構造ドメイン境界予測法の開発を支援する構造ドメイン・データベースの作成		
黒田研究室	学籍番号：01251026	熊谷 勇太

【緒言】 分子量の大きなタンパク質の中には、ドメインを複数持つものがある。そのドメイン領域をアミノ酸配列から予測できれば、ドメインごとに切り分けることで NMR 構造解析や機能解析等が容易に行えるようになる。その予測法としては、Pfam などに登録されている既知のドメインとの配列類似性に基づくものが従来から存在するが、この手法では新規のドメイン配列を予測することができない。そこで、配列特徴の識別に基づく手法（ニューラルネットワークなど）を開発することが求められている。しかし、ドメイン領域は長く、多様性に富む複雑な領域のため、配列特徴を抽出するのは難しい。そのためドメイン境界の配列を予測する研究が近年進められている。そこで我々はドメインの中でも特に、他と明確に区別できる独立した構造単位を持つドメイン（構造ドメイン）と、そのドメイン境界に存在する短く単純なループ領域（ドメインリンカー）に着目した。ドメイン領域に比べて短いドメインリンカーを予測対象とすることで、長い配列を認識する際の技術的な問題が解決され、ドメイン境界の位置から間接的に構造ドメイン領域の予測が可能になる。本研究では、ドメインリンカー予測において学習データとして使用することを目的とした構造ドメイン・データベースを作成し、その特徴について調べた。

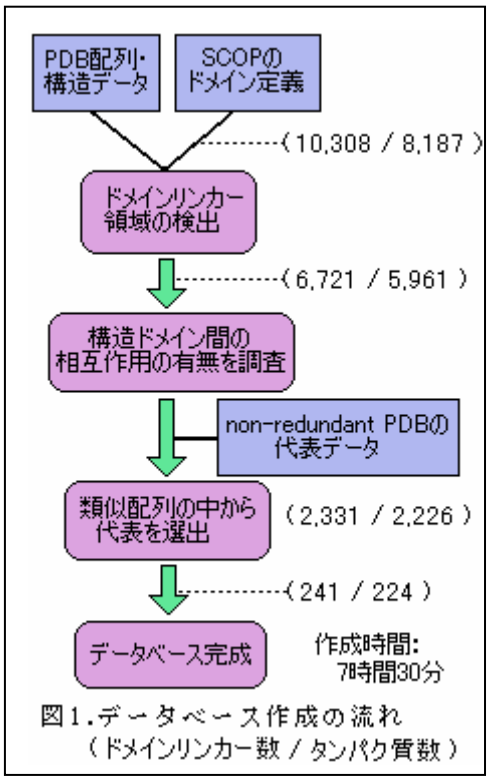

図1.データベース作成の流れ
（ドメインリンカー数 / タンパク質数）

【方法】作成の過程は図 1 のような流れで行う。まず、SCOP によるドメイン定義の境界付近で、特定の構造をもたない 4 残基以上の領域をドメインリンカーと定義する。そして、構造ドメイン間で水素結合や疎水性相互作用をもつ配列を除外する。また、配列特徴抽出の学習データは非冗長であるべきなので、一定以上の類似性を持つ配列の中から代表を 1 つ選ぶ。この作成過程は Linux 上で動作する C/C++プログラムとして自動化されており、入力データを代えることで任意のドメインリンカーデータベースを作成できる。

【結果と考察】 この過程を実行することにより、224 のタンパク質から 241 個のドメインリンカー領域を得た。また、ドメイン定義として CATH を用いた場合、148 のタンパク質から 161 個のドメインリンカー領域を得た。SCOP と CATH の登録データの違いにより、31 個の類似性のない新しい配列を得た。これらのデータセット内部において配列同士の Identity は 30%以下に抑えられており、十分に非冗長であるといえる。また、SCOP など既存のドメイン定義に依存しないドメインデータセット作成法の開発も試みた。手法としては、構造データを各残基で切断して表面積を計算し、その増減値からドメインを仮定義する。このドメイン仮定義を用いて上記と同様の過程を実行したところ、53 のタンパク質から 58 個のドメインリンカー領域を得た。このうち 36 個は SCOP を用いたときと全く同じ領域であったが、残りは SCOP の定義から外れたものであった。既存のドメイン定義に依存しない自作のドメイン定義の正確性には改良の余地を残すが、データセット全体としては、ドメイン境界の予測法を開発するための学習データセットに用いるに足るものであると思われる。