[14], T-Coffee[16], MAFFT[11], MUSCLE[7], Probcons[6] などが開発されている．これらの手法は一般的に各サイトの残基を独立したものとして扱っており，たとえばアミノ酸配列に対するアライメントであればアミノ酸置換行列をもとにした配列間尺度を利用する．これらの手法は，配列一致率が 40%以上である相同配列に対しては大変よい結果が得られるが，配列一致率がこの値以下である相同配列に対しては満足な結果が得られていない[4]．

Anfinsen のドグマ[3] として知られているように，少なくとも球形タンパクにおいてはその高次構造はそのタンパクを構築するアミノ酸配列により決定づけられていることが知られている．タンパク質は構造的にはアミノ酸のポリマーであるが，一部のタンパク質は自己組織化やシャペロンの影響により α ヘリックスや β シートといった特定の立体構造をとるように自動的に折りたたまれ，全体としては決まった構造をとる．この現象のことをフォールディングといい，タンパク質はフォールディングされることで，酵素などとしての特有の機能を発揮するとされる．つまり，配列を構成するアミノ酸の種類および前後のアミノ酸とのつながりに高次構造を決定する要因があると思われる．そしてこのことは，前後のアミノ酸の情報，つまり配列から得られる情報を含めアライメントを行うことで，タンパク質の立体構造的な対応をより正確に反映したアライメントを得られることを示唆する．こうした考えの下，我々は Transition quantity と呼ぶ量を定め，これを用いたアライメント法として MTRAP 法を提案した[8),9]．本報告では，この手法によるアライメント精度について，HOMSTRAD (version November 1, 2008)[13),20]，PREFAB4[7] といったデータベースを用いた場合の結果を示し，既存の手法と精度面で比較，検証する．

2. Transition Quantity を用いた配列間尺度

最初にいくつか記号を定義する．Ω をすべてのアミノ酸の集合，Ω^* を Ω とギャップ”$*$”による集合：$\Omega^* \equiv \Omega \cup \{*\}$ とする．Ω の要素を残基と呼び Ω^* の要素をシンボルと呼ぶ．Ω の直積を $\Gamma \equiv \Omega \times \Omega$ とし，同様に $\Gamma^* \equiv \Omega^* \times \Omega^*$ とする．

ここで，配列長 n の 2 つの配列，$A = a_1 a_2 \cdots a_n$ と $B = b_1 b_2 \cdots b_n$，$a_i, b_j \in \Omega^*$ について考える．この配列を $u_1 u_2 \cdots u_n$, $u_i = (a_i, b_i) \in \Gamma^*$ とも表記することにする．以下，u_i をサイトと呼ぶ．

配列間に何らかの関連性がある場合と，配列間に何の関連性もない場合との尤度比はオッ

†1 東京理科大学
Tokyo University of Science

ズ比と呼ばれる．

$$R(A,B)=\frac{p(A;B)}{p(A)\,p(B)}=\frac{p(a_1,a_2,\cdots,a_n;b_1,b_2,\cdots,b_n)}{p(a_1,a_2,\cdots,a_n)\,p(b_1,b_2,\cdots,b_n)}. \tag{1}$$

ここで，$p(a)$ は a の生起する確率をあらわし，$p(a;b)$ は同時確率を表す．式 1 を簡単なものにするにあたり，置換は位置独立に起き，サイト間での相関もないものと仮定する．つまり，$p(A)=\prod p(a)$, $p(B)=\prod p(b)$ and $p(A,B)=\prod p(a,b)$．このとき，式 1 の対数をとったものは独立した項を加算したものとして表せ, これは対数オッズ比と呼ばれる．

$$\log\frac{p(A;B)}{p(A)\,p(B)}=\sum_i s(a_i,b_i), \tag{2}$$

ここで，

$$s(a,b)=\log\frac{p(a;b)}{p(a)\,p(b)} \tag{3}$$

はシンボル a,b が何の関連性もなく生起する場合と何かしら関連性を持って生起する場合との対数尤度比である．この $s(a,b)$ はスコアと呼ばれ，$\mathrm{S}=(s(a,b))$ は置換行列と呼ばれる．ペアワイズアライメントで用いられる配列間尺度は一般的にこれらの量（式 2 および式 3）を用いて定義される[1)]．

ここで，置換行列の各要素を正規化したもの（0 から 1 の間の値を取るようにしたもの）を正規化置換行列として新たに定義し，これを用いて対数オッズ比の別表現として差異 $d_{\mathrm{sub}}(A,B)$ を新たに定める．まず，正規化のための関数 $f_s:[s_{\min},s_{\max}]\mapsto\mathrm{R}$ を次のように定める．

$$f_s(x)\equiv\frac{s_{\max}-x}{s_{\max}-s_{\min}},\ 0\le f_s(x)\le 1, \tag{4}$$

$$s_{\max}\equiv\max\left\{\max_{u\in\Gamma}\{\mathrm{S}(u)\},\ \text{gap cost}\right\},$$

$$s_{\min}\equiv\min\left\{\min_{u\in\Gamma}\{\mathrm{S}(u)\},\ \text{gap cost}\right\}.$$

この関数を用い，スコア $s(a,b)$ を正規化したものを $\tilde{s}(a,b)\equiv f_s(s(a,b)),a,b\in\Omega$ とする．また，正規化置換行列を $\mathrm{M}=(\tilde{s}(a,b))$ と定める．このとき，配列 A,B 間の差異は

$$d_{\mathrm{sub}}(A,B)=\sum_i\tilde{s}(a_i,b_i). \tag{5}$$

と表される．差異 $d_{\mathrm{sub}}(A,B)$ は配列 A,B が同じときに 0 となる．

この加算的で扱いやすい配列間差異は，上述したように置換は位置独立に起き，サイト間での相関もないものとの仮定のもとで導かれたものである．これに対し我々は，サイト間の相関を考慮する配列間差異を新たに提案した[9)]．

我々の提案手法では，既存の配列間尺度 $R(A,B)$ にサイト間推移の効果を加えた，次の尺度を新たに考える．

$$R_{\mathrm{our}}(A,B)=R(A,B)^{1-\varepsilon}\,R_t(A,B)^{\varepsilon}, \tag{6}$$

ここで，

$$R_t(A,B)\equiv\prod_{i=1}^{n-1}p(u_{i+1}\setminus u_i) \tag{7}$$

はサイト間推移の効果を表し，ε はその混合比率を表す．

加算的な差異を導出するにあたり，Transition quantity と呼ぶ正規化した推移量 $\tilde{t}(u_i,u_{i+1})$ を次のように定義する．

$$\tilde{t}(u_i,u_{i+1})\equiv f_t(t(u_i,u_{i+1});u_i),$$

$$t(u_i,u_{i+1})\equiv\log p(u_{i+1}\setminus u_i), \tag{8}$$

$$f_t(x;u)=\begin{cases}\dfrac{-x}{\max_{v\in\Gamma^*}\{-t(u,v)\}}, & \text{if } x>0\\ 1, & \text{otherwise}\end{cases} \tag{9}$$

サイト間推移を下にした配列間差異は Transition quantity の和として次のように与えられる．

$$d_{\mathrm{trans}}(A,B)=\sum_{i=1}^{n-1}\tilde{t}(u_i,u_{i+1}). \tag{10}$$

これら 2 つの差異 d_{sub} と d_{trans} を合わせた差異を次のように定義する．

$$d_{\mathrm{MTRAP}}(A,B)=(1-\varepsilon)\,d_{\mathrm{sub}}(A,B)+\varepsilon d_{\mathrm{trans}}(A,B). \tag{11}$$

これは既存のスコアリングシステムと同様に加算的で扱いやすい配列間尺度である．我々の開発した MTRAP 法[8)] は，この尺度を用いた動的計画法[17)] によるアライメント法である（図 1）．

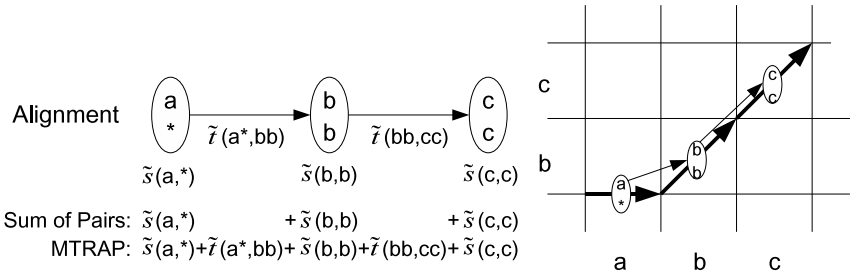

図 1 MTRAP 法

Vol.2010-BIO-22 No.3
2010/7/28

Transition quantity は PAM[5] や BLOSUM[10] といったアミノ酸置換行列と同様に既知アライメントデータセット上での統計を下にその値を求める．本論文では，SABmark データベース（version 1.63）[22] 上の superfamilies サブセットの配列すべてを用いて文献[9] と同様の手法により値を求めた．

2.1 アライメント精度の検証

この節では，我々の手法を含む各アライメント法により生成されたアライメントの精度を検証する方法について述べる．

タンパクの立体構造上の対応がそろうように整列化されたアミノ酸配列群を構造アライメント（Structural Alignment）と呼び，構造アライメントが登録されているデータベースを構造アライメントデータベースと呼ぶ．構造アライメントデータベースには現在，HOMSTRAD，PREFAB 4.0 といったものが存在する．ここで，これらのデータベースを用いアライメント精度の検証を行うことを考える．

構造アライメントデータベース上のアライメントを正しく対応がそろえられたアライメントである仮定し，これを便宜的にリファレンスアライメントと呼ぶことにする．また，リファレンスアライメントの元となるアライメント前の配列群に対し，各アライメント法を適用し構築したアライメントをテストアライメントと呼ぶことにする．この２つのアライメントを比較することで各アライメント構築法の評価を行う．具体的には次の手順となる．

(1) 構造アライメントデータベースからアライメントを取得し，これをリファレンスアライメントとする
(2) リファレンスアライメントからギャップを取り除いた配列群を作成する
(3) (2) で作成した配列群に対し，評価したいアライメント構築法でアライメントを作成する．これがテストアライメントとなる．
(4) リファレンスアライメントとテストアライメントを下で述べる指標 Q Score を用いて比較する
(5) 以上の (1) から (4) の作業を構造アライメントデータベースに登録されているデータすべてに対し行う．
(6) 以上の (1) から (5) の作業を比較したい手法それぞれにおいて行う．

本論文において利用した構造アライメントデータベースの詳細は以下の通り．

HOMSTRAD

HOMSTRAD(HOMologous STRucture Alignment Database) は立体構造既知な相同タンパクを用いた構造アライメントデータベースである[13]．随時データベースの内容が更新される．そのため，本論文では 2008 年７月１日時点でのデータベースを利用した．利用したデータはペアワイズアライメント 630 個であり，これら 630 個の配列一致率の分布は表 1 に示した通りとなる．

表 1 HOMSTRAD

0 ≤%ID< 20	20 ≤%ID< 40	40 ≤%ID< 60	60 ≤%ID< 80	80 ≤%ID≤ 100	ALL
87	273	160	83	27	630

表中の値は，2008 年 7 月 1 日時点での HOMSTRAD におけるペアワイズアライメント全 630 個のそれぞれの %ID（配列一致率）範囲における個数を表す．

PREFAB4

PREFAB4 はアライメント構築法のひとつである MUSCLE の作者らがアライメント法の評価のために作成したアライメントデータベースである[7]．本論文ではバージョン 4 にあたる PREFAB4 を用いた．利用したデータはペアワイズアライメント 1682 個であり，これらリファレンスアライメントの配列一致率の分布は表 2 の通りである．

表 2 PREFAB4

0 ≤%ID< 15	15 ≤%ID< 30	30 ≤%ID< 45	ALL
423	917	148	1682

表中の値は，PREFAB4 におけるペアワイズアライメント全 1682 個のそれぞれの%ID（配列一致率）範囲における個数を表す．

リファレンスアライメントとテストアライメントを比較するにあたり，指標 Q Score[7] を用いた．Q Score の定義を以下に示す．

Q Score とは， テストアライメントにおける残基ペアがリファレンスアライメント上において同じ列に存在しペアをつくる割合を表す．数式による定義は次の通り．

長さが L である N 本の配列から構成されるテストアライメント $\{s_1, \cdots, s_N\}$ が与えられ，$a_{ik} \in \Omega^*$ を配列 s_i における k 番目のシンボルとする．配列 s_i 上のシンボル a_{ik} と対応するリファレンスアライメント上のシンボルの列番号を I_{ik} とする．ただし，$a_{ik} = *$ の

Vol.2010-BIO-22 No.3
2010/7/28

ときは $I_{ik} = 0$ ．このとき，Q Score は以下のように与えられる．

$$\text{Q Score} = \frac{\sum_{k=1}^{L}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\Delta_{a_{ik},a_{jk}}\delta_{I_{ik},I_{jk}}}{\sum_{k=1}^{L}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\Delta_{a_{ik},a_{jk}}},$$

$$\Delta_{\mathrm{x,y}} = \begin{cases} 1, & x \neq * \quad \text{and} \quad y \neq * \\ 0, & x = * \quad \text{or} \quad y = * \end{cases}.$$

2.2 各種アライメント構築法との精度の比較

上述の構造アライメントデータベース HOMSTRAD，PREFAB4 を用い，MTRAP のアライメント精度に関して次の一般的に用いられる７つの手法：Needle，ClustalW2，MAFFT，T-Coffee，DIALIGN，MUSCLE，Probcons と比較を行った．各手法の詳細は以下の通り．

(1) Needle: Needle は Needleman-Wunsch アルゴリズム[15] によりグローバルペアワイズアライメントを行う EMBOSS パッケージ[19] のプログラムである．BLOSUM62 アミノ酸置換行列をデフォルトのアミノ酸置換行列として用いる．EMBOSS ver. 5.0.0 を用いた．

(2) ClustalW2: ClustalW2[12),21)] は累進法を実装した代表的なプログラムである．彼らの論文には明記されていないが，ClustalW2 は入力配列の情報を下に指定したシリーズの中からアミノ酸置換行列を選択し用いるアルゴリズムを実装している．GONNET アミノ酸置換行列群をデフォルトのアミノ酸置換行列として用いる．ClustalW2 ver. 2.0.9 を用いた．

(3) MAFFT: MAFFT[11] はフーリエ変換を用いた高速なアルゴリズムを実装するプログラムであり，ver. 6.240 を用いた．

(4) T-Coffee: T-Coffee[16] はマルチプルアライメント構築時の目的関数として配列一致率を下にしたものを利用する累進法によるマルチプルアライメント構築のための手法及びその手法を実装したプログラムの名称である．アルゴリズムの詳細は**??**節を参照のこと．現在，累進法に分類されるアルゴリズムの中では最高水準の精度を有するとされる．Ver. 5.30 を用いた．

(5) DIALIGN: DIALIGN[14] は segment-to-segment アプローチによる手法を用いたプログラムであり，ver. 2.2.1 を用いた．

(6) MUSCLE: MUSCLE[7] は Log-Expectation を用いた手法を用いたプログラムであり，ver. 3.7 を用いた．

(7) Probcons: Probcons[6] は Probabilistic Consistency を用いたプログラムであり，ver. 1.12 を用いた．

これらのプログラムは基本的にそれぞれのデフォルトパラメータを用いた．

提案手法の評価は次の２つを比較することで可能となる．

(1) 提案手法によるアライメントとリファレンスアライメントとの間の Q Score

(2) 上にあげた各手法によるアライメントとリファレンスアライメントとの間の Q Score

各構造アライメントデータベース上の複数のデータに対しそれぞれの手法における Q Score の計算を行うことで，これら２手法の比較を行った．

2.3 結果と考察

表３は MTRAP と代表的なグローバルアライメントプログラムである Needle，ClustalW2 との HOMSTRAD を用いた比較結果である．各手法による配列アライメントと HOMSTRAD 上の全 630 個のアライメントとの類似性は指標 Q score により測った．HOMSTRAD 上の構造アライメントを正しいアライメントだとすると，MTRAP は他の２つの手法にくらべ全範囲にわたって良い傾向を示すことがわかる．たとえば，MTRAP は 80%以上の精度（e.g., PAM250 や BLOSUM622 で 0.817）を有するのに対し，Needle や ClustalW2 は 80%未満の精度（e.g.，Needle は PAM250 で 0.768，BLOSUM62 で 0.768）となる（表 3）．それ以上に重要な点として，配列一致率が 30%未満のデータに対し，MTRAP はア

表 3 HOMSTRAD を用いた場合における MTRAP 法とその他の手法との比較

Matrix	Sequence identity （%）			
Method	0-15% (25)	15-30% (207)	30-45% (173)	ALL (630)
PAM250				
MTRAP	**0.421**	**0.655**	**0.874**	**0.817**
Needle	0.226	0.548	0.837	0.763
ClustalW2	0.234	0.528	0.817	0.747
BLOSUM62				
MTRAP	**0.410**	**0.653**	**0.878**	**0.817**
Needle	0.223	0.556	0.843	0.768
ClustalW2	0.276	0.585	0.861	0.784
GONNET250*				
MTRAP	**0.412**	**0.659**	**0.879**	**0.819**
ClustalW2	0.313	0.619	0.867	0.800

表中の値は HOMSTRAD における各配列一致率範囲での，平均 Q Score 値を表す．括弧内の数字は各配列一致率におけるアライメント数を表す．太字は各配列一致率および各アミノ酸置換行列における一番良い値を表す．
*Needle は GONNET アミノ酸置換行列をサポートしない．

ライメント精度を大変よく改善している点があげられる．例えば，PAM250 行列を用いた MTRAP では配列一致率が 0-15%のデータに対し 0.421，15-30%のデータに対し 0.655 といった精度が得られるのに対し，同様に PAM250 行列を用いた ClustalW2 では配列一致率が 0-15%のデータに対し 0.234，15-30%のデータに対し 0.528 といった精度にとどまる．

正解とする構造アライメントデータベースとして HOMSTRAD のほかに，PREFAB4 を用いた検証も行った．ここでは，PREFAB4 上の全 1682 個のデータを用い，HOMSTRAD と同様指標 Q Score による評価を行った．図 2 は他の各プログラム（Needle および ClustalW2）における平均 Q Score 値に対する MTRAP の平均 Q Score 値の比を，用いたアミノ酸置換行列ごとにプロットしたものである．配列一致率が 60%以上のデータではこれら 3 つの手法はどれも，どのアミノ酸置換行列を用いた場合においてもほぼ等しいアライメント精度を示す．しかし配列一致率が 0-60%のデータでは，その値がひくいほど MTRAP が他に比べ高いアライメント精度を有することがわかる．特に配列一致率が 0-20%のデータに対して，MTRAP は Needle の 1.5〜2.3 倍の平均 Q Score 値をとり，ClustalW2 に対しても PAM 行列の利用時に 1.4 倍，BLOSUM 行列で 1.3 倍，GONNET 行列で 1.1〜1.2 倍の値をとっている．

以上の HOMSTRAD，PREFAB4 を用いた検証の結果，MTRAP 法は配列類似性の低い相同配列に対するアライメントで効果を発揮することが分かった．また，どのアミノ酸置換行列を用いた場合も明らかな改善がみられることから，配列間差異（式 11）は既存のアミノ酸置換行列のみを用いた配列間尺度（式 2; Sum of pairs）に対し， より良くタンパクを構成するアミノ酸配列の生物学的特徴をとらえるといえる．

次に，一般的に用いられているアライメントプログラムである，T-Coffee，MAFFT，DIALIGN，MUSCLE，ClustalW2，Probcons との精度の比較を上記 2 つのデータベースを用いて行った結果を表 4 および 5 に示す．各プログラムはその作者の推奨するデフォルトパラメータで実行した．どちらのデータベースを用いた場合においても，MTRAP 法は一般的に精度を改善することが見て取れる．特に，配列一致率が 30%以下の配列に対し明らかな精度の改善を示し，配列一致率が 30%以下の配列に対してはほかの手法にくらべ 4〜10%の改善が見られた．

3. 結　　論

Transition quantity を用いたアライメント法として MTRAP 法を開発し，様々なアライメントプログラムと精度の比較を行った．その結果，特に配列一致率の低い相同配列に対す

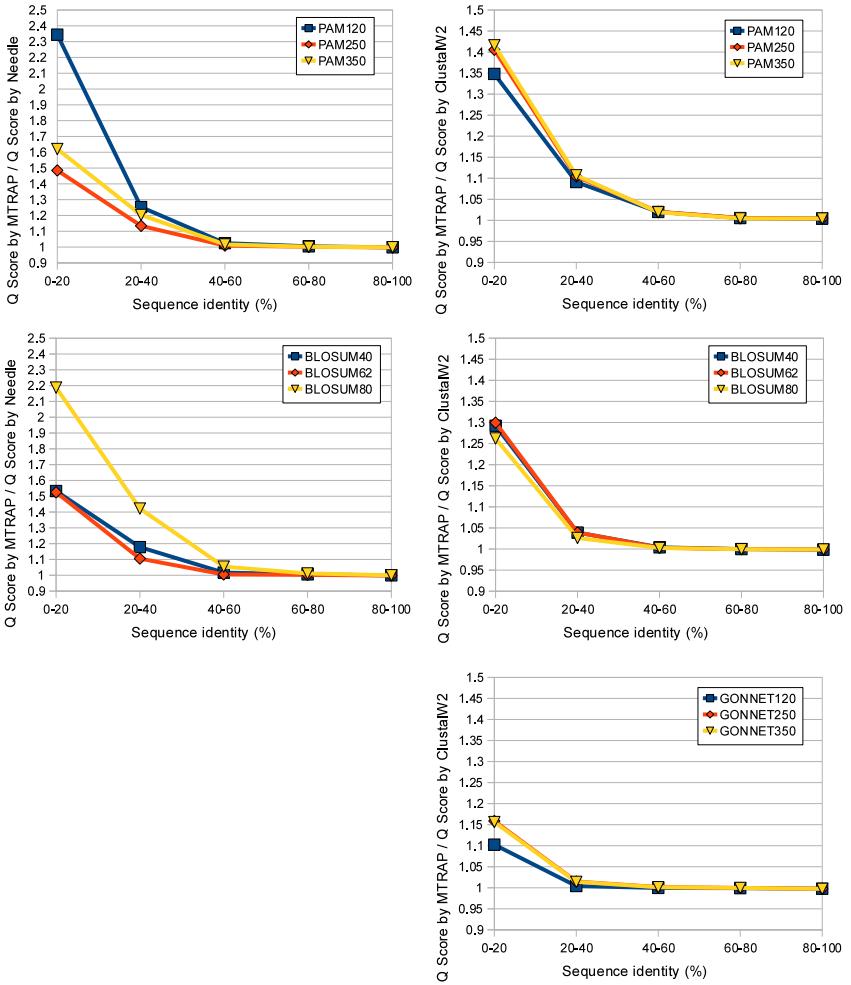

図 2 平均 Q Score の比率：左の 2 つの図は MTRAP の Needle に対する平均 Q Score の比を表し，右の 3 つの図は MTRAP の ClustalW2 に対する平均 Q Score の比を表す．それぞれの折れ線は図中に示されたアミノ酸置換行列を利用した場合の結果を表す．

表 4 PREFAB4 を用いた場合における MTRAP 法とその他の手法との精度の比較

Method	PREFAB 4.0			
	0-15%(423)	15-30%(917)	30-45%(148)	All(1682)
MTRAP[a]	**0.248**	**0.674**	**0.877**	**0.615**
MAFFT	0.170	0.671	0.860	0.568
DIALIGN[b]	0.133	0.556	0.814	0.518
MUSCLE	0.205	0.632	0.867	0.581
ClustalW2	0.199	0.644	0.859	0.586
Probcons	0.204	0.647	0.875	0.590
T-Coffee	0.198	0.642	0.872	0.585

表中の値は PREFAB4 における各配列一致率範囲での，平均 Q Score 値を表す．括弧内の数字は各配列一致率におけるアライメント数を表す．太字は各配列一致率における一番良い値を表す．

[a]MTRAP は GONNET250 アミノ酸置換行列を用いた．

[b]DIALIGN はいくつかのデータでエラーを起こしたため，正常に計算できたものだけでの平均 Q Score を求めた．

表 5 HOMSTRAD を用いた場合における MTRAP 法とその他の手法との精度の比較

Method	HOMSTRAD			
	0-15%(25)	15-30%(207)	30-45%(173)	All(630)
MTRAP[a]	**0.412**	**0.659**	0.879	**0.819**
MAFFT	0.309	0.610	0.863	0.796
DIALIGN[b]	0.216	0.546	0.825	0.760
MUSCLE	0.337	0.625	0.868	0.802
ClustalW2	0.313	0.619	0.867	0.800
Probcons	0.344	0.650	**0.884**	0.816
T-Coffee	0.341	0.634	0.872	0.809

表中の値は PREFAB4 における各配列一致率範囲での，平均 Q Score 値を表す．各種表記は表 4 と同様である．

るアライメントで効果を発揮することを確認した．

マルチプルアライメントを構築するプログラムはペアワイズアライメントを下にしたものが一般的である．よって，本提案手法をマルチプルアライメント構築時にも用いることで，これらのマルチプルアライメント法はより高精度なものとなることが期待される．

参 考 文 献

1) S.F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Bd*, 219:555–565, 1991.
2) S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
3) C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, Jul 1973.
4) G.Blackshields, I.M. Wallace, M.Larkin, and D.G. Higgins. Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biology*, 6(4):321–339, 2006.
5) M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5(3):345–352, 1978.
6) C.B. Do, M.S.P. Mahabhashyam, M. Brudno, and S. Batzoglou. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330, 2005.
7) R.C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792–1797, 2004.
8) Toshihide Hara, Keiko Sato, and Masanori Ohya. Mtrap: pairwise sequence alignment algorithm by a new measure based on transition probability between two consecutive pairs of residues. *BMC Bioinformatics*, 11:235, 2010.
9) Toshihide Hara, Keiko Sato, and Masanori Ohya. Significant improvement of sequence alignment can be done by considering transition probability between two consecutive pairs of residues. *QP-PQ: Quantum Probability and White Noise Analysis (Quantum Bio-Informatics III)*, 26:443–452, 2010.
10) S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89:10915–10919, Nov 1992.
11) K.Katoh, K.Misawa, K.Kuma, and T.Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30:3059–3066, Jul 2002.
12) M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H.McWilliam, F.Valentin, I.M. Wallace, A.Wilm, R.Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*,

Vol.2010-BIO-22 No.3
2010/7/28

23:2947–2948, Nov 2007.
13) K.Mizuguchi, C.M. Deane, T.L. Blundell, and J.P. Overington. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, 7:2469–2471, Nov 1998.
14) B.Morgenstern. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, Mar 1999.
15) S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, Mar 1970.
16) C.Notredame, D.G. Higgins, and J.Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302:205–217, Sep 2000.
17) M.Ohya and Y.Uesaka. Amino acid sequences and DP matching:a new method of alignment, Information Sciences. *Information Sciences*, 63:139–151, 1992.
18) W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.
19) P.Rice, I.Longden, and A.Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16:276–277, Jun 2000.
20) L.A. Stebbings and K. Mizuguchi. HOMSTRAD: recent developments of the homologous protein structure alignment database. *Nucleic acids research*, 32(Database Issue):D203, 2004.
21) 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 Acids Res.*, 22:4673–4680, Nov 1994.
22) I.VanWalle, I.Lasters, and L.Wyns. SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, 21(7):1267, 2005.