1W-4

遺伝子配列解析の研究 *

島田公敬 †
芝浦工業大学 電子情報システム学科 学部４年‡

1 はじめに

近年, 分子生物学が急速に進展をしている. 既にインフルエンザ菌, 大腸菌, 酵母など 20 種類以上の微生物の DNA 配列が決定し, 多細胞生物である線虫, さらにはヒトの DNA 配列も決定した [1].

ここで DNA 配列とタンパク質の関係について図 1 に簡単に示す. まず染色体から抽出された DNA の情報が, いったん mRNA に転写される.mRNA 上には, アミノ酸の並び方がコドンという形で暗号化されている. そしてリボソームは mRNA を取り込む. アミノ酸を運ぶ役目である tRNA はその中で mRNA のコドンを認識する. リボソームの酵素作用によって隣り合ったアミノ酸がペプチド結合をし, ペプチドさらにはタンパク質となる.

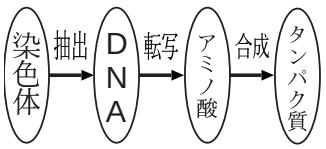

図 1: DNA とタンパク質の関係

しかし,DNA 配列がわかったからといってタンパク質の機能がわかったということにはならない. つまり,DNA 配列の「意味」を解読することが重要なテーマの一つとなる.

このテーマにおいて大量かつ多種類の情報を関連づける情報解析の必要性が増加している [2]. 一般にこのような情報解析手法の研究や, 情報解析により生物学の研究を行う学問分野は分子生物情報学と呼ばれている.

本研究は, 分子生物情報学において広く使用されている BLAST を自己組織化マップを用いて拡張し, 従来生物学者によって行われていた修正を半自動化するものである.

2 本研究とその背景

分子生物情報学において最重要技術である配列比較に注目する.

配列比較とは, 機能未知の遺伝子の DNA 配列を既知配列に似た配列があるかどうか検索することで推定を行うことである.「配列が似ていればタンパク質の立体構造も似ている」→「タンパク質の立体構造が似ていれば機能も似ている」という性質がある. またこの性質は,「似ている配列は同じ祖先から分岐してきたものであり, 機能は保存されているに違いない」という考えからも正当化される [3].

このような配列の類似性の判定において, アラインメントと呼ばれる行列を計算することが基本となっている. 現在, 動的計算法やハッシュ表による FASTA,BLAST などのツールが開発されている. 数十, 数百本の配列アラインメントの場合, 反復法, 隠れマルコフモデルなどを用いた解法が研究されている [2]. 本研究では, この BLAST を扱うこととする.

3 本研究について

3.1 BLAST とは

相同性検索は, 動的計画法を用いたアラインメントにより最適解を求めることができる. しかし, 配列の長さを m, 本数を n としてときに O(mn) オーダの計算時間が必要であり, 配列データベースが増大するにつれて, その計算コストは非常に大くなる. そのため, 実際には近似法が用いられる. 実際によく用いられる相同性検索アルゴリズムの代表例として,FASTA や BLAST(Basic Local Alignment Search Tool) がある.BLAST は, 配列類似性検索を FASTA よりも早く同程度の検出感度で実行する新しい方法として 1990 年に S.Altschul らによって開発された.[3]

BLAST のアルゴリズムの流れは, 以下図 2 に示す.BLAST のアルゴリズムは, 大きく分けて 3 つのステップに分かれている.

1. 検索配列（query sequence）からデータベース検索用の文字列リスト（neighborhood word list）作成

問合せ配列を長さ k の短い断片（ワード）に区切る. スコア行列（例え BLOSUM 行列, 塩基配列の場合, 一致は+5, 不一致は-4 で計算）を用いてそれと閾値以上のスコアでマッチするワード群を求め, そのリストを作成する. アミノ酸配列では k=3, 塩基配列では k=11 または 12 が用いられる.

2. 文字列リストの文字列をデータベース中で探索

完全に一致した文字列だけを探すようにして, 完全一致だけを許すことにより, 検索時間の劇的な短縮を実現する.

3. 見つけた文字列で相同性の高い領域の範囲を決定

配列の上流と下流にギャップを挿入しながらのばし,

*Research on gene sequence analysis
†Kimitaka Shimada
‡Department of Electronic Information Systems,Faculty of Systems Engineering,Shibaura Institute of Tehnology

相同性スコアが局地的に最大になるように相同領域の範囲を決定する (HSP を探す作業). この作業は, 相同領域の範囲を広げてもそれ以上スコアが高くならないところまで続けられる. こうして発見された相同な部分配列,HSP は BLAST の最終的な出力になる.

BLAST アルゴリズムの利点としては, はじめに同一の短い配列を探し, ダイナミックプログラムミング法でこれらのワードをアラインメントに連結していくことで, 2 つの配列の非常に高速なアラインメントを実現していることである. 欠点としては, ヒューリスティックなコンピュータプログラムであるため, 得られた解が最適解であることは保証されない. そのため, 検索後に生物分野の研究者の修正の必要性などがあげられる.

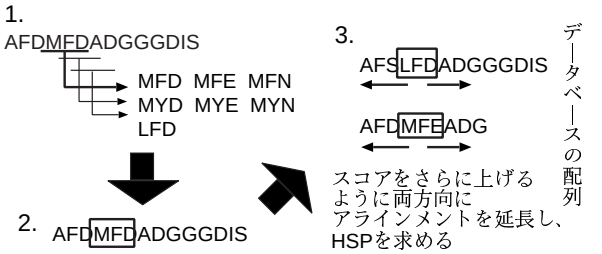

図 2: BLAST のアルゴリズム

3.2 自己組織化マップとは

自己組織化マップとは,T.Kohonen によって提案された教師なしニューラルネットの一種である. 特徴は, 次のようになる.

1. 二次元格子状にニューロンが配置されたネットワークで, 入力データによって競合学習モデルによる組織化がなされる.
2. 学習の終了段階では, 入力データのうち特徴がよく似たもの同士が近くに配置され, 似ていないもの同士が遠くに配置されるような分類フィルタとして機能するようになる
3. 入力データの次元数にかかわらず二次元格子上に配置されるため, 有効な次元削減技術の一つである.
4. 入力データのマップにおける位置関係をみるだけで, その類似性を直感的に把握することが可能である.

3.3 本システムの概要

本研究として, 上で述べた BLAST の欠点を自己組織化マップによって解決できるのではないかと考えた. つまり,BLAST を自己組織化マップによって拡張する. このシステムの概観を以下の図 3 に示す.

BLAST で出力した結果を自己組織化マップにより修正する. 自己組織化マップでの修正を BLAST にフィードバックさせることで,BLAST の検索感度を向上させることができるのではないかと予測される.

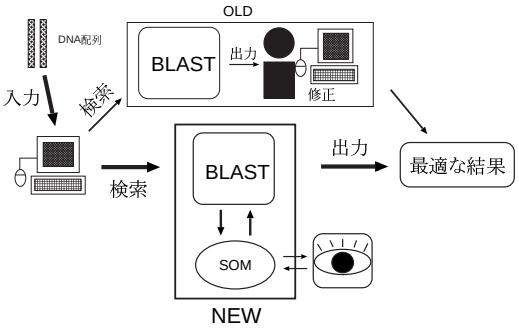

図 3: 本システムの概観

4 実験方法

本研究で製作したシステムを用いて,DNA 配列の相同性検索を行なう. その計算速度と計算精度について比較検証を行なう.

1. 計算速度については, 本研究で拡張した BLAST と従来の BLAST とを比較する. 比較方法としては, 同じ検索用 DNA 配列を同じデータベースより検索をし, その時にかかった時間 (計算速度) を比較する. 検索用 DNA 配列の長さによる変化も確認する.
2. 計算精度ついては, 本研究で拡張した BLAST と計算速度は遅いが計算精度が良いとされている SSEACH とを比較する. 比較方法としては, 同じ検索用 DNA 配列を同じデータベースより検査し, その計算精度を比較する. 検索用 DNA 配列の長さによる変化も確認する.

5 今後の課題

現状は実験を始めたばかりであるが, 自己組織化マップのフィードバックを BLAST のどこにどのようにかけるかで, 出力結果が変化することは確認できた. 修正の学習方法としては,BLAST より計算速度は遅いが検出感度が高いアルゴリズムでアラインメントした結果に近づけるようにしている. 今後の課題としては, この自己組織化マップのフィードバックを BLAST のどこにどのようにかけるかが問題となる.

また上に述べた方法で実験を行い, 半自動化によるより精度のよい解を効率よく得られるかどうかを検証する.

謝辞

本研究に関し, 貴重な御意見を頂いた指導教員である相場亮教授に深く感謝致します.

参考文献

[1] 米国バイオテクノロジー情報センター GenBank http://www.ncbi.nih.gov/
[2] 浅井 潔:配列情報と確率モデル, 人工知能学会誌,Vol.15,No.1,2000
[3] バイオインフォマティクス ゲノム配列から機能解析へ: 監訳 岡崎康司/坊農秀雅