ライフサイエンスデータベース統合推進事業
統合データ解析トライアル
研究開発課題
「大規模なタンパク質データ解析のための高速な局所配列特徴抽出法の開発」

研究開発終了報告書

研究開発期間： 平成25年9月～平成26年1月

研究代表者：蝦名　鉄平
（（独）理化学研究所　脳科学総合研究センター、研究員）

§1 研究開発のねらい

本研究開発では、巨大なタンパク質データベースにも適用可能な高速なクラスタリング法を利用し、特定の構造に対応するタンパク質の局所配列の特徴を迅速に抽出するための手法を開発する。

ヒトを含む多様な生物種のゲノムが明らかとなった今、タンパク質のアミノ酸配列とその構造との対応を明らかにする事は、ゲノムから推定される新規タンパク質の機能・構造の予測につながる重要な研究課題である。これまで、特定のタンパク質構造に対応する局所配列の特徴を抽出するために様々な方法が提案されてきた。しかし、従来の方法を巨大なデータベースに適用するためには多くの計算資源が必要となる。そのため、個々のユーザが、対象とする構造に対応する局所配列の特徴を近年の巨大な配列データベースから抽出する事は困難になってきており、今後さらに増加するであろうタンパク質データベースにも対応しうる手法の開発が求められている。

従来の特徴抽出法が巨大なデータベースに適用困難な理由として、その計算量がおおよそデータ数の二乗に比例して大きくなる事が挙げられる。そこで、本研究開発課題では、(A)対象とするデータを近年開発された高速なクラスタリング法、BOOL により分類し、(B)冗長性を排除する事によって高速な特徴抽出を可能にする一連のシステムを開発する。開発期間内には、タンパク質のループ構造をモデルデータとして、それ対応する局所配列の特徴が提案方法で抽出可能かを検証する。

§2 研究成果

研究開発期間には、BOOL による局所配列データのクラスタリング法を開発し、分類結果をもとにデータの冗長性を排除する方法の検討と冗長性を排除したデータから特徴選択が可能かどうかを見積もった。

はじめに、タンパク質の構造、配列データを PDBj から取得し、得られた配列の各アミノ酸残基について位置特異的スコア行列(PSSM)を求め、20 種類のアミノ酸残基の保存度を要素として持つ 20 次元ベクトルデータを作成した。次に、この 20 次元のベクトルデータを用い、BOOL のパラメータ最適化を行った。BOOL は(1)N次元ベクトルデータ $X=[x_0, x_1, \ldots x_i \ldots x_n]$ の各要素 x_i を 2 進符号化によって k 段階で離散化し、各々のベクトルに対してラベル付けを行う。

(2)次に、離散化されたそれぞれのベクトルデータの距離を、$D = \sum_{i}^{N} x'i - y'i$ ($x'i$、$y'i$ は離散化された N次元ベクトル X、Yの要素)と定義し、$D \leqq L$ のとき、Xと Yを同じクラスに分類する(別紙　図1)。 パラメータ kと Lの組み合わせを最適化するため、それぞれ k=2-20、L=0-20 の範囲で変化させた。最適なパラメータとして、(1)複数のベクトルデータを含むクラスタが多く作成され、(2)単一のデータからなるクラスタを最小にする kと Lの組み合わせを求めた。得られたパラメータは、$k = 2$、 $L = 0$ で、約 750,000 のベクトルデータが 83,154 個のクラスに分類された。この内、複数のデータからなるクラスは 34,533 個で単一のデータしか含まないものは 48,621 個であった。また、計算にかかる時間は一つの組み合わせにつき 1 分未満であり、BOOL によって従来法(単一連結法で約 700 分)よりも非常に高速なクラスタリングが可能である事を確認した。

次に、クラスタリングの結果から各クラスの代表ベクトルデータを選出し、これらのデータからループ構造を形成する局所配列の特徴が抽出可能かを見積もった。 各二次構造のベクトルデータについて、分類されたクラス内で平均ベクトルを求め、それぞれの平均に最も近いベクトルデータを代表データとして選出した(別紙　図1)。各二次構造に対応する代表データについて、そのベクトル要素を確認したところ、Gly や Pro、Asp や Glu の保存度に顕著な違いが見られた(別紙　図 2)。また、得られたベクトルデータを用いてサポートベクターマシンの学習を行い、全ベクトルデータを予測した時の予測効率を算出した。結果、全データを学習・予測した場合と同程度の予測効率となった。また、学習時間は代表データを用いた場合で 3 分だったのに対し、全データを利用する場合には 190 分であった (別紙　図 3)。

これらの結果は、本課題で開発した手法を用いる事で、従来法と同程度の予測効率ではあるものの、大規模なデータベースにも適用可能な、高速な局所配列特徴抽出が可能となる事を示唆する。本研究で開発したソフトウェアはソースコードを http://domserv.lab.tuat.ac.jp/ebina.html で公開している。公開するツール群には、クラスタリング・代表データ選出のためのプログラムの他に、任意のベクトルデータに対するパラメータ最適化プログラムと簡単な解析用プログラムを同梱しており、各ユーザが作成したデータに対して多角的にデータの解析を進める事ができるようにした。

§３ 研究開発計画および計画に対する達成状況

（1）達成状況

開発計画のうち、BOOL の最適化については自動化も含め達成したと考えているが、複数残基の特徴を含むベクトルデータに BOOL が適用可能かどうかについては、現在も検討を続けている。

（２） ツールの将来性への展望

本開発研究期間には、タンパク質のループ構造をモデルデータとして BOOL の最適化を行い、ループ構造を形成する残基のアミノ酸保存度の特徴が冗長性を排除したデータからも検出可能である事を示した。今後は各残基位置でのアミノ酸保存度だけでなく、その周辺残基の特徴も含めたより多くの要素を持つベクトルデータに対するクラスタリング法および代表データの選出法を確立していく。また、ループ構造に限らず、他の二次構造やディスオーダー領域などを対象に同様の解析を行い、提案手法の一般化を進める。さらに、今回提案したクラスタリング法を利用し、ユーザが指定する任意の局所構造と類似の「特徴」を有するデータを PDBj から探索したり、その特徴を可視化したりするためのソフトウェアを提供する予定でいる。これらのツール群を利用する事で、将来、一般ユーザがそれぞれ対象とするタンパク質局所構造の「特徴」を容易に、かつ迅速に解析する事が可能になると思われる。

§４ 研究参加者

氏名	所属	役職	研究開発項目	参加時期
○蝦名鉄平	(独)理化学研究所　脳科学総合研究センター	研究員	ツールの開発および Web 公開	H25.10-H26.1

§5 成果発表等

(1)原著論文発表　(国内(和文)誌　0 件、国際（欧文）誌　0 件)

(2)その他の著作物(総説、書籍など)

なし

(3)国際学会発表及び主要な国内学会発表

① 招待講演　　(国内会議　0 件、国際会議　0 件)
② 口頭発表　　(国内会議　0 件、国際会議　0 件)
③ ポスター発表　　(国内会議　0 件、国際会議　0 件)

(4)知財出願

① 国内出願 (0 件)
② 海外出願 (0 件)
③ その他の知的財産権

なし

(5)受賞・報道等

なし

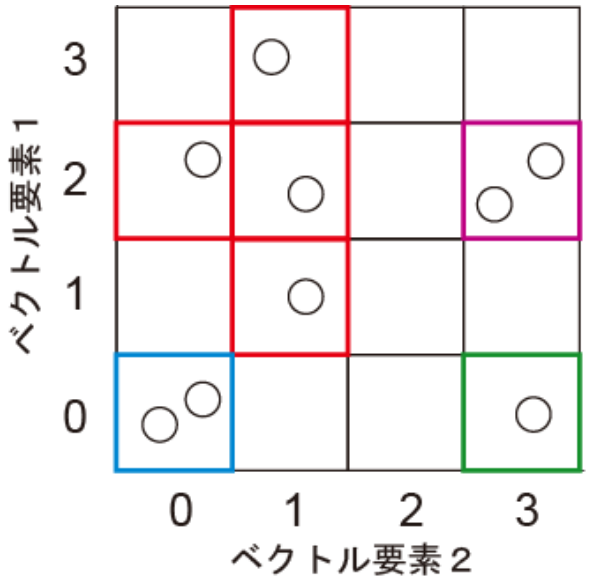
§6 自己評価

近年の構造プロテオミクスの発展にともない、解析対象となるタンパク質データベースは急激に肥大化している。一方、このような巨大なデータベースから情報を抽出するためには膨大な計算資源が必要なため、個々のユーザが、それぞれ対象とするタンパク質構造や配列の特徴を見積もる事は近い将来困難になる事が予想された。本研究開発課題ではこの問題を解決するために、一般ユーザが、個々の PC でも実行可能な代表データ選出法を開発し、代表データを用いる事で非常に高速な配列特徴の検出が可能である事を示した。今後、開発したツール群が構造バイオインフォマティクスの分野で広く活用される事を期待している。

以上

図1. BOOLの概要

BOOL (K=4, L=1の例)



1. ベクトル要素を*k*段階に離散化

2. 近傍の枠に入っているベクトルを同じクラスとして分類する

代表データの選出

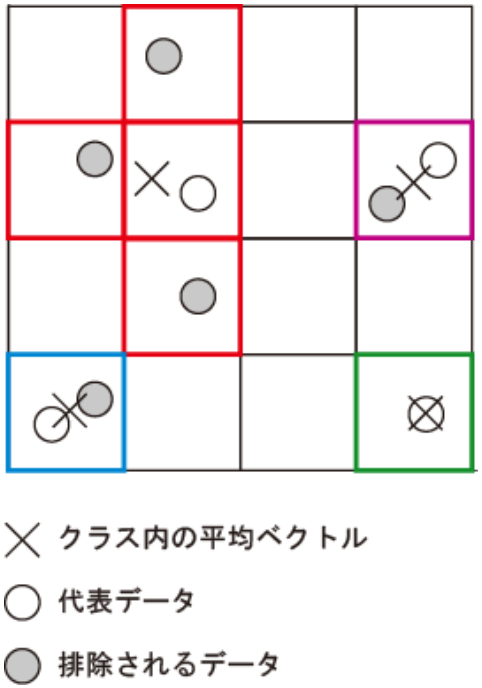
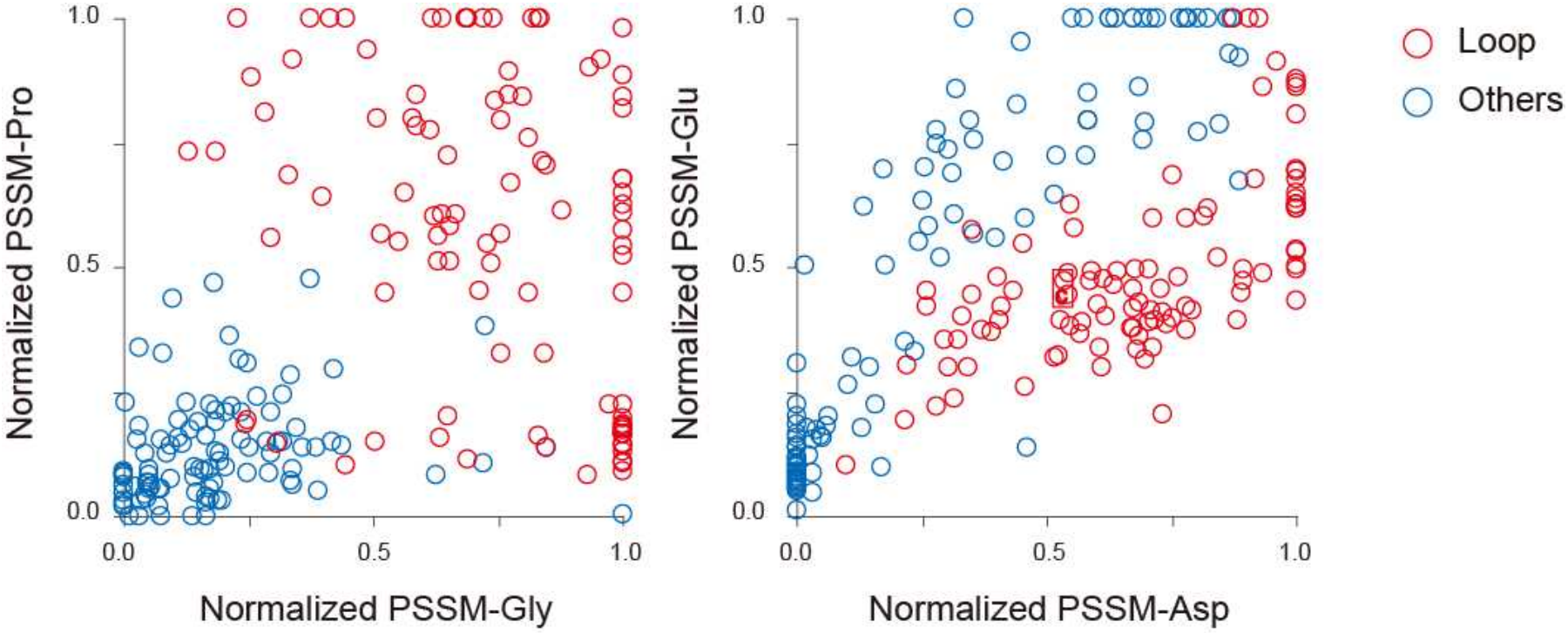


図2. 代表データのベクトル要素



※各二次構造のベクトルデータから100個を選出して図示している

図3. SVMの学習時間とデータ数の関係