既知のアミノ酸配列モチーフの改変による新規モチーフの探索

角 伸哉　小栗栖 太郎　前田 幹夫
岩大院情報工学†

1 はじめに

ヒトゲノムの概要配列決定を受け、ポストゲノム研究が本格化してきた。近年、情報科学と生命科学が融合した、生命情報科学分野が注目を集めている。

タンパク質は、20 種類のアミノ酸が遺伝子情報に従って、鎖状に結合した高分子である。このアミノ酸の一次元的な並びをタンパク質アミノ酸配列と呼び、アミノ酸をアルファベットに対応させた文字列として表現できる。また、類似の機能や立体構造を持つタンパク質集団に見られる局所的な共通のアミノ酸配列パターンを、アミノ酸配列モチーフと呼ぶ。モチーフの表現方法は様々であるが、正規表現で表す形式が知られている。いくつかの例では、モチーフがタンパク質の機能や立体構造と強く関連していることが実験的にも示されている。

従来のモチーフ抽出法は、各々の専門家が類似のタンパク質のアミノ酸配列を多数集めて、アライメントを作成するというものであった。本研究では、既知のモチーフの系統的改変により、新規モチーフの探索を行った。

分類	モチーフ例	パターン
翻訳後修飾	N-グリコシル化	N-{P}-[ST]-{P}
ドメイン	細胞接着	R-G-D
	ATP結合	[AG]-x(4)-G-K-[ST]

表 1. モチーフの正規表現

表 1 にモチーフの例を示す。各々のアルファベットがアミノ酸に対応している。ただし、x は任意のアミノ酸を表す。{P}は P 以外のアミノ酸という意味であり、[ST] は S か T のどちらかが当てはまることを意味している。x(4) は、任意のアミノ酸が 4 個連続するという意味である。

Searching protein sequence motifs by changing known patterns.
† Shinya Kado, Tarou Ogurusu, Mikio Maeda
† Graduate school of Computer and Infomation Science, Iwate Univ.

2 使用したデータベース

本研究では、モチーフデータベース PROSITE とタンパク質アミノ酸配列データベース PIR を使用した。PIR データベースでは、タンパク質のアミノ酸配列同士の類似性を考慮したスーパーファミリー分類が為されているのが特徴である。しかし、昨今のバイオテクノロジーの進展によるデータの急激な増加のため、未分類のタンパク質が多数存在しているのが実状である。本研究では、登録されているタンパク質のうち、スーパーファミリーに分類されているもののみ (216,912 種類) を使用した。

3 モチーフ候補の生成方法

3.1 拡張パターンの生成

まず、モチーフデータベース PROSITE に登録されている既知のモチーフの拡張を行う。始めに、この拡張を行う際の基本的な考え方を示す。

モチーフには特定の 1 つのアミノ酸が当てはまる部分と、[ ]や{ }などで表されるような数種類のアミノ酸を許す部分がある。一般に、タンパク質の機能に重要なアミノ酸は、生物の進化の過程でも保存される可能性が高いと考えられている。よって、モチーフの特定の 1 つのアミノ酸のみが当てはまる部分は特にタンパク質の機能上重要である可能性がある。このことから、その部分を他のアミノ酸に置き換えることにより、既知のモチーフとは異なった機能を持つ別のモチーフが得られるのではないかと考えた。なお、タンパク質の機能と立体構造には強い相関が見られる。そこで、数種類のアミノ酸を許す部分は主に立体構造の保持に関連し、間接的に機能に関与するものと考えた。

以上の考えに従い、図 1 のようにモチーフの中で特定の 1 種類のアミノ酸のみが許される部分をワイルドカード文字 X に変換し、拡張パターンとした。この結果、PROSITE に登録されているモチーフから、変換可能かつ後述するアミノ酸配列との照合に適した拡張パターン 1,125 種を得た。

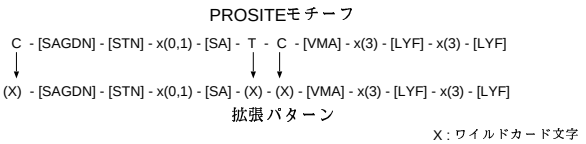

図 1. 拡張パターンの生成

図 1 に拡張パターン生成の例を示す。モチーフの中で特定の 1 種類のアミノ酸のみを許す部分をワイルドカード文字 X に変換する。X は 20 種類の任意のアミノ酸を許す。この例では、3 箇所でアミノ酸を X に変換しているので、1 つのモチーフを 20^3 通りのパターンに拡張することができる。

3.2 データベースとの照合

次に、得られた拡張パターンを用いて、PIR に収められたタンパク質アミノ酸配列との照合を行い、パターンと合致するタンパク質を調べた。このとき、前述のワイルドカード文字 X の位置に対応するアミノ酸を記録し、タンパク質と合致するこれらのパターンを特に識別パターンと呼ぶ。

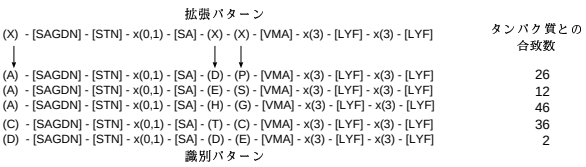

図 2. 識別パターンの例

図 2 に、拡張パターンから得られる識別パターンの一部を示す。右側の数値は各々の識別パターンのタンパク質との合致数である。

3.3 モチーフ候補の選別

拡張パターンと合致したタンパク質を識別パターン毎にグループ化し、そのタンパク質が属するスーパーファミリーを元に、モチーフ候補の選別を行った。パターンに合致したタンパク質と既存のスーパーファミリーとの一致度が高い場合、このパターンをモチーフ候補として採用した。

ただし、PIR にはスーパーファミリーに分類されてはいるものの、そのスーパーファミリーに属するタンパク質が 1 種類のみといった、実質未分類に近いようなものも混在しており、それらの影響を無視するため以下のような条件を使用した。

条件 1：あるスーパーファミリーに属するタンパク質の **90%**以上が拡張パターンと合致しており、かつ、そのスーパーファミリーに **3** 種類以上のタンパク質が属していること。

条件 2：条件 **1** を満たしたタンパク質が、拡張パターンと合致したタンパク質全体の **90%**以上を占めており、かつ、そのタンパク質が **10** 種類以上存在すること。

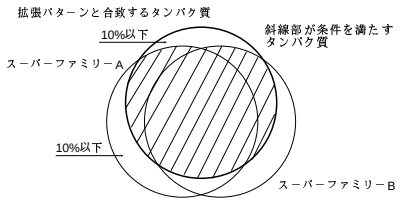

図 3 . モチーフ選別のための条件判定

4 結果

モチーフデータベース PROSITE に登録されている既知のモチーフ (1,125 種) から、モチーフ候補 (11,494 種) を生成した。

さらに、それらの中から複数のスーパーファミリーを含むモチーフを検出できた。これは従来のマルチプルアライメントによるモチーフ抽出法では検出不可能である。すなわち、タンパク質の機能やアミノ酸配列の類似性が乏しく、モチーフの抽出が困難なタンパク質集団から、共通のアミノ酸配列パターンを検出したことになる。

5 考察

従来のモチーフ抽出の手法では、多数の類似タンパク質のアミノ酸配列のマルチプルアライメントを行う必要があった。本研究では、事前に多数の類似タンパク質を収集する必要が無い。また、従来では困難であった複数のスーパーファミリーに共通に見られるモチーフの抽出が可能である。これらのモチーフは、そのスーパーファミリーに共通の機能と関連している可能性を示唆する。

6 おわりに

本研究では既知のアミノ酸配列モチーフの改変により、新規モチーフの探索を行った。得られたモチーフ候補の中から、いかに生物学的に意義のある情報を引き出すかが今後の課題である。