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　４月２７日現在

機関番号：１７１０２
研究種目：挑戦的萌芽研究
研究期間：2010～2012
課題番号：22654016
研究課題名（和文）　タンパク質の構造解析に対する位相幾何学的研究

研究課題名（英文）　Algebraic Topological Approaches to Protein Structure Analysis

研究代表者
平岡　裕章（HIRAOKA YASUAKI）
九州大学・マスフォアインダストリ研究所・准教授
研究者番号：10432709

研究成果の概要（和文）：

本研究ではタンパク質構造解析の諸問題に対して位相幾何学的な研究手法の開発を行った．まずタンパク質の柔らかさに関する物性指標である圧縮率を，データベース（Protein Data Bank）が提供しているデータを用いて定量化することに成功した．パーシステントホモロジー群という位相幾何学の道具をもちいて，圧縮率に影響を及ぼすと思われている特徴的な空洞の情報を抽出することにより，このような成果を得ることができた．またパーシステント図の空間に定まる距離関数をもちいてタンパク質の新たな幾何学的構造に基づく分類法も提案した．

研究成果の概要（英文）：

In this research, we developed some approaches based on algebraic topology to several problems of protein structure analysis. We succeeded in developing a new quantification of protein compressibility by using data in Protein Data Bank. The key tool is persistent homology and persistent diagrams, which enable us to detect information on cavities with certain properties. In addition, we also proposed a technique to classify proteins from their geometric structure by means of distance functions on persistent diagrams.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	900,000	0	900,000
２０１１年度	500,000	150,000	650,000
２０１２年度	800,000	240,000	1,040,000
年度			
年度			
総　計	2,200,000	390,000	2,590,000

研究分野：数物系科学
科研費の分科・細目：数学・数学一般（含確率論・統計数学）
キーワード：タンパク質，圧縮率，パーシステントホモロジー群，アルファ複体

１．　研究開始当初の背景

タンパク質の機能や物性は幾何学的な立体構造と密接な関係にあることが知られている．ここでタンパク質の立体構造はX線結晶解析実験により多くの研究グループにより調べられており，現在データベースとして公開されている．代表的なデータベースとしてはProtein Data Bankがある．ここには各タンパク質を構成している原子の空間座標からなるデータが用意されている．よってこれらのデータからタンパク質の機能や物性を調べる解析手法を開発する必要がある．これまで開発されている解析手法の中には大域的な幾何構造を特徴付けるものも存在するが，タンパク質内部に存在する穴を効率的かつ網羅的に取り扱う手法は確立されていなかった．

２．　研究の目的

タンパク質の物性を特徴付ける量の一つに圧縮率がある．これはタンパク質の外圧に対する変形度合いを測るものであり，機能とも密接に関係する重要な物性値である．この圧縮率はタンパク質内部に存在するある特徴的な空洞に関係することがこれまでの実験研究により示唆されているが，その明確な特徴付けはまだ未解決であった．これらの実験研究からの考察は共同研究者の泉俊輔を中心にこれまで詳細に調べられてきた背景がある．またタンパク質圧縮率の実験自体精密な測定を必要とするため，既存のデータベースを用いて大まかな予測を与えることが望まれている．そこで本研究ではパーシステントホモロジー群をもちいたProtein Data Bankのデータ解析手法を開発することで，タンパク質圧縮率と相関を持つ定量化を開発し，圧縮率の数学的な特徴付けを行った．

３．　研究の方法

Protein Data Bankのデータはタンパク質を構成している原子の種類と中心座標からなる．よって原子をファンデルワールス球体でモデル化すれば，タンパク質は球の和集合として表現できる．一方でタンパク質内部の空洞はこのタンパク質のファンデルワールスモデルの２次ホモロジー群と関係している．そこでまず球の和集合のホモロジー群を計算する必要があるが，これは対応する脈体であるアルファ複体（単体複体の一種）を導出することで計算を行った．脈対定理によればこのアルファ複体と球の和集合はホモトピー同値となるため，この操作でトポロジカルな情報は完全に保たれる．さらに単体複体のホモロジー群は高速数値計算が可能であることから，原子数が数千のオーダーとなるタンパク質に対しても効率的にホモロジー群を計算することが可能となる．

ここでアルファ複体について簡単に説明を行っておく．その構成法にはまず３次元空間に配置された有限点データ（タンパク質の問題では各原子の中心座標が定める点集合）からボロノイ図を構成し，そこから誘導されるドロネー図を用意する．この準備のもとでアルファ複体は点有限点データにある半径の球を配置し，その球とボロノイ領域の共通部分からなる部分集合族が定める脈体として構成される．ここで部分集合族{B_i ｜ i=1,,,,n}が定める脈体とは単体複体(V,S)であって次で定義される：
V = 1,,,,n（頂点集合）
単体|i_0,,,i_k|がSに属する必要充分条件はk+1個の部分集合B_{i_0},,,B_{i_k}に共通部分が存在する．
本研究ではここで導出したアルファ複体の具体的な数値計算方法については汎用ソフトウェアCGALを使うとともに，細部の改良を行いやすくするための独自の数値計算ソフトウェアの開発も行った．

またタンパク質のファンデルワールスモデルにはトポロジカルなノイズ（サイズの小さな空洞）が多く含まれることになる．パーシステントホモロジー群を使う利点の一つに，このトポロジカルなノイズと本質的な空洞を識別することが可能である．

ここで本研究を遂行するうえでで最も大事な方法であるパーシステントホモロジー群，パーシステント図について簡単に解説をおこなう．パーシステントホモロジー群は空間構造のフィルトレーションに対して定まる代数的な概念である．例えば単体複体のフィルトレーションなどが考えられるが，本研究のようなタンパク質の問題ではアルファ複体のフィルトレーションを考えることになる．この際フィルトレーションの入れ方には幾つかの方法があるが，例えば各原子の球体モデルの半径の増大列を考える方法が考えられる．ある半径で単体が脈体に現れれば，半径を大きくした脈体にも現れることからフィルトレーションが自然に導入できる．このフィルトレーションの各スライスに通常の単体ホモロジー群が定まるが，このホモロジー群の列とその間の包含写像が誘導する誘導準同型写像の組からパーシステントホモロジー群と呼ばれる対象が得られる．体係数のホモロジー群の設定では（本研究ではこ

の状況で常に考えるが) Quiver の表現の立場からは，ここで得られたパーシステント加群は A_n 型 Quiver 上の表現と見なせる．そこでこのパーシステントホモロジー群の直既約分解を考えることができるが，各直既約は正のルート系と一対一の対応にあり，A_n 型の Quiver の場合はホモロジー群の発生時刻と消滅時刻の情報からルートは一意に定まる．この発生時刻と消滅時刻の対を R^2 空間にプロットしたものをパーシステント図と呼ぶ．

最終的に Protein Data Bank のデータからタンパク質圧縮率と相関を持つ定量化は，各データから上記の手法によってパーシステント図を計算し，そこから圧縮率に関係することが示唆されている幾何学的性質を有する生成元を数え上げることで導出を試みた．ここで考察を行った幾何的性質とは以下で与えられる．

i) トポロジカルノイズの除去．つまりパーシステント図において発生時刻と消滅時刻の差が小さい生成元は除去する．

ii) キャビティの構成においてその境界領域の原子密度が疎なものを抽出．これは原子密度が疎なものはより外圧に対して変形の自由度を持っているという考察に基づく．パーシステント図からは，２次の生成元であって発生時刻が遅いものが対応するため，発生時刻が早い生成元は除去するように対応する．

iii) キャビティの軸方向の効果を反映する．これはキャビティの断面に対して軸方向に長いものはより大きい空間構造を内部に有しているという考察に基づく．パーシステント図からは，２次パーシステント図の生成元の個数に対して１次パーシステント図の生成元の個数を割ることにより，その効果を抽出した．

４． 研究成果

実験によって圧縮率が調べられているタンパク質に対して，本研究で提案しタパーシステントホモロジー群を用いた定量化との相関を調べた．まず Protein Data Bank 上のデータからパーシステント図を自動的に計算するプログラムの開発と汎用化を意識したソフトウェアの開発までをおこなった．これにより Protein Data Bank のデータのパーシステント図を容易にかつ高速に計算することが可能となった．
これらのソフトウェアの開発をもとに，タンパク質圧縮率の解析へ適用し，非常に明確な線型相関関係を満たすことを確認できた．これにより圧縮率が未知のタンパク質に対して初期推測を行うことを可能にした．またタンパク質圧縮率に影響を及ぼすと推測されていた幾何構造を数学的に明確に示すことにも成功した．これらの研究成果は論文としてまとめられ，現在投稿中である．

また各タンパク質に対して特徴的な生成元がパーシステント図から抽出できることになるが，この生成元が元のタンパク質の空間構造の中でどこに位置するものなのかを調べる理論・アルゴリズム・ソフトウェアの開発を行った．理論的にはここでの問題設定は，与えられたホモロジー群の生成元から最小生成元を求める問題が該当し，離散最適化問題を解くことになる．本研究ではこの離散最適化問題を線型計画問題へ軟化させるプロセスを定式化し，圧縮センシングの議論を用いて具体的に最小生成元を構成するアルゴリズムを提示した．具体的なソフトウェアの開発の際には線型計画法の数値計算ソフトウェアである CPLEX をホモロジー群計算ソフトウェアの CHomP をその基本的な部分で併用している．これにより圧縮率に十分寄与する空間構造の抽出をはじめとして，パーシステント図を用いた情報抽出の際に幾何のレベルまで情報を引き戻すことが可能となった．

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）

〔雑誌論文〕（計 １件）

①平岡裕章，パーシステントホモロジー群--離散データのトポロジカル解析--，COE Lecture Note Vol. 46: Kyushu University, 63-73, 2013.

〔学会発表〕（計 ２件）

①平岡裕章，パーシステントホモロジー群のタンパク質構造解析への応用，2012 年度日本数学会年会，応用数学分科会，3 月 22 日，京都大学.

②平岡裕章，計算トポロジーの応用，日本鉄鋼協会 3D4D シンポジウム，２０１２年３月，横浜国立大学.

〔図書〕（計　1件）

①平岡裕章,タンパク質構造とトポロジー:パーシステントホモロジー群入門,シリーズ「現象を解明する数学」,共立出版（近刊).

〔産業財産権〕

○出願状況（計　0件）

〔その他〕

６．研究組織

(1)研究代表者

平岡　裕章（Yasuaki Hiraoka）
九州大学・マスフォアインダストリ研究所・准教授
研究者番号：10432709

(2)研究分担者

泉　俊輔（Shunsuke Izumi）
広島大学・数理分子生命理学専攻・教授
研究者番号：90203116

大西　勇（Isamu Onishi）
広島大学・数理分子生命理学専攻・准教授
研究者番号：30262372