ICOT Technical Memorandom: TM-1238

TM-1238

ルールベースに基づく
蛋白質配列解析システムの試み

廣澤　誠、石川　幹人、星田　昌紀

November, 1992

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

ルールベースに基づく蛋白質配列解析システムの試み

A tiny step toward Rule-based a Protein Sequence Analysis System

広沢 誠、石川 幹人、星田 昌紀

Makoto Hirosawa, Masato Ishikawa, Masaki Hoshida

(財) 新世代コンピュータ技術開発機構 (ICOT)

Institute for New Generation Computer Technology (ICOT)

Abstract

We will present a paradigm that makes multiple sequence alignment by knowledge.

The technology of multiple alignment of protein is important for protein sequence analysis. So far, many alignment algorithm have been developed. However, they produce just temprorary alignment which biologists must refine to produce biologically meaningful alignment.

We interviewed alignment experts and extracted knowledge from them and analyze them. The knowledge was essentially know-how to find possiible motifs in the temporary alignment and knowledge on motifs.

Based on this analysis, we formulated alignment system with *an aligner* and *an inteligent refiner* which modifies alignment produced by aligner. The intellgent refiner refines the alignment using rules stored in *a refinement rule base* according to the prioriy of the rules. And Some rules consult *biological knowledge base* which contains motifs and so on.

1 はじめに

蛋白質の機能／構造予測や、生物種の進化系統樹を行うために、蛋白質配列の類似性を解析することは必須であり、この中でもマルチプルアライメントは重要な技術である。従来は、アライメントに割り当てられた評価値を最適化する手法が採られていたが、これは不十分なものである。これを、解決するために我々は、知識処理を用いたアライメントシステムを定式化し、一部を実現した。以下、順を追ってシステムを説明する。

2 背景

従来、マルチプルアライメントは、すべての過程において生物学者が経験に頼って行われていた。しかし、これは、労力がいる仕事なので部分的に計算機が導入されることが多くなった。 現在、彼らが行っている典型的な手法は第1図（上）に示すような2段階のものである。

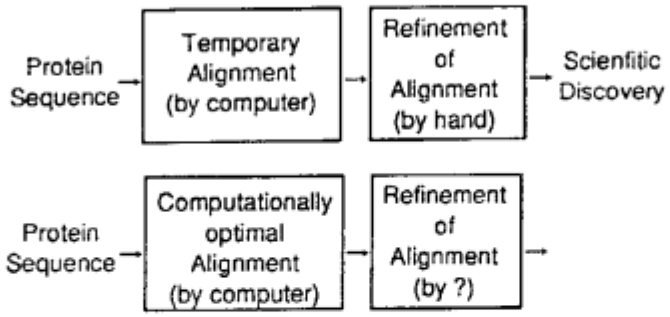


図 1: Multiple Alignemnt using computer [Biologist(upper),Computer Scientist(lower). Computer Scientists can make scientific discovery only if they can cooperate with biologist.]

まず、第1段階では、計算機を用いて何らかのアライメントを作成する。しかしながら、このアライメントは、配列に含まれているモチーフ、または保存配列について見当をつけるために作られたものであり、必ずしも最終的なアライメントではない。 第2段階では、前段階で作られたアライメントをモチーフなどの生物学的知識を用いて徐々に

修正していき最終的なアライメントを作成する。そして、このアライメントを解析することにより生物学的発見をする。

しかし、蛋白質配列の決定技術が著しく進歩したために、実験生物学者がアライメントするべき蛋白質配列は増えてきた。また、データベースに登録された配列を解析することにより生物学的研究をする計算機生物学者が行うべきアライメントの回数も増えてきた。このため、生物学者が行うべきアライメントの仕事量は増加し、第2段階の修正過程にも計算機による助けが必要になった。

さて、最近は、遺伝子治療や遺伝子診断などが身近な言葉になるにつれ、計算機科学者の中にも遺伝子を解析することにより科学的発見をしたいという人たちが徐々ではあるが現れてきた。しかし、彼らは第1段階として、計算機的に定義された、最適、または、準最適なアライメントを求めるプログラムを開発することができる（計算機的に最適なアライメントの定義については [Hirosawa 1993] を参照のこと）。また、計算機科学者は、このプログラムを実行できる計算機環境に恵まれている（第1図（下））。しかしながら、計算機的に最適なアライメントは生物学的に意味のあるアライメントではないので、生物学者の場合と同様に修正過程が必要である。

```
17.6    : -------ILDF--HE-KLL-HPGIQKTTKLF--GET-YY-FPNSQLLIQNIINECSICNLAKTEHRNTDMPTKTT
M-MULV  : -------LLDF-LHQ---LTHLSFSKMKALLERSHSPYYMLNRDRTL-KNITETCKACAQVNASKSAVKQGTR--
RSV     : VADSQATFQAYPLREAKDL-HTALHIGPRAL--SKA-CN-ISMQQA--REVVQTCPHCNSAPALEAGVN------
   (Evaluation value = 161)
```

図 2: Computationally optimal alignment

```
17.6    : -------ILDF--------HEKLLHVGIQKTTKLF--GET-YY-FPNSQLLIQNIINECSICNLAKTEHRNTDMPTKTT
M-MULV  : -------LLDF-------LHQ-LTHLSFSKMKALLERSHSPYYMLNRDRTL-KNITETCKACAQVNASKSAVKQGTR--
RSV     : VADSQATFQAYPLREAKDLHT ALHIGPRAL--SKA-----CN-ISMQQA--REVVQTCPHCNSAPALEAGVN------
   (Evaluation value = 156)
```

図 3: Biologically optimal alignment [The alignment captures Zinc Finger Motif (two H and two C).]

修正過程が必要な例を第2図と第3図とに示す。第2図のアライメントがレトロウイルスのある蛋白質配列に対しての計算機的に最適なアライメント（評価値は 161）であり、第3図が生物学的に意味のあるアライメントのひとつである（評価値は 156）。後者は、モチーフとして、二つの "H" と、二つの "C" からなる Zinc Finger を持っている。これに対して、前者は、後者より評価値が良いが、Zinc Finger のモチーフを持っていない。したがって、計算機的に最適なアライメントを修正する過程が必要である。

しかし、この修正過程を行うには、生物学的知識が必要であるのでこの過程は生物学者に依頼せざるをえない。このため、生物学者と協力関係がある環境にある計算機科学者のみ生物学的発見をすることができる。したがって、計算機科学者に対して生物学的発見をする機会を提供する配列解析システムは、修正過程も含めた全過程が自動化されたものである必要がある。

このように、アライメントの全過程を自動化したシステムは生物学者にとっても、生命科学に興味のある計算機科学者にも有用である。我々は、知識処理を用いて上記のようなシステムを構築している。次の章においてシステムの概要を説明する。

3 システムの概要

我々の配列解析システムの概要を第4図に示す。このシステムは、アライメントするべき配列、または、既にある手法によりアライメントされた配列を入力とする（アライメントされた配列は、生物学者によりあまり時間をかけないでされたものである場合もあるし、計算機によりされたものである場合もある）。そして、出力として、生物学的意味のあるアライメント、発見されたモチーフとその説明、さらに、モチーフの候補として検出された保存配列がシステムの出力となる。

配列がアライメントされてない場合には Aligner により最適、または、準最適なアライメントが作成される。Aligner は、アライメントされるべき配列が3本の場合には、Dynamic Programming[Needleman and Wunsh 1978; Murata 1985] により、最適なアライメントを作成する。配列が4本の以上場合には、ツリーベース反復改善法 [Hoshida

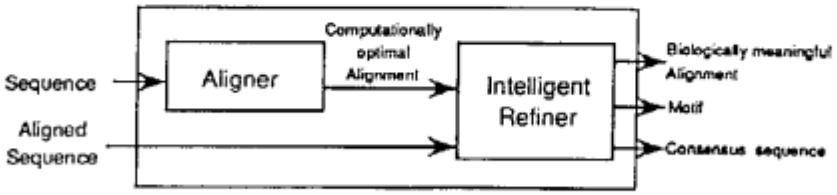


図 4: Overview of the system

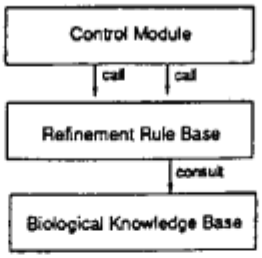


図 5: Structure of Inteligent Refiner

1992] などにより、準最適なアライメントを作成する。準最適なアライメントを作成する手法は、他のアルゴリズムを適用しても同様に可能であるので、

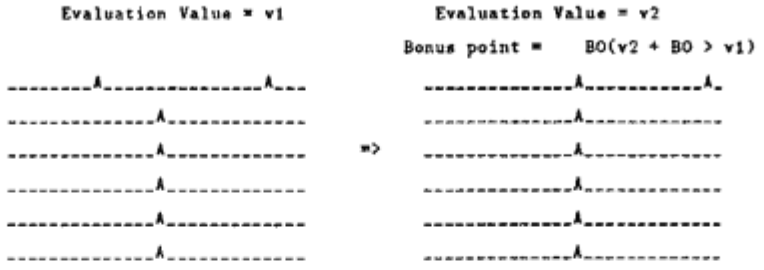
この作成されたアライメント、または、システムに入力されたアライメントは、*Intellignent Refiner* に送られ、ここで生物学的意味のあるアライメントが作成される。Intellignent Refiner は、登録されているルールを 優先度が高い順に次々に適用し、徐々にアライメントの品質をあげていく。次の章では、Intellignent Refiner について説明する。

4 Intellignent Refiner

Intellignent Refiner の構成を第5図に示す。*Control Module* は、*Refinement Rule Base* に登録されているルールを次々に発火させ、モチーフや保存配列を発見していくことにより、徐々に品質の高いアライメントを作成する。ルールは必要であれば *Biological Knowledge Base* を参照する。これにはモチーフ情報がなどが含まれている。

Refinment Rule Base に登録されているルールは、基本的に、与えられたアライメントの中に、少し修正しすれば、モチーフ、または、保存配列が生じそうな部分を見つけ出す。そして、モチーフを形成するという制約の基に計算機的に最適なアライメントを求める。そして、このように求めたアライメントの評価値に、モチーフ、または、保存配列を捕らえているというボーナスポイントを足し合わせた値が、以前の評価値よりも良い場合は、このアライメントの修正を受け入れる。例えば、第6図("-" は "A" 以外のアミノ酸を意味する) では、"A" を縦に一列に並べる方法には、一番上の配列の右の "A" を用いる場合と右の "A" を用いる場合が考えられるが、前者が用いられている。

このように、ルールを適用することにより、アライメントの制約、すなわち、モチーフ、または、保存配列が見つけ出されていく。



第6図: Refinement of alignment [Computaionally opitmal alignment is searched on the constraint that the alignment should have aligned "A".]

Refinment Rule Base には、現在、10個のルールが登録されている。これらは、アライメントの専門家とのインタビューを参考にして作成した。ここでは、2つのルールのみを第7図に示す。ルール2は、第2図のアライメントから、生物学的に意味のあるアライメントを作成するために用いられる。ルール1については、後に記述する。

Rule 1

IF あるモチーフ (m_i) がアライメントで特定され **AND**

モチーフ m_i を持つ蛋白質が他のモチーフ m_j を持っていれば

THEN Motif_finding routine が呼び出され、m_j を特定する。

Rule 2

IF Biological Knowledge Base に登録されているあるモチーフの中に、同一種類のアミノ酸 x が（x_i and x_j）存在し、**AND**
この2つのアミノ酸 x_i、x_j の間に他の保存アミノ酸が存在せず **AND**
refinement されるべきアライメントの中に、x が一部の配列を除き存在するカラム c_i と x が全ての配列に存在するカラム c_j があれば、

THEN Modification routine が呼び出され、以下の制約の下でアライメントを修正する（制約：c_i において x が存在しない配列を s_t とし、c_j の配列 s_t に対応するアミノ酸を $x_{j,t}$ とした時、$x_{j,t}$ を c_i に揃える）。

第7図：Alinment Rule Base

第8図に、Biological Knowledge Base の一部を示す。現状は実験レベルなので登録されている知識はまだ少ない。モチーフの表現は、Prosite[Bairosh 1991] に準拠している。モチーフに対しては、そのモチーフの名前、または、そのモチーフを持つ蛋白質の名前が記述されている。さらに、tyrocine kinase が kinase のサブクラスであるというような蛋白質の階層関係が記述されている。

```
motif(name, zinc_finger, ''H-X(3,5)-H-X(10,25)-C-X(3,5)-C'').
motif(protein, kinase,''[LIV]-G-X-G-[FY]-[SG]-X-[LIV]'').
motif(protein, kinase(tyrosine), ''[LIVMFYC]-X-[HY]-X-D-[LIVMFY]-K-X(2)-N-[LIVMFC](3)'').
upper_concept(kinase(tyrosine), kinase).
motif(protein,Protein,Motif)  :-    upper_concept(Protein,X), motif(protein,X,Motif).
```

第8図：Biological Knowledge Base

さて、Alignment Rule Base のルール1は、ある蛋白質のモチーフが見つかった時、その蛋白質が持つ、他のモチーフを探せというものである。この階層関係を用いると、tyrocine kinase のモチーフが見つかった時、上位概念である kinase に登録されているモチーフを探すこともできるし、逆に、kinase のモチーフが見つかった時、下位概念である tyrocine kinase に登録されているモチーフを探すこともできる

5 適用例

第2図のアライメントに、Intelligent Refiner を適用した。第7図に示してある Rule 2 などが適用され、第3に示されている生物学的に意味のあるアライメントが作成された。これは、Zinc Finger motif を捕らえている。

謝辞

九州大学の隈啓一さん、岩部直行さんには、実際にアライメントする過程を見せていただき、また、たびかさなる質問にも答えていただきました。ここに深く感謝いたします。

References

[Bairoch 1991] Bairoch,A. Prosite : A dictionary of protein site ans pattern : User manual Release 7.00, May 1991.

[Hirosawa 1993] Hirosawa,M, Ishikawa,M, Hoshida,M. "Formulation of Protein Sequence Analysis using Knowledge" *Proceedings of 26th Hawaii Internation Conference on System Science*

[Hoshida *et al.* 1992] 星田、石川、広沢、戸谷 "並列反復法による蛋白質アライメント" 情報処理学会情報基礎研究会ゲノム特集

[Murata 1985] Murata,M. "Simulteneous comparison of three protein sequences" *Proc. Natl. Acad. Sci. USA Vol. 32,* 1985, pp.3073-3077.

[Needleman and Wunsch 1970] Needleman,S.B. and Wunsch,C.D. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins." *J. of Mol. Biol.*, 48, 443-453.