学位（博士）論文の要旨

論文提出者	工学府博士後期課程　生命工学　専攻 平成　20　年度入学 学籍番号　08831101　氏名　蝦名　鉄平　印		
主指導教員 氏　名	黒田　裕	副指導教員 氏　名	養王田 正文
論 文 題 目	機械学習法を用いた構造ドメイン領域とドメイン境界領域の解析および予測手法の開発 Analysis and prediction of structural domains and their boundaries using machine learning approaches		

論文要旨（2000 字程度）

生物学的に重要なタンパク質は多くの場合、その構造的・機能的単位であるドメインを多数含む事が知られている。ドメインの中でも構造的に独立で、それが単独で発現・精製可能な「構造ドメイン」は組み換え大腸菌などによる発現・精製がタンパク質全体と比べ容易である。そこで、近年のタンパク質構造解析の場においてはまず、対象のタンパク質中に存在する構造ドメインを計算的手法により予測し、実験的に検証、さらに構造ドメイン単独での立体構造・機能解析の結果を基にタンパク質全体の構造や機能を推定する手法が広く用いられる。構造ドメインを予測する手法の中で、現在最も広く利用されているのは配列相同性に基づく手法である。多くの場合、構造ドメインはそのアミノ酸配列が様々な種間で保存されているため、既存の構造ドメインデータベースから類似配列を検索する事で構造ドメインの予測を行う事が可能である。しかしながら、類似配列に基づく手法は機能・構造未知の「新規タンパク質」に対しては適用困難であり、アミノ酸配列の類似性によらない構造ドメイン予測手法の開発が求められてきた。
このような背景から本研究では「機械学習法を用いた構造ドメイン領域とその境界領域の解析および予測手法の開発」を行った。

本論文は５章から構成されている。

第一章「緒言」では機能・構造未知であり、既存の構造データベース内に類似配列を持たない「新規タンパク質」における構造ドメインの予測法について総括した。また、これら予測手法を開発する際に必要となる構造ドメインデータベースについて述べた。
第二章「構造ドメインデータベースの作成」では現在一般に用いられている構造ドメインの定義を改善し、新しく構造ドメインのデータベースを作成した。また、データベースを評価する基準として「類似構造ドメイン（配列長がほぼ等しい類似配列の中に単独で構造を持つタンパク質が存在するドメイン)」を定義した。我々のデータベース内における類似構造ドメイン割合は他の構造ドメインデータベースと比較し５%以上向上している事から、本データベースが構造ドメイン同定手法の予測効率向上に大きく貢献する事が期待される。

第三章「アミノ酸配列パターンを用いたドメインリンカー領域予測手法の開発」では機械学習法の一種である SVM (Support Vector Machine)が構造ドメイン間のループ領域、ドメインリンカー領域の予測法に適用可能である事を示した。また、配列長依存的なドメインリンカー領域のアミノ酸配列パターンを予測に利用する SVM-Joint はランダム法、二次構造によるドメイン境界予測法および統計的手法によるドメインリンカー予測法と比較し高い予測効率を示した。

第四章「最適なパラメータ群を用いたドメインリンカー領域予測手法の開発」では計算的手法によりドメインリンカー領域の特徴抽出を行い、ドメインリンカー領域を予測する際に最適となるパラメータを同定、SVM によるドメインリンカー領域の予測に応用した。開発した最適パラメータ同定法により、３，０００個の候補から２５個のパラメータを同定した。得られたパラメータの多くは二次構造、またはタンパク質中の核残基位置における Pro および親水性アミノ酸の保存度に関連するものであった。このパラメータを用いて構築した新規ドメインリンカー予測手法、DROP (Domain linker pREdicion using OPtimal features) はパラメータを選出しない場合と比べ、その予測効率を大きく向上させた。これらの結果から、ドメインリンカー領域はドメイン内のループ領域と比較し、強固で揺らぎの少ない構造を持つ事でその両端に存在するドメインのフォールディング過程における相互作用形成を阻害し、独立した構造ドメイン形成に貢献している事が示唆された
また、新規タンパク質に含まれる構造ドメイン領域予測において、DROP が他の予測手法より高い予測精度を示した事から、新規タンパク質の構造ドメイン予測における本提案手法の有用性が示されたと言える。

最後に、第五章「結論」では得られた成果を要約し、本研究の意義を述べた。

ABSTRACT

Domains are functional and structural unit of protein, and large proteins often consist of many domains. A structural domain can fold in isolation, and is usually easier to be characterized by biophysical methods than entire proteins. Computer aided/assisted approaches, where domains are first predicted and the predictions are then experimentally assessed, are thus being actively investigated. Methods that can predict structural domains without using sequence similarity to a domain cataloged in the reference databases are particularly useful, as they may lead to the discovery of "novel" domains, which are preferred targets of proteomics projects. In chapter 1, I review recent structural domain detection and prediction methods and discussed their possibilities and limitations for practical applications.

In chapter 2, I develop a new method for preparing a dataset of structural domains. Our new method refined the definition of “structural domain” as given in various domain dataset, and selected proteins with autonomously foldable domains. The refined dataset is expected to provide more accurate information of structural domains than existing domain datasets and therefore improve the prediction performances of computational structural domain detection methods developed using

them.

In chapter 3, I tested the ability of the support vector machine (SVM), which is a machine learning method used in the diverse area of proteomics, for detecting the sequence patterns of domain linkers in a protein. SVM-Joint, which assumes both long and short domain linker characteristics in its prediction, exhibited the higher prediction performances when compared with random guess and, longest loop prediction and statistical domain linker prediction. Our predictor detect loop regions between two structural domains (domain linkers) first, and it turn assign the location of the domain regions, because this strategy takes advantage of the local nature of domain linker sequence characteristics.

In chapter 4, I identified the optimal feature combination for distinguishing linkers from domains using the random forest method implemented with stepwise selection method, and constructed an SVM domain linker predictor, DROP (Domain linker pRedicion using OPtimal features) using the optimal features. Our feature selection method enables to identify the feature combination in realistic computational time, even though the initial feature candidate set contains a number of features. DROP's performances were superior to previously developed domain linker predictors trained without systematic optimization of the features. DROP's performances for novel protein targets were also higher than those of CASP8 servers, indicating its efficiency as a domain dissection tool for novel multi-domain proteins.