相互エントロピーを用いたアライメントの改良

○池正人　谷田貝甲児　佐藤圭子　大矢雅則
東京理科大学　理工学部

Improvement of Sequence Alignment Based on Mutual Entropy

Masato IKE, Kouji YATAGAI, Keiko SATOH, Masanori OHYA

Abstract

We improve the algorithm to align amino acid sequences of identical 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 consists 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 geometrical 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 geometrical structure to obtain the biological alignment.

1. はじめに

遺伝情報の解析を行う上でアライメントという操作は非常に重要な操作である．その操作により，生物が進化の過程で置換，欠如，挿入という変化をしたアミノ酸を見分けられ，進化の過程を示す系統樹を作成すること等に役立てられるからである．これら操作後の解析のずれを無くすために，この操作の正確性を向上させることが必要不可欠である．

本研究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(文献[1])を取り上げ，その改良を行った．従来のアルゴリズムでは配列間の相同性を表す，ある“距離”を求めることを目的としていたが，我々は以下のような点に着目した．

1) アライメント結果で得られた複数の結果の中に生物学的見地からみたアライメントの結果（これを真のアライメント結果とする）が含まれているか
2) 相互エントロピー(文献[2])を用いて，数多く得られた結果の中から真のアライメント結果である可能性が高いものを特定することが可能かどうか

以上 2 点について正しいアミノ酸配列が求められるか，について解析を行った．

2. アライメント

本節ではアライメントについて紹介する．このアルゴリズムは 2 本のアミノ酸配列が与えられている時にギャップ(*)を挿入することによってそれらの相同性を明らかにする操作のことである．例えば以下のアミノ酸配列 $\mathcal{A}$，$\mathcal{B}$ が与えられたとする．$\mathcal{B}$ のMとPの間にギャップ(*)が挿入されることにより，新たなアミノ酸配列 $\mathcal{A}'$，$\mathcal{B}'$ を得ることができる．

$$\mathcal{A}:\mathrm{MNPQY} \qquad \mathcal{A}':\mathrm{MNPQY}$$
$$\mathcal{B}:\mathrm{MPQR} \longrightarrow \mathcal{B}':\mathrm{M*PQR}$$

このような操作をアライメントという．これを数学的なアライメントとして表すと以下のように記述される．
アミノ酸間 a，b に対する距離 $d(a,b)$ を導入する．

$$d(a,b)=\begin{cases}0\ (a=b)\\ 1\ \big((a\neq b)\text{かつ}(a\neq *\text{かつ}b\neq *)\big)\\ 2\big((a\neq b)\text{かつ}(a=*\text{または}b=*)\big)\end{cases}$$

そしてアミノ酸配列間の距離$d'(a_1a_2\cdots,b_1b_2\cdots)$は以下のように与えられる.

$$d'(a_1a_2\cdots,b_1b_2\cdots)\equiv\sum_{i=1}^{\infty}d(a_i,b_i)$$

この数学的な観点からのアライメントでは，ギャップを挿入することによって得られる$d'(a_1a_2\cdots,b_1b_2\cdots)$の最小値を求めることを表す．上記の例で求められる最小距離は 3 となる.

本論文ではこれらの定義に基づき，最小距離を求める時間において利点を持つ MOU-アライメント(文献[1])を導入し，解析を行った.

3. MOU-アライメントの改良

例として以下の配列を用意する．アライメントを行うと最小距離 3 を与える 2 つの配列が出てくる.

$\mathcal{A}$:MPQYSA　　$\mathcal{A}'$:MPQYSA　　$\mathcal{A}'$:MPQYSA

$\mathcal{B}$:MPYYA　→　$\mathcal{B}'$:MP＊YYA　　$\mathcal{B}'$:MPYY＊A

そこで我々は第 1 章で挙げた点に着目した．このような状況をふまえ，本研究では一回の数値解析により最小距離となる配列を全て出力できるようにプログラムを変え，それを元に解析を行った.

本研究で用いたアミノ酸配列のデータは Web サイト(文献[3])に記載されている種々の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．このデータを生物学的見地からみたアライメント結果としている.

4. 解析

複数の結果を出力する解析において，Web サイト(文献[3])に記載してある globin のデータで解析を行った．そして MOU-アライメントからの最小距離と生物学的見地からみたアライメント結果の距離を比較すると，一致または MOU-アライメントの方が短くなった．最小距離が一致した結果に真のアライメント結果が含まれていることが分かった．そのデータを下表 1 に示す.

PDB ID	アミノ酸配列	PDB ID	アミノ酸配列	PDB ID	アミノ酸配列
2HHB(β)	VLSPADKTNVKAAWGKVGAHAGEYGA	2PGH(β)	VLSAADKANVKAAWGKVGGQAGAHGA	2MHB(β)	VLSAADKTNVKAAWSKVGGHAGEYGAE
	EALERMFLSFPTTKTYFPHFDLSHGSA		EALERMFLGFPTTKTYFPHFNLSHGSD		ALERMFLGFPTTKTYFPHFDLSHGSAQ
	QVKGHGKKVADALTNAVAHVDDMPN		QVKAHGQKVADALTKAVGHLDDLPGAL		VKAHGKKVGDALTLAVGHLDDLPGALS
	ALSALSDLHAHKLRVDPVNFKLLSHCL		SALSDLHAHKLRVDPVNFKLLSHCLLVT		DLSNLHAHKLRVDPVNFKLLSHCLLSTL
	LVTLAAHLPAEFTPAVHASLDKFLASV		LAAHHPDDFNPSVHASLDKFLANVSTV		AVHLPNDFTPAVHASLDKFLSSVSTVLT
2HHB(α)	VHLTPEEKSAVTALWGKVNVDEVGGE	2PGH(α)	VHLSAEEKEAVLGLWGKVNVDEVGGEA	2MHB(α)	VQLSGEEKAAVLALWDKVNEEEVGGEA
	ALGRLLVVYPWTQRFFESFGDLSTPD		LGRLLVVYPWTQRFFESFGDLSNADAV		LGRLLVVYPWTQRFFDSFGDLSNPGAV
	AVMGNPKVKAHGKKVLGAFSDGLAHL		MGNPKVKAHGKKVLQSFSDGLKHLDNL		MGNPKVKAHGKKVLHSFGEGVHHLDNL
	DNLKGTFATLSELHCDKLHVDPENFRL		KGTFAKLSELHCDQLHVDPENFRLLGN		KGTFAALSELHCDKLHVDPENFRLLGNV
	LGNVLVCVLAHHFGKEFTPPVQAAYQ		VIVVVLARRLGHDFNPNVQAAFQKVVA		LVVVLARHFGKDFTPELQASYQKVVAG

表 1 最小距離が一致したデータ

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

上記のデータから求められた配列の組の本数とその距離も表 2 に示す.

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

表 2 それぞれのデータにおける配列の組の本数とその最小距離

こうして得られた複数のアミノ酸配列に対して真のアライメント結果を求めることになるが，本論文では，アミノ酸配列の保持している情報がどのくらい関連しているものであるのかを示す相互エントロピー(文献[2])という値を指標として採用するものとした．その結果の一例として 2HHB のデータを示す．

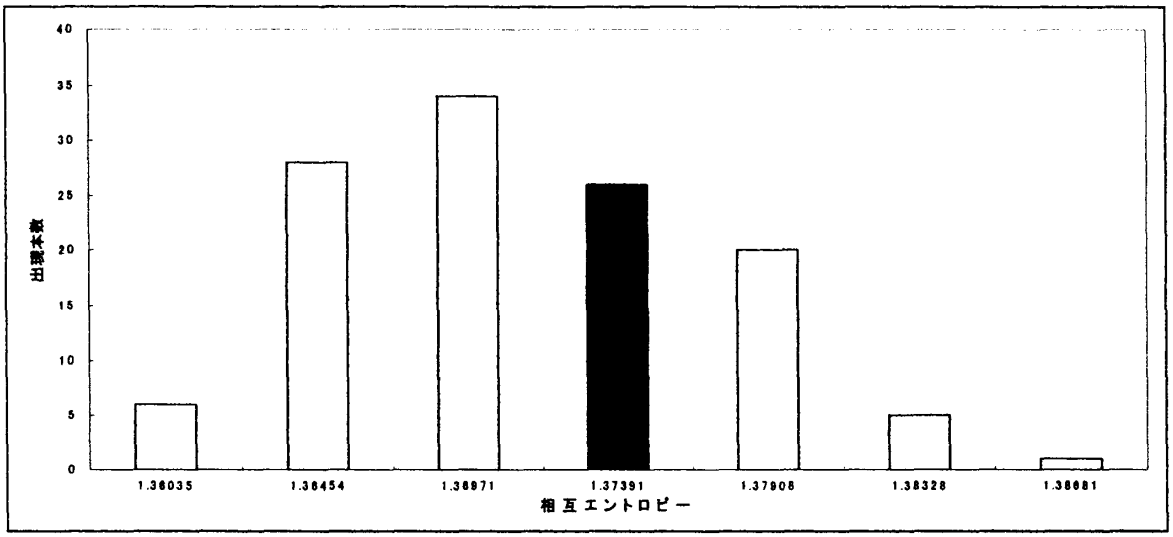

図 1 2HHB の相互エントロピーによる分類

これらのデータを元に，最小距離が一致した 3 組について相互エントロピーの平均値を算出し，生物学的見地からみたアライメント結果の相互エントロピー値と比較してみた．これを以下の表 3 に示した．

PDB ID	A	B	C	A-C
2HHB	1.371216	0.0060747	1.37391	0.002695
2MHB	1.363516	0.0149739	1.36354	0.0000248
2PGH	1.341087	0.0171877	1.34092	0.0001689

表 3 相互エントロピーの平均値と
生物学的見地からみたアライメント結果の相互エントロピー値との比較
（A:相互エントロピーの平均値　B:相互エントロピーの標準偏差　C:生物学的見地からみたアライメント結果の相互エントロピー値）

表 3 から分かるようにそれぞれのデータにおいて相互エントロピーの平均値と最も近かったものは，生物学的な見地からみたアライメント結果から求められた相互エントロピーの数値のグループであった．それぞれのデータにおける相互エントロピーの平均値を出せば，その値との誤差が最も小さい相互エントロピー値の配列結果の中に真のアライメント結果が含まれている，ということである．
これにより立体構造を考慮せずに真のアライメント結果のグループを特定することができる．

5. 考察

1) ヘモグロビンでの配列の組のうち，3 組は生物学的見地からみたアライメント結果と同じ距離が得られた．

一方で，生物学的な見地からみたアライメント結果と MOU-アライメントの結果の最小距離が異なることが多く出た．これにより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（文献[4]）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．

2) 最小距離を与える複数の配列の組各々に対し，相互エントロピーを計算した．その値の等しい組同士でグループを作り，全グループの相互エントロピーの平均値を算出した．その結果，求めた平均値と最も近いグループに生物学的見地からみたアライメント結果が含まれていた．

これに関して我々の例においては，有効であることを示している。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PAMマトリクスやたんぱく質の機能などの情報を用いてさらなる精度を向上させることが今後の課題である．

参考文献

[1] M.Ohya; S.Miyazaki; Y.Ohshima: “A new method of Alignment of Amino Acid Sequences”, Viva Origino 17, pp.139-151, 1989.
[2] http://www-cryst.bioc.cam.ac.uk/homstrad/
[3] 大矢　雅則; 渡邊　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, 67p., 牧野書店
[4] D.T.Jones; W.R.Taylor; J.M.Thornton: “The Rapid Generation of Mutation Date Matrices from Protain Sequences”, CABIOS 8 No.3, pp.275-282, 1992.