I-025

ドロネー四面体分割を用いたタンパク質構造比較
Applying Delauny Tessellation to Protein Structure Comparison

寺本　やえみ†　甲藤　二郎†　輪湖　博‡
Yaemi Teramoto　Jiro Katto　Hiroshi Wako

１．はじめに

生命活動を維持するタンパク質の機能はその構造と深い関連性を持っており、タンパク質構造に関する知見を得、アミノ酸配列から構造を予測決定することが、バイオインフォマティクスの究極の目的の一つとされている。

実験的に解明されたタンパク質立体構造を格納するProtein Data Bank(PDB)データベースのエントリー数は現在２万５千を超えており、それら大量の構造データをもとにしたタンパク質構造理解が重要な課題となっている。

本稿では、タンパク質構造をコード化して表し、コードを用いてタンパク質構造の特徴を捉えた結果を示す。

２．タンパク質構造のコード化

２．１　ドロネー四面体分割

空間に存在する点群において、各点同士の隣接を一意に決定する方法に Voronoi 図がある。Voronoi 図では、N 点からなる点群において、ある点 a に対し、点群内の他の全ての点との距離の垂直二等分線（N-1 本）を引き、それによって形成される点 a を含む小空間のうち最も体積の小さいものを点 a に属する空間と定める。この Voronoi 図において、属する小空間の境界を共有する点同士を「空間上隣接する」と考える。３次元空間においては Voronoi 図で隣接と判断された点同士を結ぶことによって、四面体が形成される。これをドロネー四面体と呼ぶ。

ドロネー四面体分割をタンパク質構造理解に適用する大きな利点として、原子同士の空間上の隣接を一意に定めることが可能であることと、タンパク質が隙間なく隣接した四面体によって表されることが挙げられる。

２．２　タンパク質構造のドロネーコード

タンパク質の立体構造をアミノ酸配列の主鎖を形成するＣα原子の座標で代表し、その座標群をドロネー四面体分割する。これによって、タンパク質は隙間なく隣接した四面体のネットワークとして表現される[1]。本研究では、Qhull プログラム（http://www.qhull.org/）を用いてドロネー分割を行った。

各四面体に付与するコードは、それに隣接する４つの四面体を含む最大８つの頂点に存在するアミノ酸の、配列上の並び順と、配列上近接しているか・離れているか、の情報を含む、A-Hの英大小文字８字で表される。A-Hで表される８頂点の位置関係を図１に示す。コードの付与される四面体の頂点はA-Dの英字で表され、周囲に隣接する四面体の頂点はE-Hの英字で表される。コード付けは以下のルールによる。８頂点を配列上連続ブロックの長い順に並べ、中心の四面体に存在する頂点をコードの左側から順番にA、B、C、Dと定める。これによって、E-Hは必然的に決定する。各ブロックの境目は大文字小文字の切り替わりで表す。例えば、「ABHCdgfE」というコードでは、ＡＢＨＣ、ｄｇｆ、の２つのセグメントでアミノ酸が配列上並んで近接しており、Ｅは単独で離れていることが示される。

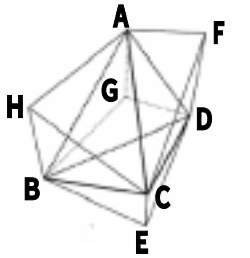

図１　頂点A-Hの位置関係

３．タンパク質構造分類データベース

SCOP（http://scop.mrc-lmb.cam.ac.uk/scop/）[2]は、ＰＤＢのタンパク質立体構造データを、構造やそれに基づく生化学的な機能の特徴、タンパク質同士の進化的関連性によって階層的に分類したデータベースである。

タンパク質構造は、αhelix と呼ばれるらせん構造と、βsheet と呼ばれるシート状の構造の２種の二次構造の空間配置で特徴づけられる。ＳＣＯＰでは、タンパク質構造を、Class（構成する二次構造の種類）、Fold（二次構造の空間的配置）、Superfamily（進化的関連）、Family（機能的関連）、の順に階層的に分類する。例えば、Class は、allα、allβ、α/β（αhelix とβsheet が交互に出現）、α+β（αhelix 領域とβsheet 領域が分離）そしてその他の、大きく５つに分類される。

４．コードによる構造特徴の表現

allα、allβ、α/β、α+βの各クラスに属するタンパク質群でコードの頻度の平均値をとり、クラスごとにコード頻度プロファイルを作成した。

４．１　データセット

プロファイル作成に用いたデータセットは以下の条件である。

①配列類似度２５%以下
(PDBselect: http://www.cmbi.kun.nl/gv/pdbsel/)
②Ｘ線構造解析によるもの
③解像度 3.0Å未満
④残基数 100 以上 300 以下

各 Class のプロファイル作成用データを表１に示す。

表１　プロファイル作成用データ

	データ数	平均残基数	平均コード付け四面体数	平均コード種類数
Class α	95	153.5	579.5	324.6
Class β	130	158	504.8	381.1
Class α/β	137	208.2	764.2	510.5
Class α+β	139	165.1	567.9	403.3

†早稲田大学大学院　理工学研究科
‡早稲田大学　社会科学部

４．２　クラスプロファイル

表１のデータから作成した各 Class のコードプロファイルを表２．１、表２．２、表２．３、表２．４に示す。平均頻度 0.5 未満のコードはプロファイルから除いた。

表２．１　プロファイル Class all α

コード	出現頻度
FHABCDEG	14.102487
ABEHCDfG	3.601822
ABEHCDfg	3.346075
ABEHCDgF	2.507966
ABEHCDFg	1.26454
GABEHCDf	1.125226
ABEHCDgf	1.046145
FGABEHCD	0.738101
AHBCEDgF	0.545418
AHBCegdF	0.5395
GABEHCDF	0.52109
ABEHCDFG	0.517358
FHABCDEg	0.511824

表２．２　プロファイル Class all β

コード	出現頻度
FHABCDEG	1.99113
FGABehcd	1.614553
HFABGEcd	1.186919
ABEHcdfg	1.172439
FHABEGcd	1.147016
AFHbcdEG	0.770966
ABCdegFH	0.738131
FHAbcdEG	0.685636
GABCedfH	0.622878
FHAbcdGe	0.576604
ABHEcdgF	0.521156
FGABEHCD	0.506581

表２．３　プロファイル Class α/β

コード	出現頻度
FHABCDEG	8.232273
ABEHCDfG	1.927617
ABEHCDfg	1.862465
ABEHCDgF	1.706255
FHABcdeg	0.760897
ABEHCDFg	0.745422
ABEGfhcd	0.736343
GFABHEcd	0.712877
FGABEHcd	0.703831
ABEHCDgf	0.700623
GABEHCDf	0.659473
FGABEHCD	0.54962

表２．４　プロファイル Class α+β

コード	出現頻度
FHABCDEG	7.012942
ABEHCDfg	1.564431
ABEHCDfG	1.542994
ABEHCDgF	1.466282
FGABehcd	0.893674
ABEHcdfg	0.67068
FGABEHCD	0.647966
ABEHCDFg	0.625361
FHABEGcd	0.605392
GABEHCDf	0.579679
HFABGEcd	0.577815
ABEHCDgf	0.565334

５．実験

作成したプロファイルを用いて、実験データタンパク質がどのクラスに属するかの判定実験を行った。判定に用いるスコアを１式に示す。

$$Score = \frac{\sum_{N_p} \left| \frac{f_i - f_p}{f_p} \right|}{N_p} \qquad (1)$$

１式において、N_p はプロファイルのコード数、f_p はプロファイルにおける各コードの頻度、f_i は実験データにおける各コードの頻度である。実験データに対し、１式のスコアを各クラスのプロファイルを用いて計算し、最もスコアの小さいクラスにそのタンパク質が属すると判断する。

５．２　実験データセット

実験に用いたデータは以下の条件である。

①配列類似度９０％以下（PDBselect）
②プロファイル用データを含まない
③プロファイル用データ②～④の条件を満たす

各 Class の実験用データを表３に示す。

表３　実験データ

実験データ	データ数	平均残基数	平均コード付け四面体数	平均コード種類数
Class all α	46	160.2	604.9	338.4
Class all β	48	165	530.6	393.4
Class α / β	49	195.3	711.9	475.4
Class α+β	49	152.6	514.4	364.6

５．３　実験結果

表１に実験結果（クラスに属することの正判定率（true positive）・クラスに属さないことの正判定率（true negative））を示す。

表４　実験結果

	true positive (%)	true negative (%)
Class all α	93.5	88.4
Class all β	70.8	95.9
Class α / β	73.5	92.4
Class α+β	61.2	89.6

表２のプロファイルから、FHABCDEG は、α helix を特徴的に表すコードであること、α helix に現れるコードはブロックがひとつながりになっている確率が高いこと、β sheet に表れるコードはブロック数が比較的多い傾向があること、などがわかる。プロファイルでは、Class all β のコード組成は他クラスとの特徴の違いが明らかだが、Class all β の判定率はあまり高くない。この原因として、β sheet は比較的多様なコードによって表され、特徴的なコードが現れにくいことが考えられる。

クラス判定方法にはさらなる検討が必要だが、プロファイルと実験から、ドロネーコードが二次構造に基づいたタンパク質立体構造の特徴を表現することが示された。

６．まとめ

本稿では、タンパク質立体構造のコード化手法と、それによって構造特徴を捉えた実験結果を示した。

今後は、より低階層の構造グループにおけるコードの特徴や、各コードにおけるアミノ酸空間配置の特徴の収集・検討を進め、コードを用いてタンパク質構造の特徴を緻密に捉え、タンパク質立体構造理解に努める。タンパク質構造の比較検討を行うことは、機能や進化過程が未知のタンパク質への知見にもつながると期待される。

参考文献

[1]Hiroshi Wako, Takahisa Yamato, “Novel method to detect a motif of local structures in different protein conformations”, Protein Engineering, vol.11, no.11, pp.981-990, 1998.

[2]Loredana Lo Conte, Stevev E.Brenner, Tim J.P.Hubbard, Cryus Chothia, Alexey G.Murzin “SCOP database in 2002: refinements accommodate structural genomics”, Nucleic Acids Res.,vol.30, no.1, pp.264-267, 2002.

[3]Anne Poupon, “Voronoi and Voronoi-related tessellations in studies of protein structure and interaction”, Current Opinion in Structural Biology, vol.14, pp.233-241, 2004.