様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月３１日現在

機関番号：13904
研究種目：基盤研究(C)
研究期間：2010～2012
課題番号：22500130
研究課題名（和文）　近傍フラグメントスペクトル表現に基づくタンパク質構造データマイニングに関する研究
研究課題名（英文）　Protein structural data mining based on the Neighborhood Fragment Spectra representation
研究代表者
加藤　博明（KATO　HIROAKI）
豊橋技術科学大学・大学院工学研究科・講師
研究者番号：30303704

研究成果の概要（和文）：
本研究では、分子の構造特徴を定量的に記述するための新たな方法として近傍フラグメントスペクトル表現(NFS)を提案するとともに、分子構造の高次縮約表現を基礎として、巨大で複雑なタンパク質のアミノ酸配列ならびに立体構造データの構造類似性解析を試みた。タンパク質複合体も対象とした立体構造の自動分類も含めたその手法の確立は、タンパク質構造データマイニングにおける一つの重要な要素技術をなすものであり、その意義は極めて大きい。

研究成果の概要（英文）：
In the present work, the author has proposed the Neighborhood Fragment Spectra (NFS) method to describe structural feature of a given molecule.　It has applied for sequence and structural similarity analysis of proteins using the reduced representation of complex and large protein structures.　The developed method including an automated classification of quaternary structures is one of the key technologies for protein structural data mining.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,200,000	360,000	1,560,000
2011 年度	900,000	270,000	1,170,000
2012 年度	700,000	210,000	910,000
年度			
年度			
総　計	2,800,000	840,000	3,640,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：分子構造情報処理・三次元構造類似性・タンパク質構造分類・データマイニング・分子生命情報学・分子構造データベース・タンパク質モチーフ・近傍フラグメント

１．研究開始当初の背景

(1) タンパク質は主たる遺伝情報の最終的な発現系となる生体高分子であり、その三次元構造と機能との間には密接な関係があることはよく知られている事実である。特にモチーフと呼ばれるタンパク質構造中に特定の配置で存在する局所構造特徴は、遺伝子配列の中でもよく保存されている部分であると考えられる。従って、タンパク質のモチーフ構造探索、あるいは広い意味での共通構造特徴の探索はタンパク質の構造-機能解析だけでなく、遺伝情報解析においても極めて重要な問題の一つである。

(2) 一方、ポストゲノム計画の進展、並びにタンパク質構造決定技術の進歩に伴い立体構造のデータは急速に増加しており、その構造データベースはタンパク質の構造と機能との関係解明など分子生物学上の新たな知識獲得のための基本要素としてその重要性はますます高まっている。しかし、タンパク質構造の巨大さや複雑さ、さらには近年の急激なデータ数の増大から、手動によるモチーフの検索やその特徴解析はほとんど不可能となっている。そのため、これらのデータベースを有効に活用し、三次元構造特徴の系統的な解析（タンパク質構造データマイニング）を行なうための方法論の確立、並びに有効なコンピュータツールの開発が切望されている。

(3) 筆者らはこれまでに、三次元分子構造特徴解析に基づく知識発見の視点から、アミノ酸配列レベルのモチーフデータベース PROSITE に登録されている配列パターンに注目し、これに対応する三次元部分構造情報を網羅的に集積した三次元モチーフ辞書の構築を試みた。また、グラフ論的な部分構造検索技法を基礎とした三次元モチーフ構造検索アルゴリズム、さらには質問構造の設定を要求しない複数タンパク質間の三次元共通構造特徴（新規モチーフ候補部位）の自動認識のためのシステムの開発を進めてきた。

２．研究の目的

(1) 本研究課題では、これらの成果をもとに、モチーフの情報も内包したタンパク質構造全体の類似性を定量的に評価するための新たな方法として、注目する原子の周辺環境に注目した近傍フラグメントスペクトル（NFS: Neighborhood Fragment Spectra）表現を提案する。

(2) また、従来の局所的なドメイン、あるいは１本のポリペプチド鎖で構成される三次構造レベルだけではなく、複数の鎖が会合する四次構造やタンパク質複合体を対象とした、より複雑で高度な生体高分子の構造-機能相関解析のための分子構造データマイニング手法の確立を目指す。

３．研究の方法

(1) まず最初に、有機低分子を対象として、筆者らが先に提案したジオメトリカルフラグメントスペクトル（GFS）法を参考に、分子の立体構造を反映した新たな表現方法を定義する。GFS とは、化学物質の三次元構造式から、(結合に関係なく) 原子の可能なすべての組み合わせを列挙し、その数値的な特徴づけにもとづいて化学物質の立体的な構造プロフィールを多次元ベクトルとして表現するものである。GFS は分子のグローバル（全域的）な立体構造特徴に注目したものであり、また分子のサイズ（すなわち構成原子数）が大きくなると、そのフラグメントの列挙には組み合わせ論的な問題が発生する。本研究課題ではタンパク質複合体（四次構造）の会合領域の構造特徴探索など、より広い応用を視野に、分子の空間的な局所構造特徴を効果的に記述するための方法について検討した。

(2) 与えられた分子構造データから、原子間の結合情報（隣接関係）に注目したトポロジカル距離(DT)、および三次元座標情報に基づくジオメトリカル距離(DG)の２種類の距離行列を生成し、それに対応するエッジ重み付き完全グラフを考える。次に、①各ノードの近傍に位置するノード集合（近傍フラグメント）を列挙する。これは、注目する原子から、指定した距離の半径 r の球内に位置する原子（団）を探索することに対応する。②近傍フラグメントに対し、特徴量 W（例えばフラグメントの構成原子数）を計算する。③この特徴量 W を元の分子の注目原子の重みとしてラベル付けすることにより、構造特徴を可視化することができる。さらに、④同じ W を持つフラグメントを数え上げ、ヒストグラムを生成する（図１）。

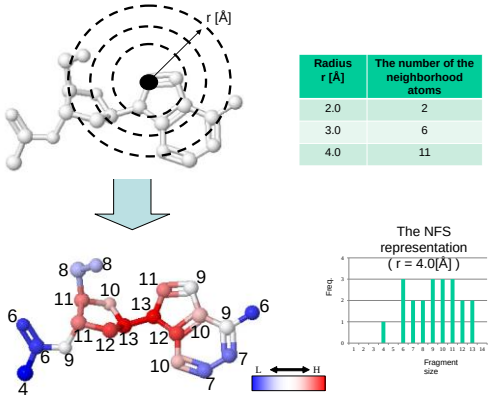

図１　注目する原子の三次元近傍フラグメントと NFS 生成の概念図.

本研究では、このヒストグラムを近傍フラグメントスペクトル(NFS)と定義した。

(3) タンパク質の構造類似性評価への応用に際しては、最初に、有機低分子（分子グラフ）のトポロジカル距離(DT)（原子間の結合関係の情報）の考えを、タンパク質のアミノ酸配列に基づく距離に対応づけた。すなわち、注目するアミノ酸残基から前後 m 残基内の部分配列（ペプチドフラグメント）を切り出し、これをその注目残基の近傍フラグメント[DT=m]と定義した。例えば、グリシン残基を中心とする前後 1 残基からなるフラグメント（トリプレット）は全部で 400 種類定義でき、与えられたアミノ酸配列中での、各トリプレットの有無の情報を 400 ビットのフィンガープリントとして記述することができる（図２）。

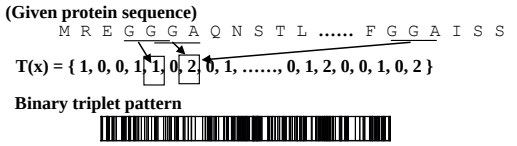

図２　アミノ酸配列の近傍フラグメント表現例（グリシン中心のトリプレットパターン）.

(4) 一方、三次元座標情報に基づくジオメトリカル距離(DG)の算出に際しては、与えられたタンパク質立体構造データから、構成アミノ酸残基をそれぞれひとつの仮想原子とみなし、対応するアルファ炭素の座標を用いて近似して表現した。これらの点の集合から、指定した距離の半径 rÅの球内に位置するアミノ酸残基を探索し、近傍フラグメント[DG=r]と定義した。抽出された近傍フラグメントは、そのフラグメントサイズと頻度の情報に基づき多次元パターンベクトルとして表現した。

(5) さらに、複数の鎖が会合する四次構造やタンパク質複合体を対象として、その会合領域パターンに注目した立体構造の自動分類への応用についても検討した。近傍原子（アミノ酸残基）の情報だけではなく、四次構造を構成しているサブユニットの帰属情報を重み付けることにより、サブユニット間の会合領域など、構造的に重要な役割を果たしていると思われる部位の特徴を強調して表現することができる。

４．研究成果

(1) ある分子構造から生成した NFS は、多次元ベクトル空間上の一つの点とみなすことができる。従って、ある二つの分子構造の類似度（相違度）は、それに対応する多次元ベクトル空間上での二点間の距離（例えば、ユークリッド距離）で定義することができる（図３）。この表現をもとに、クエリー（リファレンス）化合物とデータベース内の各化合物構造とのペアワイズ比較を行ない、類似度の値でソートして順位付けすることで構造類似性検索を実現した。結晶構造データベースから抽出したテストデータセットを例に検索実験を試みた結果、ユーザが指定する距離の種別や、そのしきい値に応じて、様々な視点からの構造類似性評価が可能であることが確認できた。

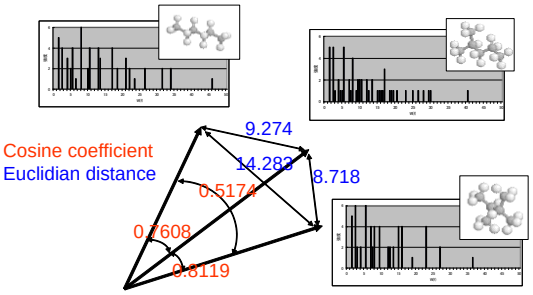

図３　NFS 表現に基づく分子の構造類似性比較の概念図.

(2) アミノ酸配列レベルの近傍フラグメント、すなわち注目残基とその前後残基から構成されるペプチドフラグメントを網羅的に集積し、新規配列モチーフ候補発見のためのデータベースを構築した（図４）。

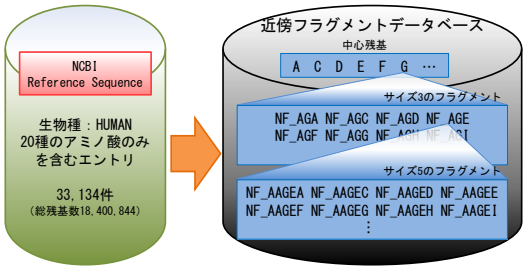

図４　構築したアミノ酸配列近傍フラグメントデータベースの概念図.

NCBI RefSeq データベースから抽出した 33,134 件（総残基数 18,400,844）の配列データをもとに階層的な近傍フラグメントデータベースの構築を行なった。その結果、近傍距離１のフラグメント（トリプレット）は、組み合わせ可能なフラグメントパターン 20^3 ＝8000 の全てのパターンが登録（すなわち，少なくとも１回は出現）されていることが分かった。一方、近傍距離２のフラグメント（クインテット）では、3,200,000 のうち 2,348,055 パターンのフラグメントが登録されていた。これらに対し、フラグメントの出現確率やその共起確率に基づく特徴解析を試みた結果、モチーフに関連する興味深い特徴パターンを見出すことができた（図５）。

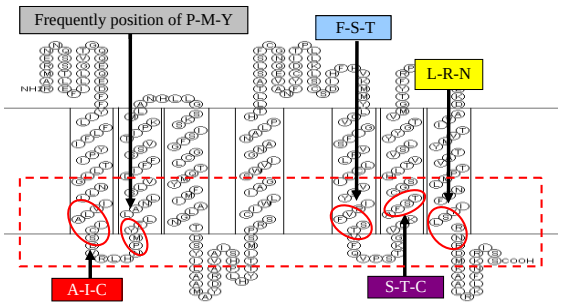

図５　匂い受容体タンパク質中の特徴的な配列近傍フラグメント（トリプレット）の例.

(3) 一方、三次元座標情報に基づく近傍フラグメントスペクトル表現をもとに、タンパク質立体構造の類似性評価への応用を試みた。図６に示すように、近傍距離のしきい値を変化させることにより、例えば二次構造要素の違いやその折りたたみパターンなど、異なる視点による特徴解析が可能である。テストデータセットを例に検索実験を試みた結果、クエリーと同様の構造特徴を持つタンパク質を類似度上位に検索できたことを確認した（図７）。また、これら特徴空間における分子同士の関係ネットワークの系統樹表現を用いた可視化についても併せて検討を行なった。

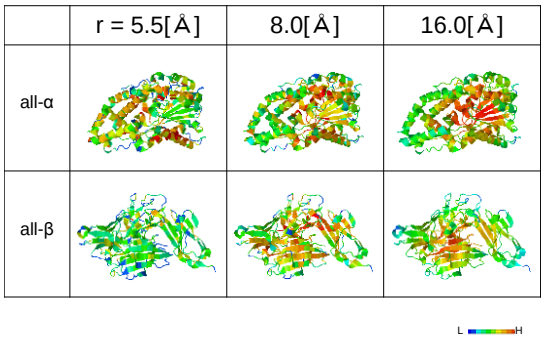

図６　指定した近傍距離による三次元近傍フラグメント表現.

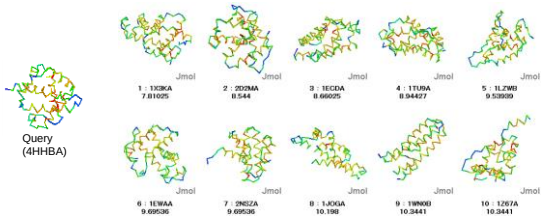

図７　NFS 表現に基づくタンパク質三次元構造類似性検索の結果例.

(4) 従来の近傍フラグメントサイズ（構成残基数）の情報だけでなく、その物理化学的特性値情報の利用について検討した。具体的には、近傍フラグメントを構成する各アミノ酸残基の特性値の総和をそのフラグメントの重みと定義した。例えば、疎水性インデックスで重み付けした結果を元の分子構造に重ね合わせて表示することにより、分子表面と内部の違いや、多量体構造における周辺環境の違いを可視化することができた。また、複合体を形成しているサブユニットの帰属情報を重み付けることにより、その会合領域を可視化し（図８）、その会合パターンに基づき構造クラス分類を行なうことができた。

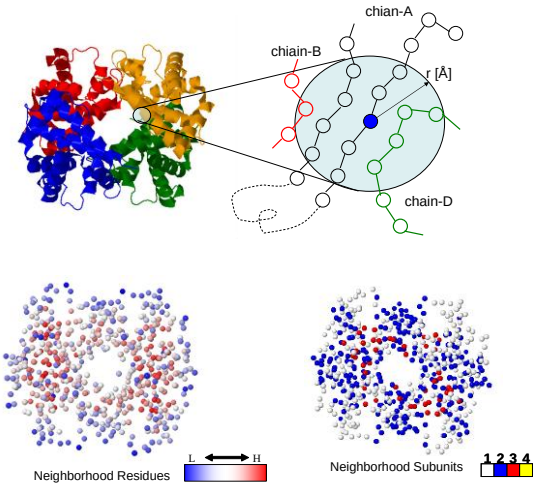

図８　近傍フラグメント表現に基づくタンパク質四次構造の特徴解析.

(5) 医農薬品開発など、新規有用物質の候補構造探索やリスク評価における特性予測問題では、トポロジカル（二次元的）な構造情報だけでなく、その立体構造を考慮したより詳細な構造特徴解析が極めて重要な意味を持つと考えられる。本研究で提案した近傍フラグメントスペクトル(NFS)は、分子の構造特徴を反映した新規の構造プロファイル表現であり、例えば分子構造中の活性部位周辺の構造特徴に注目した構造類似性検索など、従来のものとは異なる視点からの特徴探索を実現できる。

一方、その構造的自由度の高さから系統的な取り扱いが困難で、これまであまり注目されていなかったランダムコイルと呼ばれる領域を含めて、タンパク質の特徴的な構造パターンをもとに分子全体の構造類似性評価を試みた。ポリペプチド鎖の会合によって形成される複雑な機能部位の同定など、従来とは異なる視点による、より合理的な構造特徴解析を可能とした。データベースに対する構造類似性検索や立体構造の自動分類も含めたその手法の確立は、タンパク質構造情報解析における一つの重要な要素技術をなすものであり、その意義は極めて大きい。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計９件）
[1] 永松晃一，加藤博明，近傍フラグメント表現に基づくタンパク質配列特徴解析システムの開発，第 40 回構造活性相関シンポジウム，平成 24 年 11 月 29 日，岡崎市図書館交流プラザ（愛知県岡崎市）
[2] 森本孝朗，加藤博明，遺伝的アルゴリズムを用いた分子ライブラリデザインシステムの開発，第 40 回構造活性相関シンポジウム，平成 24 年 11 月 29 日，岡崎市図書館交流プラザ（愛知県岡崎市）
[3] Hiroaki Kato, Koichi Nagamatsu and Masaaki Iitsuka, Structural similarity search of molecules using the Neighborhood Fragment Spectra, Joint Conference on Informatics in Biology, Medicine and Pharmacology 2012, 平成 24 年 10 月 15 日，タワーホール船堀（東京都江戸川区）
[4] Hiroaki Kato, and Masaaki Iitsuka, Three-dimensional protein structural data mining based on the NFS representation, The 9th China-Japan Joint Symposium on Drug Design and Development, 平成 24 年 9 月 22 日，Guilin Bravo Hotel（中国・桂林市）
[5] 檜山綾乃，加藤博明，系統樹表現に基づく分子の関係ネットワークの可視化システムの開発，第 39 回構造活性相関シンポジウム，平成 23 年 11 月 28 日，東京理科大学薬学部（千葉県野田市）
[6] Hiroaki Kato, Chisato Morishita, and Sachie Hakamata, Development of structural feature analysis program for proteins based on the triplet pattern, The 2010 Annual Conference of the Japanese Society for Bioinformatics, 平成 22 年 12 月 13 日，九州大学医学部（福岡市）
[7] 松田貴人，加藤博明，構造活性相関研究のための分子の三次元特徴フラグメント辞書の作成，第 33 回情報化学討論会，平成 22 年 10 月 30 日，徳島大学工学部（徳島市）
[8] 家村享明，檜山綾乃，加藤博明，アミノ酸残基の周辺環境に注目した GPCR の配列特徴解析システムの開発，第 38 回構造活性相関シンポジウム，平成 22 年 10 月 30 日，徳島大学工学部（徳島市）
[9] Hiroaki Kato, Three-dimensional structural data mining of proteins based on Geometrical Fragment Spectra representation, 18th European Symposium on Quantitative Structure-Activity Relationships, 平成 22 年 9 月 20 日，Rodos Palace（ギリシア）

６．研究組織
(1)研究代表者
加藤　博明（KATO HIROAKI）
豊橋技術科学大学大学院・工学研究科・講師
研究者番号：30303704