様式 C-19

科学研究費補助金研究成果報告書

平成２２年　３月２５日現在

研究種目：若手研究(B)
研究期間：2007～2009
課題番号：19700139
研究課題名（和文）
ジオメトリカルフラグメントスペクトル表現に基づくタンパク質の三次元構造類似性探索
研究課題名（英文）
Three-dimensional structural similarity search of proteins based on Geometrical Fragment Spectra
研究代表者
加藤　博明（KATO HIROAKI）
豊橋技術科学大学・工学部・講師
研究者番号：30303704

研究成果の概要（和文）：
本研究では、分子の三次元構造特徴を定量的に記述する新たな方法としてジオメトリカルフラグメントスペクトル（GFS）を提案した。この表現をもとに、匂い分子に対する構造類似性検索を試みた結果、トポロジカルな構造特徴は異なるものの、同じ匂い活性を持つ構造を見出すことができた。また、グリシン残基の空間配置に注目したタンパク質三次元構造の縮約表現を基礎として、タンパク質立体構造の類似性評価への応用を試み、その有用性を示した。

研究成果の概要（英文）：
In the present work, to describe 3D structural feature of a molecule, Geometrical Fragment Spectra (GFS) method has been proposed. Based on the GFS representation, structure similarity search was carried out for odorant molecules. As a result, the authors successfully found the molecules that have similar odor but quite different topological structure with a query molecule. We also tried to apply the method to structure similarity searching for proteins using a reduced representation of protein structures. The results show that the present approach is quite useful for 3D structural data mining for proteins.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,200,000	0	1,200,000
2008 年度	1,100,000	330,000	1,430,000
2009 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,200,000	600,000	3,800,000

研究分野：分子情報工学・バイオインフォマティクス
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・構造活性相関・生体生命情報学・タンパク質モチーフ・化学グラフ・構造データマイニング・タンパク質構造分類

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、分子の三次元構造特徴を定量的に記述する新たな方法としてジオメトリカルフラグメントスペクトル（GFS）を提案する。

(2) また、モチーフの情報も内包したタンパク質構造全体の漠然とした類似性を評価するために、グリシン残基の空間配置に注目した高次の構造縮約表現を導入するとともに、GFS 表現に基づくタンパク質構造類似性探索と立体構造の自動分類への応用を目指す。

３．研究の方法
(1) 最初に、有機低分子を対象として、トポロジカルフラグメントスペクトル（TFS）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。TFS とは、化学物質の構造式から可能な部分構造（フラグメント）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トポロジカル（二次元的）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。分子の立体構造を記述する方法としては、TFS のフラグメントに、その立体構造を反映した特徴量を重み付ける方法も考えられる。しかし、本研究課題では生体高分子の三次元構造類似性探索など、より広い応用を考え、トポロジカルな構造情報に依存しないフラグメントの生成と、その利用について検討した。

(2) 本研究では、三次元幾何情報を含めた化合物分子の構造を、その構成原子をノードとし、それらの間の幾何学的関係（原子間ユークリッド距離）の情報をエッジ上に重み付けた、エッジ重み付き完全グラフとして表現する。次に、①与えられた三次元構造式に対応する上記のグラフから、可能なすべての部分グラフ(フラグメント)を列挙する。これは、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することに対応する。②それぞれのフラグメント f に対し、フラグメント特徴量 W(f)を計算する。③同じ W(f)を持つフラグメントを数え上げ、ヒストグラムを作る。このヒストグラムを、ジオメトリカルフラグメントスペクトル(GFS)と定義する(図１)。フラグメント特徴量 W(f)としては、そのフラグメントの三次元構造を反映した値、例えば、全ての原子間距離の総和として定義される 3D Wiener Number などを用いる。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。そのため、本研究では列挙するフラグメントのサイズ（フラグメントの構成ノード数）の範囲（上限と下限）をユーザが指定できるように工夫した。

(3) ある一つの分子構造から生成した GFS は、TFS と同様に、多次元ベクトル空間上のひとつの点とみなすことができる。したがって、ある二つの分子構造の類似度は、それに対応する多次元ベクトル空間上での二点間の距

離で定義することができる。類似度の評価には、例えばユークリッド距離や Cosine 係数を用いることができる（図２）。

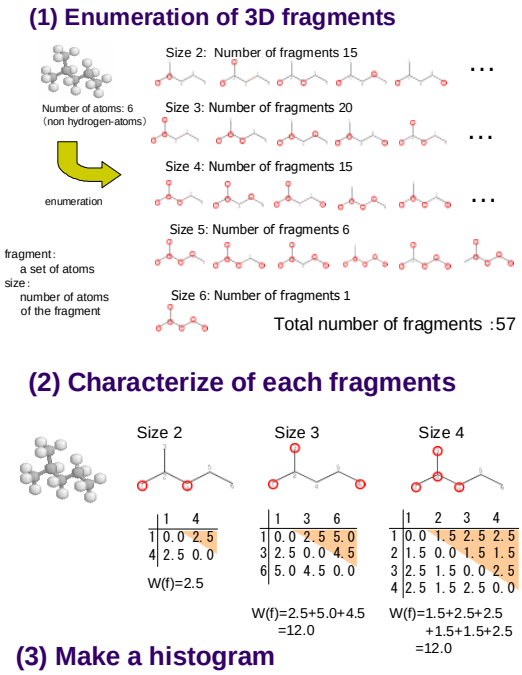

図１　ジオメトリカルフラグメントスペクトル（GFS）の生成手順.

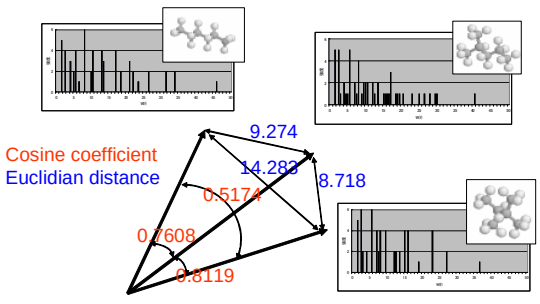

図２　GFS に基づく分子の立体構造の比較.

(4) 上記で得られた知見をもとに、タンパク質立体構造の類似性評価への応用を試みる。まず最初に、筆者らが先に構築した三次元モチーフ辞書に登録されている配列モチーフに対応する三次元部分構造を例に、その構成アミノ酸残基をそれぞれ一つの仮想原子とみなし、GFS 表現の生成と、構造類似性評価の妥当性について検討を行なう。

(5) タンパク質全体構造への適用に際しては、タンパク質構造中のある特定アミノ酸残基、例えば、その構造が最も単純で（すなわち構造的な自由度が最も高く）、かつ出現頻度が比較的高いグリシン残基に注目し、全体構造をその中に含まれるグリシン残基に対応する点の集合として表現する（図３）。このような表現をもとに、同様にグリシン残基間距離（対応するアルファ炭素間の原子間距離）の情報をもとに、GFS を生成する。このような表現を行なうことにより、フラグメント列挙に関わる組み合わせ爆発を防ぐとともに、主鎖の向きを変えるターン領域など、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可能になる。

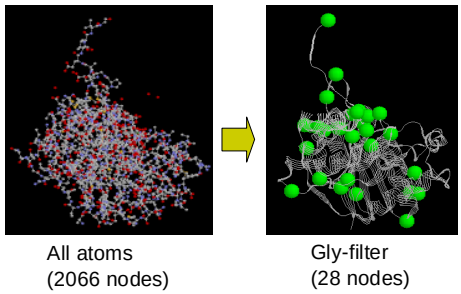

図３　グリシンフィルタ縮約表現.

４．研究成果

(1) 当研究室で作成した匂い-構造式データベースから抽出した匂い分子（1066 件）を用いて、GFS による構造類似性探索を行なった。図４の分子の GFS をクエリーとして探索したときの類似度上位 10 件の構造を図５に示す。図５で、各構造の上に順位と類似度（Cosine 係数）を、下に各構造の ID 番号を示す。また、順位の後につけられた「*」は、その構造がクエリーと同様、Musky の匂い活性を持つことを表す。なお、1066 件中、Musky の匂い活性を持つ構造は 90 件ある。これらの結果から、上位 10 件中 9 件がクエリー構造と同じく、Musky の匂い活性を持つことがわかる。特に、トポロジカルな構造情報がクエリー構造と大きく異なる 9 位の構造も Musky の匂い活性を持つことは興味深い。なお、TFS/W を用いて同様の実験を行なうと、9 位の二つの構造の順位はともに 981 位であった。以上

のように、GFS を用いると、従来の TFS とは異なる視点からの構造類似性探索が可能である。

(2) 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もまた極めて重要な意味を持つと考えられる。また、合成技術等の進歩に伴い蓄積された大量の構造データを背景としたバーチャルスクリーニングでは、種々の構造表現・類似性尺度に基づくデータベース検索結果を総合的に評価するデータフュージョンテクニックが提案されている。本研究で提案した GFS は、分子の立体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、従来のものとは異なる視点からの特徴探索を実現するものである。また、これを従来の TFS、あるいはフィンガープリント表現等による類似性検索結果と総合評価することで、より多彩な候補構造探索を可能とした。

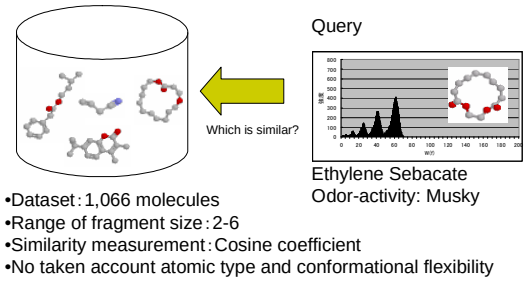

図４　GFS 表現に基づく構造類似性探索.

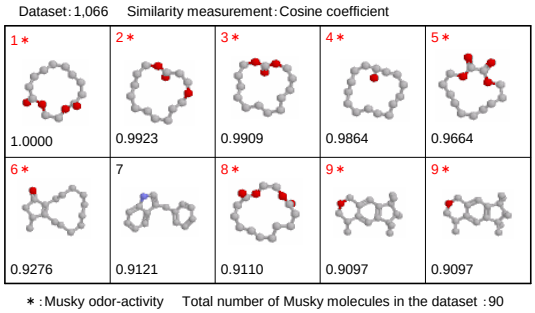

図５　匂い分子の構造類似性探索結果.

（1066 件中類似度上位 10 件）

(3) 筆者らが先に開発した三次元モチーフ辞書システムから抽出したタンパク質部分構造（EF-hand モチーフのアミノ酸配列パターンに対応する連続する 13 残基）8 件に対して GFS を生成し、スパニングツリーを用いた構造クラスタリングを試みた。ここでは、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。アミノ酸の種類は区別しないものとする。

クラスタリングの結果を図６に示す。各構造の下に分子全体のトポグラフィックインデックスである 3D Wiener Number (3DW)の値を示す。なお PDB-ID: 3RUBS と 1GAI 中のモチーフ構造は、EF-hand モチーフの機能を持たないノイズ成分であることがわかっている。本実験では、EF-hand モチーフの機能を持つ構造（クラス 1）とノイズ成分（クラス 2 と 3）とを分類することができた。一方、分子全体の 3DW を見ると、1GAI の構造の値は他のものと大きく異なるが、3RUBS の構造の値はクラスタ 1 の構造のものとほとんど変わらない。以上より、GFS は、従来の 3DW と比べて分子の立体構造をよりよく表現できていると考えられる。

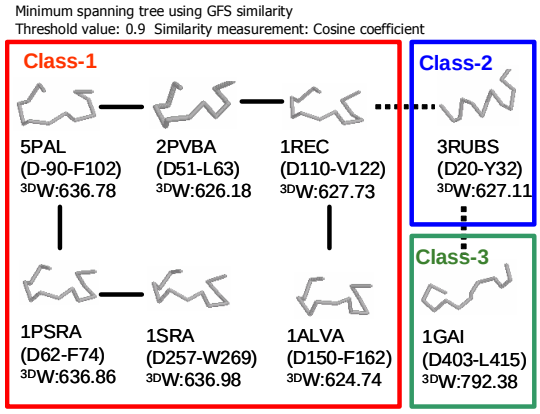

図６　GFS 表現に基づく EF-hand モチーフの構造クラスタリング結果.

(4) GFS によるタンパク質全体構造同士の構造比較（三次元構造類似性評価）を行なうためには、より高次の構造縮約表現の導入が不可欠である。本研究では、グリシンフィルタ縮約表現（図３）を適用し、GFS の生成を行なう。最新の PDB (Protein Data Bank)から調整した 1,300 鎖 (1,206 タンパク質) からなる実験用データセットに対し、それぞれグリシンフィルタで縮約表現し、距離の離散化幅 1.0Å、フラグメントサイズ 2～4 の条件のもと GFS データベースを作成した。

ここでは、図３で示したニトロゲナーゼ(1NIPA)をクエリーとして GFS に基づく三次元部分構造検索を試みた。類似性尺度 Cosine 係数を用いて構造類似性検索を試みたところ、類似度 0.950 以上のものが、1,300 件中 100 件見出された。図７（中央）に示すタンパク質分子(1CP2A)は、クエリー分子(1NIPA)と同じタンパク質ファミリーに属しており、そのアミノ酸配列や折り畳みのパターンが

類似している。従って、グリシン残基の空間的位置も比較的類似していることが予想され、このことは図から視覚的にも確認できる。よって、この検索結果は妥当であると考えられる。また、興味深い結果として、図７（左側）に示すDNA結合タンパク質(1GCB)が見つかった。1GCBはクエリー分子とはその全体構造のサイズが大きく異なることが分かる。さらに、その折り畳みのパターンも全く異なる。しかしながら、二つのタンパク質分子のグリシン残基の空間的配置に注目すると、非常に類似していることが分かった。これらの結果は、タンパク質三次元構造特徴探索における本法の有用性を強く示唆するものである。

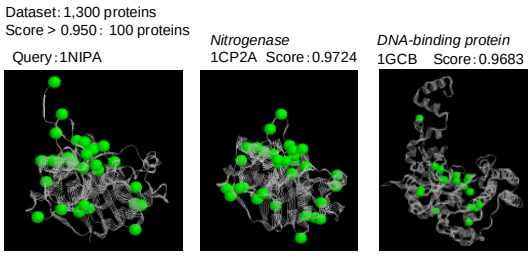

図７　GFS表現に基づくタンパク質三次元構造類似性探索結果（例）．

(5) タンパク質構造はアミノ酸配列の類似性や進化的な関係をもとにしたファミリーやスーパーファミリーと呼ばれる分類の次の階層として、折り畳み構造（フォールド）レベルによる分類が行なわれている。しかしながら、これらは主に二次構造セグメントの折り畳みパターンに注目したものである。本研究は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、タンパク質構造全体の漠然とした三次元構造類似性評価を試みた。データベースに対する三次元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８件）

[1] Chisato Morishita, and Hiroaki Kato, Structural feature analysis of human olfactory receptors based on the triplet pattern, 第37回構造活性相関シンポジウム, 平成21年11月12日, 北里大学薬学部（東京都港区）
[2] 中谷光裕, 加藤博明, 重心からの距離情報に基づく分子の三次元構造特徴解析, 第32回情報化学討論会, 平成21年10月30日, 常盤工業会館（宇部市）
[3] 吉廣伸也, 加藤博明, サブユニット間の会合領域に注目したタンパク質四次構造の自動分類, 第32回情報化学討論会, 平成21年10月30日, 常盤工業会館（宇部市）
[4] Hiroaki Kato, Shigeru Yoshida, and Yoshimasa Takahashi, Three-dimensional structural data mining based on Geometrical Fragment Spectra, The 8th China-Japan Joint Symposium on Drug Design and Development, 平成20年10月3日, 神戸国際会議場（神戸市）
[5] Yuuji Kouga, and Hiroaki Kato, Structural feature analysis of transmembrane helices in human olfactory receptors, 第36回構造活性相関シンポジウム, 平成20年10月2日, 神戸国際会議場(神戸市)
[6] Hiroyuki Ogawa, and Hiroaki Kato, Automated classification of quaternary structure of protein based on domain pattern, 第36回構造活性相関シンポジウム, 平成20年10月2日, 神戸国際会議場（神戸市）
[7] 田中裕貴, 加藤博明, ヘテロツリー表現に基づく分子の三次元構造類似性探索, 第31回情報化学討論会, 平成20年11月14日, 東京大学山上会館（東京都文京区）
[8] Hiroaki Kato, Shigeru Yoshida, and Yoshimasa Takahashi, Three-dimensional structural similarity search of molecules based on Geometrical Fragment Spectra, 2007 Annual Meeting of CBI Society, 平成19年10月3日, 広島大学（東広島市）

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・工学部・講師
研究者番号：30303704

以上