タンパク質の配列解析プログラム

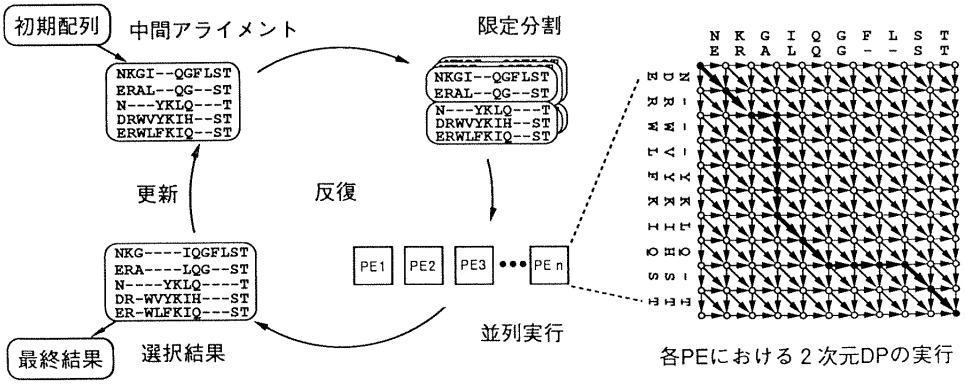
―― 並列反復改善法によるマルチプルアライメント ――

概要

タンパク質は２０種類のアミノ酸が鎖状に連なってできており、その構造や機能はアミノ酸の配列から決定される。タンパク質の配列解析は、タンパク質の機能や構造の予測、またはタンパク質同士の進化的関係の把握を行う上で重要な研究課題である。我々は**並列反復改善法**を用いて、配列解析の代表的な問題である**マルチプルアライメント**の解決に取り組んでいる。一般に、この種の組合せ最適化問題を解決するには莫大な計算量が必要であるが、我々の開発した並列反復改善法によるシステムは、PIM 上で実行することによって、生物分野の研究者が実際に扱う規模の問題を、高速かつ高品質に解決する。

特徴

- **反復改善法:** 部分的な改善を反復的に行い、その結果として、より高品質のアライメント結果を効果的に得ることを可能とした手法である。
- **並列実行による速度向上:** 組合せ問題における探索木の枝の評価を、多くの要素プロセッサを使って並列に行うことで、マルチプルアライメントが高速に行える。
- **探索木の枝刈り:** 探索木の枝刈りを効果的に行うヒューリスティクスである、**限定分割手法**を導入し、この組合せ問題を現実的な時間内で解決している。



並列反復改善法

タンパク質の配列解析

タンパク質は２０種類のアミノ酸から構成されており、各アミノ酸はそれぞれ異なるアルファベット１文字で表される。ひとつのタンパク質は平均２００個程度のアミノ酸が連なっており、２０種類の文字からなる配列として表現される。アミノ酸は対水和性、極性、大きさなどの性質を持っており、タンパク質の構造や機能は、アミノ酸の順序によって決定されると考えられている。

タンパク質配列を決定する実験技術はすでに確立されているため、２万以上の配列が、文字列として特定されており、日々その数は増加し続けている。また、タンパク質の構造も次第に特定されてきてはいるものの、それには大変な困難が伴うため、１つのタンパク質の構造決定に１年程度を必要とするのが現状である。そのため、構造が決定されたタンパク質の数は、今のところ極めて少ない。

そこで、配列情報からタンパク質の構造や機能を推察する手法の開発が、期待されている。アミノ酸配列が似ているタンパク質は、類似の構造や機能をとりやすいので、配列解析によって、未知の構造や機能へのアプローチが行える。配列の類似性解析の代表的な手法に、マルチプルアライメントがある。

次に、６本のタンパク質配列の一部分をマルチプルアライメントした例を示す。

```
---------------HEKLLHPGIQKTTKLF-GET---YYFPNSQLLIQNIINECSICNLAKTEHRNTDM--P-TKTT
-------------LHQ-LTHLSFSKMKALLERSHSPYYMLNRDRTL-KNITETCKAC--AQVNASKSAVKQG-TR--
-PVLQ---LSPA-ELHS-FTHCG---QTAL--TLQ----GATTTEA--SNILRSCHAC---RGGNPQHQMPRGHI---
QATFQAYPLREAKDLHT-ALHIG---PRAL--SKA---CNISMQQA--REVVQTCPHC------NSAPALEAG-VN--
--ISD--PIHEATQAHT-LHHLN---AHTL--RLL---YKITREQA--RDIVKACKQC---VVATPVPHL--G-VN--
--ILT--ALESAQESHA-LHHQN---AAAL--RFQ---FHITREQA--REIVKLCPNC---PDWGSAPQL--G-VN--
```

各配列の文字のひとつひとつがアミノ酸を表している。例えば、最上段左端に見られるHEKLは、それぞれヒスチジン、グルタミン酸、リシン、ロイシンである。ところどころにある“-”は、ギャップと呼ばれている。このギャップを配列中に入れることで、各カラムに同じアミノ酸か、もしくは性質の似通ったアミノ酸が並ぶようにした配列群が、マルチプルアライメントである。この例には、H....HやC..Cなどの共通文字が、縦に並んでいる箇所が見られる。このような箇所は、タンパク質の構造や機能のうえで重要である可能性が高く、一般に**配列モチーフ**と呼ばれている。こうしたモチーフは、突然変異や自然淘汰といった進化の過程を経ても、共通にずっと保存されてきた部分と考えられている。

ダイナミックプログラミングによるアライメント

ダイナミックプログラミング（以下DP）は、最適なアライメントを見つける基本的な手法である。このDPはN次元ネットワーク上の最適経路探索とみなすことができる。２組の配列群が与えられたときには、まず矢で連結された多数のノードからなる２次元のネットワークを作る（図１）。各々の矢にはスコアが与

えられており、ネットワーク上で、左上から右下への矢のスコアの総計を最大にする経路を探索するのである。この例では、黒丸のノードを通る矢の経路が最適であり、その経路が、最適アライメントに対応している。矢のスコアには比較される文字、または文字群の間の類似度が反映されている。タンパク質配列のアライメントの場合、スコアにはDayhoffの数値テーブルが最も一般的に使用されている。これは、アミノ酸の突然変異率を統計的に解析した結果から得られた。

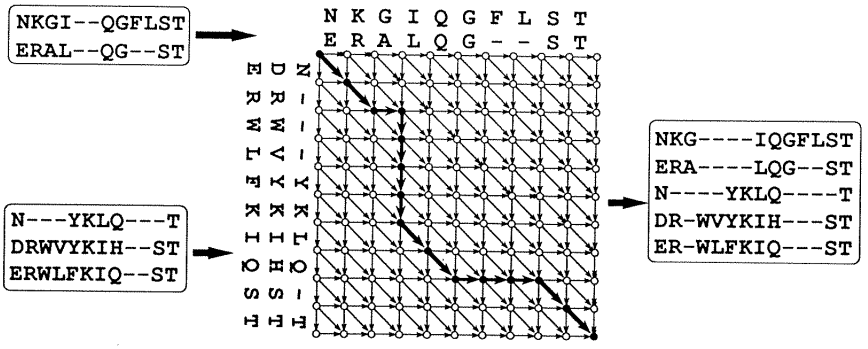


図1: 2次元DPの処理

理論的にはN次元のDPを用いることで、N組の配列群間の最適なアライメントが得られるのであるが、配列群の数が増えると指数的に計算量が増大する。現在でも、3組を越える配列群間のアライメントに対しては、現実的に計算機上での実行が困難である。従来から、生物分野の研究者は、アライメントされた配列群を結合することでマルチプルアライメントを作成してきた。**ツリーベース法**と呼ばれる従来の典型的な方法は、2次元DPを用いて、似ている配列から順に、次々と並べ合わせていくというものであった。

この従来のアルゴリズムは短時間で実行が済むのだが、結果のアライメントは、質という面では十分ではなかった。そこで、マルチプルアライメントは多くの場合、熟練した研究者によって人手で行われている。今日でも、たくさんの配列をアライメントするのは、研究者にとって非常に大きな負荷となっている。

並列反復改善法によるマルチプルアライメント

我々は、計算機による高品質の解を高速に算出する、**並列反復改善法**を用いたアライメントシステムを開発した。この並列改善法のアルゴリズムは、**反復改善法**をもとにしている。まず最初に、この反復改善法について解説し、その後、我々の並列反復改善法について説明する。

反復改善法では、次のような反復戦略（図2）を採っている。まず、初期状態にある配列群をランダムに2つのグループに分割する。そして、2次元DPを用いて、分割されたグループ間でのアライメントを行う。こうして得られた結果を、

最初の処理に戻して、再び分割から繰り返す。このように、ランダムに２つに分割されたグループ間での２次元 DP を繰返すと、全体のアライメントの質が徐々に改善されていく。

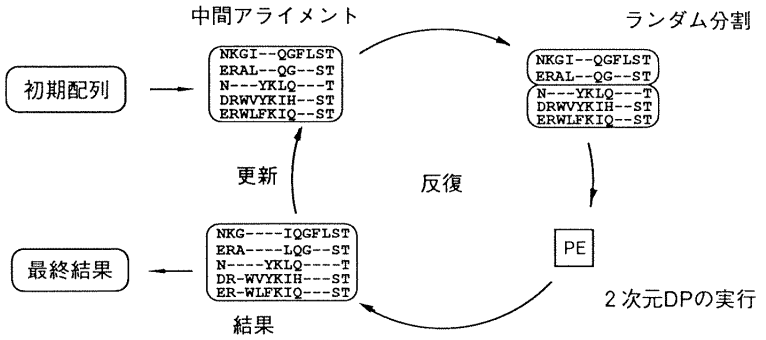


図 2: 反復改善法のアルゴリズム

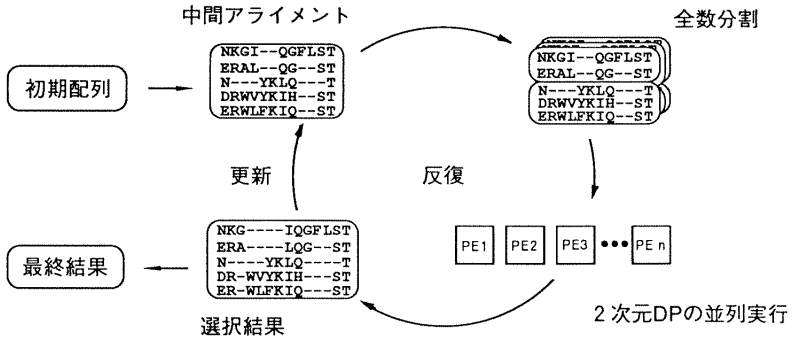


図 3: 反復改善法の並列化アルゴリズム

この反復改善法では、従来のアルゴリズムで得られるアライメントに比べ、かなりスコアの高いものを得ることができる。図４における (b) が、実際に７本のアミノ酸配列を反復改善法でアライメントした際の性能をグラフ化したものである（乱数列を変えて実験を３回行った）。比較のために、ツリーベース法で得られた結果のスコアを水平線 (a) で表示している。ただし、配列数が増えると、反復改善法は収束までに多大な反復回数を要してしまう。

一般に並列計算機を利用することで実行時間の削減を期待できる。我々は以下に説明する並列反復改善法（図３）を考案した。まず、あらかじめアライメントされた配列群を２つのグループに分割する全組合せを網羅する。そして、それぞれの分割に対して、並列に２次元 DP を用いてアライメントを行う。こうして得られた中で最もスコアの良い結果を選び、それを次のアライメントとして処理の最初に戻して、分割から繰返す。図４の (c) を見ると、実行時間の点、安定して良い解が得られる点において、反復改善法に比べ、優れていることがわかる。

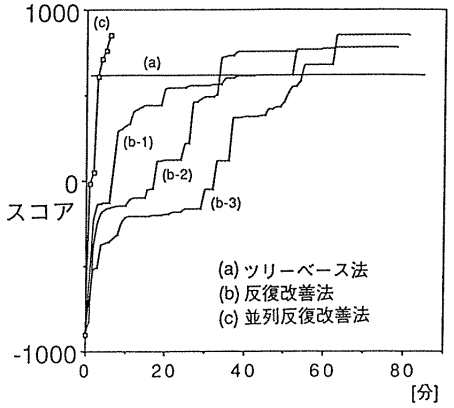


図 4: 各アルゴリズムの性能比較

さらに、より多くの配列を一度に、実用的な時間内でアライメントすることを可能にする**限定分割手法**も、同時に開発した。これは、上記の並列反復改善法を用いた実験結果を調査し、各反復ごとに選ばれたアライメントがどのように分割されていたかに注目することで得られた。2つのグループに分割する際に、片方を少ない本数にすると効果的な改善が行われるのである。この手法を並列反復改善法に組込むことで、実際に生物分野で扱われるような多くの本数の配列群を同時に扱っても、高品質なアライメントを妥当な時間内で獲得できる。

デモンストレーション

今回のデモンストレーションでは、並列反復改善法を用いて、一度に 22 本のアミノ酸配列群のアライメントを行う。256 の要素プロセッサ（PE）をもつ PIM/m 上で実行を行い、約 10 分で終了する。

なお、分割限定法（今回は、少ない方の配列群を 1 本、もしくは 2 本の場合のみに限定）を適用し、すべての分割に対して、2 次元 DP を並列に実行するために 253PE を使用している。また、各反復ごとに最良のアライメントを選択するために 1PE を使用しているため、実際には 254PE を稼働させている。

このデモンストレーションでは、反復を繰返すたびに、アライメントが改善されてゆく様子がわかりやすく表示される（図 5 参照）。それぞれの文字は、そのアミノ酸のもつ性質に従って配色を行っている。また、アライメントの下に、各カラムごとのスコアを棒グラフを用いて図示している。これは、棒が高いほど、そのカラムに性質の良く似たアミノ酸が並べられていることを意味している。特に良く似たアミノ酸がたくさん並んでいる部分は、進化の過程において保存された部分だと考えられ、タンパク質の構造や機能の点で重要な意味を持つ、配列モチーフである可能性が高い。このデモンストレーションの最終結果では、ATP に結合する機能をもつ配列モチーフを捕えることに成功している。

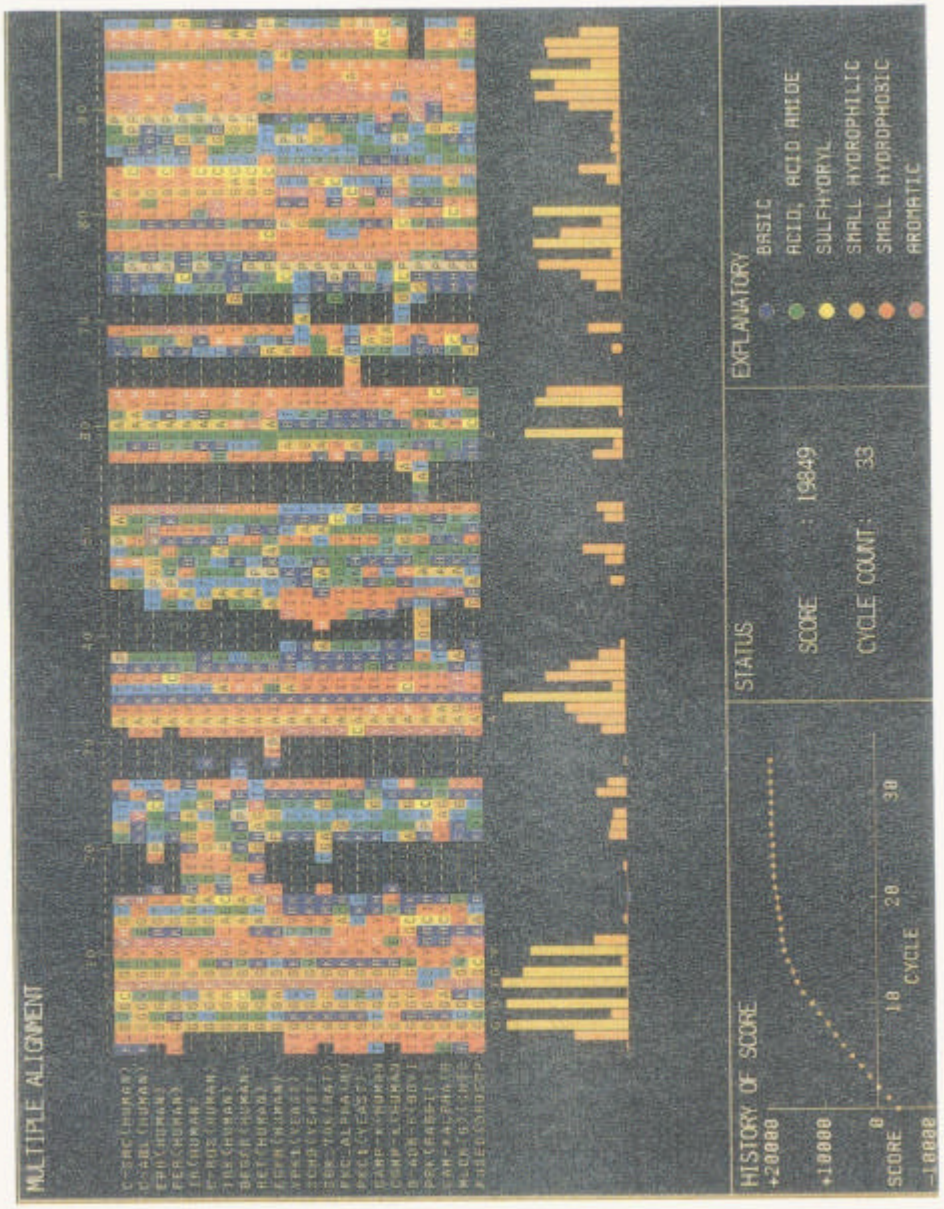


図 5: デモンストレーションの画面