生物物理 **65**（4），201-205（2025）
DOI: 10.2142/biophys.65.201
受理日：2025 年 2 月 3 日

総説

3 次元 Zernike 記述子を用いたタンパク質の構造分類

加賀谷祐輝[1]，木原大亮[1,2]

[1] Department of Biological Sciences, Purdue University, IN, USA

[2] Department of Computer Science, Purdue University, IN, USA

Proteins perform various functions in living organisms, and understanding their 3D structures is essential to understanding their functions. The 3D Zernike descriptor (3DZD) provides a compact, rotationally and translationally invariant description of protein surface shapes. These features are desirable when applying 3DZD to describe the shape of proteins. Therefore, 3DZD is widely used for comparing protein structures and for shape-based searches. In this review, we describe the properties of 3DZD and their various applications in protein structure analysis.

protein structure comparison / 3D Zernike Descriptor / 3DZD

1. はじめに

タンパク質は生命の最も基本的な構成要素であり，化学反応を触媒したり，細胞の形態を維持したりするなど，その機能は重要かつ多岐に渡る．タンパク質の機能はその立体構造に大きく依存しているため，タンパク質の形状についての理解を深めることは，その機能を知るために必要不可欠である．特に，タンパク質の機能の多くはその表面で発現するため，表面の形状に注目してタンパク質の構造を特徴づけることは意義がある．このような構造的特徴に基づいて体系的にタンパク質を整理することによって，共通点や相違点を明確にすることは，機能についての理解を深めるために重要である．

本総説では，タンパク質の立体構造を表す表現の一つとして，3D Zernike（ゼルニケ）記述子（3DZD）について紹介する．また，3DZD を使ったタンパク質の構造分類について紹介し，さらに 3DZD を応用したアプリケーションを紹介する．

2. 3DZD の性質

3DZD は，物体の 3 次元表面形状を表現するために使用される回転不変のモーメントベース記述子である[1)]．3DZD は球面調和関数を拡張したものであり，動径関数を組み込むことで，球面調和関数よりも多様な 3 次元形状を正確に表現することができる．直感的には 3DZD は，信号処理において信号を周波数成分の線型結合に分解するフーリエ変換と似たものであると捉えることもできる．

3DZD は，次のようにして求められる．はじめに，3D ゼルニケモーメントを求める．表面形状が極座標表現（r, θ, φ）を用いて表せる時，3D ゼルニケ多項式 Z_{nl}^{m} は，次のように与えられる．

$$Z_{nl}^{m}(r, \theta, \varphi) = R_{nl}(r) Y_{l}^{m}(\theta, \varphi)$$

ここで，$Y_{l}^{m}(\theta, \varphi)$ は複素球面調和関数である．n は正の整数であり，3D 多項式の次数（order）である．また，l と m はそれぞれ位数（degree）と次数（order）を表す整数であり，$-l \le m \le l$，$0 \le l \le n$，かつ（$n-l$）が偶数であるという条件を満たす．$R_{nl}(r)$は Canterakis によって定義された動径関数であり，半径情報を直接基底関数に組み込む[2)]．これにより，球面調和関数が単位球や星形形状しか表現できないという制約を解消することができる．

関数 $f(x)$によってモデル化された 3D オブジェクトの 3D ゼルニケモーメントは，前の 3D ゼルニケ多項式で展開された時の係数として次のように定義される．

$$\Omega_{nl}^{m} = \frac{3}{4\pi} \int_{|x| \le 1} f(\boldsymbol{x}) \overline{Z_{nl}^{m}(\boldsymbol{x})}\, d\boldsymbol{x}$$

3DZD は，このモーメントのノルム $||\Omega_{nl}^{m}||$ を計算することによって得られる．得られる 3DZD は，n が偶数

Applications of 3D Zernike Descriptors in Protein Structure Comparison

Yuki KAGAYA[1] and Daisuke KIHARA[1,2]

[1] Department of Biological Sciences, Purdue University, USA

[2] Department of Computer Science, Purdue University, USA

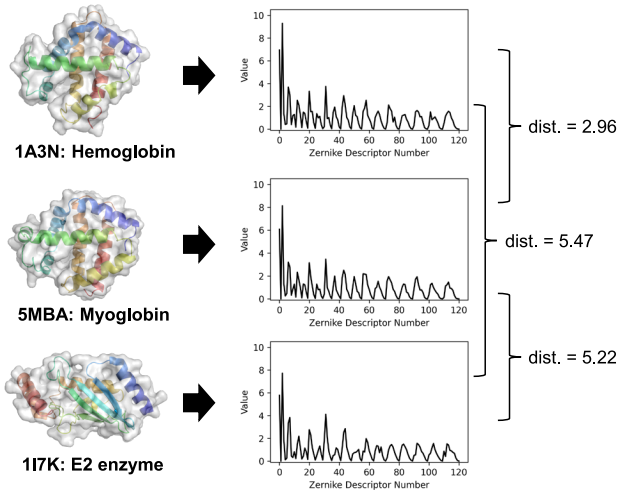


図 1
タンパク質の 3DZD の比較．タンパク質の立体構造（虹色）から計算された表面形状（灰色）は，3DZD に変換できる．3DZD のベクトルはユークリッド距離を用いて比較することができる（図中の dist.）．似たタンパク質（ここでは Hemoglobin と Myoglobin）から得られた 3DZD 間の距離は，異なるタンパク質と比較して近くなる．

ならば$(n/2+1)^2$，奇数ならば$(n+1)(n+3)/4$ の次元の実数ベクトルとして得られる．n の値を調節することで，どの程度の細かい特徴まで扱うかを変化させることができ，n の値が大きければ大きいほど，表面の精密な構造をエンコードできる．タンパク質の 3DZD を求める際には，$n=20$ がよく用いられるが，これによって得られる 3DZD の次元数は 121 となる．

タンパク質はたくさんのアミノ酸が一本に繋がった分子であり，その立体構造は構成する原子のユークリッド座標の集合として表現される．このため，タンパク質の 3DZD を計算するためには，分子表面をまずボクセル化（格子状に分割）して，各ボクセルは表面か否かのバイナリとして表される．このように離散化された表面を用いることで，前の節で紹介した式に従って 3DZD を計算することができる．**図 1** には，いくつかのタンパク質を使ってこの手順を示した．異なる形状のタンパク質からは異なる 3DZD が得られ，3DZD はユークリッド距離や相関係数などを利用して比較することができる．

3. タンパク質の立体構造比較と 3DZD

3DZD は，タンパク質の立体構造を表現する際，特に表面形状を球面調和関数で展開する場合と比較して，優れた特徴を持つ．最大の利点は，球や星形形状ではない複雑な表面形状を扱うことができることである．タンパク質の表面は複雑に入り組んだ形状をしている傾向にあるが，3DZD はこのような形状を適切に表現することができる．また，3DZD は回転や並進に対して不変であるという特徴も持つ．つまり，同じタンパク質を空間内でどのように回転させても，同一の 3DZD 表現が得られる．球面調和関数では回転によって異なる表現が得られるため，回転不変性を持たせるためには，対象となる形状の姿勢を事前に正規化する必要がある．しかし，多くのタンパク質は特定の方向性を持たないか，形状の方向的特徴が少ないため，姿勢の正規化はしばしば困難である．一方で，球面調和関数の係数の大きさだけを使えば回転不変な特徴量が得られるが，元の形状を再現するための情報が一部失われるため，元の構造の違いを十分に反映することができない．この他にも，3DZD は球面調和関数と比べて，よりコンパクトな表現を可能にし，同じ次数で展開した場合にベクトルの長さを短くできる．これにより，事前に計算した 3DZD を効率的に保存することができる．また，3DZD の比較ではユークリッド距離や相関係数を利用できるため，立体形状を直接比較する場合に比べ，計算コストを大幅に削減できる．

タンパク質の立体構造を比較する手法には，3DZD 以外にも様々な方法が存在する．一般的に，二つの構造をアラインメントして，二乗平均平方根誤差（RMSD）や TM-score[3]を計算する方法がよく用いられる．アラインメントフリーの手法としては，DaliLite[4] が残基間の距離パターンを比較し，二つのタンパク質構造を評価できる．これらの手法と比較しても，3DZD による比較は非常に高速であり，大量の構造データの比較に適している．特に近年，実験構造データの増加に加え，AlphaFold などの機械学習手法から得られる大量の構造データを効率的に比較する必要性が高まっているため，3DZD はさらに注目を集めている．

4. 3DZD によるタンパク質の 3 次元形状のマッピング

タンパク質の立体構造は実験的に解き明かされ，その結果得られた構造は Protein Data Bank（PDB）データベースに登録・公開されている[5]．現在，20 万件を超える大量の構造が利用可能であるが，このような多くのデータから得られる知識を俯瞰し体系的に理解するためには，それらを適切に分類し，その類似点と相違点を浮き彫りにすることが必要である．

このため，我々が発表した Han らの論文[6]では，3DZD を利用して立体構造が既知のタンパク質の単鎖および複合体の形状のマッピングを行った．この研究では，まず PDB から解像度などの条件を満たす構造の

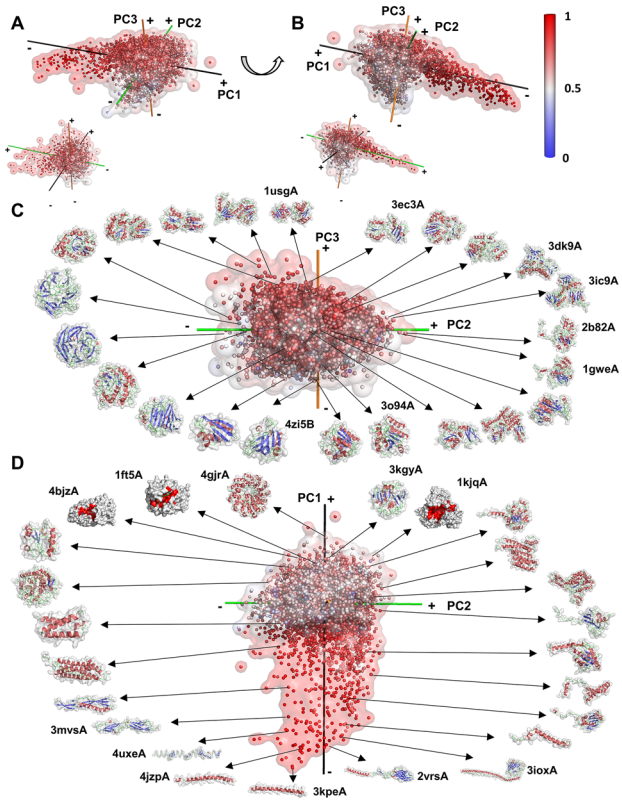


図 2
単鎖タンパク質の 3D 形状空間．各点はタンパク質を表す．色はタンパク質の離心率を表し，0（青色）に近いほど球形に近い形状を持つ．（A, B）単鎖タンパク質の 3D 形状空間．A と B は異なる角度から見た同じ図である．（C, D）は，この空間上でのタンパク質形状の具体例を示している．（図は文献 6の Fig. 1 より引用．Creative Commons License によってライセンスされた．）

うち，各チェーン単位でアミノ酸配列同士のペアごとの配列類似性が 25%以下になるように 6,841 個のタンパク質を収集した．これらには，実際には複合体を形成するタンパク質も含まれる．また，実際に機能する単位でもタンパク質を分類するために，生物学的単位が複合体を構成している 5,326 個のタンパク質複合体を，同様に互いに 25%以下の配列類似性になるように選択した．次に，これらのタンパク質それぞれについてその表面形状の 3DZD を計算し，各タンパク質について 121 次元の 3DZD を得た．これらの 3DZD を用いて主成分分析（PCA）を行い，最初の三つの主成分を使って各タンパク質を 3 次元空間に射影した．

得られた単鎖タンパク質形状の空間マッピングを**図 2** に示す．単鎖タンパク質形状では，離心率が小さい球状のタンパク質（**図 2** では青色の点）はほとんど存在しなかった．タンパク質の表面構造を特徴づける要因を調べるため，PCA のそれぞれの軸がどのような要因と関連しているかを調査した．第一主成分である PC1 の軸に沿って，青から赤までの異なる離心率のタンパク質が分布していることが観察できることから，タンパク質形状の離心率が単鎖タンパク質を形状的に特徴づける主な要因であることがわかる．また，PC2 と PC3 についても調査した結果，これらはタンパク質が持つドメインの数やタンパク質鎖の長さと相関していることが明らかとなった．

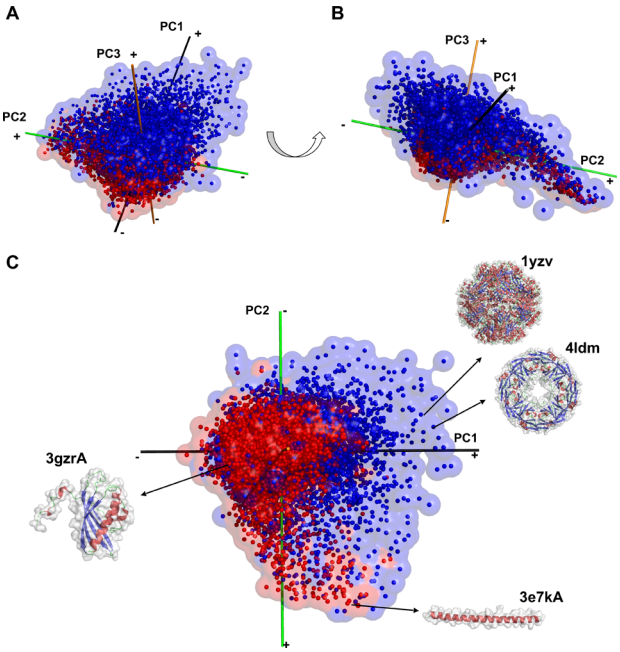


図 3
単鎖タンパク質とタンパク質複合体の形状空間の重ね合わせ．赤は単鎖タンパク質，青はタンパク質複合体の構造を表す．（A）から（C）はそれぞれ別の角度からこの形状空間を表示している．（C）では，単鎖と複合体タンパク質の具体例を示す．（図は 文献 6の Fig. 7 より引用．Creative Commons License によってライセンスされた．）

次に，タンパク質複合体の形状空間マッピングを単鎖タンパク質の場合と同じように実行した．また，この結果のマッピングを単鎖タンパク質のマッピングと重ね合わせて比較することによって，その違いについて考察した．**図 3** の重ね合わせた分布を観察すると，タンパク質複合体（青）は，単鎖タンパク質（赤）より広範な空間に分布していることが観察できる．これは，タンパク質複合体が単鎖タンパク質よりも多様な形状をとることができることを示している．タンパク質形状の多様性はその機能の多様性に関連するため，タンパク質が複合体を形成することで実現可能な構造と機能の範囲が広がることを示唆している．

5. 3DZD を用いた全 PDB の高速検索

我々は，3DZD の類似性に基づいてタンパク質の立体構造を比較し，高速なリアルタイム検索ができる

3D-Surfer を開発した[7]．PDB ID か PDB ファイルを入力として，PDB で公開されている立体構造全体か，AlphaFold[8]によって予測された立体構造のデータベースである AlphaFold Database[9]の一部である約 100 万の立体構造から検索できる．3D-Surfer は，タンパク質のフォールド分類ができるニューラルネットワークを用いることで，3DZD の単純な比較より高精度なデータベース検索を実現している[10]．検索は，3DZD の単純比較を用いると数秒，機械学習を利用した検索は 1 分程度であり，立体構造類似性検索ができるソフトウェアでは最速である．3D-Surfer による検索は，我々の“Web サーバー”から利用できる．

同様の検索は，電子顕微鏡の 3 次元マップを検索するためにも利用することができる．私たちの開発した EM-Surfer[11]は，電顕のマップのデータベースである EMDB[12]の ID または 3 次元マップを入力として，EMDB のマップを 3DZD の類似性を使って高速に検索することができる．

6. 3DZD による相補的な表面構造・部分構造の検索

タンパク質の相互作用部位などの相補的な分子表面は裏表を考えなければ類似した表面の形状を持つため，3DZD を用いて相補性を定量的に比べることができる．このことを用い，我々のタンパク質のドッキングのソフトウェアである LZerD[13]では，ドッキングするタンパク質の表面を直径 6 Å の球によって切り出したパッチで表現し，二つのタンパク質から相補的なパッチの組を探すことでドッキング構造の構築を行う（**図 4A**）．3DZD が回転不変であるという性質から，実際に重ね合わせて相補性を調べる必要がなく，高速かつ網羅的に表面形状を比較できる．LZerD は，ドッキング予測

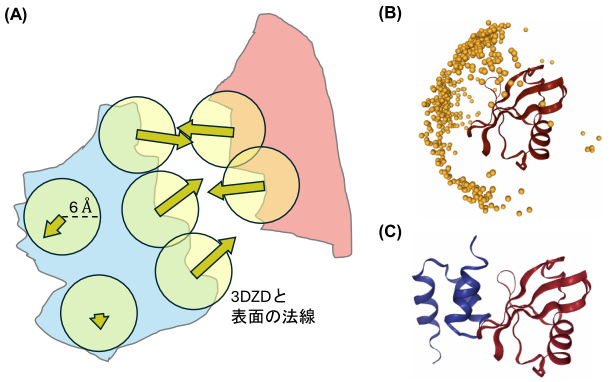


図 4
LZerD によるタンパク質ドッキング．（A）LZerD によるパッチの比較の概念図．（B, C）ドッキングの例．（B）黄色の点は上位 500 構造の重心を示す．（C）最上位のドッキング構造．

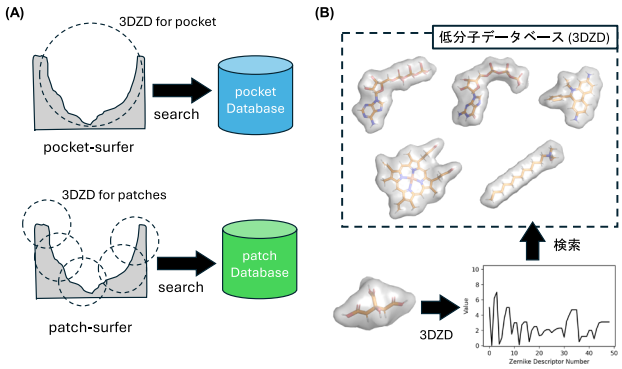


図 5
低分子結合ポケットの検索と低分子リガンド検索．（A）Pocket-Surfer と Patch-Surfer の概念図．ポケットを 3DZD として記述する方法が異なる．（B）3DZD による形状の似た低分子の検索の例．

手法を比較するコンテストの CAPRI で，非常に優れた性能を示している[14]．LZerD は，Web サーバーとしても提供されていて，誰でも利用することができる[15]．**図 4B, C** では，Web サーバーでの LZerD ドッキングの例を示した．

Pocket-Surfer[16]と Patch-Surfer[17]は，結合ポケットを既知の結合ポケットと 3DZD を用いて比較することで，そのポケットに合う低分子を探索することができる．二つの手法は形状を 3DZD として表現して比較を行うが，Pocket-Surfer では結合ポケットの全体形状を扱うのに対して，Patch-Surfer では結合ポケットを小さなパッチの集合として扱う（**図 5A**）．パッチ表現を使うと，ポケットの全体形状が異なっていても，ポケット内の対応する領域を識別できる．Patch-Surfer は，PDB から選択されたポケットと低分子の非冗長なデータセットを用いたベンチマークでは，他の手法と比較して高い精度を示した．

PL-PatchSurfer[18),19)]では，ポケットとリガンドの両方の局所表面パッチの相補性を使って，ポケットに結合するリガンドを検索し順位付けする．表面表現は立体配座のわずかな違いに対する感度が低いため，3DZD と組み合わせて高速に検索することができる．

低分子リガンド同士の形状比較や検索も，3DZD を用いることで高速かつ精度よく行うことができる[20]（**図 5B**）．このようなタンパク質の結合ポケットやそこに結合する低分子の探索は，創薬におけるバーチャルスクリーニングのために利用することができる．

7. おわりに

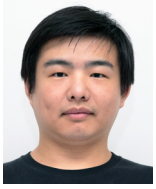
本総説では，まずタンパク質の立体構造を 3DZD で表現する方法と，3DZD を用いた比較からタンパク質

の立体構造を空間マッピングし分類した研究について紹介した．また，3DZD を使った応用として，タンパク質の立体構造や電子密度マップを高速に比較し検索するアプリケーションを紹介した．これらのアプリケーションは，ウェブサーバーや実行可能パッケージとして利用可能であり，我々の研究室 Web ページの“ソフトウェアリスト”に一覧されている．

3DZD は，タンパク質の立体構造などの表面形状の特徴を効率的な実数値ベクトルで表すための優れた手法の一つである．3DZD は回転不変であるという性質があるため，二つの表面を正しい方向に回転させたり，実際に位置合わせをしたりすることなく，表面形状を比較してその類似性や相補性を見出すことができる．実験的に解決された立体構造の蓄積に加えて，AlphaFold をはじめとした最先端の予測手法が与える大量の立体構造が持つ可能性を最大限に活用するためには，これらの構造の全体ないしは部分を高速かつ正確に比較・分類するための手法が必要不可欠である．3DZD はそのための手法の一つとして，今後さらに活用されることが期待される．

文　献

1) Novotni, M., Klein, R. (2003) In Proceedings of the eighth ACM symposium on Solid modeling and application. 216-225. DOI: 10.1145/781606.781639.
2) Canterakis, N. (1999) In 11th Scandinavian Conf. on Image Analysis. 85-93.
3) Zhang, Y., Skolnick, J. (2004) Proteins **57**, 702-710. DOI: 10.1002/prot.20264.
4) Holm, L. (2019) Bioinformatics **35**, 5326-5327. DOI: 10.1093/bioinformatics/btz536.
5) wwPDB consortium (2019) Nucleic Acids Res. **47**, D520-D528. DOI: 10.1093/nar/gky949.
6) Han, X. *et al.* (2019) PLOS Comput. Biol. **15**, e1006969. DOI: 10.1371/journal.pcbi.1006969.
7) La, D. *et al.* (2009) Bioinformatics **25**, 2843-2844. DOI: 10.1093/bioinformatics/btp542.
8) Jumper, J. *et al.* (2021) Nature **596**, 583-589. DOI: 10.1038/s41586-021-03819-2.
9) Varadi, M. *et al.* (2022) Nucleic Acids Res. **50**, D439-D444. DOI: 10.1093/nar/gkab1061.
10) Aderinwale, T. *et al.* (2022) Commun. Biol. **5**, 316. DOI: 10.1038/s42003-022-03261-8.
11) Sael, L., Kihara, D. (2010) BMC Bioinformatics **11**, S2. DOI: 10.1186/1471-2105-11-S11-S2.
12) The wwPDB Consortium (2024) Nucleic Acids Res. **52**, D456-D465. DOI: 10.1093/nar/gkad1019.
13) Venkatraman, V. *et al.* (2009) BMC Bioinformatics **10**, 407. DOI: 10.1186/1471-2105-10-407.
14) Christoffer, C. *et al.* (2020) Proteins **88**, 948-961. DOI: 10.1002/prot.25850.
15) Christoffer, C. *et al.* (2021) Nucleic Acids Res. **49**, W359-W365. DOI: 10.1093/nar/gkab336.
16) Chikhi, R. *et al.* (2010) Proteins **78**, 2007-2028. DOI: 10.1002/prot.22715.
17) Zhu, X. *et al.* (2015) Bioinformatics **31**, 707-713. DOI: 10.1093/bioinformatics/btu724.
18) Hu, B. *et al.* (2014) Int. J. Mol. Sci. **15**, 15122-15145. DOI: 10.3390/ijms150915122.
19) Shin, W. H. *et al.* (2016) J. Chem. Inf. Model. **56**, 1676-1691. DOI: 10.1021/acs.jcim.6b00163.
20) Venkatraman, V. *et al.* (2009) J. Cheminform. **1**, 19. DOI: 10.1186/1758-2946-1-19.



加賀谷祐輝

加賀谷祐輝（かがや　ゆうき）
Purdue University 博士研究員
2021 年東北大学大学院情報科学研究科博士後期課程修了，博士（情報科学），同年から現職．
研究内容：バイオインフォマティクス
連絡先：Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA
E-mail: ykagaya@purdue.edu
URL: https://kiharalab.org/

木原大亮

木原大亮（きはら　だいすけ）
Purdue University 教授
1994 年東京大学教養学部基礎科学科第一卒業（卒業研究は深田吉孝研），1996 年，1999 年に京都大学大学院理学研究科生物物理学科，金久實研で修士，博士を取得．同年から 2003 年まで Jeffrey Skolnick 教授の下でポスドク，2003 年 8 月に Purdue 大学生物科学科，計算科学科（併任）アシスタントプロフェッサー，2009 年に准教授（テニュア）昇進，2014 年から現職．
研究内容：タンパク質のバイオインフォマティクス．構造予測，機能予測の手法などを開発．最近では電子顕微鏡の像からのモデリング手法の開発に従事
連絡先：Department of Biological Sciences, Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
E-mail: dkihara@purdue.edu
URL：同上