

単独で構造を維持する新規ドメインデータベースの構築		
黒田研究室	学籍番号：10251003	井出 宗一郎

[背景・目的]

プロテオミクス研究において巨大タンパク質解析は、発現・精製などの実験条件の検討が困難である。そのため、比較的分子量が小さく単独での発現・精製が容易であるタンパク質の構成単位「構造ドメイン」を同定し、これを解析する手法が広く用いられている。この手法において、信頼性の高い構造ドメインデータベースが重要であり、より精確かつ自動的な同定法が求められている。そこで本研究では、単独での構造安定性を重視した定量的なドメイン定義の作成を目的とし、原子間距離情報から構造ドメインの自動的な同定法を開発し、その結果をデータベース化した。

[方法]

Autonomously Foldable Domain (AFD)：PDB に全長の類似配列を持つドメインを AFD と定義した。そして、AFD は実験的に単独での構造維持が確認されたと考え、最適化における正解とした。**AFD データセットの構築**：ヘテロマーを除いた全 PDB を解析対象とした。全長タンパク質に対して 30% 以上と 50% 以上の相同性を持つ PDB 断片配列をそれぞれ AFD 候補とした。例えば図 1A の水色で示したドメインは他の全長 PDB と 50% 以上の配列相同性があり、その PDB と構造アライメントした結果構造も一致している (図 1B)。**ドメイン境界同定法**：先行研究のドメイン構造安定化において重要な役割を果たしていると考えられる相互作用は、主鎖同士、主鎖と側鎖、側鎖同士の水素結合と疎水性クラスタの 4 種である。そこで、全ての相互作用の数が閾値を下回る領域がドメイン境界領域であると仮定した。**閾値の最適化**：各閾値を網羅的に変動させ、正しく同定された AFD の割合を調べ、最適な閾値を求めた。最適化の指標には同定の正確性を表す Precision と網羅性を表す Recall の調和平均である F 値 (F factor) を用いた。

[結果・考察]

本手法で得られた 30% 以上及び 50% 以上の相同性を持つ AFD 候補の数、その中で SCOP・CATH と Domain Parser に類似配列が登録されていた数、そして新規に発見された AFD 候補の数を以下に示した(表 1)。この 2 つの AFD 候補を含むタンパク質の構造を確認したところ、30% 以上の相同性を持つ群は二次構造を持つ領域にドメイン境界が存在するもの、近傍にある別のドメイン構造が不安定だと思われるものが数多く存在していたため (図 1C) 不適であるとし、そのようなドメインが少ない 50%以上の相同性を持つ新規マルチ AFD タンパク質群をデータセットとして採用した。相互作用数の閾値による同定結果を網羅的に調べたところ、「水素結合 (主鎖-主鎖) 9 個、水素結合 (主鎖-側鎖) 4 個、水素結合 (側鎖-側鎖) 2 個、疎水性クラスタ 0 個」の場合に評価の指標である F 値が最も高くなった。この閾値を用いて、SCOP や CATH、Domain Parser を元に同定した AFD のデータセットで検定を行ったところ、同定された境界の約 62.7% のみがデータセットと一致していた。これは最適化用のデータセットにいくつかの相互作用数分布が異常なタンパク質が含まれていたためと思われる。

表 1 相同性 30% と 50% における AFD 数

	代表配列	SCOP・CATH	DomainParser	新規AFD
30%	1909	134	319	1559
50%	1436	163	390	1020

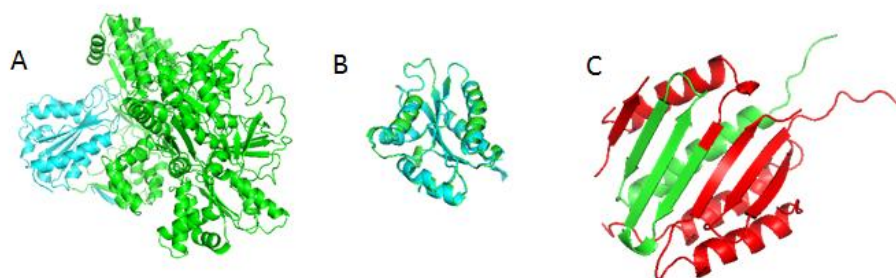


図 1 A:理想的な AFD 構造 B:AFD と PDB の構造アライメント結果 C:ドメイン境界が不適な AFD