ICOT Technical Memorandom: TM-1133

TM-1133

3次元ダイナミックプログラミングを活用した蛋白質のアライメントシステム

広沢　誠、星田　昌紀、石川　幹人、
戸谷　智之

November, 1991

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

3次元ダイナミックプログラミングを活用した蛋白質のアライメントシステム[1]

広沢 誠，星田 昌紀，石川 幹人，戸谷智之[2]
(財) 新世代コンピュータ技術開発機構 (ICOT)[3]

1 はじめに

蛋白質配列の類似性を解析する典型的な手法であるマルチプルアライメントは、蛋白質の機能／構造予測や、生物種の進化系統樹の作成などに利用される、貴重な技術である。従来、マルチプルアライメントは、おもに生物学者の経験に頼って行われていたが、近年、蛋白質配列が次から次へと決定され、計算機によるマルチプルアライメントの導入が不可欠となっている。

これまで行われてきた計算機によるマルチプルアライメントは、そのほとんどが2次元のダイナミックプログラミング（以下ＤＰと呼ぶ）[1]によって得られた、配列2本のアライメント結果を組み合わせたものである[2]。こうした方法は、計算量が少ないという点では優れているが、類似性の低い蛋白質間のアライメントの品質は良くないという欠点がある。これに対して、我々が開発した統合アライメントシステム[3]は、3次元ＤＰを基本とするものであり、類似性の低い蛋白質配列間のアライメントの品質を大幅に改善するものである。3次元ＤＰを使用することによる計算量の増大は、並列推論マシン Multi-PSI を利用して計算時間の短縮を図った。

本論文は、統合アライメントシステムに使われている要素技術のうち、とくに3次元アライメントの組み合わせ方法を中心に紹介するものである。そして、統合アライメントシステムの適用例について具体的に示す。

2 統合アライメントシステム

我々が開発した統合アライメントシステムは、(1)“初期アライメントの作成”、(2)“アライメントの洗練化”、(3)“アライメント評価”の3モジュールにより構成される。

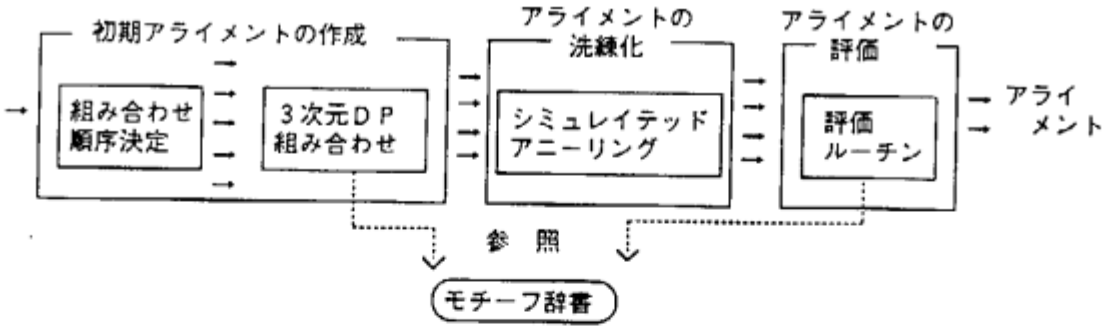

図1: 統合アライメントシステム

“初期アライメントの作成”のモジュールは、“組み合わせ順序決定”と“3次元ＤＰ組み合わせ”という2つのサブモジュールにより構成される。このモジュールでは、3次元ＤＰを適用して得られた3本のアライメント結果を、複数組み合わせてマルチプルアライメントを作成していく。しかしながら、作成されるマルチプルアライメントの品質は、3次元ＤＰの結果の組み合わせ順序により異なってくる。このため、まず、配列間距離解析を行って適切な組み合わせ順序を複数決定する（“組み合わせ順序決定”）。次に、各組み合わせ順序に対して、その指定順序で、3次元ＤＰを適用し、結果を組み合わせ、初期アライメントの作成を行う（“3次元ＤＰ組み合わせ”）。

“アライメントの洗練化”のモジュールは、“初期アライメントの作成”で求めた各初期アライメントを、シミュレーテッドアニーリング（以下ＳＡ）を用いて、一定時間洗練化する。ＳＡは、アライメントの評価尺度（Dayhoff Score）がより良くなるように、試行錯誤的な探索を行う[4]。その際、各初期アライメントを、ＳＡによる探索の初期状態として使用し、最良のアライメントへの収束を早めている。我々の採用しているＳＡの手法は、並列推論マシンを利用して

[1]Multiple Alignment System for Protein Sequences employing 3-dimensional Dynamic Programming

[2]Makoto Hirosawa, Masaki Hoshida, Masato Ishikawa, Tomoyuki Toya

[3]Institute for New Generation Computer Technology (ICOT)

いる。我々は、各プロセッサに異なる温度を割り当てる温度並列ＳＡを採用した。温度並列ＳＡでは、プロセッサ間で解交換を行うことで、いつでもその時点で最良の解が最低温度のプロセッサに得られる。そのため、与えられた時間において、準最適な結果が手軽に得られる。

“アライメント評価”は、“アライメントの洗練化”が済んだ複数のマルチプルアライメントの内から、優れているアライメントをいくつか選択し、ユーザに提示するモジュールである。選択基準は、Dayhoff Score が中心的になるが、全配列にわたり同一の文字が縦に並んでいるカラムがどの程度あるか、それがモチーフであるかなどの、生物学的なノウハウも含めて総合的に判断される。モチーフとは、蛋白質の構造や機能の点で重要な意味をもつ部分にみられる、特徴的な配列パターンである。本システムは、モチーフ辞書（実験段階）を装備しており、既知のモチーフの情報がマルチプルアライメントに反映できるようになっている。

さて“初期アライメントの作成”を行わず、始めからＳＡを用いてアライメントを行うことが理論的には可能である。しかし、ランダムな状態から満足のいくアライメント状態まで、すべてＳＡで行うのは現実的でなく、問題によってはかなりの時間が必要となる。けれどもこの時間は、初期状態にアライメントがある程度済んでいる状態を採用することで、大幅に削減できる。したがって、最適解に近い初期アライメントを見つけ出すことは非常に重要な課題である。このための注意深い解析処理を“初期アライメントの作成”が行っている。次章では、このモジュールを詳しく解説する。

3 “初期アライメントの作成”

“初期アライメントの作成”では、“組み合わせ順序決定”において、蛋白質配列間距離解析を用いて３次元ＤＰの結果の組み合わせ順序を決定し、この順序に従い“３次元ＤＰ組み合わせ”によりアライメントを作成する。説明の都合上、始めに“３次元ＤＰ組み合わせ”を述べ、その後、“組み合わせ順序決定”の概略を説明する。“組み合わせ順序決定”の詳細については、文献 [5] も参照されたい。

3.1 “３次元ＤＰ組み合わせ”

３次元ＤＰは、配列３本のアライメントを同時に行い、評価尺度のうえで最適なアライメントを求めるよく知られた手法である。計算量は配列の長さの３乗のオーダーであり、計算時間が多くかかるので、これまであまり本格的には使われてこなかったが、最近は計算機の処理能力が向上し、実現可能になった。しかし、４次元以上のＤＰは、とても実用的な時間では動作しない。そのため、４本以上のマルチプルアライメントには、３本以下のアライメントを複数組み合わせる必要がある。その際、従来のように２次元ＤＰの結果を組み合わせるよりも、次のようにして、３次元ＤＰの結果を組み合わせる方が効果的である [6]。

まず、配列の３本組のうち、２本が共通になるように２組を選び、それぞれ３次元ＤＰする。その結果を２本の共通配列を手がかりに組み合わせて、４本のアライメントをつくる。そして、アライメントされた４本のうち２本が共通になるような３本組を、新たに１組選び３次元ＤＰする。その結果と先の４本のアライメントとを、２本の共通の配列を手がかりに組み合わせて、５本のアライメントをつくる。以下これを繰り返す。

例を用いて説明しよう。配列０から配列５までの配列６本があるとする。そして、{{4,5,3}, {4,5,2}, {4,5,1}, {4,1,0}} のような、組み合わせ順序が指定されたとする（この順序の決定法は次節で述べる）。最初に３次元ＤＰを４回 ({4,5,3}, {4,5,2}, {4,5,1}, {4,1,0}) 行い、３本のアライメントを４組求めてしまう（ここで {4,5,3} は、配列４、配列５、配列３のマルチプルアライメントを意味する）。この後、組み合わせ手法により、{4,5,3} と {4,5,2} から共通配列である配列４と配列５を利用して {4,5,3,2} をつくる。そして、{4,5,3,2} と {4,5,1} より {4,5,3,2,1} を作る。最後に、{4,5,3,2,1} と {4,1,0} より {4,5,3,2,1,0} を作り、６本のマルチプルアライメントが完成する。

個々の組み合わせ手法は、ｎ本のアライメントと３本のアライメントより、(n+1) 本のアライメントを作るものである。以下では、とくに３本のアライメント２組より、４本のアライメントを作る手順について、実例に沿って説明する。

```
Step0   HTLV-1  : LLQAIAHLGKPSYINT       Step1   HTLV-1  : LLQAIAHL.GK.PSY.INT
        HTLV-2  : VLQAISLLGKPLHINT              HTLV-2  : VLQAISLL.GK.PLH.INT
        BLV     : LLEAIVHLGRPKKLNT              BLV     : LLEAIVHL.GR.PKK.LNT
                                                          1111100001111111111
        HTLV-2  : VLQAI-SLLGK-PLH-INT           HTLV-2  : VLQAI-SLLGK-PLH-INT
        BLV     : LLEAIVHL-GR-PKK-LNT           BLV     : LLEAIVHL-GR-PKK-LNT
        HIV     : FL--L-KLAGRWPVKTIHT           HIV     : FL--L-KLAGRWPVKTIHT

Step2   HTLV-1  : LLQAIAHL.GK.PSY.INT   Step3   HTLV-1  : LLQAIAHL-GK-PSY-INT
        HTLV-2  : VLQAISLL.GK.PLH.INT   (Result) HTLV-2  : VLQAISLL-GK-PLH-INT
        BLV     : LLEAIVHL.GR.PKK.LNT            BLV     : LLEAIVHL-GR-PKK-LNT
        HIV     : FL--L-KLAGRWPVKTIHT            HIV     : FL--L-KLAGRWPVKTIHT
                  1111100001111111111                      1111100001111111111
```

上の図は、３次元ＤＰの結果である、{HTLV-1,HTLV-2,BLV} と {HTLV-2,BLV,HIV} とのアライメント２組 (Step0) を組み合わせて、４本のアライメント {HTLV-1,HTLV-2,BLV,HIV} (Step3) を作る過程を示したものである。まず、

Step1 では、{HTLV-1,HTLV-2,BLV} に含まれる HTLV-2 と BLV のアライメントと、{HTLV-2,BLV,HIV} に含まれる HTLV-2 と BLV のアライメントの一致部分を見つけ出す。図では、一致領域に'1'、矛盾領域に'0'を割り当てている（この図で、'.'はカラムを合わせるために挿入したギャップであり、元からあるギャップは'-'で示されている）。'0'が並んだ矛盾領域では、2組に共通な配列である HTLV-2 と BLV とのアライメントが、上の組で [SLL,VHL] である一方、下の組で [-SLL,VHL-] となって異なり、矛盾を示している。

次に、Step2 では、一致領域と矛盾領域に分けて、アライメント {HTLV-1,HTLV-2,BLV,HIV} を作っていく。一致領域のアライメントを作るのは簡単である。左側の一致領域では、上の組の [LLQAI,VLQAI,LLEAI] と下の組の [VLQAI,LLEAI,FL--L] より、[LLQAI,VLQAI,LLEAI,FL--L] というアライメントを作る。右側の一致領域でも同様である。矛盾領域の上の組の [AHL,SLL,VHL] と下の組の [-SLL,VHL-,-KLA] では、以下の4種の配列群間の2次元DPを行い、得られた部分アライメントのうち、Dayhoff Score が最も良いものを矛盾領域のアライメントとする。

1. 上の組の [AHL,SLL,VHL] と，下の組の [-KLA] のアライメント.
2. 上の組の [AHL,SLL] と，下の組の [VHL-,-KLA] のアライメント.
3. 上の組の [AHL,VHL] と，下の組の [-SLL,-KLA] のアライメント.
4. 上の組の [AHL] と，下の組の [-SLL,VHL-,-KLA] のアライメント.

この例では、1と2のアライメント結果は同一で、[AHL-,SLL-,VHL-,-KLA] となり、それが最良の評価値を与える。そこで、それをその領域のアライメントとして、図の Step2 のような結果を得る。最後に、'.'を'-'に変えて最終的なアライメントとなる（Step3）。一致領域、矛盾領域という情報はアライメントの評価にも使える。一致領域のアライメントは矛盾領域のアライメントに比較して信頼性が良いとみなせるからである。

3.2 "組み合わせ順序決定"

配列がn本ある場合に、2本ずつを共通にして、3本の組で組み合わせる順序は $C_{n,3}(\prod_3^{l=n-1} C_{l,2})(n-3)!$ 通りある。例えば、6本の場合には 28800 通りの膨大な組み合わせがある。この組み合わせから最終的に品質の良いアライメントを作ると予想される、少数の組み合わせを選び出す必要がある。このために配列間距離解析を用いる。

まず、配列 i と配列 j の類似性距離 S(i,j) を、すべてのペアについて求める。現在は、S(i,j) として、配列 i と配列 j を2次元DPでアラインした時のコスト Cost(i,j) を逆符号にしたものを利用している。そして、3つの配列 i と j と k が何如に近いかの尺度 R(i,j,k) を、S(i,j)、S(i,j)、S(i,j) を足し合わせたものと定義し、すべての三つ組についてこの値を求める。この R(i,j,k) を、配列 i と j と k に3次元DPを適用したときの、アライメントの信頼性と考える。

次に、上で求めた R(i,j,k) を用いて、3次元DPの結果を組み合わせる順序を決定していく。なるべく信頼性の高い三つ組について、3次元DPを行い、信頼性の高い組から、その結果を組み合わせていくように組み合わせ順序を決める。つまり、始めに、信頼性 R(i,j,k) が最大の三つ組を選択し、次に、前の三つ組と組み合わせ可能な三つ組のうち、最も信頼性の高い三つ組を選択する。そしてまた、すでに選択されている三つ組群と組み合わせ可能な三つ組のうち、最も信頼性の高い三つ組を選ぶ。これを、すべての配列が選択された三つ組群に含まれるまで、繰り返す。

信頼性の値がそれほど違わない三つ組が複数ある場合には、組み合わせ順序として複数の候補を決定する。それらは、統合システムの後の過程において優劣が決められる。

4 統合システムの適用結果例

図2に示した配列を、統合システムを用いてアライメントを行った。これは、レトロウイルス等の endonuclease の一部分の配列である。まず、"組み合わせ順序決定"により組み合わせ順序を7つ決定した。この7つに対してアライメントを求め、アラインメントの Dayhoff Score を基準として5つの初期アライメントを選択した。さらに"アライメントの洗練化"を、SAでもって行い、最終的に5つのアライメントを求めた。SAは各配列に対して1時間行った。

```
Copia17.6(0): ILDFHEKLLHPGIQKTTKLFGETYYFPNSQLLIQNIINECSICNLAKTEHRNTDMPTKTT
M-MULV   (1): LLDFLHQLTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVNASKSAVKQGTR
HTLV     (2): LTDALLITPVLQLSPAELHSFTECGQTALTLQGATTTEASNILRSCHACRGGNPQHQMPRGHI
RSV      (3): VADSQATFQAYPLREAKDLHTALHIGPRALSKACNISMQQAREVVQTCPHCNSAPALEAGVN
MMTV     (4): ISDPIHEATQAHTLHHLNAHTLRLLYKITREQARDIVKACKQCVVATPVPHLGVN
SMRV     (5): ILTALESAQESHALHHQNAAALRFQFHITREQAREIVKLCPNCPDWGSAPQLGVN
```

図2: アライメントをするべき配列

最後に"アライメント評価"において、アラインメントの Dayhoff Score と、全配列に同一である文字（保存配列）の個数の評価を行った（表1）。そして、アラインメントの Dayhoff Score が最良であり、保存配列の文字数が最多であるという理由で、表1の上から2番目のアライメントを最適なものとして選択した。

このアライメントは、H..H.....C..C という Zinc Finger モチーフを捕らえている。Zinc Finger は蛋白質がDNAに結合する機能を持つ。このように、生物学的に重要な蛋白質配列の部位を特定できるアライメントを作成できることは、このシステムの適用結果を、蛋白質の機能／構造予測に利用できる可能性を示唆している。

組み合わせ順序	Cost	Dayhoff Score (SA前)	Dayhoff Score (SA後)	保存配列	モチーフ
{5,4,3},{5,4,2},{5,4,1},{5,4,0}	10	626	687	H C..C	
{5,4,3},{5,4,2},{5,4,1},{4,1,0}	11	703	709	H...H C..C	Zinc Finger
{5,4,3},{5,4,2},{5,4,1},{4,3,0}	12	673	702	H C..C	
{5,4,3},{5,4,2},{5,4,0},{5,4,1}	12	626	687	H C..C	
{5,4,3},{5,4,2},{5,4,1},{5,1,0}	13	667	703	H C..C	
{5,4,3},{5,4,2},{5,4,0},{4,3,1}	13	551	641	H C..C	
{5,4,3},{5,4,1},{5,4,2},{5,4,0}	13	626	687	H C..C	

表 1: アライメントの評価

```
Copia17.6: -------ILD--F-------HEKLLHPGIQKTTK-LF--GET-YY-FPNSQLLIQNIINECSICNL-AKT-EHR--N-TDMPTKTT
Mu-MULV  : -------LLD-FL-------HQ-LTHLSFSKM-KALLERSHSPYYMLNRDRTL-KNITETCKACAQ-VNA-SKS--A-VKQGTR--
HTLV     : LTDALL-ITP-VLQLSPAELHS-FTHCGQTAL-T-LQ--------GATTTEA--SNILRSCHACRG-GNPQHQMPRGHI-------
RSV      : VADSQATFQAYPLR-EAKDLHT-ALHIGPRAL-S-KA--------CNISMQQA--REVVQTCPHC---NSA-PALEAG-VN------
MMTV     : -------ISD-PIH-EATQAHT-LHHLNAHTL-R-LL--------YKITREQA--RDIVKACKQCVV-ATPVPHL--G-VN------
SMRV     : -------ILT-ALE-SAQESHA-LHHQNAAAL-R-FQ--------FHITREQA--REIVKLCPNCPDWGSA-PQL--G-VN------
```

図 3: 統合アライメントシステムによる最適アライメント

5 おわりに

3次元DPを基本として蛋白質配列のマルチプルアライメントを求める、統合アライメントシステムについて紹介した。そして、このシステムで重要な役目を果たしている“初期アライメントの作成”について詳細に述べた。また、このシステムの適用例を示し、その結果、生物学的に重要な配列部位を特定できたことを述べた。今後は、多数本の長い配列の処理に挑戦すると同時に、さらに類似性の低い蛋白質配列も扱えるようにシステムを改良していく予定である。

文部省科学研究費補助金重点領域研究「ゲノム情報」の班員の方々からは、多くの助言や批判をいただきました。ここに感謝の意を表します。また、本研究の機会を与えていただいた、ICOT の古川康一所次長、および内田俊一研究部長にお礼申し上げます。

参考文献

[1] Needleman and Wunsch "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins", in J. Molecular Biology 48, 1970, pp443-453

[2] Geoffrey J.Barton "Protein Multiple Alignment and Flexible Pattern Matching" in Methods in Enzymology Volume 183 Academic Press, 1990, pp.626-645

[3] 石川、星田、広沢、戸谷、鬼塚、新田、金久 :“並列推論マシンを用いたタンパク質の配列解析”, 情報処理学会 情報学基礎研究会報告 23-2, 1991

[4] 戸谷、星田、石川、新田、金久 :“並列シミュレーテッドアニーリングを用いたマルチプルアライメント”, 情報処理学会第 43 回全国大会論文集, 1991

[5] 広沢、星田、石川 :“蛋白質配列間距離解析を用いた 蛋白質の相同性解析システム” 情報処理学会第 43 回全国大会論文集, 1991

[6] 石川、星田、広沢、戸谷、新田 :“3次元ダイナミックプログラミングを用いたタンパク質の配列解析” 情報処理学会第 43 回全国大会論文集, 1991