論文

相互エントロピーを用いたアライメントの改良
Improvement of Sequence Alignment Based on Mutual Entropy

池 正人 *　佐藤 圭子　谷田貝 甲児　大矢 雅則

Masato IKE　Keiko SATO
Koji YATAGAI　and　Masanori OHYA

我々は，タンパク質の分析を行う上で基本的な操作であるアライメントアルゴリズムの改良を行った．2 本の配列を対象とするペアワイズアライメントにおいて，今までは，配列間の類似性を表す距離を求め，その最小値をとる複数の結果から無作為に一組の配列を決定していた．本研究では，最小値を与える複数の結果すべてを考え，その配列の組各々に対し，相互エントロピーを計算した．そして，その値の等しい組でグループを作り，全グループの相互エントロピーの平均値を求めることで，結果の絞込みを行った．ヘモグロビンのアミノ酸配列を用いてアライメントしてみたところ，相互エントロピーの平均値と最も近い値をもつグループの中に，生物学的な立体構造を考慮したアライメント結果が含まれていることがわかった．この結果は，タンパク質の立体構造などを考慮することなく，生物学的なアライメント結果のグループを特定することができることを示している．

We improve the algorithm to align amino acid sequences of protein which is one of the most fundamental operations studying the analysis of genome. In pair-wise alignment, one chooses one aligned pair (i.e., two sequences) without special reasons from several aligned pairs (the number of these pairs is often very large) giving the same smallest values to the difference properly defined between two sequences. In this paper, we compute the mutual entropy for several such pairs having the same difference, and we classify the pairs into some groups such that the same group consist of the pairs having the same value of the mutual entropy, then we finally compute the mean value of the mutual entropy over the whole groups. As a consequence, we can observe the following interesting fact for some proteins that the aligned pair obtained by usual alignment with 3D protein structure (we call such a alignment the biological alignment here) is in the group having the value of the mutual entropy closest to the mean value of the mutual entropy. From the above observation we conclude that our method using the alignment (MOU-alignment) and the mutual entropy makes us possible to find the biological alignment, that is, we do not need to know the 3D structure to obtain the biological alignment.

キーワード：アミノ酸，アライメント，相互エントロピー
amino acid, alignment, mutual entropy

1　はじめに

生物はタンパク質から成っており，それを構成するのが 20 種類のアミノ酸である．タンパク質には 1 次構造から 4 次構造までの 4 つの構造が存在する．そのようなアミノ酸配列を解析するとき，最初に行われる重要な操作がアライメントと呼ばれる操作である．アライメントとは，アミノ酸配列の並び方を整えることによって，複数の配列間に存在する関係を明確に浮かび上がらせる目的で行われる操作である．この操作を行うことによって配列の比較や分子進化系統樹作成等に役立てら

* 東京理科大学理工学部情報科学科
Tokyo University of Science, Department of Information Sciences
E-mail:j6302601@ed.noda.tus.ac.jp

れる．そのため，この操作を少しでも正しく行うことは，後々の解析にとって非常に重要なことである．

本論文では，数あるアライメントアルゴリズムの中から 2 本の配列を対象にしたペアワイズアライメントの 1 つである MOU-アライメント[1]を取り上げ，その改良を行った．従来の一次構造のみから行うアライメントアルゴリズムでは，そもそも配列間の類似性，すなわちある距離を求めることを目標としており，数値計算によって得られた配列自体は必ずしも生物学的に有意なものであるわけではなかった．そこで我々は，1 回の数値解析につき，配列間の距離が最小となる 1 組の配列を決定していた従来のアルゴリズムに対し，さらに生物学的に有意な配列を求めるため，同じ距離をもつアライメント結果複数を得た後，2 つの配列間における情報のやりとりの精度を表す相互エントロピー[2]を用いて絞込みを行うアルゴリズムに改良した．具体的にはその複数の組それぞれに対して相互エントロピーを求め，その値の等しい組ごとにグループ分けをし，全グループの相互エントロピーの平均値を算出した．このアライメントアルゴリズムで生物学的なアミノ酸配列を絞ることができるかを調べた．

2　アライメント

2.1　アライメントの基本操作

いま 2 つの同じタンパク質を表すアミノ酸配列があるとする．

a: MPQRSTVWPYT
b: MNPQRYSTWQYT

しかしながら，このままの状態では 2 つの配列間の類似性は見えてこない．なぜなら配列 a，b は，長い年月を経過して行われてきた生物の進化の過程において，アミノ酸の挿入・欠落・置換による変異が与えられて存在するものと考えられるからである．実際に挿入・欠落があったと考えられるところに ∗ を挿入すると

a′: M ∗ PQR ∗ STVWPYT
b′: MNPQRYST ∗ WQYT

となり，アミノ酸配列 a，b 間の類似性が浮かび上がってくる．a′ からみれば b′ では N と Y が挿入され，V が欠落し，P から Q に置換されているとみなすことができる．このように ∗ を挿入してアミノ酸配列間の類似性を明らかにするように並び替える操作をアライメントという．特にアミノ酸配列を 2 本ずつアライメントする方法をペアワイズアライメント，N 本のアミノ酸配列を一度にアライメントする方法をマルチプルアライメントという．

本論文ではペアワイズアライメントについて扱っており，それには Needleman，S.B.，Wunsch, C.D. らによる NW-アライメント[3]，Sellers, P.H. による S-アライメント[4]，Ohya, M.，Uesaka, Y. らによる OU-アライメント[5]などのアルゴリズムがある．これらの計算効率は全て同じである[1]．これらを効率の面から改良したものが Ohya, M.，Miyazaki，S.，Ohsima，Y. らによる MOU-アライメント[1]である．

2.2　MOU-アライメント

ここでは，MOU-アライメントのアルゴリズムについて説明する．そのために，まず，3 つの同等なアライメントのうち OU-アライメントについて説明する．これはパターンマッチングの方法のひとつである動的計画法から直接導出されたもので，数学的に最も一般的かつ単純なものである．

以下，次の a と b を，それぞれ ∗ を含まない 2 本のアミノ酸配列とする．

a: $a_1 a_2 \cdots a_m$
b: $b_1 b_2 \cdots b_n$

なお，この配列は有限列 $a_1 a_2 \cdots a_m$ に ∗ を無限個加えた無限列 $a_1 a_2 \cdots a_m ***\cdots$ と同

一視する．このアライメントでは，2 つのアミノ酸配列 a，b の両端と各アミノ酸の間に予め $*$ を挿入した配列を $\mathsf{a}^+ = a_0^+ a_1^+ \cdots a_{2m}^+$ $\mathsf{b}^+ = b_0^+ b_1^+ \cdots b_{2n}^+$ とし，a^+，b^+ を x 軸，y 軸上に配置した格子グラフ (図 1，ただし a_i^+ は $x = i$，b_j^+ は $y = j$ に対応させる) を考える．

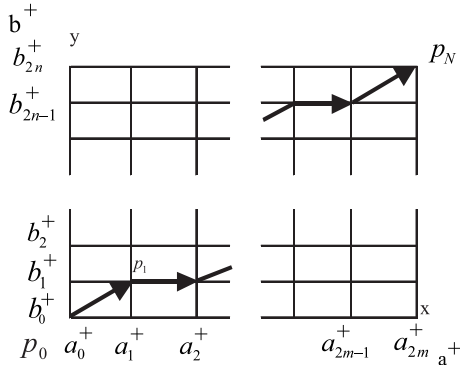

図 1　OU-アライメントの格子グラフ

さて a^+，b^+ の適当な対応を見つけることは, この格子グラフ上で点 $P_0 = (0, 0)$ と $P_N = (2m, 2n)$ を結ぶ，最適なパスを見つけることである．

ところで点 P_0 と点 P_N を結ぶ最適なパスを選ぶということは，そのパスを通して考えられる a^+，b^+ の間のある距離を最小にすることを意味する．そこでこの a^+，b^+ 間の距離を

$$D_{OU}(\mathsf{a}^+, \mathsf{b}^+) = \min \left\{ \sum_{k=0}^{N} d(a_{i_k}^+, b_{j_k}^+); \right. \\ \left. P = (P_0, P_1, \cdots, P_N) \in S \right\}$$

但し $P_k = (i_k, j_k),\ N = 1, 2, \cdots$

とおく．ここで S は P_0 から P_N への全てのパスの集合であり，d は，ここでは,

$$d(a, b) = \begin{cases} 0 & (a = b) \\ 1 & ((a \neq b) \text{ かつ } (a \neq * \text{ かつ } b \neq *)) \\ w & ((a \neq b) \text{ かつ } (a = * \text{ または } b = *)) \end{cases}$$

ととることにする．ここでの w は，アミノ酸と $*$ との間の距離を示す値として用いられる重み (ウエイト) と呼ばれる数値である．この w の値が小さいほど $*$ の挿入が発生しやすく大きいほど発生しにくい．一般的には値を $0.5 \leq w \leq 2.0$ の範囲に設定する場合が多いが，数々の生物学的解釈により 0 以上のいろいろな値をとる．ちなみに，本論文におけるウエイトの値は 2.0 を用いるものとし，すべての計算においてこの値で統一した．以下にこのアライメントを行う方法を説明する．

アミノ酸配列 a^+，b^+ 間の距離，すなわち $(0, 0)$ から $(2m, 2n)$ までの距離を

$$D(i, j) = D_{OU}(a_0^+ a_1^+ a_2^+ \cdots a_i^+, b_0^+ b_1^+ b_2^+ \cdots b_j^+) \\ 0 \leq i \leq 2m, \quad 0 \leq j \leq 2n$$

とすると，$D(i, j)$ は次の制限をもとに求められる．

$$D(i, j) \\ = \begin{cases} D(i-1, j) + d(a_i^+, b_j^+) \\ \qquad (i, j) = (\text{奇数}, \text{奇数}) \\ \min \left\{ \begin{array}{l} D(i-1, j) \\ D(i-1, j-1) \end{array} \right\} + d(a_i^+, b_j^+) \\ \qquad (i, j) = (\text{奇数}, \text{偶数}) \\ \min \left\{ \begin{array}{l} D(i-1, j-1) \\ D(i, j-1) \end{array} \right\} + d(a_i^+, b_j^+) \\ \qquad (i, j) = (\text{偶数}, \text{奇数}) \\ \min \left\{ \begin{array}{l} D(i-1, j) \\ D(i-1, j-1) \\ D(i, j-1) \end{array} \right\} + d(a_i^+, b_j^+) \\ \qquad (i, j) = (\text{偶数}, \text{偶数}) \end{cases}$$

また上記の式は以下のように簡単にできる．

$$
D(i,j)
= \begin{cases}
D(i-1,j) + d(a_i^+, b_j^+) \\
\qquad (i,j) = (\text{奇数}, \text{奇数}) \\
D(i-1,j) + d(a_i^+, b_j^+) \\
\qquad (i,j) = (\text{奇数}, \text{偶数}) \\
D(i,j-1) + d(a_i^+, b_j^+) \\
\qquad (i,j) = (\text{偶数}, \text{奇数}) \\
\min \left\{ \begin{array}{l} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{array} \right\} + d(a_i^+, b_j^+) \\
\qquad (i,j) = (\text{偶数}, \text{偶数})
\end{cases}
$$

これを用いると $(i,j) = (\text{偶数}, \text{偶数})$ の場合は，以下のように表すことができる．

$$
D(i,j)
= \min \left\{ \begin{array}{l} D(i-2,j) + w \\ D(i-2,j-2) + d(a_{i-1}^+, b_{j-1}^+) \\ D(i,j-2) + w \end{array} \right\}
$$
$$
(i,j) = (\text{偶数}, \text{偶数})
$$

この式から，$(i,j) = (\text{偶数}, \text{偶数})$ における $D(i,j)$ は $(i',j') = (\text{偶数}, \text{偶数})$ における $D(i',j')$ より直接求められていることが分かる．よって，実際に計算機上でアライメントを行う時には $P_k = (i_k, j_k)$ $(i_k = \text{偶数}, j_k = \text{偶数})$ のみを考えればよく，新たに $*$ を挿入した列を考える必要はない．

さらにアライメント結果に影響を与えない格子点をアライメント実行前に削除して計算量を抑え，計算効率を上げたものが MOU-アライメントである．我々が実際に長さ m, n のアミノ酸配列をアライメントすると，ほとんどの場合最適なパスとして対角線 $(0,0)$，(m,n) 付近を通ることがわかる．

そこで MOU-アライメントは格子グラフの要素を減らす関数を以下のように定めることでその計算を省略することができる．

実際，以下の命題を証明することができる．

＜命題 2.1 ＞
長さ m, n $(m \geq n)$ の 2 つのアミノ酸配列を，a，b とし，アライメント結果 a'，b'(長さ p) が得られたとする．

$$
\begin{aligned}
\mathsf{a}\colon a_1 \cdots a_n a_{n+1} \cdots a_m &\rightarrow \mathsf{a}'\colon a'_1 a'_2 \cdots a'_p \\
\mathsf{b}\colon b_1 \cdots b_n &\rightarrow \mathsf{b}'\colon b'_1 b'_2 \cdots b'_p
\end{aligned}
$$

このとき，G を以下のように定める．

$$
G \equiv \sum_{i=1}^{n} \beta(a_i, b_i)
$$

$$
\beta(a,b) = \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases}
$$

G はアライメントされる前のアミノ酸配列において異なっているアミノ酸対の数である．また，アライメント結果のアミノ酸配列 a'，b' に挿入される $*$ の個数を $\#(\mathsf{a})$，$\#(\mathsf{b})$ とする．このとき

$$
0 \leq \#(\mathsf{a}) \leq G
$$

$$
m - n \leq \#(\mathsf{b}) \leq G + (m-n)
$$

が成り立つ．(証明は文献[1]を参照)

この命題は，パスが図 2 の V と (m,n) の線分より下に $G+1$ 以上，$(0,0)$ と V′ の線分より左に $G+1$ 以上離れないということである．すなわち，格子グラフの要素を減らす関数 f，g を

$$
f\colon y = x - \{G + (m-n)\}
$$
$$
g\colon y = x + G
$$

で定めると (図 2)

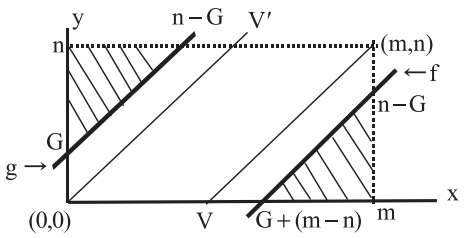

図 2　f と g

上記のグラフは，このアライメントが，斜線部分 (f，g 上を含まず) にある格子点に対応する要素 $\{(x,y) \mid y > g(x)$ あるいは $y < f(x)\}$ の計算を，省略することができることを示している．

実際に次のアミノ酸配列 a と b をアライメントする．

$$\mathsf{a}: \mathrm{MNPQY}$$
$$\mathsf{b}: \mathrm{MPQR}$$

まず格子グラフの要素を減らす関数 f，g を求めると，$m=5$，$n=4$，$G=3$ より

$$f: y = x - 4$$
$$g: y = x + 3$$

となる．そして各々の配列の先頭に $*$ を挿入する．

$$\mathsf{a}: *\mathrm{MNPQY}$$
$$\mathsf{b}: *\mathrm{MPQR}$$

この a，b を x 軸，y 軸上に配置した格子グラフを考え，各格子点における距離を求める (図 3)．例えば, 格子点 $(3,2)$ における $D(3,2)$ は

$$\begin{aligned} D(3,2) &= D(2,1) + d(\mathrm{P},\mathrm{P}) \\ &\vdots \\ &= d(*,*) + d(\mathrm{M},\mathrm{M}) + d(\mathrm{N},*) + d(\mathrm{P},\mathrm{P}) \\ &= 0+0+2+0 = 2 \end{aligned}$$

となる．

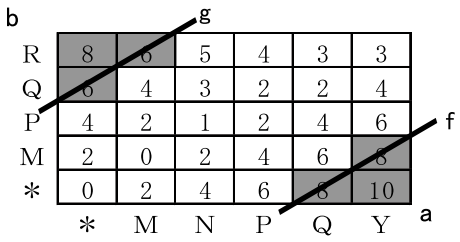

図 3　計算後の格子グラフ
(黒い部分は計算省略)

算出された距離の値に基づいて，(m,n) 地点から始まって $(0,0)$ 地点まで至るパスのうち，最小のものを帰納的に求めていくトレースバックを行うと，最終的なアライメント結果が得られる．実際に $(5,4)$ から $(0,0)$ までトレースバックを行うと

$$\begin{aligned} D(5,4) &= D(4,3) + d(\mathrm{Y},\mathrm{R}) \\ &\vdots \\ &= d(\mathrm{M},\mathrm{M}) + d(\mathrm{N},*) + d(\mathrm{P},\mathrm{P}) \\ &\quad + d(\mathrm{Q},\mathrm{Q}) + d(\mathrm{Y},\mathrm{R}) \end{aligned}$$

となる．これを格子グラフで表すと，次に示す図 4 の網掛け部分になる．

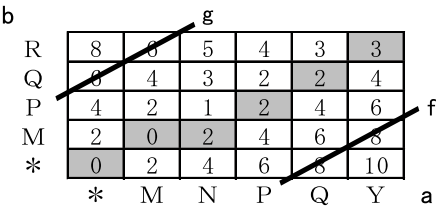

図 4　アライメント結果

このとき得られるアライメント結果は，

$$\mathsf{a}': \mathrm{MNPQY}$$
$$\mathsf{b}': \mathrm{M}*\mathrm{PQR}$$

となり，アミノ酸配列 a′，b′ 間の距離は，

$$D(5,4) = 3$$

となっている．

このようにして得られたアミノ酸配列は，実際に距離を最小にする a′ と b′ となっていることが容易にわかる．

3　MOU-アライメントの改良

MOU-アライメントを行うと，最小距離が等しくなる複数の結果が得られる場合が多い．しかし，今まではそのような場合，無作為に得られる 1 組の配列をアライメント結果として決定していた．したがって，そのアライメント結果は生物学的な構造を考慮した配列であるとは限らなかった．そこで，本研究では，同じ最小距離をもつ複数結果を全て求めた後，結果の絞込みを行うよう，アライメントの改良を行った．

結果の絞込みには，相互エントロピーを用いた．生物の進化の過程で，蓄積した変異を

捉えるということは，DNA の塩基配列やタンパク質のアミノ酸配列のもつ情報量がどれだけ正しく伝達されたかを知ることにつながる．相互エントロピーは，配列間の情報のやりとりの精度を表すもので，我々は，この量を用いて配列間の類縁度を測ることができると考えている．

ここで，アライメントされた 2 本のアミノ酸配列 a と b の完全事象系を考える．a と b において出現するアミノ酸の出現確率を，それぞれ p_i，q_j $(1 \leq i, j \leq 20)$ とし，$*$ の出現確率を p_0，q_0 とする．このとき，完全事象系は次のようになる．

$$\begin{bmatrix} \mathsf{a} \\ p \end{bmatrix} = \begin{bmatrix} * & A & C & \cdots & W & Y \\ p_0 & p_1 & p_2 & \cdots & p_{19} & p_{20} \end{bmatrix}$$

$$\begin{bmatrix} \mathsf{b} \\ q \end{bmatrix} = \begin{bmatrix} * & A & C & \cdots & W & Y \\ q_0 & q_1 & q_2 & \cdots & q_{19} & q_{20} \end{bmatrix}$$

次に，a と b の各アミノ酸と $*$ が同時に出現する確率を r_{ij} $(0 \leq i, j \leq 20)$ とすると完全複合事象系は次のように書き表すことができる．

$$\begin{bmatrix} \mathsf{a} \times \mathsf{b} \\ r \end{bmatrix} = \begin{bmatrix} ** & *A & *C & \cdots & YW & YY \\ r_{00} & r_{01} & r_{02} & \cdots & q_{2019} & q_{2020} \end{bmatrix}$$

完全事象系，完全複合事象系が設定されると，アミノ酸配列のエントロピー $S(\mathsf{a})$ 及び相互エントロピー $I(\mathsf{a}, \mathsf{b})$ は，次のように計算できる．

$$S(\mathsf{a}) = -\sum_{i=0}^{20} p_i \log p_i$$

$$I(\mathsf{a}, \mathsf{b}) = \sum_{i,j} r_{ij} \log \frac{r_{ij}}{p_i q_j}$$

$S(\mathsf{a})$ は，系 (a, p) の持っている情報量を表しており，$I(\mathsf{a}, \mathsf{b})$ は b を知ることによって得られる a の情報量を表している．相互エントロピーは，その値が大きくなるほど 2 本のアミノ酸配列間の類似性が大きいといえる．

我々は，MOU-アライメントをもとに，複数結果を求め，上記の相互エントロピーを用いて，結果の絞込みを行うアライメントアルゴリズムに改良した．

4　改良 MOU-アライメントを用いた解析

改良した MOU-アライメント (以下，改良アライメントと呼ぶ) で次の 1)，2) の解析を行った．

1) 最小距離が等しくなる複数の結果の中に，生物学的な立体構造を考慮したアライメント結果 (Web サイト[6]に記載されているもの) が含まれているか．
2) 数多く得られた結果の中から，生物学的な立体構造を考慮したアライメント結果である可能性が高いものを，相互エントロピーを用いて特定することが可能かどうか．

1) は改良アライメントが，従来の目的，「アミノ酸配列間の距離を求める」ということについて成功しているかどうかの検証である．また 2) は 1) が成功している場合において，距離のみならず，更に生物学的な立体構造を考慮したアライメント結果を絞りこむことが可能かという新たな問題点に立っている．

以下の計算においては，解析データとして Web サイト[6]に記載されている生物における様々なタンパク質のアミノ酸配列を用いた．このサイトに記載されているデータは，タンパク質を分析することによって得られたアミノ酸配列を，タンパク質の構造情報などに基づいてアライメントされたものである．このデータを生物学的な立体構造を考慮したアライメント結果としている．

4.1　1) の解析

各生物のヘモグロビン α 鎖と β 鎖のアミノ酸配列を用いてアライメントを行った．この対象となった生物の中には，得られた複数結果の最小距離が生物学的な立体構造を考慮したアライメント結果よりも小さくなるものがあった．この場合，改良アライメントによって得られた結果の中には，生物学的な立体構造を考慮したアライメント結果と同一な配列

は存在しなかったことになる．このことについては，タンパク質の構造などを情報論的に扱い，MOU-アライメントをさらに改良し，この距離の差と組数がどのように変わるのか，解析してみる必要がある．

一方，表 1 に示した 3 種においては，改良アライメントから出された最小距離と生物学的な立体構造を考慮したアライメント結果の距離が等しく，改良アライメントの複数結果の中に生物学的な立体構造を考慮したアライメント結果が含まれた．表 2 は，これらのデータに対して改良アライメントを施して得られた配列の組数とその最小距離を示している．

表 1　3 種におけるヘモグロビン α 鎖，β 鎖のアミノ酸配列

PDB ID	アミノ酸配列
2HHB(α)	VHLTPEEKSAVTALWGKVNVDEVGGEALGRL LVVYPWTQRFFESFGDLSTPDAVMGNPKVK AHGKKVLGAFSDGLAHLDNLKGTFATLSELH CDKLHVDPENFRLLGNVLVCVLAHHFGKEFT PPVQAAYQKVVAGVANALA
2HHB(β)	VLSPADKTNVKAAWGKVGAHAGEYGAEALER MFLSFPTTKTYFPHFDLSHGSAQVKGHGKK VADALTNAVAHVDDMPNALSALSDLHAHKL RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVH ASLDKFLASVSTVLTSKYR
2PGH(α)	VHLSAEEKEAVLGLWGKVNVDEVGGEALGRL LVVYPWTQRFFESFGDLSNADAVMGNPKVK AHGKKVLQSFSDGLKHLDNLKGTFAKLSELH CDQLHVDPENFRLLGNVIVVVLARRLGHDFN PNVQAAFQKVVAGVANALA
2PGH(β)	VLSAADKANVKAAWGKVGGQAGAHGAEALE RMFLGFPTTKTYFPHFNLSHGSDQVKAHGQ KVADALTKAVGHLDDLPGALSALSDLHAHKL RVDPVNFKLLSHCLLVTLAAHHPDDFNPSVH ASLDKFLANVSTVLTSKYR
2MHB(α)	VQLSGEEKAAVLALWDKVNEEEVGGEALGRL LVVYPWTQRFFDSFGDLSNPGAVMGNPKVK AHGKKVLHSFGEGVHHLDNLKGTFAALSELH CDKLHVDPENFRLLGNVLVVVLARHFGKDFT PELQASYQKVVAGVANALA
2MHB(β)	VLSAADKTNVKAAWSKVGGHAGEYGAEALER MFLGFPTTKTYFPHFDLSHGSAQVKAHGKK VGDALTLAVGHLDDLPGALSDLSNLHAHKLR VDPVNFKLLSHCLLSTLAVHLPNDFTPAVHA SLDKFLSSVSTVLTSKYR

（2HHB･･･ヒトのヘモグロビン
2PGH･･･イノシシのヘモグロビン
2MHB･･･ウマのヘモグロビン）

表 2　表 1 のデータにおける組数とそれらの距離

PDB ID	結果組数	距離
2HHB	120	93.0
2PGH	336	96.0
2MHB	820	92.0

4.2　2) の解析

表 1 の 3 種それぞれにおいて，最小距離を与える複数の結果の配列の組各々に対し，相互エントロピーを計算した．そして計算した相互エントロピーの等しい組ごとにグループを作り，各グループの組数をカウントした (図 5, 図 6，図 7)．横軸には相互エントロピーの値をとり，縦軸は相互エントロピーの値が等しい配列の組数を表している．それぞれのグラフに示した棒グラフのうち，塗りつぶされたものは，生物学的な立体構造を考慮したアライメント結果から算出した相互エントロピーと等しいことを示している．

グループ分けをした全てのグループの相互エントロピーの平均値を各生物ごとに算出し，生物学的な立体構造を考慮したアライメント

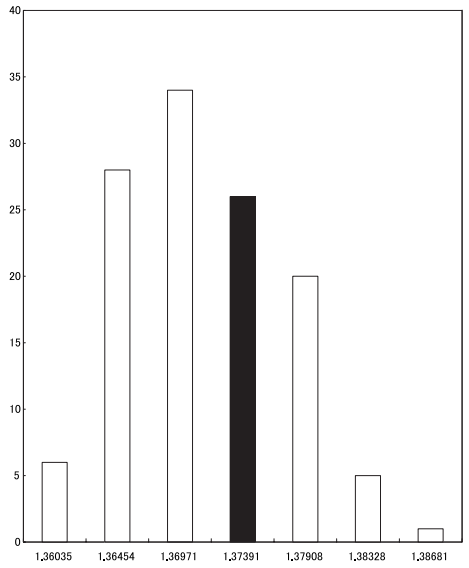

図 5　2HHB の相互エントロピーによる分類
(縦軸：出現組数 横軸：相互エントロピー)

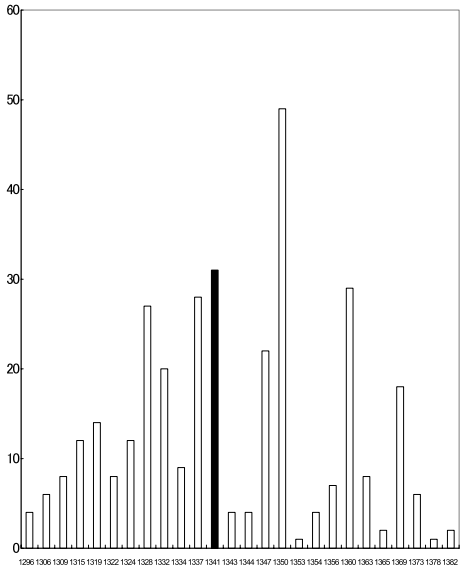

図 6　2PGH の相互エントロピーによる分類
(縦軸：出現組数 横軸：相互エントロピー)

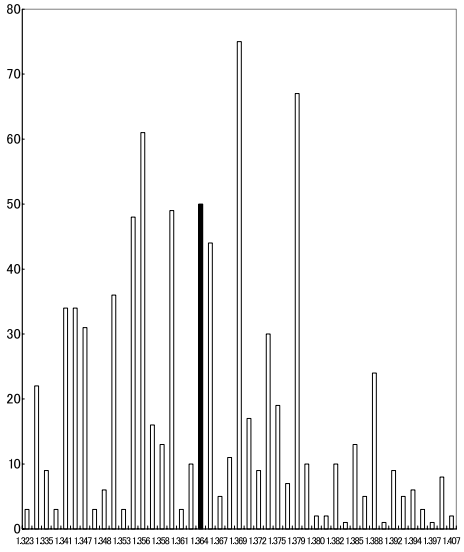

図 7　2MHB の相互エントロピーによる分類
(縦軸：出現組数 横軸：相互エントロピー)

結果の相互エントロピー値と比較した．(表 3)

その結果，相互エントロピーの平均値 (A) は，生物学的な立体構造を考慮したアライメント結果の相互エントロピー値 (C) に限りなく近い．言い換えれば，配列間の距離が最小な複数結果をもとに相互エントロピーの平均値を求めれば，その値と最も近いグループの中に生物学的なアライメント結果が含まれていることがわかる．

これにより，立体構造を考慮せずに生物学的なアライメント結果のグループを特定することができると考えられる．

表 3　相互エントロピーの平均値と生物学的な立体構造を考慮したアライメント結果の相互エントロピー値との比較

PDB ID	A	B	C	|A − C|
2HHB	1.371216	0.0060747	1.37391	0.002695
2PGH	1.341087	0.0171877	1.34092	0.0001689
2MHB	1.163516	0.0149739	1.36354	0.0000248

A: 改良アライメントにおける相互エントロピーの平均値
B: 改良アライメントにおける相互エントロピーの標準偏差
C: 生物学的な立体構造を考慮したアライメント結果の相互エントロピー値

5　まとめ

最後にヘモグロビン α 鎖，β 鎖のアミノ酸配列を用いて，改良アライメントを行った本研究のまとめを述べる．

ある生物では，改良アライメントで得られた最小距離が，生物学的な立体構造を考慮したアライメント結果より小さい値を示した．これについてはすべてのアミノ酸間の距離を 1 とするだけでなく，距離に変化をつけるような工夫を加えることで精度を高める必要がある．このような工夫を取り入れることで現在よりも細かな最小距離の算出が可能となり，最小距離を与える複数の結果にも影響してくるだろう．

我々は，その工夫の 1 つの指標として，PAM マトリクス[7]を用いた解析を 3 本以上のアミノ酸配列で行うマルチプルアライメントに関して進めている．PAM マトリクスとは，20

種類のアミノ酸それぞれに対して変わりやすさ，または変わりにくさを数値で表したものである．この数値が大きいほど変異が起きやすく，小さいほど変異が起きにくいことを表している．

一方，改良アライメントで得られた配列間の最小距離と生物学的な立体構造を考慮したアライメント結果の距離が一致した場合，その最小距離を与える複数の結果の中には，生物学的に有意とされるアライメント結果が含まれる．従って，我々は，複数結果各々に対し，相互エントロピーを計算し，その値の等しい組ごとにグループ分けをし，全グループの相互エントロピーの平均値を算出した．この改良アライメントでアライメントしたところ，生物学的に正しいとされるアライメント結果が属するグループを特定することができた．このアルゴリズムの有効性を確かめるために，他のサンプルによる検証，またはPAMマトリクスやタンパク質の機能などの情報を考慮してさらなる精度を向上させることを現在試みている．

参考文献

[1] Ohya, M.; Miyazaki, S.; Ohshima,Y.: “A new method of Alignment of Amino Acid Sequences”, Viva Origino, 17, pp.139–151, 1989.

[2] 大矢 雅則; 渡邉 昇: 「量子通信理論の基礎 量子情報から光通信へ」, 数理情報科学シリーズ 17, 牧野書店, 291p., 1998.

[3] Needleman, S. B.; Wunsch, C. D.: “A General Method Applicable to Search for Similarities in the Amino Acid Sequence of Two Proteins”, J. Mol. Biol., 48, pp.443–453, 1970.

[4] Sellers, P.H.: “On the Theory and Computation of Evolutionary Distance”, SIAM Journal of Math., Vol.26, No.4, pp.787–793, 1974.

[5] Ohya, M.; Uesaka, Y.: “Amino Acid Sequences and DP Matching”, Information Science, 63, pp.139–151, 1992.

[6] “HOMSTRAD” http://www-cryst.bioc.cam.ac.uk/homstrad/

[7] Jones, D.T.; Taylor, W.R.; Thornton, J.M.: “The Rapid Generation of Mutation Date Matrices from Protain Sequences”, CABIOS, 8, No.3, pp.275–282, 1992.

（2003 年 5 月 20 日受付）
（2003 年 7 月 8 日採択）