L0021-7	単独での構造の維持が期待される「構造ドメイン」の自動的な同定法開発					
	氏名	井出　宗一郎	主査	黒田	副査	平田・小関・吉野・中村暢

[背景・目的]

プロテオミクス研究において、巨大タンパク質は解析が困難であることが多い。そこで単独で構造を維持する「構造ドメイン」という単位にタンパク質を分割し、迅速な構造・機能解析を可能にする手法が広く用いられている。そのため、タンパク質の立体構造を基にした「単独で構造を維持するドメイン」(Independently Structural Domain, ISD) の定量的な定義とデータベース化は求められている。代表的なドメインデータベースとして、SCOP や CATH が挙げられる。両者はドメイン同定の一部を目視で行うという手法のため構造データベースに比して更新に時間を要する。さらに、両者とも構造類似性と進化的関係に主眼を置いており、構造維持に重きを置いた定量的な定義が成されていない。先行研究において開発されたデータベース IS-Dom は定量的な評価を用いて ISD を評価しているものの、SCOP と CATH にドメイン定義が依存しており、更新速度に依然として構造数の増加より遅い。

そこで本研究では ISD は周囲のドメインと接触が少ない構造単位と考え、他のドメインデータベースに依存せずにタンパク質の構造(原子座標)から ISD を自動的に同定する手法を開発した。

[手法]

ドメイン境界決定法　まず、主鎖同士（MM)、主鎖と側鎖（MS)、側鎖同士（SS）の水素結合、及び疎水性クラスタ（HydroPhobic Cluster,HPC）という 4 種の接触数を調べた。これら4種の接触についてそれぞれ閾値を設け、4種の接触数が閾値を下回るような領域を切断可能領域とした。そして切断可能領域を挟み込むように存在する非切断可能領域を ISD とした(図 1)。

閾値の最適化　IS-Dom で ISD と定義された SCOP に登録されている多ドメインタンパク質を選出した。さらにその中で目視によって明確にドメインとリンカーが確認された 76 タンパク質を「最適化データセット」として用いた。全正答の中で同定がされた割合（sensitivity）と全ての同定された ISD 中の正答の割合（precision）の積を指標として接触数の閾値を最適化した。

ドメイン同定法の評価　最適化した閾値を用いて、PDB に存在する全長タンパク質(237,914 タンパク質)に対して構造ドメインの同定を行った。その結果を、他のドメイン同定法及び SCOP に登録されているドメインと比較した。

[結果および考察]

閾値の最適化　MM:7、MS:9、SS:10、HPC:4 の閾値において sensitivity と precision がそれぞれ 97.5%、88.8%になり、積が最大となった。

ドメイン同定法の評価　PDB 上の全タンパク質に対する ISD 同定の結果、89,399 ドメインが検出され、14,174 ドメインが実験的に単独での構造確認がなされているドメイン(AFD) であった。この結果は他のドメイン同定ツールと比較しても総数や検出率において優れていた (表 1)。

代表配列比較　本研究により同定された AFD の代表配列 671 ドメインと、SCOP に登録されている AFD 代表配列 350 ドメインと比較した結果、SCOP 定義の 74.3%と重複があった(図 2)。また、SCOP では同定されていない ISD が多く存在する点、工程を完全に自動化できる点が本同定法の特長であり、構造決定されたタンパク質を即座にドメイン同定可能なシステムと言える。

表 1. 手法ごとの構造ドメイン同定結果

	This Method	DomainParser	DOMAK	SCOP
ドメイン数	89,399	128,929	118,193	37,867
AFD 数	14,174	17,521	12,928	6,252
AFD 率	15.9%	13.6%	10.9%	16.50%

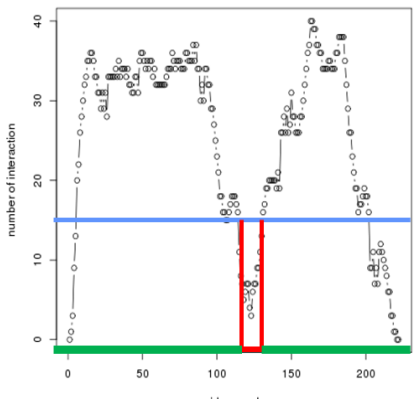

図 1.各残基番号における接触数のグラフ
青：閾値、赤：切断可能領域、緑：ISD

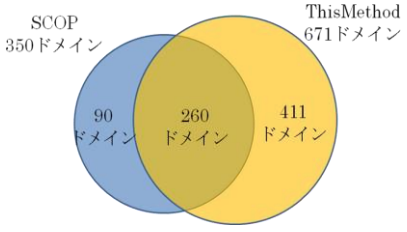

図 2.SCOP と本手法の
AFD 代表配列の重複数