Journal of Computer Aided Chemistry , Vol.7, 57-68 (2006) **ISSN 1345-8647**

タンパク質の機能解析を目指した三角形IDに基づくタンパク質三次元モチーフの検討

丸山英俊 [a,b]*、海尻賢二 [b]

a キッセイ薬品工業株式会社　中央研究所（〒399-8304　長野県安曇野市穂高柏原4365-1）
b 信州大学 工学部 情報工学科（〒380-8553　長野県長野市若里4-17-1）

(May 9,2006 ; July 10,2006)

タンパク質の機能解析において、配列モチーフは重要な情報である。配列モチーフとは、機能を代表するアミノ酸部分配列であり、タンパク質ファミリにおいて非常に良く保存されている。未知タンパク質のアミノ酸配列中から配列モチーフを見つけ出すことで、そのタンパク質の機能を類推することができるのである。しかしアミノ酸配列に基づく配列モチーフ解析には限界があり、タンパク質立体構造からの特徴抽出の方法論が望まれている。タンパク質立体構造からの特徴抽出として三次元モチーフの研究が行われている。著者らは、新規な三角形IDに基づくタンパク質立体構造ホモロジー検索の可能性について検討を行ってきている。三角形IDとは、辺の長さに幅を持たせた三頂点20種類のアミノ酸の組み合わせ8000個の数値IDである。タンパク質のアミノ酸Cαの位置関係を三角形に見立て、該当する三角形IDにすべて数値変換し、その三角形IDを比較する方法である。その三角形IDに基づくタンパク質立体構造ホモロジー検索を行い、有効性を示した。本論文では、この三角形IDに基づく近傍三角形IDを用いて、タンパク質ファミリの三次元的特徴、すなわち三次元モチーフの作成を試みた。同じファミリ中から任意に選択したいくつかのタンパク質の共通した近傍三角形IDを抽出し、三次元モチーフとした。具体的には、医薬品のターゲットとして注目され、また比較的立体構造データが多いプロテアーゼファミリに着目し、それら各ファミリの三次元モチーフの作成を行った。作成した三次元モチーフを用いて、プロテアーゼファミリ検証用データセットの機能解析を試みた。その結果、プロテアーゼの各ファミリへ帰属を可能とする類似度、選択性の高い三次元モチーフの作成ができた。三角形IDに基づく三次元モチーフを用いたタンパク質の機能解析の可能性を示した。

キーワード: 三次元モチーフ，三角形ID，近傍三角形ID，タンパク質三次元構造，PDB

* *hidetoshi_maruyama@pharm.kissei.co.jp*

１． はじめに

ポストゲノム研究で、タンパク質の構造や機能（生理活性）解析が盛んに行われている。その研究の有力解析ツールとしてタンパク質X線結晶構造解析が、国際的に盛んに行われている。タンパク質X線結晶構造解析とは、タンパク質を原子レベルで解析し、その立体構造を解明し、機能を調べることである。その代表的な存在であるタンパク質立体構造データベースProtein Data Bank (PDB)[1]には、2006年6月現在、3万6千件以上の登録があり、ここ数年でその登録数が飛躍的に増加している。

タンパク質の機能は、アミノ酸配列が三次元的に折れ畳まれることによってできる立体構造により決定されている。アミノ酸配列上では離れて位置するアミノ酸においても、折り畳まれた結果、立体的空間において近接し協同的に働いて、タンパク質の機能を発現させている。

一般的に機能解析として最初に行うことは、タンパク質を構成するアミノ酸配列やその遺伝子情報である核酸配列を基とした機能既知配列データベースでの配列比較である。この時、配列の類似性の高いものが見つかれば、同様の類似機能を持つと予測できる。具体的には、機能未知タンパク質を表現するアミノ酸配列や遺伝子塩基配列を検索式として、BLAST (Basic Local Alignment Search Tool)[2]ホモロジー検索が広く行われている。またタンパク質の機能的に重要な配列部分は進化の過程で保存されている場合が多く、この領域をドメインと呼び、更にドメイン中に見られる比較的小さな部分配列パターンはモチーフと呼ばれている。このドメインやモチーフの検索も配列に基づくホモロジー検索と同様に機能解析時に行われている。

タンパク質の機能とは、配列全体に対して、比較的短い配列部分で、その機能を発現していると言え、その機能が類似しているタンパク質群をタンパク質ファミリと呼んでいる。タンパク質ファミリとは、配列全体のホモロジーが低くても、その機能的配列部分であるドメインやモチーフは比較的保存され、その部分のホモロジーは高い。

一般的にモチーフの抽出は、機能が関連する複数のタンパク質のアミノ酸配列を同時に解析するマルチプルアラインメントと呼ばれる手法が用いられる。モチーフを集めたデータベースとしてはPROSITE[3]やタンパク質ファミリデータベースPfam[4]などがある。

しかしアミノ酸配列に基づく配列モチーフ解析には限界があり、タンパク質立体構造からの特徴抽出の方法論が望まれている[5]。このためタンパク質立体構造からの特徴抽出として三次元モチーフの研究が行われている。三次元構造既知のタンパク質構造情報を対象に配列モチーフを検索し、その対応する三次元モチーフを集積[5]や、またタンパク質構造中のある特定のアミノ酸残基の空間配置に注目した縮約表現を導入した三次元モチーフ解析[6]などが試みられている。しかし全アミノ酸に基づく立体的特徴抽出においては、情報量の増大が問題になり、また一部特定アミノ酸に基づく立体的特徴抽出においては、アミノ酸の情報が欠如する。全アミノ酸の情報を利用しながら、同時に情報量を減らす技術がこの問題を解決する鍵となる。

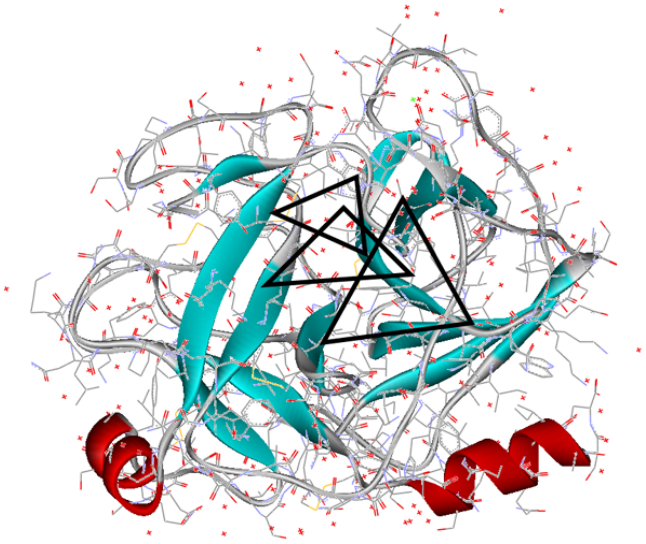

Figure 1. Schematic diagram of Triangle ID.

著者らは、Figure 1の様な新規な三角形IDに基づくタンパク質立体構造ホモロジー検索の可能性について検討を行ってきている[7]。三角形IDとは、辺の長さに幅を持たせた三頂点20種類のアミノ酸の組み合わせの8000個の数値IDである。タンパク質のアミノ酸Cαの位置関係を三角形に見立て、該当する三角形IDにすべて数値変換し、その三角形IDを比較する方法である。タンパク質全体では近傍三角形IDを用いると、基質結合周辺では三角形IDを用いると、タンパク質立体構造ホモロジー検索で良い結果が得られることを報告している[7]。

本論文では、タンパク質全体を解析するため、近傍三角形IDを用いた。そして着目したタンパク質ファミリの中で選択タンパク質に共通な近傍三角形ID群の抽出を行い、その数値ID群をそのファミリの三次元モチーフとして定義した。作成した三次元モチーフを用いて機能解析を行った。

２． 実験方法

医薬品のターゲットとして注目されており、また比較的タンパク質立体構造データが多いことから、三次元モチーフ作成や検証用データセットとしてプロテアーゼに着目した。Pfam分類を利用して各プロテアーゼファミリ、具体的にはセリンプロテアーゼファミリ、システインプロテアーゼファミリ、アスパルチックプロテアーゼファミリ、メタロプロテアーゼファミリを選択した。それらファミリの共通の近傍三角形IDを抽出し、それを三次元モチーフとした。作成した三次元モチーフと検証用データセットを用いて、タンパク質の機能解析の可能性を試みた。

２.１ 三角形ID

タンパク質の３つアミノ酸の空間的配置を、予め用意した三頂点にアミノ酸を持つ三角形のIDで数値化した。アミノ酸の種類は20種であるので、基本的にこのような三角形のIDは20×20×20の8000個となり、自然数により数値化する。これを“三角形ID”と呼ぶ。アミノ酸をアルファベット順に番号付けをした。これをアミノ酸コードとし、アラニン（ALA）からバリン（VAL）までを1から20の番号とした。

三角形IDを以下のように定義する。
・三角形IDの各ノードはタンパク質に含まれるアミノ酸である。
・各ノード間の距離は与えられた上限及び下限の範囲内であるものとする。
・上下限の範囲内にある限り、辺の長さの相違は考慮しない。
・三角形IDは取りうるすべての３つアミノ酸コードの組みで表現する。

なお三角形IDに変換する時、アミノ酸コードが昇順になるように三角形の方向を一定にし、三角形IDに変換した。従って厳密には、8000個のIDの中で、使用しないIDが出てくるが、処理の単純化、高速化を考え、IDは1から8000までの数値を使用した。

２.２ 近傍三角形ID

前稿[7]において、タンパク質全体の場合、検証用データセットに対して三角形IDに基づくタンパク質立体構造ホモロジー検索を行った結果、三角形IDでは良い結果が得られなかった。これはタンパク質中の三角形IDの分散状態が考慮できていないと考察した。この問題を解決するため、近傍の２つの三角形IDの立体空間的な関係を表現するため、近傍三角形IDを考え出した。近傍三角形IDを用いて上記実験を行った結果、良好な結果が得られた。このため、今回はタンパク質全体を用いることから近傍三角形IDを使用した。

近傍三角形IDとは、ある距離範囲内に存在する２つの三角形IDを組み合わせたIDであり、三角形ID－三角形IDと言う表現である。距離範囲は、三角形IDの重心間距離とした。

三角形ID同士の重心間距離を計算し、ある一定の距離の範囲にある２つの三角形IDを抽出し、２つの三角形IDを組み合わせて新たなID“近傍三角形ID”を作成した（Figure 2）。

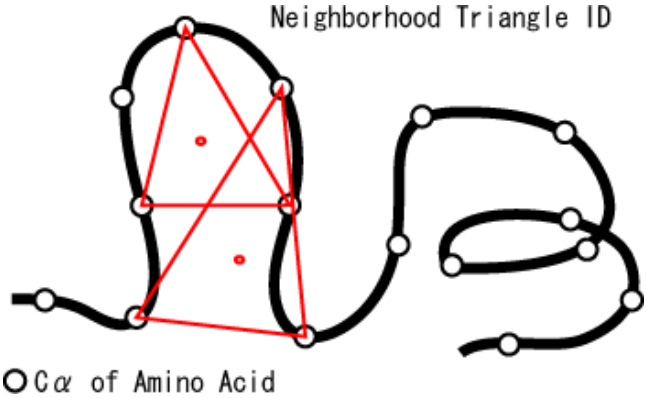

Figure 2. Schematic diagram of Neighborhood Triangle ID.

２.３ 三角形ID、近傍三角形IDの辺の長さと重心距離の値

今回は、前稿[7]の実験結果と比較的多くの近傍三角形IDの抽出を考慮して、タンパク質全体において三角形IDの辺の下限を3Å、上限を9Åに、三角形IDの重心距離を1.5Åに固定した。

２.４ 三次元モチーフ作成

一般的にモチーフとは、機能に関連するアミノ酸の部分配列を意味する。しかし本論文では、この三角形IDに基づく近傍三角形ID、すなわちアミノ酸6残基の構造の集合体を、三次元モチーフと呼ぶことにする

今回、着目したプロテアーゼとは、タンパク質を切断する機能を有する生体内で重要な役割を果たしている酵素である。プロテアーゼは、その酵素触媒機構の違いによってセリンプロテアーゼファミリ、システインプロテアーゼファミリ、アスパルチ

ックプロテアーゼファミリ、メタロプロテアーゼファミリの４つのファミリに分類できる[8]。

このタンパク質のファミリ分類について、タンパク質公共データベースであるPfamを利用した。Pfamは、アミノ酸配列に基づくクラスタリングによってファミリ分類が行われ、データベースが構築されている。このPfam分類に基づく各ファミリの中から任意にいくつかのタンパク質を選択し、そのタンパク質の立体構造データをPDBから入手した。

三次元モチーフの作成に用いたPDBデータセットをTable 1に示す。選択した各タンパク質のPDBデータを用いて、自作プログラムにより近傍三角形ID群を作り出した。つぎに各ファミリ内のタンパク質に共通な近傍三角形ID群を今回新たに作成した自作プログラムで抽出し、三次元モチーフとした。

２.４.１　セリンプロテアーゼの三次元モチーフ

当初、セリンプロテアーゼの三次元モチーフの作成において、試験的にいくつか試みてみたが、1個の三次元モチーフで検証用データセットを満足する三次元モチーフを得ることができなかった。試行錯誤の結果、セリンプロテアーゼの中でも機能的に近いグループごとで三次元モチーフを作成することによって、比較的類似度、選択性の高いモチーフを作ることに成功した。そこで三次元モチーフ作成のため、機能別のいくつかの組み合わせで検討を行った。

血液凝固に関わるグループの三次元モチーフSerine Protease1 (SP1)、その血液凝固に関わるグループに消化酵素トリプシンを加えた三次元モチーフSerine Protease2 (SP2)、免疫，炎症に関わるグループの三次元モチーフSerine Protease3 (SP3)、血液凝固に関わるグループと炎症に関わるのグループを加えた三次元モチーフSerine Protease4 (SP4)の機能別三次元モチーフの作成を試みた。

セリンプロテアーゼの三次元モチーフ作成に際し、Pfam PF00089に属するタンパク質1種類につき２個の任意のデータを用いた。ただしトリプシンについては、1個のデータを使用した。

Table 1に作成した三次元モチーフのPDBデータセットを示す。三次元モチーフSerine Protease1で具体的な作成方法を示す。Factor Xa (1C5M, 1P0S)、Factor XI (1XX9, 1XXF)、Thrombin (1A2C, 1D4P)のPDBデータを用いて近傍三角形IDを前稿[7]の自作プログラムで生成した。全てのタンパク質に共通な近傍三角形IDを今回新たに自作したプログラムを用いて抽出した。このように作成したSerine Protease1の三次元モチーフ、すなわち共通の近傍三角形IDの数は723個であった。同様に作成したSerine Protease2の数は462個、Serine Protease3の数は、474個、Serine Protease4の数は201個であった。

２.４.２　システインプロテアーゼの三次元モチーフ

システインプロテアーゼの三次元モチーフ作成に際し、Pfam PF00112とPF00656に属するタンパク質を、それぞれ１種類につき2個の任意のデータを用いた。Table 1に作成した三次元モチーフのPDBデータセットを示す。Pfam PF00112の三次元モチーフCysteine Protease1 (CP1)の数は453個、Pfam PF00656の三次元モチーフCysteine Protease2 (CP2)の数は、444個であった。

２.４.３　アスパルチックプロテアーゼの三次元モチーフ

アスパルチックプロテアーゼの三次元モチーフ作成に際し、Pfam PF00026に属するタンパク質を1種類につき2個の任意のデータを用いた。Table 1に作成した三次元モチーフのPDBデータセットを示す。三次元モチーフAspartic Protease (AP)の数は733個であった。

２.４.４　メタロプロテアーゼの三次元モチーフ

メタロプロテアーゼの三次元モチーフ作成に際し、Pfam PF00413に属するタンパク質を1種類につき２個の任意のデータを用いた。Table 1に作成した三次元モチーフのPDBデータセットを示す。三次元モチーフMMP M10 (MMP)の数は509個であった。

Table 1 Data sets for 3D motifs

3D Motif Name	Protein Name (PDB)
Serine Protease1 (SP1)	Factor Xa (1C5M, 1P0S) Factor XI (1XX9, 1XXF) Thrombin (1A2C, 1D4P)
Serine Protease2 (SP2)	Factor Xa (1C5M, 1P0S) Factor XI (1XX9, 1XXF) Thrombin (1A2C, 1D4P) Trypsin I (1TRN)
Serine Protease3 (SP3)	Neutrophil Elastase (1B0F, 1PPG) Granzyme B (1FQ3, 1IAU) Cathepsin G (1AU8, 1T32)
Serine Protease4 (SP4)	Factor Xa (1C5M, 1P0S) Factor XI (1XX9, 1XXF) Thrombin (1A2C, 1D4P) Trypsin I (1TRN) Cathepsin G (1AU8, 1T32)
Cysteine Protease1 (CP1)	Cathepsin K (1ATK, 1BGO) Cathepsin B (1PBH, 1GMY) Cathepsin S (1GLO, 1NQC)
Cysteine Protease2 (CP2)	Caspase-1 (1RWK, 1SC4) Caspase-3 (1PAU, 1RHU) Caspase-7 (1GQF, 1SHL)
Aspartic Protease (AP)	Beta-secretase 1 (1TQF, 1XS7) Cathepsin D (1LYA, 1LYW) Renin (1BIL, 2REN)
MMP M10 (MMP)	MMP-3 (1B3D, 2USN) MMP-8 (1A85, 1JJ9) MMP-13 (1YOU, 830C)

２.５　検証用データセット

検証用データセットとしてPfamで登録されている同じファミリで、２．４で記載した三次元モチーフ作成に使用していない、なおかつ触媒部位を含むPDBデータを任意に選択した。

Table 2は､各ファミリの検証用データセットを示した。セリンプロテアーゼ25個、システインプロテアーゼ8個、アスパルチックプロテアーゼ3個、メタロプロテアーゼ8個である。

Table 2 Data sets for test

Family Name	Protein Name (PDB)
Serine Protease	Thrombin (1A2C) Factor VIIa (1CVW) factor IX (1RFN) Kallikrein-1 (1SPJ) Kallikrein-6 (1LO6) Azurocidin (1A7S) Chymase (1KLT) Tryptase Beta-2 (1A0L) Myeloblastin (1FUJ) Activated Protein C (1AUT) Urokinase-Type Plasminogen Activator (1FV9) Complement Factor B (1RRK) Complement Factor D (1DIC) Plasminogen (1L4D) Hepsin (1P57) Matriptase (1EAX) Granzyme K (1MZA) Alpha-Tryptase (1LTO) Trypsin III (1H4W) Complement C1r (1MD7) Complement C1s (1ELV) Granzyme A (1ORF) HGF Activator (1YC0) Hepatocyte Growth Factor (1SHY) Serine Protease HTRA2 Mitochondrial (1LCY) Mannan-Binding Lectin Serine Protease 2 (1Q3X)
Cysteine Protease	Caspase-2 (1PYO) Caspase-8 (1I4E) Caspase-9 (1NW9) Dipeptidyl Peptidase I (1K3B) Cathepsin F (1M6D) Cathepsin L2 (1FH0) Cathepsin Z (1DEU) Cathepsin L (1CS8)
Aspartic Protease	Pepsin A (1FLH) Gastricsin (1HTR) Cathepsin E (1TZS)
MMP	MMP-1 (1SU3) MMP-2 (1CK7) MMP-7 (1MMP) MMP-9 (1L6J) MMP-10 (1Q3A) MMP-12 (1JIZ) MMP-14 (1BQQ) MMP-16 (1RM8)

２.６　三次元モチーフの類似度と選択性

三次元モチーフを用いた検査対象のタンパク質のファミリ帰属時において、その帰属のための類似性指標として類似度を定義した。類似度とは、下記の式で表せる。三次元モチーフID群（a）と検査対象のタンパク質の近傍三角形ID群とを比較し、その一致数（b）を三次元モチーフ（a）で割った値を百分率で表し、類似性指標として用いた。

類似度（%） = b / a × 100

a: 三次元モチーフの数
b: 三次元モチーフと検査対象タンパク質の近傍三角形IDとの一致数

類似度が100%とは、検査対象のタンパク質の近傍三角形ID群中に三次元モチーフが完全に包含されたことである。その三次元モチーフのタンパク質ファミリと全く同じ立体的特徴を持つこと示唆し、そのファミリに帰属する可能性が高いことを意味する。

また立体的特徴と言うのは、そのファミリに特有なものであり、他のファミリにその特徴がない、または少ないことを意味する。従ってファミリに帰属させる条件として、類似度が高いばかりでなく、更に他のファミリに対する選択性の議論が必要である。選択性が高いとは、類似度の高いファミリの三次元モチーフと比較して、別ファミリの三次元モチーフに対しては類似度が低いことを意味する。

２.７　プログラム

タンパク質の立体構造情報を近傍三角形IDに変換生成するプログラムとその変換生成された近傍三角形IDを比較するプログラムは、前稿[7]で作成したものを使用した。また今回新たにタンパク質ファミリ共通の近傍三角形IDを抽出するプログラムをC言語で作成した。それらプログラムをペンティアム４　2.6GHz、1Gメモリのパーソナルコンピュータで実行した。

３．結果と考察

３．１　三次元モチーフの類似度と選択性評価

作成した三次元モチーフの類似度と選択性について検証用データセットを用いて調べた。Table 3は、検証用データセット（縦）に対する作成したプロテアーゼの各三次元モチーフ（横）の類似度を調べた結果である。今回、試行錯誤の結果、比較的類似度が高く、なおかつ選択性の高い三次元モチーフを得ることができた。三次元モチーフ作成時のタンパク質の組み合わせで、選択性を高くできる可能性が示唆された。しかし今回、選択性の出る明確な根拠を見出すことができなかった。

３．１．１　セリンプロテアーゼについて

セリンプロテアーゼでは、４つの三次元モチーフの作成を行った。PDBに登録され、任意に選択したヒトのセリンプロテアーゼについて、その類似度と選択性について調べた。

Table 3において作成したセリンプロテアーゼの三次元モチーフSerine Protease1,2,3,4は、検証用セリンプロテアーゼデータセットにおいて類似度が比較的高かった。

先ず、三次元モチーフの作成に用いたThrombin(1A2C)について、考察してみた。1A2Cを使用した三次元モチーフSerine Protease1,2,4において、それぞれの三次元モチーフにおける類似度は100%であり、作成した三次元モチーフの精度の高さが証明できた。

Serine Protease4は、検証用セリンプロテアーゼデータセット内において、比較的類似度が高い結果になった。しかしSerine Protease4は、他の検証用プロテアーゼデータセットにおいても類似度が高く、セリンプロテアーゼに対する選択性が低かった。例えば、検証用セリンプロテアーゼでは、平均87.8%と高い類似度を示したが、MMPのMMP-2の90.5%やCaspase-9の82.6%などセリンプロテアーゼファミリ以外でも高い類似度を示し、選択性の低いことがわかった。

Serine Protease1は、検証用セリンプロテアーゼにおいて平均66.2%と必ずしも類似度が高くなかった。例えば、Azurocidinの43.3%、HTRA2の45.5%、Complement Factor Dの51.7%など類似度が低いタンパク質があり、また全般的にSerine Protease2よりも類似度も低かった。Serine Protease2は、セリンプロテアーゼ全般に対して、平均76.6%と良い類似度であった。しかし一部Azurocidinの51.5%、Chymaseの74.0%、Myeloblastinの67.7%、Complement Factor Dの66.2%、HTRA2の53.0%については、類似度は低かった。別のSerine Protease3においては、Azurocidinの71.9%、Chymaseの84.0%、Myeloblastinの87.8%、Complement Factor Dの67.5%、HTRA2の62.0%であ

り、類似度、選択性の高い結果が得られた。

従ってセリンプロテアーゼにおいて、三次元モチーフSerine Protease2、Serine Protease3を用いることで類似度が高く、なおかつ選択性の高い結果が得られることがわかった。

３．１．２　システインプロテアーゼについて

システインプロテアーゼでは、２つの三次元モチーフを作成した。PDBに登録されている任意に選択したヒトのシステインプロテアーゼについて、その類似度、選択性について調べた。Table 3は、システインプロテアーゼの各三次元モチーフの類似度を調べた結果である。PF00656に属するシステインプロテアーゼは、三次元モチーフCysteine Protease2で平均73%と類似度が高く、選択性も高かった。またPF00112に属するシステインプロテアーゼは、三次元モチーフCysteine Protease1において、平均76%と類似度が高く、選択性も高かった。

３．１．３　アスパルチックプロテアーゼについて

アスパルチックプロテアーゼでは、１つの三次元モチーフを作成した。PDBに登録されている任意に選択したヒトのアスパルチックプロテアーゼについて、その類似度、選択性について調べた。Table 3は、プロテアーゼの各三次元モチーフの類似度を調べた結果である。三次元モチーフAspartic Proteaseは、平均76.6%と類似度が高く、また選択性ともに高い結果が得られた。

３．１．４　メタロプロテアーゼについて

メタロプロテアーゼでは、１つの三次元モチーフを作成した。PDBに登録されている任意に選択したヒトのメタロプロテアーゼについて、その類似度、選択性について調べた。Table 3は、メタロプロテアーゼの各三次元モチーフの類似度を調べた結果である。三次元モチーフMMP M10は、平均78.4%と類似度が比較的高く、選択性があり、良い結果が得られた。

３．２　三次元モチーフに基づくタンパク質の機能解析

３．１の結果を用いて、タンパク質の機能解析を行った。６つの三次元モチーフの類似度に基づき、順位付けを行い、その一位のものを識別値とした。すなわち一位のものをその三次元モチーフのタンパク質ファミリへの帰属とした。

Table 3の選択した６つの三次元モチーフを用いて、検証用データセットに対する類似度に基づく順位結果をTable 4に示す。一位のものにハイライトを行った。今回作成した三次元モチーフと検証用データセットにおいては完璧にファミリへの帰属ができた。また作成した各三次元モチーフは帰属するファミリ以外には一位になることもなく、選択性も高い結果となった。

検証用データセットのセリンプロテアーゼにおいては、Azurocidin、Chymase、Myeloblastin、Complement Factor D、HTRA2においてSerine Protease3で帰属ができ、それ以外のセリンプロテアーゼにおいてSerine Protease2で帰属ができた。

検証用データセットのシステインプロテアーゼにおいて、PF00656に属するシステインプロテアーゼはCysteine Protease2で帰属ができ、PF00112に属するシステインプロテアーゼではCysteine Protease1で帰属ができた。

検証用データセットのアスパルチックプロテアーゼにおいて、Aspartic Proteaseで帰属ができた。

検証用データセットのメタロプロテアーゼにおいて、MMP M10で帰属できた。

これにより検証に用いたプロテアーゼファミリデータセットにおいて、類似度が高く、なおかつ比較的選択性の高い三次元モチーフを作成できた。今回、新規な三角形IDを基にした近傍三角形IDを用いた三次元モチーフによるタンパク質の機能解析の可能性を証明できた。

３．３　作成時間、検索時間

セリンプロテアーゼの三次元モチーフSerine Protease1の作成時間について検討した。作成条件が、タンパク質全体において三角形IDの辺の下限を3Å、上限を9Åに、三角形IDの重心距離を1.5Åに固定し、6個のタンパク質の近傍三角形IDデータ群を作成した時間は、約13秒であった。三次元モチーフ抽出時間は、1秒以下であった。

またTable 3で作成した三次元モチーフ8個に対して、Human Factor VIIa (1CVW)の三次元モチーフ検索時間は、1秒以下であった。

Pfamの種々の生物種におけるファミリ数は、2005年12月までのバージョンで、8183個の登録である。上記実験結果から、三次元モチーフ8個で1秒の検索時間と定義すれば、ファミリ8183個で約17分となり、比較的有効時間内で検索できる可能性を示唆

した。ただし現在、三次元モチーフ作成には、人手による試行錯誤が必要であり、今後、この作成についても簡便化による作成時間短縮が必要である。

４．まとめ

著者らは、新規な三角形IDに基づくタンパク質立体構造ホモロジー検索の可能性について検討を行ってきている。その三角形IDに基づくタンパク質立体構造ホモロジー検索を行い、有効性を示した。本論文では、この三角形IDに基づく近傍三角形IDを用いて、タンパク質ファミリの三次元的特徴、すなわち三次元モチーフの作成を試みた。同じファミリ中から任意に選択したいくつかのタンパク質の共通した近傍三角形IDを抽出し、三次元モチーフを作成した。具体的には、医薬品のターゲットとして注目され、また比較的立体構造データが多いプロテアーゼファミリに着目し、それら各ファミリの三次元モチーフの作成を行った。作成した三次元モチーフを用いて、プロテアーゼファミリ検証用データセットの機能解析を試みた。その結果、プロテアーゼの各ファミリへ帰属を可能とする類似度、選択性の高い三次元モチーフの作成ができた。近傍三角形IDに基づく三次元モチーフを用いたタンパク質の機能解析の可能性を示した。

今回、三次元モチーフ作成において、試行錯誤で任意に選択したいくつかのタンパク質に基づいて三次元モチーフを作成した。従って三次元モチーフ作成するにあたり、選択したタンパク質の種類の組み合わせや選択数が最適かどうかは不明である。今後、最適な三次元モチーフ作成や実データに適応させるにあたり、各ファミリ内におけるタンパク質の種類の組み合わせや個数について更なる検討が必要である。

Table 3 Results of 3D motif similarity.

Family		Protein Name	Number of Residues	SP1	SP2	SP3	SP4	CP1	CP2	AP	MMP
Serine Protease		Thrombin	297	100.0%	100.0%	70.5%	100.0%	72.6%	68.2%	60.6%	48.1%
		Factor VIIa	311	74.7%	86.8%	70.7%	92.0%	70.9%	61.9%	57.7%	41.7%
		Factor IX	235	78.4%	81.4%	63.5%	87.6%	67.3%	58.3%	57.3%	39.7%
		Kallikrein-1	234	57.3%	68.4%	67.3%	85.1%	54.3%	59.7%	51.2%	37.9%
		Kallikrein-6	217	64.3%	76.0%	55.7%	81.1%	57.6%	50.9%	45.8%	40.5%
		Azurocidin	221	43.3%	51.5%	71.9%	100.0%	50.8%	53.6%	47.1%	42.8%
		Chymase	226	66.7%	74.0%	84.0%	89.6%	57.6%	63.7%	53.6%	43.0%
		Tryptase Beta-2	244	72.6%	79.0%	67.3%	89.1%	62.7%	53.4%	51.6%	42.0%
		Myeloblastin	221	61.4%	67.7%	87.8%	91.5%	51.9%	55.0%	50.8%	43.2%
		Activated Protein C	337	77.0%	83.1%	69.2%	92.0%	68.2%	64.6%	58.8%	47.3%
		uPA	245	65.8%	78.6%	59.3%	80.6%	64.2%	63.5%	57.0%	39.1%
		Complement Factor B	480	71.2%	79.9%	70.9%	88.6%	73.1%	75.5%	69.4%	52.7%
		Complement Factor D	228	51.7%	66.2%	67.5%	83.6%	53.6%	47.7%	41.7%	35.2%
		Plasminogen	361	70.0%	76.4%	74.1%	90.5%	67.1%	69.6%	59.3%	46.2%
		Hepsin	356	87.3%	90.3%	77.6%	96.5%	76.6%	70.3%	63.4%	48.9%
		Matriptase	241	64.2%	78.8%	67.5%	88.6%	62.5%	58.6%	55.8%	42.0%
		Granzyme K	240	64.3%	72.3%	69.8%	83.6%	59.2%	59.0%	52.7%	37.7%
		Alpha-Tryptase	243	73.7%	82.0%	66.0%	92.0%	62.7%	54.5%	52.3%	42.2%
		Trypsin III	219	58.8%	84.4%	60.3%	88.6%	65.8%	62.8%	52.3%	40.3%
		Complement C1r	304	66.7%	71.4%	63.3%	82.1%	67.5%	60.8%	56.3%	46.6%
		Complement C1s	303	73.6%	80.5%	68.1%	87.6%	73.7%	64.9%	63.4%	46.0%
		Granzyme A	232	65.0%	74.0%	71.1%	89.6%	58.5%	58.8%	54.2%	37.7%
		HGF Activator	313	68.0%	82.0%	73.6%	88.6%	75.5%	62.8%	57.3%	48.3%
		HGF	228	58.4%	71.0%	61.8%	81.1%	64.7%	56.5%	52.8%	52.8%
		HTRA2	296	45.5%	53.0%	62.0%	74.1%	50.6%	55.6%	54.3%	44.8%
		MASP-2	321	75.8%	81.2%	68.6%	92.0%	72.2%	67.1%	65.5%	49.3%
Cysteine Protease	PF 00656	Caspase-2	527	55.3%	65.6%	56.5%	78.6%	64.0%	80.0%	57.0%	47.2%
		Caspase-8	293	41.6%	53.0%	43.2%	56.7%	57.2%	58.8%	50.9%	35.4%
		Caspase-9	329	57.4%	69.0%	62.4%	82.6%	72.6%	80.6%	63.7%	52.8%
	PF00112	Dipeptidyl Peptidase I	352	53.7%	65.6%	58.4%	74.6%	78.6%	69.8%	64.3%	46.8%
		Cathepsin F	214	47.4%	61.7%	52.1%	73.1%	70.9%	52.5%	51.2%	36.7%
		Cathepsin L2	221	54.4%	67.5%	51.5%	75.1%	79.7%	59.2%	55.3%	38.9%
		Cathepsin Z	275	55.0%	66.7%	59.5%	77.6%	71.3%	58.3%	55.8%	41.3%
		Cathepsin L	312	54.4%	67.5%	51.5%	71.6%	79.7%	59.2%	55.3%	38.9%
Aspartic Protease		Pepsin A	326	47.2%	57.8%	56.3%	76.6%	58.9%	49.5%	74.6%	40.9%
		Gastricsin	372	50.8%	60.8%	57.8%	76.6%	64.2%	63.1%	77.8%	45.0%
		Cathepsin E	322	50.3%	62.6%	58.2%	80.1%	60.7%	59.5%	77.5%	45.0%
MMP M10		MMP-1	415	53.5%	64.9%	63.7%	80.1%	65.1%	71.8%	67.0%	86.1%
		MMP-2	619	64.3%	76.2%	72.2%	90.5%	78.1%	78.6%	76.5%	79.4%
		MMP-7	166	37.3%	47.8%	46.2%	58.2%	46.1%	47.7%	48.7%	75.8%
		MMP-9	405	53.3%	62.8%	61.8%	78.6%	66.2%	65.5%	62.2%	80.4%
		MMP-10	156	31.5%	38.7%	35.9%	47.3%	43.7%	47.3%	43.5%	88.6%
		MMP-12	166	37.2%	45.7%	43.2%	58.7%	44.8%	49.3%	46.5%	76.4%
		MMP-14	358	54.5%	64.3%	57.0%	71.6%	63.4%	69.8%	59.9%	72.3%
		MMP-16	169	36.0%	45.7%	43.5%	55.2%	40.8%	45.7%	49.5%	67.8%

Table 4 Ranking of 3D motif similarity using selected 6 motifs.

Family		Protein Name	SP2	SP3	CP1	CP2	AP	MMP
Serine Protease		Factor VIIa	1	3	2	4	5	6
		Factor IX	1	3	2	4	5	6
		Kallikrein-1	1	2	4	3	5	6
		Kallikrein-6	1	3	2	4	5	6
		Azurocidin	3	1	4	2	5	6
		Chymase	2	1	4	3	5	6
		Tryptase Beta-2	1	2	3	4	5	6
		Myeloblastin	2	1	4	3	5	6
		Activated Protein C	1	2	3	4	5	6
		uPA	1	2	3	4	5	6
		Complement Factor B	1	4	3	2	5	6
		Complement Factor D	2	1	3	4	5	6
		Plasminogen	1	2	4	3	5	6
		Hepsin	1	2	3	4	5	6
		Matriptase	1	2	3	4	5	6
		Granzyme K	1	2	3	4	5	6
		Alpha-Tryptase	1	2	3	4	5	6
		Trypsin III	1	4	2	3	5	6
		Complement C1r	1	3	2	4	5	6
		Complement C1s	1	3	2	4	5	6
		Granzyme A	1	2	4	3	5	6
		HGF Activator	1	3	2	4	5	6
		HGF	1	3	2	4	5	6
		HTRA2	4	1	5	2	3	6
		MASP-2	1	3	2	4	5	6
Cysteine Protease	PF 00656	Caspase-2	2	5	3	1	4	6
		Caspase-8	3	5	2	1	4	6
		Caspase-9	3	5	2	1	4	6
	PF00112	Dipeptidyl Peptidase I	3	5	1	2	4	6
		Cathepsin F	2	4	1	3	5	6
		Cathepsin L2	2	5	1	3	4	6
		Cathepsin Z	2	5	1	3	4	6
		Cathepsin L	2	5	1	3	4	6
Aspartic Protease		Pepsin A	3	4	2	5	1	6
		Gastricsin	4	5	2	3	1	6
		Cathepsin E	2	5	3	4	1	6
MMP M10		MMP-1	5	6	4	2	3	1
		MMP-2	5	6	3	2	4	1
		MMP-7	3	5	6	4	2	1
		MMP-9	4	6	2	3	5	1
		MMP-10	5	6	3	2	4	1
		MMP-12	4	6	5	2	3	1
		MMP-14	3	6	4	2	5	1
		MMP-16	4	5	6	3	2	1

参考文献

[1] Protein Data Bank, (URL =http://www.rcsb.org/pdb/)
[2] Altschul SF et al., Nucleic Acids Res., 25,3389-402(1997)
[3] PROSITE, (URL=http://expasy.nhri.org.tw/prosite/)
[4] Pfam, (URL=http://www.sanger.ac.uk/Software/Pfam/)
[5] 加藤博明 他, J. Comput. Chem. Jpn., Vol.3 (4), pp.137-144 (2004).
[6] 近松信一 他，第31回構造活性相関シンポジウム，KP18, pp.113-114 (2003)
[7] 丸山英俊、海尻賢二, J. Comput. Aided Chem., Vol. 6, 44-56 (2005)
[8] Carl Branden et al.，タンパク質の構造入門　第二版, Newton Press (2000)

Study of Protein Structure 3D Motifs Based on the Triangle ID for Protein Function Analysis

Hidetoshi Maruyama [a,b*], Kenji Kaijiri [b]

a Kissei Pharmaceutical Co., LTD. (4365-1, Hotaka Kashiwabara, Azumino city, Nagano pref., 399-8304)
b Shinshu Uinvercity Faculty of Engineering (4-17-1 Wakasato, Nagano city, Nagano pref., 380-8553)

In the former work, we proposed a novel 3D protein structural homology search algorithm based on the Triangle ID comparison method. In that work, we focused a triangle structure consisting of three amino acids and called it as Triangle ID. We assumed that proteins can be characterized by using these Triangle IDs. To prove the validness of this assumption, we developed the homology search tool, did several experiment based on the sample data sets, and showed the validness of our assumption and the scalability of our method. On the other hand, identification of 3D characteristics of protein is required, and we assumed that our Triangle ID method can be used for this purpose. In this study, we propose 3D protein structure clustering by using 3D motifs based on the Triangle ID. The 3D motifs were extracted from the common Triangle IDs which have the same feature and belong to the same protein families. We defined the selectivity criteria, did several experiments, and showed the effectiveness of our proposed approach. We selected protease families as our experiment target, because they are attracted the attention as drug target proteins. Our method opens the possibility of the efficient protein function analysis by 3D motifs based on the Triangle ID.

keywords: 3D motif, Triangle ID, Neighborhood Triangle ID , 3D protein structure, PDB

[*] *hidetoshi_maruyama@pharm.kissei.co.jp*