データ圧縮空間を用いた遺伝子の分類と機能予測

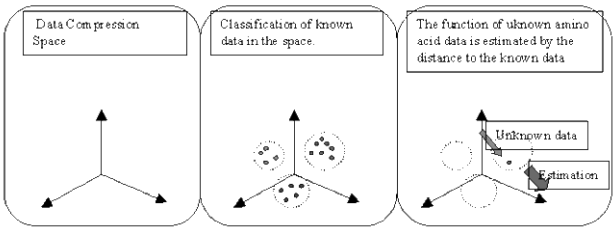
●菅原　研

電気通信大学大学院情報システム学研究科

〈研究の目的と進め方〉

現在、タンパク質の構造・機能をDNA配列・アミノ酸配列から正確に予測する技術は確立されておらず、配列の全体的な類似性に基づいて遺伝子の機能を予測するホモロジー検索と、機能的に似ている配列に共通に出現するパターンを使って機能を推定するモチーフ検索により機能推定がなされている。ホモロジー検索のための手法として、DP、BLAST、FASTAなどのアルゴリズムがあげられる。本研究ではホモロジー検索の観点から、辞書式圧縮法によるデータ圧縮空間テキストデータ圧縮法を用いてタンパク質の機能推定を行う。様々なデータ圧縮法のうち、Ziv-Lempel符号アルゴリズムでは、元データを圧縮しながら辞書を作成するが、この手法は適応型辞書法を用いると十分長い記号列に対して効果的な圧縮が得られることが分かっている。圧縮されたファイルから圧縮率というスカラー量と圧縮辞書が得られる。構造が似ている部分が多数含まれるデータは同一の辞書による圧縮で同様の圧縮率が得られるという点が本研究のキーポイントである。

本研究ではこのデータ圧縮率とデータ構造の類似性に着目し、高速にゲノム構造の分類をおこなうことを試みる。こうした背景の中、ホモロジー検索の観点から、辞書式圧縮法によるデータ圧縮空間（渡辺,1997）という概念を用いて配列情報の類似性判別を試みることを目的とする。また、明確なクラスタリングが可能となる辞書の自動生成の可能性も探る。



〈研究開始時の研究計画〉

・適切な圧縮法の探索

変形LZW圧縮アルゴリズムならびに階層文法抽出データ圧縮法SEQUITURなどを主軸として様々な圧縮手法の中から有効なタンパク質特徴空間を形成できるものを探求する。現在まで上記圧縮法をベースとして様々な圧縮法が提唱されているが、それぞれ長所・短所があり、一概にどの手法が有効であるかを即決することができないため、ある程度試行錯誤が必要となる。本研究では塩基配列ならびにアミノ酸配列を同様の手法にて分類・推測することを試みる。

・分類・検索性能の向上

計算時間の短縮化を図るため、市販のパーソナルコンピュータをクラスタリングし、分類・検索性能の向上を試みる。ただし、クラスタリングにはプログラミングも含めてやや時間がかかることが予想されるため、次年度にわたって長期的におこなうものとし、当年度は並列処理のためのライブラリの開発に重きを置くものとする。

〈研究期間の成果〉

本研究ではLZWを用いて圧縮をおこなった。LZWとは、出てきた単語を登録し、以後出てきた同じ単語を登録先ポインタに置き換えて圧縮を行う手法である。n個のテキストデータ列を圧縮すると、n個の辞書が得られる。このn個の辞書を用いて既知データを圧縮すると、このデータはn次元ベクトルとして表現される。構造が似ているデータ列は原理的にｎ次元空間内で近いところに位置するため、クラスタ解析により既知データの分類・未知データの機能推測が可能となる。

本研究では、インターロイキン、７回膜貫通型膜タンパク、カドヘリンの３つのグループのデータを用いて各種検討を行った。結果は次の２点にまとめられる。

①各タンパクのアミノ酸配列を、６つの基底辞書を用いて６次元ベクトル化し、グループごとにクラスタリングできることを示した。

②遺伝的アルゴリズムの導入により、グループが明確にクラスタ化できるような仮想辞書を作り出すことができた。

〈今後の課題〉

本研究ではデータとしてアミノ酸配列を用いたが、単なるテキスト列と見なして処理しているため、塩基配列に対しても同様の分類は可能であると考えられる。また生成された辞書を解析することでモチーフが得られる可能性がある。これらが本研究を通じて得られた今後の重要課題である。

〈研究期間の全成果公表リスト〉

(1) K.Sugawara and T.Watanabe, "Classification and Function Prediction of Protein by using Data Compression", Artificial Life and Robotics, Vol.6, No. pp.185-190, 2002.

(2) K.Sugawara, S.Chiba, T.Watanabe, "Classification and Function Estimation of Protein by using Data Compression and Genetic Algorithms", Proc. of the 2001 Congress on Evolutionary Computation, pp.839-844, 2001.

(3) 千葉慎二、菅原　研、渡辺俊典、"遺伝的アルゴリズムによるタンパクの分類と機能予測"、情報処理学会東北支部研究会、2001.

(4) K.Sugawara, T.Watanabe, "Classification and Function Prediction of the Protein by using Data Compression", Proc. of the Sixth Int'l Symp. on Artificial Life and Robotics, pp.246-249, 2001.

(5) 菅原　研、渡辺俊典、"データ圧縮空間を用いたタンパクの分類と機能予測"、平成12年電気関係学会関西支部連合大会講演論文集、pp.284, 2000.

(6) 千葉慎二、菅原　研、渡辺俊典、"データ圧縮空間を用いたタンパクの分類と機能予測"、システム・情報

部門シンポジウム2000講演論文集、pp.191-194, 2000.
(7) 菅原　研、渡辺俊典、"データ圧縮空間によるタンパクの分類と機能予測"、平成12年度電気関係学会北海道支部連合大会講演論文集、pp.307-308, 2000.