様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　　１　日現在

機関番号：　１２１０２
研究種目：　挑戦的萌芽研究
研究期間：　2013～2014
課題番号：　２５５４００６２

研究課題名（和文）マルチモーダル多視点画像を用いたタンパク質立体構造の解析

研究課題名（英文）3D protein analysis based on the multiple view images

研究代表者
福井　和広（Fukui, Kazuhiro）

筑波大学・システム情報系・教授

研究者番号：４０３７５４２３

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：タンパク質の高度な構造解析を目指して，３次元形状だけではなく，電荷や親水性などの情報も反映できるマルチモーダル構造類似度を提案した．比較する２つのタンパク質の3次元モデルから生成した多視点画像セットからそれぞれ部分空間を生成し，両者の成す正準角を構造類似度とする．回転不変LBP特徴とグラスマン判別分析（GDA）の組合せから成る提案法と，代表的なアライメントベース法との性能比較を行った結果，提案法の高い識別性能(69.43%)を確認できた．さらに提案フレームワークを実装したタンパク質解析サーバー(View-based Protein Comparison (VPC) system)を構築した．

研究成果の概要（英文）：In this research project, we have proposed a new similarity measure for protein structure comparison that uses a set of multi-view visualization images of 3D protein structures. The advantage of our method is that distinctive structural information can be embedded in the visualization. In this approach, each set of multi-view images is represented by a subspace, while applying LBP feature extraction to the image set. The similarity between two protein structures is then characterized by the distance between two points which are corresponding to different two subspaces, on a Grassman manifold. The proposed method was evaluated by classification experiments on 7 protein classes based on SCOP (alpha, beta, alpha/beta, alpha+beta, multidomain protein, membrane, and small protein), using 700 randomly selected proteins. In the experiments, our method significantly outperformed the CE and FATCAT alignment methods. In addition, we have constructed a view-based protein comparison (VPC) system.

研究分野：　コンピュータビジョン

キーワード：　タンパク質　３次元構造　多視点画像　相互部分空間法　画像認識

１．研究開始当初の背景

タンパク質がどのような機能を持つかはその立体構造と密接に関係しており，立体構造データベース(PDB,SCOP など)の重要性は増大している．代表的な立体構造情報データベース SCOP は専門家の目視分類により構築されているために，更新頻度が遅く，エラーや修正が多く，このような人手によるデータ整備は既に限界に来ている．

立体構造データベースの構築・運用において，基盤となるのが立体構造同士の類似度を測る方法である．良く知られた構造アライメント法では，まず並進・回転により２つの構造をできるだけ一致させた後，対応する中心炭素原子間の距離から類似度(RMSD)を測る．この方法は現在，標準的に使われているが，予め全体形状のアライメント(位置合わせ)が必要であり，局所解の存在による計算の不安定性や，構成分子数が異なるタンパク質ペアには適用できないという問題もあった．さらに形状と同様に重要な情報：電荷，親水性，疎水性などを，構造類似度に直に反映させることが難しかった．

２．研究の目的

先に述べたように，タンパク質の立体構造の類似度を測る方法として，構成分子の３次元位置合わせが基本と見なされており，これに基づく様々な方法が提案されている．しかしながら，いずれも表現能力，安定性の点で課題があった．

本研究ではこれらの問題を解決する新しいタンパク質立体構造の類似度を測るマルチモーダル構造類度を提案し，SCOP データベースを用いた３次元タンパク質のクラス識別実験によりその有効性を検証する．更に提案するマルチモーダル構造類似度に基づく識別エンジンを搭載したタンパク質解析サーバーを構築する．

３．研究の方法

(1)基本アイデア

本研究ではこれらの問題を解決するために，従来とはまったく異なるアプローチをとる．すなわち，比較するタンパク質ＡとＢの３次元モデル情報から，それぞれ多視点画像セットを生成し，２つの分布の類似度に基づいてＡとＢ立体構造類似度を測る(図１)．ここで各画像セットの分布は対応するタンパク質の立体構造を反映しているので，両分布の形状類似度を立体構造類似度と見なせる．さらに各画像セットの分布を主成分分析により部分空間Ｐと Q で表すと，構造類似度は２つの部分空間Ｐと Q の成す正準角 θ で効率的に測ることができる．この正準角は解析的に求まるので，構造類似度は一意に安定して求まる．

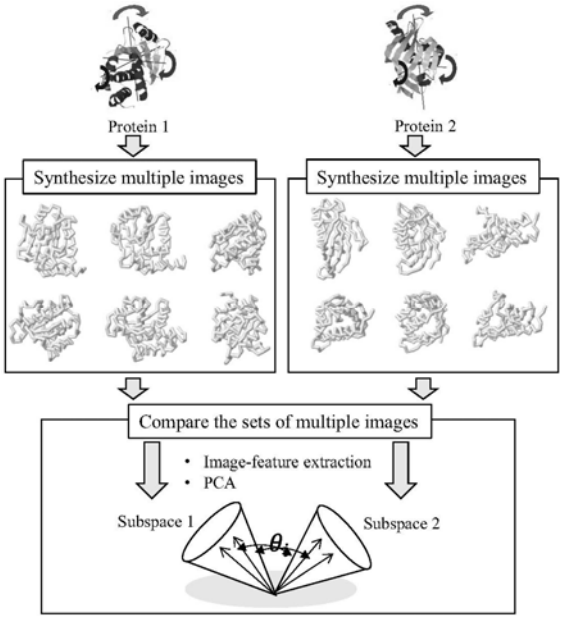

図 1

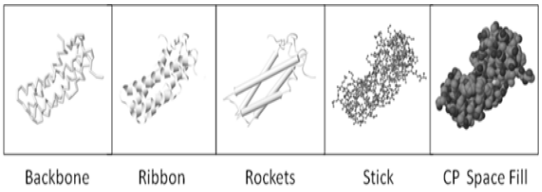

図 2

更に本アイデアに基づいて，構造情報だけはなく，表面特性：アミノ酸配列情報，２次構造，親水性，疎水性，電荷などを反映した記述能力の高い構造類似度を定義する．これは，反映させたい表面特性を，カラー情報，粒子サイズ，局所構造の違いとして埋め込んだマルチモーダル多視点画像を生成し，それを用いて構造類似度を求めることで実現する．代表的なビューアーである Webmol などで使われている表示方式：図２に示したバックボーン，リボン，ロケット，スティック，空間充填はいずれも利用能であるが，２次構造(α，βシートなど)情報まで含んでいるリボン，ロケットや，構成分子の種類で色分けした空間充填を用いると，骨格の構造情報だけではなく，局所構造情報まで反映したマルチモーダル構造類似度を定義する．このようにして記述能力の高い構造類似度を用いることで，タンパク質の識別・分類の精度を大きく改善できる．

本方法の主な利点を再度，整理すると以下のとなる．(a)従来法では必要であったアライメント(位置合わせ)が不要となり，処理がシンプルになり安定性が増す．(b)多視点画像セットの部分空間表現に基づいているので，部分空間法に適用できる様々な特徴抽出法

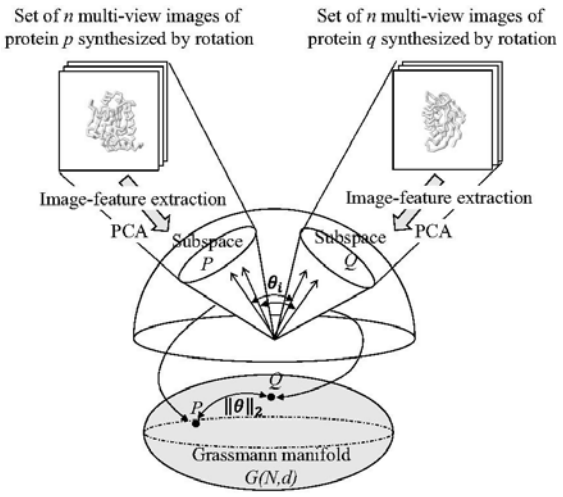

図 3

が適可能であり，是により識別性能の向上が期待できる．

(2)マルチモーダル構造類似度に基づく識別

識別処理全体の流れは以下のようになる．

学習フェーズ：識別すべきタンパク質クラス毎に上述した方法により，それぞれに対応する部分空間を生成する．これが識別の際の辞書部分空間となる．

識別フェーズ：未知タンパク質に対して辞書クラスと同様に３次元モデルから生成した多視点画像セットに対して主成分分析を適用して入力部分空間を生成する．次に，入力部分空間と学習フェーズで求めた各タンパク質クラス部分空間との正準角（マルチモーダル構造類似度）を求め，最も高く，かつ予め設定したしきい値以上の構造類似度を有するクラスに入力タンパク質を分類する．

(3)グラスマン多様体を用いた性能強化

先に述べた２つの部分空間の成す正準角に基づく識別は，グラスマン多様体上における最小距離に基づく最も単純な識別と見なすことができる(図 3)．この解釈に基づいて，より判別能力の高い判別分析(GDA)をグラスマン多様体上で適用する．これにより，単純な正準角を用いた方法(相互部分空間法(MSM))に比べて，識別性能の大幅な向上を図る．

(4)異なるタイプの類似度の統合:

機械学習の観点から，提案法と様々な従来法との組合せにより高い識別性能が期待できる．そこで，複数の異なる方法により得られた類似度（距離）を large margin nearest neighbor (LMNN)を適用することで統計的に統合することで，新たな識別に有効な類似度を求めるフレームワークを考案した．

(5)有効性の検証

マルチモーダル構造類似度の有効性を検証するために，以下の評価実験を実施した．Astral データベースから，アミノ酸列一致が20%以下の 700 個のタンパク質をランダム選択して評価データセットを構築した．このデータセットを用いて，２次構造の違いによる図４に示す７クラス（α型，β型，α/β型(両者が構造を持って混在)，α+β型(両者がランダムに混在)等）のクラス識別実験を行った．

４．研究成果

(1)マルチモーダル構造類似度の確立:

タンパク質３次元モデルから生成した多視点画像セットに対して，相互部分空間(MSM)あるいはグラスマン判別分析（GDA）を適用す

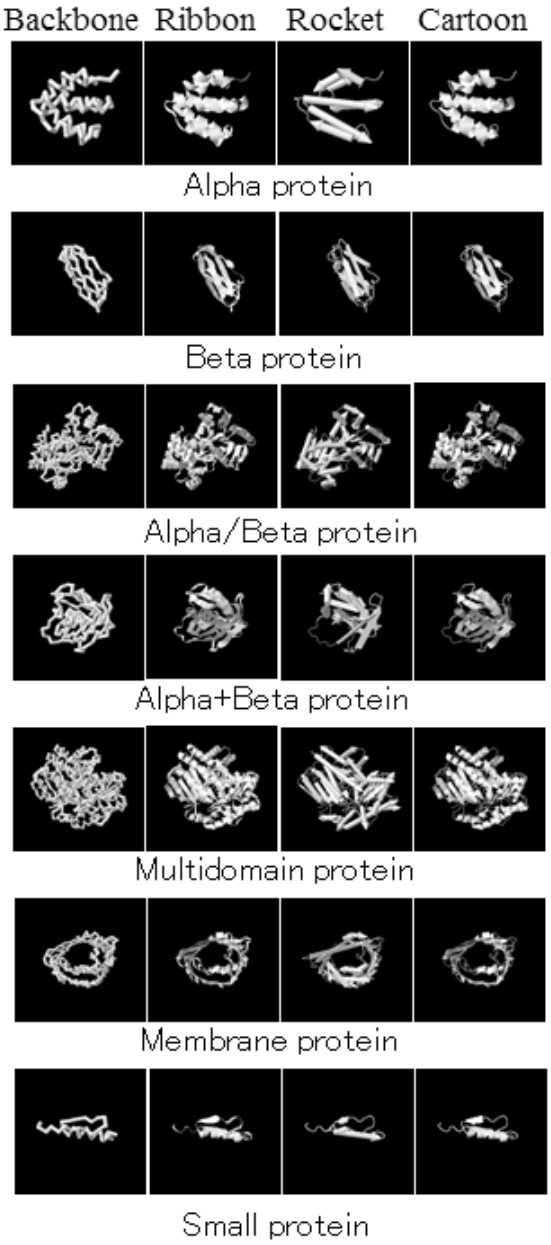

図 4

ることで，３次元形状の情報だけではなく，電荷や親水性などの情報も考慮できるマルチモーダル構造類似度を提案した．

(2) 構造類似度の有効性を検証：

まず基本的な識別性能を検証するために，提案法とタンパク質の構造解析において標準的に使われているアライメントベース法（CE+RMSD, CE+Z-score, FATCAT, TM-align+TM-score)との性能比較を行った．その結果，MSM に基づく方式： 60.86%, CEwith RMSD: 33.86%, CE with Z-score: 49.14%, FATCAT with raw score: 53.14%,TM-align with TM-score:64%に対して，提案法の識別性能は69.43%となり，その有効性を確認した．この際，様々な特徴抽出が適用可能な中，回転不変 LBP 特徴と GDA の組合せの有効であることを明らかにした．

次に，複数タイプの類似度を統合するフレームワークの有効性を 27 種類のタンパク質立体構造の識別実験で検証した．その結果，得られた 18 種類の類似度を統合して得られた類似度を用いることで，より高い識別性能が得られることを確認した．

(3)構造類似度計算サーバーの構築：

マルチモーダル構造類似度に基づく提案フレームワークを実装した計算機サーバー(View-based Protein Comparison (VPC) system)を構築した．図 5 にインタフェース画面を示す．図 6 にシステム構成の概略を示す．本サーバーは，ユーザがネット経由で比較したい２つの３次元タンパク質を指定すると，両者の３次元構造類似度を上記アルゴリズムで計算し，ユーザにフィードバックする機能を有する．

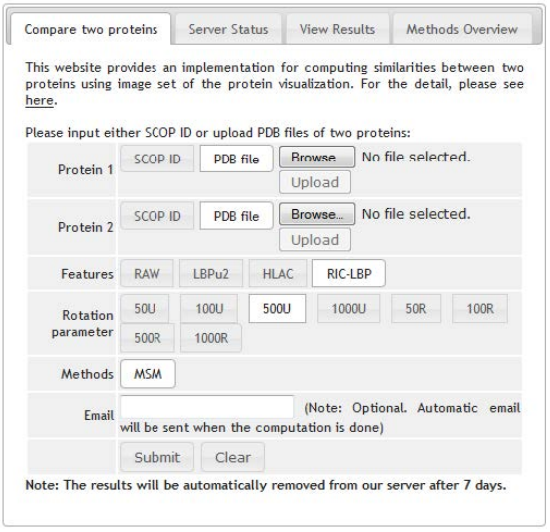

図 5

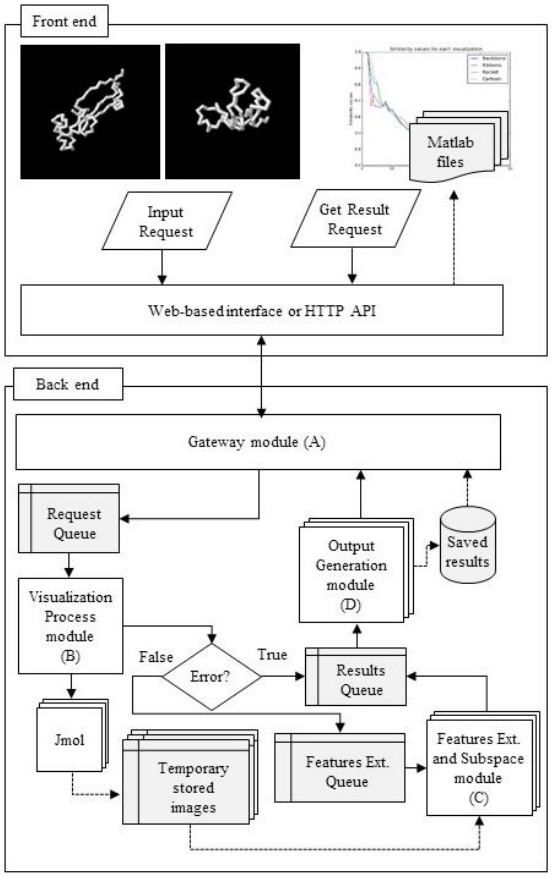

図 6

５．主な発表論文等

〔雑誌論文〕（計 1 件）

(1) Chendra Hadi Suryanto, Hideitsu Hino, Kazuhiro Fukui, ″Combination of Multiple Distance Measures for Protein Fold Classification″, ACPR2013, pp. 440-445, 2013.(査読有り)
DOI:10.1109/ACPR.2013.139

〔学会発表〕（計 1 件）

(1) Chendra Hadi Suryanto, Hideitsu Hino, Kazuhiro Fukui, ″Combination of Multiple Distance Measures for Protein Fold Classification″, ACPR2013. 2013 年 11 月 7 日, ロワジールホテル国際会議場(沖縄県)

〔その他〕

ホームページ等

View-based Protein comparison (VPC) system

http://www.cvlab.cs.tsukuba.ac.jp/~chendra/protein/

本研究課題で開発したマルチモーダル構造類似度アルゴリズムを実装したネットワーク対応のサーバーの URL である．

上記成果に加えて，研究成果をまとめた研究論文（２本）を英文ジャーナルへ投稿し，現在，査読審査中である．

６．研究組織

(1)研究代表者

福井 和広（Fukui, Kazuhiro）
筑波大学・システム情報系・教授
研究者番号：４０３７５４２３

(2)研究分担者

西郷 浩人（Saigo, Hiroto）
九州工業大学・情報工学研究院・准教授
研究者番号：９０５８６１２４