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アルゴリズム

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　長野県南安曇郡穂高町柏原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）

(Received July 7, 2005; Accepted August 25, 2005)

ポストゲノムの近年、タンパク質の立体構造データの登録数が飛躍的に増加している。タンパク質を構成しているアミノ酸の配列に基づく類似性検索では見つからない立体的に類似している遠縁ホモログが見つかっており、タンパク質の立体構造の類似性（ホモロジー）検索が求められている。我々は、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た。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。本手法の有効性を示すために、立体構造的に類似しているタンパク質のセリンプロテアーゼファミリとそれと異なるファミリのタンパク質を用いて検証用データセットを作成した。そしてセリンプロテアーゼファミリであるヒトのトロンビンを検索式に、同じセリンプロテアーゼファミリのヒトとウシのトロンビン、ファクターXa、トリプシン、そして異なるファミリのHIV-1プロテアーゼ、PTP1Bを含む58個の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。その結果、類似度の高い順から、ヒトのトロンビン、ウシのトロンビン、そして同じセリンプロテアーゼファミリであるファクターXa、トリプシンがその次ぎに続き、下位にセリンプロテアーゼと異なるHIV-1プロテアーゼ、PTP1Bとなり、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を示した。

キーワード: タンパク質立体構造ホモロジー検索，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

ヒトの遺伝情報であるゲノム配列が解読され、ポストゲノムの時代が来ている。次の興味は、タンパク質の構造や機能（生理活性）解析である。タンパク質の機能を解明することで、ゲノム創薬すなわち論理的･効率的な医薬品の開発や、オーダーメード医療など革新的な医療を提供できる可能性が出てきた。

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する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)[1]ホモロジー検索が広く行われている。しかしながら配列の類似性が低い場合においても、類似の立体構造を形成して類似する機能を保持する遠縁ホモログの存在が知られている。その遠縁ホモログを従来のBLASTなどの配列類似性から見出すことは非常に困難であり、遠縁ホモログを見つけ出す方法論が望まれている[2]。

一方、ポストゲノム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、タンパク質の構造を原子レベルで解析し、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)[3]には、2005年5月現在、約3万の登録があり、ここ数年でその登録数が飛躍的に増加している。しかしながらPDBの主な役割は、データの整理、蓄積であり、それらデータの高度利用が求められている[4]。

タンパク質の機能とは、Figure 1の様に局所的な相互作用部位（基質結合周辺）である溝に、化合物や他のタンパク質やペプチドなどが結合することにより起こるものである。その溝は、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。

タンパク質立体構造の類似性を比較する尺度として、アミノ酸原子の立体位置のずれである根平均二乗誤差（RMSD： Root Mean Square Deviation）を用いる場合が多い[5][6]。RMSDを計算するための条件として、２つのタンパク質のアミノ酸同士が対応しているかを調べるために配列アライメントを取る必要がある。通常、配列アライメントを取るためには、タンパク質同士のアミノ酸配列の類似性がある程度高くなければならないし、比較するアミノ酸同士の数が大幅に異なると配列アライメントは取れない[7]。従ってすべてのタンパク質に対するアミノ酸同士の配列アライメントを取り、自動でRMSDを計算することは非常に困難と言える。

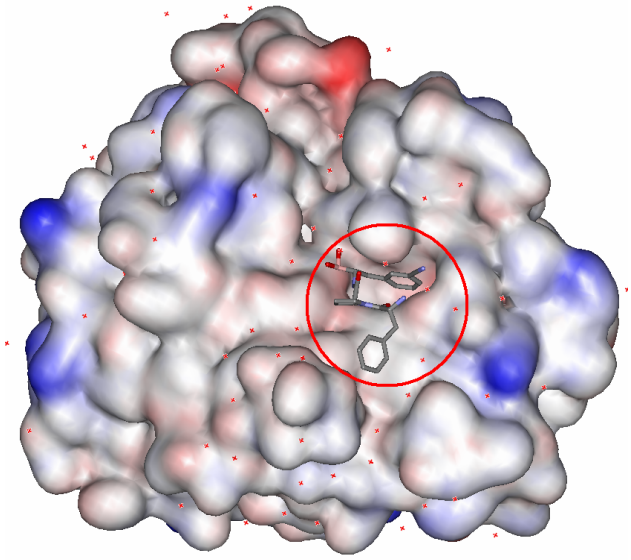

Figure 1. Bovine trypsin 3D structure (PDB: 1AUJ) and inhibitor (red circle).

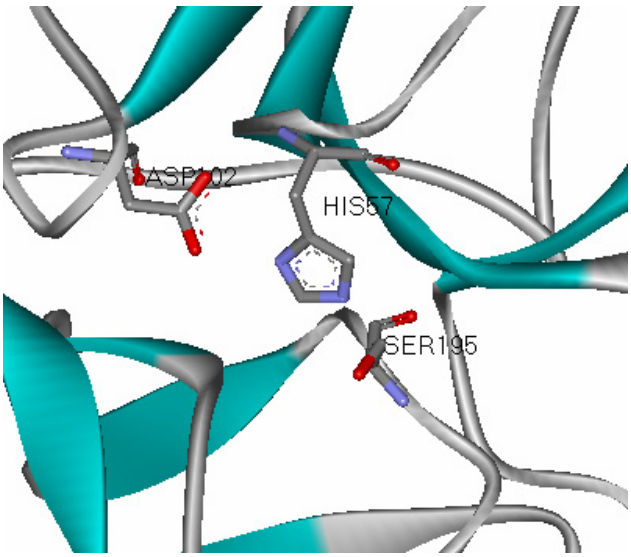

Figure 2. Catalytic triad of bovine trypsin (PDB: 1AUJ).

更にPDBデータについて言及すると、同じタンパク質でありながら異なる分解能、実験誤差、動物種、また実際に解析できたアミノ酸配列数などにより、データは均一ではなく、バラツキが存在する。そのようなバラツキのあるデータを前提として取り扱

うことができることも課題である。

このような問題を解決するために、我々は、ある種のタンパク質が持つFigure 2の様なCatalytic triadに着目した。Catalytic triadとは、三次元的空間に近接配置した３つのアミノ酸を示し、その３つのアミノ酸の近接された空間的配置によって酵素触媒機能を発現するのである。例えばセリンプロテアーゼのCatalytic triadは、ヒスチジン（HIS）、アスパラギン酸（ASP）、セリン（SER）であり、その３つのアミノ酸の立体的空間配置は、同じセリンプロテアーゼファミリ内や動物種を越えて、非常に良く保存されている。しかしこのCatalytic triadは、アミノ酸配列の順番が隣り合っているわけではなく、アミノ酸配列上では離れた位置に存在する場合がほとんどである。例えばウシのトリプシンのCatalytic triadは、ヒスチジン57番目、アスパラギン酸102番目、セリン195番目である（Figure 2）。従ってアミノ酸配列の解析からCatalytic triadを同定することは非常に困難である。

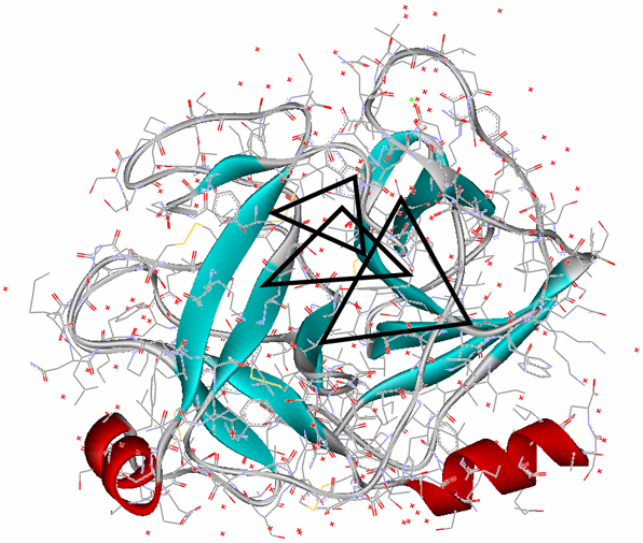

Figure 3a. Schematic diagram 1 of Triangle ID.

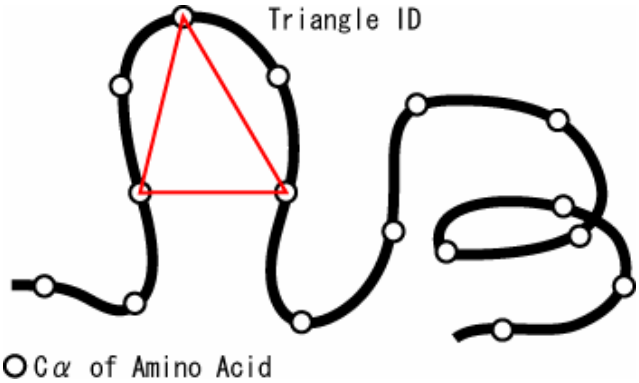

Figure 3b. Schematic diagram 2 of Triangle ID.

我々は、この三次元的空間の近接関係であるCatalytic triadを応用して、タンパク質の全配列または基質結合周辺のアミノ酸配列すべてのアミノ酸の三角形を予め用意した“三角形ID”に数値置換する方法を考えた。タンパク質の三角形IDの数値で、２つのタンパク質を比較する新規な手法である。（Figure 3a,3b）。

本論文では、検証用データセットを使って三角形IDに基づくタンパク質立体構造ホモロジー検索が有効に行えるかを評価した。すなわち、精度の高い検索が行え、検索効率（速度）が高いことを検証した。検証用データセットには、配列的、立体構造的に類似しているセリンプロテアーゼファミリのトリプシン（ヒト）、トリプシン（ウシ）、ファクターXa[8]を選び、そしてセリンプロテアーゼと異なるファミリのタンパク質のPTP1B、HIV-1プロテアーゼを選定した。

２． 三角形IDに基づくタンパク質立体構造ホモロジー検索法

タンパク質を構成するアミノ酸は、分子中心の炭素原子（Ｃα）にアミノ基（－ＮＨ２）、カルボキシル基（－ＣＯＯＨ）、 水素原子（－Ｈ）、側鎖（－Ｒ）が結合し、側鎖Ｒはアミノ酸ごとに異なり、基本的に20種類ある。本論文で提案する三角形IDに基づくタンパク質立体構造ホモロジー検索とは、アミノ酸Ｃα原子の三次元座標をそのアミノ酸の代表とし、３つの近傍アミノ酸のＣαの三角形を、三角形IDへ数値変換し、その三角形IDを比較する新しい方法である。三角形ID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を網羅的にその三角形IDに次々と変換し、タンパク質を識別するための識別子として三角形ID群で表現した。タンパク質同士の立体構造ホモロジー検索が、それぞれの三角形IDの数値群同士を比較することにより、実現できると考えた。従って、立体構造的に類似した２つのタンパク質を比較した場合、三角形ID群同士の一致が多くなる。

しかしながら三角形ID群同士の一致数が多くても、立体構造的に類似性が高くない事例が見つかった。この事は三角形IDのタンパク質における立体空間的な分散状態の相違に起因する。そこで三角形IDの分散状態を見るため、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”を作り出した。

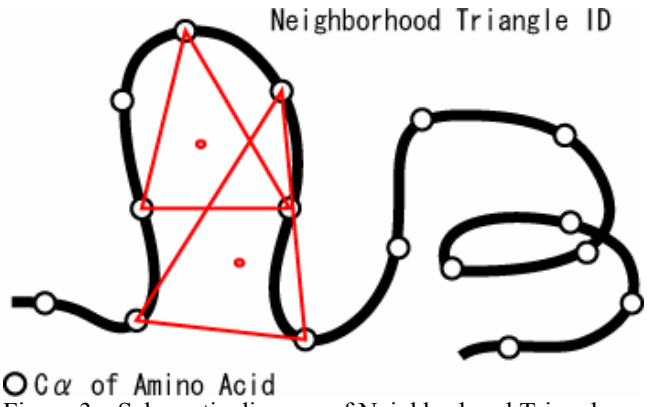

Figure 3c. Schematic diagram of Neighborhood Triangle ID.

この三角形IDまたは近傍三角形IDを基にして、調べたいタンパク質のそれらIDを検索式とし、検索対象となるタンパク質のID群をデータベース化する。検索はIDの単純な一致比較で済み、処理が非常に容易になり、タンパク質立体構造ホモロジー検索の高速化が可能となる。

また三角形IDの辺の幅の許容を持たすことにより、実験誤差やバラツキのあるデータの中から同じタンパク質の抽出や、ファミリの抽出が可能となった。

3． 実験方法

3.1 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種となり、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。

3.2 近傍三角形ID

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”を作り出した。その重心間距離について検討を行った。

3.3 類似性指標としての谷本係数

２つのタンパク質、AとBの立体構造ホモロジー検索において、類似性指標の一つである谷本係数を使用した。谷本係数は、化学構造の類似性比較に広く用いられている[9]。化学構造の部分構造などを固有のビットに変換し、その部分構造の有無をビット比較し、類似比較するものである。今回、我々はこの谷本係数を、タンパク質の立体構造の類似性を表現するために、三角形IDをビットと見なし、利用した。この谷本係数を百分率で表し、タンパク質立体構造ホモロジー検索の類似性指標に用いた。値は常に0%から100%になる。

$$\text{谷本係数（%）} T(a,b) = c / (a + b - c) \times 100$$

a: Aのビット数
b: Bのビット数
c: AとBの一致ビット数

3.4 タンパク質立体構造ホモロジー検索

検証用データセットは、タンパク質分解酵素として広く研究され、比較的数多くX線結晶解析されているセリンプロテアーゼファミリを中心に作成し、実験を行った。一般的にタンパク質には、機能的に近いタンパク質ファミリが存在する。セリンプロテアーゼファミリは、アミノ酸配列や立体構造がよく類似していると言われている。セリンプロテアーゼファミリ中から代表としてトリプシン、トロンビン、ファクターXa[8]を選択し、PDBデータベースの中から任意にデータを選択した。またセリンプロテアーゼファミリとは異なるファミリのタンパク質としてPTP1BとHIV-1プロテアーゼを任意に選択し、比較検討を行った。

具体的には、ヒトのトロンビン（PDB: 1AHT）を検索式に、トロンビン（ヒト）、トロンビン（ウシ）、トリプシン（ウシ）、ファクターXa、PTP1B、HIV-1プロテアーゼのタンパク質全体のデータセット、基質結合周辺のみの部分構造のデータセットのタンパク質立体構造ホモロジー検索を行った。

3.4.1　検索アルゴリズム

検索アルゴリズムをまとめると以下のようになる。

1）該当タンパク質から、三つのアミノ酸の組を求める。

2）そのすべての組に関して、辺の長さが上下限内のもののみ残す。

3）同一の三角形IDを識別し、8000種の三角形IDのどれが存在するかの列として表現する。

4）検索式となるタンパク質のID列と、対象するタンパク質のそれと谷本係数に基づき比較する。

3.5　検証用データセット

検証に用いたタンパク質の立体構造データファイルは、PDB[3]で入手した。

検索式は、トロンビン（ヒト）1AHTを用い、基質結合周辺は、下記に示す任意に1AHTから選択したアミノ酸20個を使用した。括弧内は各アミノ酸の配列番号である。CYS(42), HIS(57), CYS(58), TYR(94), ASN(98), LEU(99), ASP(102), ILE(174), THR(177), GLU(192), SER(195), ILE(212), VAL(213), SER(214), TRP(215), GLY(216), GLU(217), PHE(227), TYR(228), THR(229)

各データセットは、タンパク質データベースSwiss-Prot（Uni-Prot）にリンクされているPDBデータから、ほぼ同じアミノ酸セットが作れるファイルを任意に選択した。またタンパク質の基質結合周辺は、アミノ酸を任意に選択して作成した。具体的なPDBファイル名を下記に示す。

トロンビン（ヒト）データセットは、Swiss-Prot: P00734にリンクされている下記の10個を任意に選択した。（1AHT, 1BMM, 1C5N, 1DWB, 1HAI, 1K21, 1LHG, 1O2G, 1THR, 1UMA）また基質結合周辺は、上記セットから、検索式と同じ20個のアミノ酸を任意に選択して作成した。

トロンビン（ウシ）データセットは、Swiss-Prot: P00735にリンクされている下記の8個を任意に選択した。(1AVG, 1ETR, 1ETS, 1ETT, 1HRT, 1ID5, 1UVT, 1UVU）また基質結合周辺は、上記セットから、検索式と同じ20個のアミノ酸を任意に選択して作成した。

トリプシン（ウシ）データセットは、Swiss-Prot: P00760にリンクされている下記の10個を任意に選択した。(1AUJ, 1BJV, 1C1N, 1CE5, 1F2S, 1GI0, 1JIR, 1MAX, 1O3O, 1YYY）また基質結合周辺は、上記セットから下記の19個のアミノ酸を任意に選択して作成した。ALA(55), HIS(57), TYR(94), ASN(97), THR(98), LEU(99), ASP(102), GLN(175), CYS(191), GLN(192), GLY(193), ASP(194), SER(195), SER(214), TRP(215), GLY(216), SER(217), GLY(219), CYS(220)

ファクターXa（ヒト）データセットは、Swiss-Prot: P00742にリンクされている下記の10個を任意に選択した。（1C5M, 1EZQ, 1FAX, 1G2M, 1HCG, 1IOE, 1KSN, 1LPG, 1MQ5, 1NFX）また基質結合周辺は、上記セットから下記の32個のアミノ酸を任意に選択して作成した。HIS(57), GLN(61), THR(95), LYS(96), GLU(97), THR(98), TYR(99), ASP(102), ARG(143), HIS(145), PHE(174), ILE(175), MET(180), ASP(189), ALA(190), CYS(191), GLN(192), GLY(193), ASP(194), SER(195), VAL(213), SER(214), TRP(215), GLY(216), GLU(217), CYS(220), ALA(221), ARG(222), TYR(225), GLY(226), ILE(227), TYR(228)

PTP1B（ヒト）データセットは、Swiss-Prot: P18031にリンクされている下記の10個を任意に選択した。（1AAX, 1BZC, 1EEN, 1L8G, 1NL9, 1OEM, 1PH0, 1PYN, 1Q1M, 1QXK）また基質結合周辺は、上記セットから下記の16個のアミノ酸を任意に選択して作成した。ARG(24), ASP(29), TYR(46), ASP(48), VAL(49), ASP(181), PHE(182), SER(216), ALA(217), GLY(218), ILE(219), GLY(220), ARG(221), MET(258), GLN(262), THR(263)

HIV-1プロテアーゼ（HIV）データセットは、Swiss-Prot: P03369にリンクされている下記の10個を任意に選択した。（1AID, 1B6J, 1CPI, 1F7A, 1MT7, 2AID, 1KJ4, 1D4K, 1YTG, 8HVP）また基質結合周辺は、上記セットから下記の20個のアミノ酸を任意に選択して作成した。ARG(8), PRO(9), LEU(23), LEU(24), ASP(25), GLY(27), ALA(28), ASP(29), ASP(30), VAL(32), LEU(33), ILE(47), GLY(48), GLY(49), ILE(50), THR(80), PRO(81), VAL(82), ASN(83), ILE(84)

3.6　プログラム

提案するアルゴリズムを実証するために、C言語を用いて、２つのプログラムを作成した。１つは、タンパク質の立体構造情報を三角形IDまたは近傍三角形IDに変換生成するプログラムである。もう１つは、その変換生成された三角形IDまたは近傍三角形IDを比較するプログラムである。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。

４． 結果と考察

４.1　三角形IDの一辺の長さの幅検討

データのバラツキ等を考慮して、タンパク質のアミノ酸Ｃα三角形の一辺の長さに許容幅を持たせた。三辺とも同じ許容幅にし、0から1Å刻みで12Åまでの許容幅の三角形IDを作成した。ヒトのトロンビン（PDB: 1AHT）を検索式にタンパク質立体構造ホモロジー検索を行い、その三角形IDの一辺の長さの有効な上下限を算出するため、検証用データセットで実験した。

４.1.1　三角形IDの一辺の長さの上限検討

三角形IDの一辺の長さの上限を調べるため、下限を0に固定して検討を行った。タンパク質全体において、三角形IDの許容の幅を増やせば、トロンビンやそれが属するセリンプロテアーゼファミリの谷本係数は増すが、セリンプロテアーゼファミリと異なるPTP1BやHIV-1プロテアーゼへの谷本係数も増した（Figure 4a）。これは、三角形IDのタンパク質中の分散度合いに依存すると考えた。

基質結合周辺においては、データ自体が局所的なものであり、三角形IDの分散度合いに比較的影響されず、6Å～12Åの間で比較的良い結果が得られた（Figure 4b）。

タンパク質の全体、基質結合周辺ともに三角形IDの一辺の長さの上限が増加すると、三角形ID数は単純に増大した（Figure 4c）。

４.1.2　三角形IDの一辺の長さの下限検討

三角形IDの一辺の長さの下限を調べるため、上限を9Åに固定して検討を行った。タンパク質全体と基質結合周辺ともに、3Åから三角形IDの一辺の長さの下限が増えると谷本係数も減少した。特に谷本係数が上昇する特異的な数値は見出せず、下限を3Åと決めた（Figure 5a,5b）。

タンパク質の全体、基質結合周辺ともに三角形IDの一辺の長さの下限が増加すると、三角形ID数は単純に減少した（Figure 5c）。

４.2　三角形IDの分散度合い：近傍三角形IDの検討

上記結果から、タンパク質中の三角形IDの分散状態も考慮できれば、比較的結果が良いことが分かり、三角形IDの分散度合いを表現する方法の検討を行った。三角形IDの立体空間的な分散状態を見るため、近傍三角形IDを考え出した。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。

その近傍三角形IDに基づくタンパク質立体構造ホモロジー検索の検討を行った。今回、上記結果から三角形IDの辺の長さを下限3Åと上限9Åに固定して、近傍三角形IDの妥当性と、重心間距離の範囲を探った（Figure 6）。

タンパク質全体において、三角形IDのみではセリンプロテアーゼファミリと異なるPTP1BやHIV-1プロテアーゼとの分離が悪かった。しかし近傍三角形IDでは、差はわずかではあるが上位からセリンプロテアーゼファミリの順番であり、PTP1BやHIV-1プロテアーゼと分離できた（Figure 6a）。

基質結合周辺においては、トロンビンのヒトとウシの類似性は見出せたが、それ以外はほとんど検出できなかった。それは近傍三角形IDにおいて、少なくとも三角形２個分である６個のアミノ酸のデータが必要なため、基質結合周辺においてはアミノ酸の数が少なく、厳しい条件になったためと考えられる（Figure 6b）。

タンパク質全体における近傍三角形IDを構成する三角形IDの重心距離を増せば、近傍三角形ID数も増加し、1.3Åで一段と増大した（Figure 6c）。

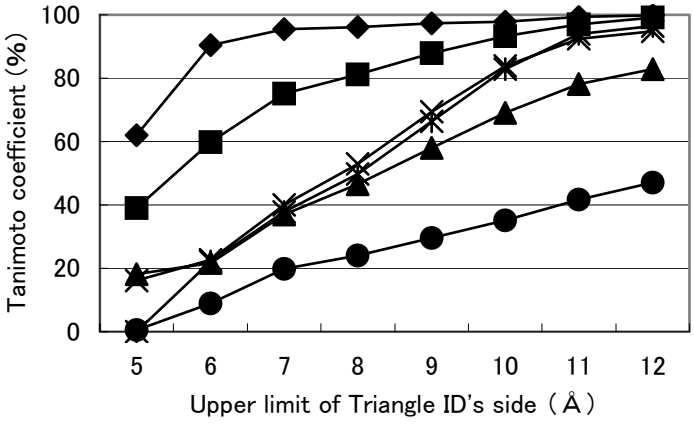

Figure 4a. Plot of Tanimoto coefficient against upper limit of Triangle ID's side on the whole protein. (◆:Thrombin (Human), ■:Thrombin (Bovine), ×:Factor Xa, ▲:Trypsin, ＊:PTP1B, ●:HIV-1protease)

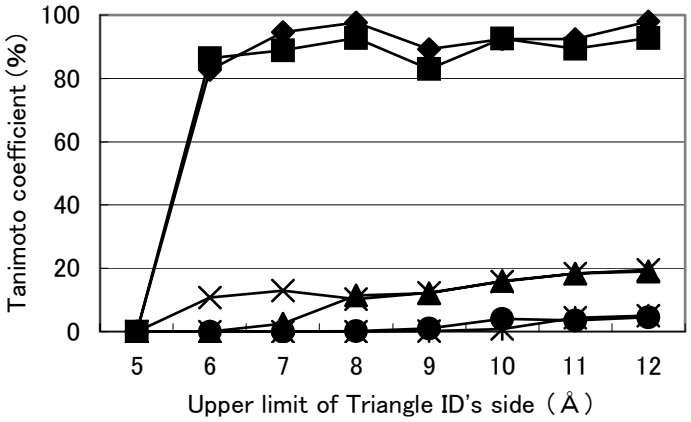

Figure 4b. Plot of Tanimoto coefficient against upper limit of Triangle ID's side on the catalytic domain. (◆:Thrombin (Human), ■:Thrombin (Bovine), ×:Factor Xa, ▲:Trypsin, ＊:PTP1B, ●:HIV-1protease)

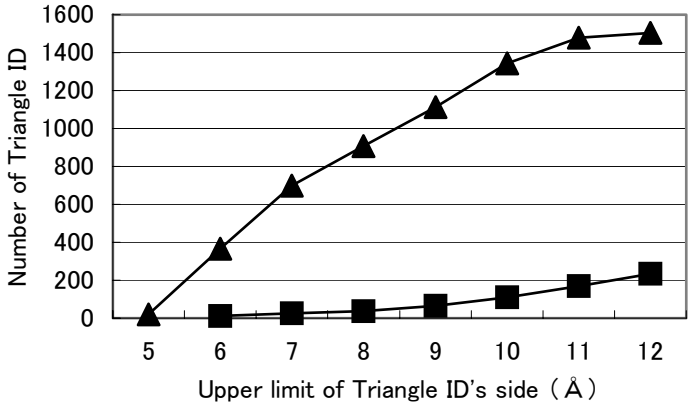

Figure 4c. Plot of number of Triangle ID against upper limit of Triangle ID's side. (■:Catalytic domain, ▲:Whole protein)

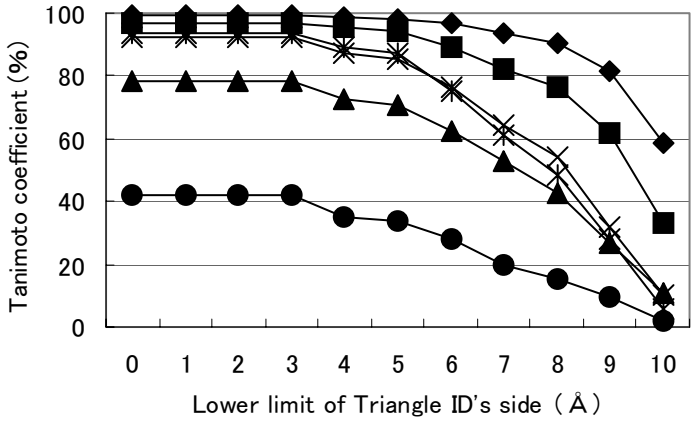

Figure 5a. Plot of Tanimoto coefficient against lower limit of Triangle ID's side on the whole protein. (◆:Thrombin (Human), ■:Thrombin (Bovine), ×:Factor Xa, ▲:Trypsin, ＊:PTP1B, ●:HIV-1protease)

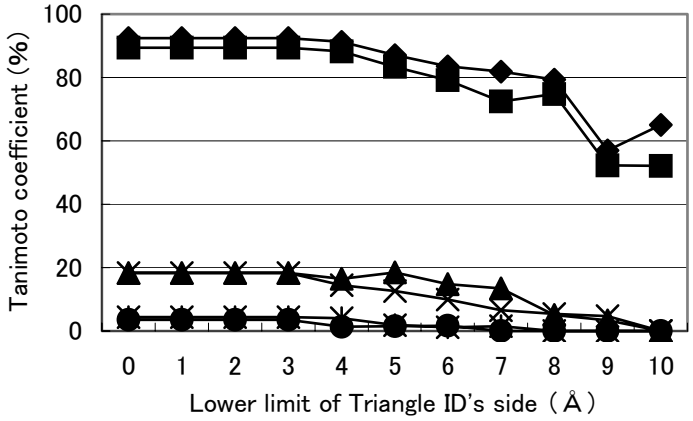

Figure 5b. Plot of Tanimoto coefficient against lower limit of Triangle ID's side on the catalytic domain. (◆:Thrombin (Human), ■:Thrombin (Bovine), ×:Factor Xa, ▲:Trypsin, ＊:PTP1B, ●:HIV-1protease)

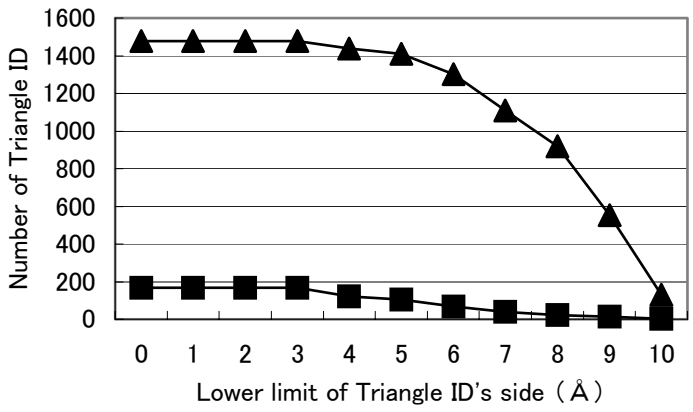

Figure 5c. Plot of number of Triangle ID against lower limit of Triangle ID's side. (■:Catalytic domain, ▲:Whole protein)

４.３　三角形ID、近傍三角形IDに基づくタンパク質立体構造ホモロジー検索

検討して得た条件を使用して、ヒトのトロンビン1AHTを検索式にし、検証用データセットに対する三角形ID、近傍三角形IDに基づくタンパク質立体構造ホモロジー検索を試みた。

タンパク質全体においては、近傍三角形IDを使用した。条件は三角形IDの辺を下限3Å、上限9Åに、近傍三角形IDを生成するために三角形IDの重心距離を1.2Åに固定した（Table 1）。結果を見ると、各タンパク質が非常に良く分離された。検索式であるヒトのトロンビンとウシのトロンビンの類似性が示され、トロンビンと同じセリンプロテアーゼファミリであるファクターXa、トリプシンが続いた。

また基質結合周辺においては、三角形IDを使用し、条件は三角形IDの辺を下限3Å、上限9Åに固定して行った（Table 2）。結果を見ると、各タンパク質は分離されているが、一部タンパク質種類が交ざる所があった。しかし概ねセリンプロテアーゼとそれ以外との分離は良かった。

４.４　作成時間、検索時間

Table 1の条件、すなわちタンパク質全体において三角形IDの辺の下限3Å、上限9Åに、三角形IDの重心距離を1.2Åに固定し、58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1分23秒であった。またトロンビンの三角形IDを検索式に、58 個のタンパク質の近傍三角形ID データ群を作成するために要した時間は、約1秒であった。

またTable 2の条件、すなわち基質結合周辺において三角形IDの辺の下限3Å、上限9Åに、58個のタンパク質の三角形IDデータ群を作成した時間は、1秒以下であった。またトロンビンの三角形IDを検索式に、58個の検証用データセットの三角形IDに対して検索に要した時間は、1秒以下であった。

三角形IDまたは近傍三角形IDに基づくタンパク質立体構造ホモロジー検索は、配列アライメントを取る必要性がないため、配列の類似性に関係なくすべてのタンパク質を対象にできる非常に大きなメリットがある。PDBデータベースには、約3万件のタンパク質立体構造情報が登録されており、今回の実験データに基づき単純に計算すると、3万件におけるデータ作成時間は約13時間、検索時間は約9分となり、比較的許容時間範囲内でのPDB全体の検索の可能性を示した。

５．まとめ

本論文では、検証用データセットを使って三角形IDに基づくタンパク質立体構造ホモロジー検索が有効に行えるかを評価した。すなわち、精度の高い検索が行え、検索効率（速度）が高いことを検証した。辺の長さに幅を持たせた20種類のアミノ酸の組み合わせで8000個の三角形IDを作成し、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。セリンプロテアーゼのヒトのトロンビンを検索式に58個の検証用データセットに対してタンパク質立体構造ホモロジー検索を行った。その結果、類似度の指標である谷本係数の高いものから、検索式のヒトのトロンビンそのもの、引き続きウシのトロンビン、そして同じセリンプロテアーゼであるファクターXa、トリプシン、そして下位にセリンプロテアーゼと異なるHIV-1プロテアーゼ、PTP1Bとなった。タンパク質全体の場合は近傍三角形IDを、また基質結合周辺の場合は三角形IDを使用すると良い結果が得られた。検索時間は、タンパク質全体、基質結合周辺ともに約1秒であった。三角形IDまたは近傍三角形IDに基づくタンパク質立体構造ホモロジー検索は、配列アライメントを取る必要性がないため、配列の類似性に関係なくすべてのタンパク質を対象にできる非常に大きなメリットがある。PDBデータベースには、約3万件のタンパク質立体構造情報が登録されており、今回の実験データに基づき単純に計算すると、3万件におけるデータ作成時間は約13時間、検索時間は約9分となり、比較的許容時間範囲内でのPDB全体の検索の可能性を示した。

今回はタンパク質のアミノ酸のCαで三角形IDを作成したが、今後、アミノ酸側鎖の官能基や性質で検討を行うことを考えている。

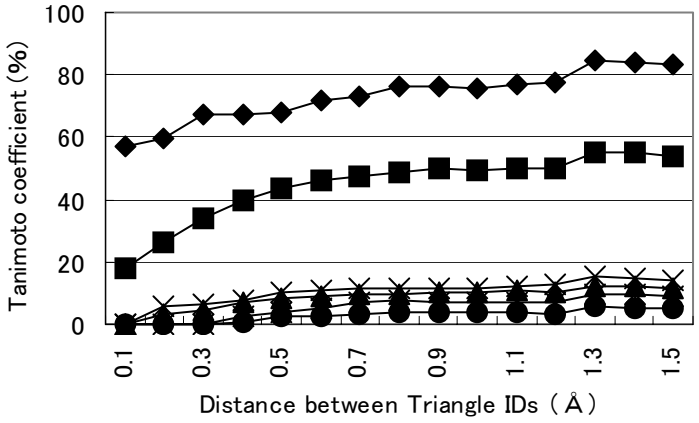

Figure 6a. Plot of Tanimoto coefficient against distance between Triangle IDs on the whole protein. (◆:Thrombin (Human), ■:Thrombin (Bovine), ×:Factor Xa, ▲:Trypsin, ＊:PTP1B, ●:HIV-1protease)

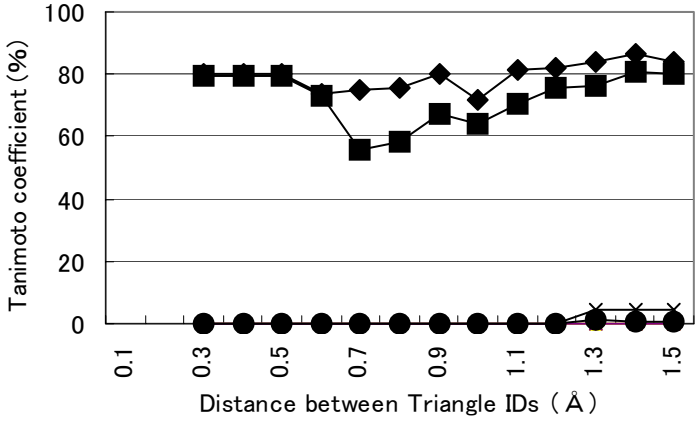

Figure 6b. Plot of of Tanimoto coefficient against distance between Triangle IDs on the catalytic domain. (◆:Thrombin (Human), ■:Thrombin (Bovine), ×:Factor Xa, ▲:Trypsin, ＊:PTP1B, ●:HIV-1protease)

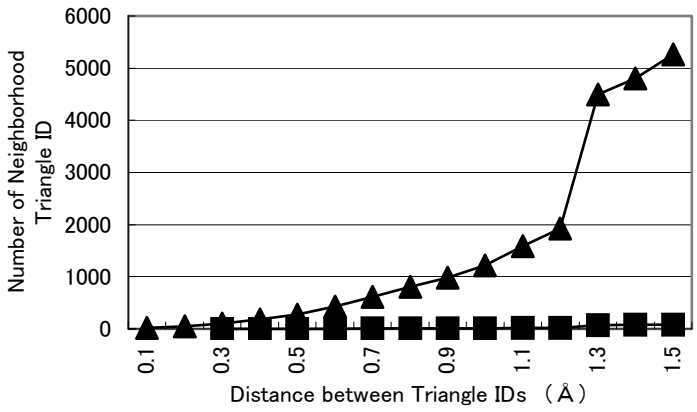

Figure 6c. Plot of number of Neighborhood Triangle ID against distance between Triangle IDs. (■:Catalytic domain, ▲:Whole protein)

Table 1 Results of 3D structural homology search based on Neighborhood Triangle ID on the whole proteins.

	Tanimoto coefficient (%)	PDB	Protein Name
1	100.0	1AHT	Thrombin (Human)
2	83.8	1K21	Thrombin (Human)
3	81.9	1THR	Thrombin (Human)
4	79.8	1O2G	Thrombin (Human)
5	77.2	1LHG	Thrombin (Human)
6	76.1	1DWB	Thrombin (Human)
7	72.1	1BMM	Thrombin (Human)
8	71.4	1UMA	Thrombin (Human)
9	70.3	1C5N	Thrombin (Human)
10	63.4	1HAI	Thrombin (Human)
11	54.5	1UVT	Thrombin (Bovine)
12	52.5	1AVG	Thrombin (Bovine)
13	52.1	1UVU	Thrombin (Bovine)
14	50.5	1ID5	Thrombin (Bovine)
15	49.5	1ETS	Thrombin (Bovine)
16	49.3	1ETT	Thrombin (Bovine)
17	47.8	1ETR	Thrombin (Bovine)
18	43.5	1HRT	Thrombin (Bovine)
19	13.6	1NFX	Factor Xa
20	13.1	1MQ5	Factor Xa
21	13.0	1IOE	Factor Xa
22	13.0	1HCG	Factor Xa
23	12.9	1KSN	Factor Xa
24	12.9	1EZQ	Factor Xa
25	12.6	1G2M	Factor Xa
26	12.6	1C5M	Factor Xa
27	12.5	1LPG	Factor Xa
28	11.8	1FAX	Factor Xa
29	10.7	1BJV	Trypsin
30	10.7	1GI0	Trypsin
31	10.5	1JIR	Trypsin
32	10.5	1AUJ	Trypsin
33	10.5	1YYY	Trypsin
34	10.5	1C1N	Trypsin
35	10.4	1CE5	Trypsin
36	10.4	1MAX	Trypsin
37	10.0	1O30	Trypsin
38	9.9	1F2S	Trypsin
39	7.3	1L8G	PTP1B
40	7.2	1Q1M	PTP1B
41	7.2	1AAX	PTP1B
42	7.1	1BZC	PTP1B
43	7.0	1NL9	PTP1B
44	6.9	1QXK	PTP1B
45	6.8	1EEN	PTP1B
46	6.7	1PYN	PTP1B
47	6.6	1PH0	PTP1B
48	6.6	1OEM	PTP1B
49	3.8	1YTG	HIV-1protease
50	3.6	1KJ4	HIV-1protease
51	3.6	1AID	HIV-1protease
52	3.5	1D4K	HIV-1protease
53	3.5	2AID	HIV-1protease
54	3.5	1CPI	HIV-1protease
55	3.5	1B6J	HIV-1protease
56	3.4	1F7A	HIV-1protease
57	3.3	8HVP	HIV-1protease
58	3.2	1MT7	HIV-1protease

Table 2 Results of 3D structural homology search based on Triangle ID on the catalytic domains.

	Tanimoto coefficient (%)	PDB	Protein Name
1	100.0	1AHT	Thrombin (Human)
2	97.1	1K21	Thrombin (Human)
3	96.5	1C5N	Thrombin (Human)
4	95.1	1UMA	Thrombin (Human)
5	87.1	1THR	Thrombin (Human)
6	86.5	1DWB	Thrombin (Human)
7	86.4	1BMM	Thrombin (Human)
8	86.0	1O2G	Thrombin (Human)
9	82.8	1UVT	Thrombin (Bovine)
10	82.8	1ETT	Thrombin (Bovine)
11	80.1	1HRT	Thrombin (Bovine)
12	79.7	1UVU	Thrombin (Bovine)
13	79.4	1ETR	Thrombin (Bovine)
14	78.3	1AVG	Thrombin (Bovine)
15	77.9	1ID5	Thrombin (Bovine)
16	74.8	1ETS	Thrombin (Bovine)
17	64.7	1HAI	Thrombin (Human)
18	63.9	1LHG	Thrombin (Human)
19	7.9	1NFX	Factor Xa
20	7.8	1KSN	Factor Xa
21	7.8	1HCG	Factor Xa

22	7.8	1C5M	Factor Xa
23	7.8	1MQ5	Factor Xa
24	7.8	1IOE	Factor Xa
25	7.7	1LPG	Factor Xa
26	7.6	1EZQ	Factor Xa
27	7.3	1YYY	Trypsin
28	7.2	1FAX	Factor Xa
29	7.2	1G2M	Factor Xa
30	5.9	1AUJ	Trypsin
31	5.9	1BJV	Trypsin
32	5.9	1C1N	Trypsin
33	5.9	1CE5	Trypsin
34	5.9	1F2S	Trypsin
35	5.9	1GI0	Trypsin
36	5.9	1JIR	Trypsin
37	5.8	1MAX	Trypsin
38	5.7	1O30	Trypsin
39	1.0	1B6J	HIV-1protease
40	1.0	1CPI	HIV-1protease
41	1.0	1D4K	HIV-1protease
42	1.0	1OEM	HIV-1protease
43	1.0	1YTG	HIV-1protease
44	1.0	1F7A	HIV-1protease
45	1.0	1AID	HIV-1protease
46	1.0	1KJ4	HIV-1protease
47	0.9	8HVP	HIV-1protease
48	0.9	2AID	HIV-1protease
49	0.9	1MT7	HIV-1protease
50	0.0	1AAX	PTP1B
51	0.0	1BZC	PTP1B
52	0.0	1EEN	PTP1B
53	0.0	1L8G	PTP1B
54	0.0	1NL9	PTP1B
55	0.0	1PH0	PTP1B
56	0.0	1PYN	PTP1B
57	0.0	1Q1M	PTP1B
58	0.0	1QXK	PTP1B

参考文献

[1] Altschul SF et al, *Nucleic Acids Res*., 25,3389-402(1997).

[2] Ben-Hur A et al, *Bioinformatics*, 19 Suppl 1:i26-33(2003).

[3] Protein Data Bank, (URL =http://www.rcsb.org/pdb/index.html).

[4] 伊藤 暢聡, 楠木 正巳, 中村 春木, *日本化学会情報化学部会誌*, Vol. 21, 20 (2003).

[5] Maiorov VN, Crippen GM, *J Mol Biol*., 235,625-34(1994).

[6] 佐藤章一, 八尾徹, *情報化学討論会・構造活性相関シンポジウム講演要旨集*, 10th-15th, 88-91(1987).

[7] Carugo O, Pongor S., *Protein Sci*., 10,1470-3(2001).

[8] Whitlow M et al, *Acta Crystallogr D Biol Crystallogr*., 55 (Pt 8),1395-404(1999).

[9] John M et al, *J. Chem. Inf. Comput. Sci*., 38, 983 -996(1998).

Novel 3D protein structural homology search algorithm based on the Triangle ID

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

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

In the post-genome era, the number of registered 3D protein structures increases dramatically. BLAST homology search tool is widely used for finding homologues. However, there are many remote homologue proteins with 3D structural similarity that BLAST cannot detect. Accordingly 3D protein structural homology search is strongly required. We propose a novel 3D protein structural homology search algorithm based on the Triangle ID comparison method. We focused a triangle structure consisting of three amino acids and called it as Triangle ID. We considered 20 kinds of amino acids, so there are 8000 kinds of Triangle IDs. 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 and made several sample data sets that consist of serine protease family and different family from serine protease. Using these data, we did several search experiments and showed the validness of our assumption and the scalability of our method. Our method opens the possibility of the efficient 3D protein structural homology search.

keywords: 3D protein structural homology search, Triangle ID, Neighborhood Triangle ID , 3D protein structure, PDB

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