TM-1212 他

情報処理学会
第45回全国大会論文集

October, 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

知識を用いた蛋白質配列解析システムの試み

広沢 誠、星田 昌紀、石川 幹人

(財)新世代コンピュータ技術開発機構

1 はじめに

蛋白質の相同性解析の技術であるマルチプル・アライメントは、蛋白質の機能・構造予測、生物種の進化系統樹の作成の際に欠かせない技術である。従来は、複数本の蛋白質配列のマルチプル・アライメントは、生物学者が経験と勘を頼りに行ってきた。しかし、蛋白質配列の決定技術が著しく進歩したために、アライメントするべき蛋白質配列のみではなく、アライメントの回数も増えてきた。このため、計算機を用いたマルチプル・アライメントが導入されつつある。

現在まで多くのマルチプル・アライメントのアルゴリズムが開発されてきている。これらは、アライメントに対して定義された評価値を最適化することを目的とするものである。配列の本数が少ない時（2 or 3）には上記の計算機的に最適なアライメントを求めることができる[Needleman and Wunsch 1978 ; Murata 1985]。それ以上の本数の場合にも理論的には最適なアライメントを求めることができるが、計算量が膨大であるので、実際には準最適なアライメントを求めるアルゴリズムが用いられる[Barton 1990; Berger and Manson 1991]。

しかしながら、上記のアルゴリズムが導き出すアライメントは、生物学的に意味のあるアライメントでは必ずしもない。我々は、原点に戻り、生物学的に意味のあるアライメントとは何かを考えた。そして、アライメントの専門家にインタビューし、どのようなアライメントを良いアライメントであると見なしているかを把握し、また、彼らがアライメントを行う時に意識、無意識に用いているルール、知識を抽出した。そして、これらを解析した結果を反映したマルチプル・アライメントシステムを試作した。

以下、計算機的に最適なアライメントが、生物学的に意味のあるアライメントではない例を示す。そして、我々のマルチプル・アライメントシステムを紹介し、このアライメントシステムが生物学的に意味のあるアライメントを作成することを示す。なお、このシステムの詳しい内容については[Hirosawa 1992]を参照していただきたい。

2 計算機的に最適なアライメントの問題点

この章では、計算機的に最適なアライメントが、生物学的に意味のあるアライメントではない例を示す。しかしながら、生物学的に意味のあるアライメントの定義をアラインされる配列に依存せずに行うことは困難である。ここでは、レトロウイルスというウイルスに特有である endonuclease という蛋白質のマルチプル・アライメントを例題にとる。なぜなら、この蛋白質の生物学的に意味のあるアライメントは、過去の研究により明らかであるからである。図1に6種類のレトロウイルスが持つ endonuclease の蛋白質配列をアライメントしたものを示す。これは、生物学的に意味のあるアライメントのひとつである。

図1のアライメントが生物学的に意味のあるアライメントである理由は、図にも示されているように、全ての配列において共通するアミノ酸として、左側に二つの "H"、右側に二つの "C" を捕らえているからである。。分子生物学では、全ての配列に共通なアミノ酸パターンなどをモチーフと呼ぶが、この図のモチーフは、Zinc Finger のモチーフと呼ばれており、生物学的に重要な機能を持つとされている。以降、例の蛋白質配列のアライメントでは、Zinc Finger のモチーフを持つものを生物学的に意味のあるアライメントと定義する。

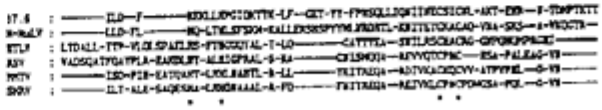

図1：生物学的に意味のあるアライメントの例

前に述べたように、計算機的に最適なアライメントを求められるのは、3本の配列のアライメントまでである。これは、Dynamic Programming という手法で行われる[Needleman and Wunsch 1978]。図1の蛋白質配列から3本の配列を選び、これに対し Dynamic Programming を適用し、計算機的に最適なアライメントを求めた。3本の配列を6本の配列から選び組み合わせは20通りあるので、20通りの計算機的に最適なアライメントが生成された。この内、生物学的に意味のあるアライメントは6通りのみであった。

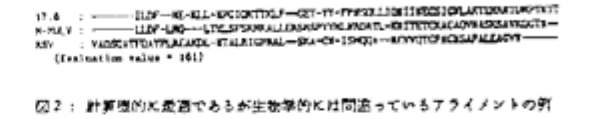

図2：計算機的に最適であるが生物学的には間違っているアライメントの例

図2に生物学的は間違っているアライメントの例を示す。この例では、RSV の配列で、2番目の共通な "H" を形成するべき "H" が、17.6 と M-MULV で、1番目の共通な "H" を形成するべき "H" と、同じカラムに並べられてしまっている。

3 Intelligent Refiner を用いたアライメントシステム

前の章で示したように、計算機的に最適なアライメントが、必ずしも、生物学的に意味のあるアライメントではない。以下、生物学的に意味のあるアライメントを作成するシステムを紹介する。我々のシステムは、*Aligner* と *Intelligent Refiner* の2つのモデュールにより構成される。

Aligner は、与えられた配列の計算機的に最適または準最適なアライメントを求める。これが Intelligent Refiner への入力となる。我々は、Aligner のアルゴリズムに[Ishikawa 1992]を採用しているが、他のアルゴリズムで作成されたアライメントを In-

Framework of protein sequence analysis with knowledge
Makoto Hirosawa, Masaki Hoshida, Masato Ishikawa
Institute for New Generation Computer Technology.

telligent Refiner とすることができるので、以降、Aligner の説明は行わない。Intelligent Refiner は、Aligner により作成されたアライメントを生物学的な知識を用いながら徐々に更新しながら、配列に含まれる重要な部位を特定していく。更新のサイクルが進むごとに、Intelligent Refiner は、配列が意味する生物学的情報をより詳細に把握していく。

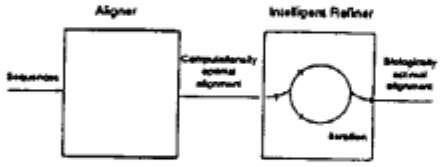

図3 : Intelligent Refiner を用いたアライメントシステム

図4 に Intelligent Refiner の構成を示す。*Refinement Rule Base* には、アライメントの更新を行う時に用いられるルールが蓄積されている。この中には、我々が生物学者から抽出したルールなどが含まれている。*Control Module* は、Refinement Rule Base に登録されているルールを実行することによりアライメントを徐々に更新していく。ルールを用いる時に必要な時には *Bilogical Knowledge Base* に蓄えられている生物学的知識を参照する。

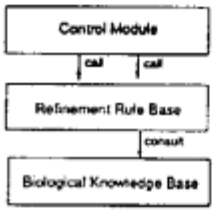

図4 : Intelligent Refiner の構成

Refinement Rule Base には現在１０個のルールが登録されている。スペースの関係でこの内、２つのルールのみを図５に示す。また、Bilogical Knowledge Base に登録されている知識の一部を図５に示す。モチーフの表現は [Bairoch 1991] に準拠している。

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_i とし、c_j の配列 s_i に対応するアミノ酸を $x_{j,i}$ とした時、$x_{j,i}$ を c_i に揃える)。

図5 : Refinement Rule Base に登録されているルール

```
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).
```

図6 : Bilogica Knowledge Base に登録されている知識

4 適用例

図2の Intelligent refiner を適用した (図7)。Rule 2 などが適用され、生物学的に意味のある (Zinc Finger を new えている) アライメントが作成された。なお、計算機的な評価値は 計算機的に最適なアライメントは 161 であったが、生物学的に意味のあるアライメントでは 156 となり、評価は悪くなった。これは、従来のアライメント手法では誤りとなる場合も、我々の手法を用いると適切なアライメントを作成できることを示している。

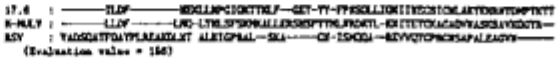

図7 : Intellignent Refiner の適用例

References

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

[Berger and Manson 1991] Berger,M. and Manson,P. A novel randomized iterative strategy for aligning multiple protein sequences. *Computer Application in the Biosciences*, 7, 1991. pp.479-484.

[Barton 1990] Barton,J.G. Protein Multiple Alignment and Flexible Pattern Matching. in *Methods in Enzymology Vol.183*, Academic Press, 626-645.

[Hirosawa *et al.* 1992] Hirosawa,M, Ishikawa,M, Hoshida,M Formulation of Protein Sequence Analysis using Knowledge 情報処理学会情報基礎研究会 ゲノム特集 (1992)

[Ishikawa *et al.* 1992] Ishikawa,M., Hoshida,M., Hirosawa,M., Toya,T. and Nitta,K. Protein Sequence Analysis by Parallel Inferrence Machine. *Proc. Int. Conf. on Fifth Generation Computer Systems 1992.*

[Murata 1985] Murata,M. (1985) Simulteneous comparison of three protein sequences *Proc. Natl. Acad. Sci. USA Vol. 82*, 1985, pp.3073-3077.

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