Vol.2012-BIO-28 No.18
2012/3/29

タンパク質におけるアミノ酸ペアの出現頻度への制約の解析

城田松之[†]　　木下賢吾[†,††]

天然タンパク質のアミノ酸配列にはその構造・機能を維持するために様々な制約が課せられている．この制約を明らかにするため，本研究ではタンパク質の配列上できまった残基間隔で出現するアミノ酸残基ペアの出現傾向について統計的に解析を行った．その結果，同じアミノ酸の局所での出現頻度は天然変性タンパク質において高いこと，また，異なるアミノ酸からなるペアの頻度はその順番に大きく影響を受けることが分かった．

Analysis of the Constraints on Amino Acid Pairs in Proteins

Matsuyuki Shirota[†]　and Kengo Kinoshita[†,††]

Various constraints are imposed on amino acid sequences of native proteins in order to maintain their structures and functions. To clarify these constraints, we analyzed the propensities of amino acid pairs at particular sequence separations. As results, we found that identical amino acid pairs in local amino acid sequence occur more frequently than random in intrinsically disordered proteins, and that the occurrence of amino acid pairs are strongly dependent on the order of the two amino acids.

1. はじめに

多くのタンパク質は，そのアミノ酸配列によって規定される天然構造に折りたたまれることで機能を果たす．それに対し，決まった構造をとらずに機能を発揮する天然変性タンパク質と呼ばれる一群が近年注目されており，これらは構造をとるタンパク質と比べて荷電性残基が多く疎水性残基が少ないという，顕著なアミノ酸組成の違いがあることが報告されている[1]．タンパク質の構造と機能の理解においては，構造をとるタンパク質と天然変性タンパク質のアミノ酸配列を比較するだけでなく，それぞれがランダムな配列と比較してどのような特徴を持つかを理解することが必要である．そこで，本研究では，配列上 1 残基から 10 残基の距離にあるアミノ酸ペアの出現頻度を比較することで，天然のタンパク質においてどのような制約が課せられているかについて解析を行った．

2. 方法

データセット中のタンパク質のアミノ酸配列において，アミノ酸 a と b が k 残基離れて出現する頻度についてのスコアを，実際の出現数とアミノ酸がランダムに分布したときの期待値との比で定義した．すなわち，

$$S(a,b,k)=\frac{\sum_p n(a,b,k \mid p)}{\sum_p n(k \mid p)\times f(a \mid p)\times f(b \mid p)}$$

ここで，タンパク質 p について，$n(a,b,k \mid p)$ は a と b が k 残基離れてみられる頻度，$n(k \mid p)$ は k 残基離れたアミノ酸ペアの数，$f(a \mid p)$ はアミノ酸 a の比率を表し，和は全てのタンパク質についてとる．1 より大きい（小さい）スコアは，そのアミノ酸ペアがその残基間隔でランダムよりも多く（少なく）現れることを意味する．

3. 結果と考察

3.1 タンパク質配列の同一アミノ酸の反復

まず, UniRef データベース[2]から配列一致度が 50%以下となる非冗長なアミノ酸配列データセットにおいて，同一アミノ酸のペアが 1~10 残基の間隔で出現する頻度を検証した．その結果，UniRef 全体では同じアミノ酸は 1~10 残基という配列上の局所で期待されるよりも出現しやすいことが分かった．一方，生物種ごとにこの比較を行ったところ，ヒトのタンパク質では同様に同じアミノ酸の反復が多かったが，大腸菌ではほぼランダムと変わらなかった．天然変性タンパク質が特徴的なアミノ酸組成を持ち，原核生物に比べて真核生物に多いということを考えると，データベースにおいてみられる同一アミノ酸の局所での反復は天然変性領域の影響が大きいと考えられた．

3.2 構造領域と天然変性領域の違い

そこで，次に，ヒトのタンパク質の配列を天然変性領域予測プログラム PrDOS[3]

[†] 東北大学大学院情報科学研究科
Graduate School of Information Sciences, Tohoku University
[††] 東北大学加齢医学研究所
Institute of Development, Aging and Cancer, Tohoku University

Vol.2012-BIO-28 No.18
2012/3/29

を用いて構造をとる領域と天然変性領域に分け，それぞれで同一アミノ酸の局所での出現しやすさを調べた．その結果，構造をとる領域では， 10 残基まで離れるとほとんどの同一残基ペアの頻度はランダムとほぼ変わらなくなるのに対し，天然変性領域では荷電性残基，極性残基については，10 残基に達しても顕著に高いスコアを示した（図 1）．これは，天然変性領域においてはアミノ酸の組成の面だけでなく，その並びにおいても荷電性･極性残基の局所での反復が重要であることを示唆する．たとえば，チロシンは 7 残基の間隔で高いスコアを示す．これは RNA-polymerase II subunit RPB1 の天然変性領域が，50 以上の YSPTSPS からなる 7 残基のリピートを持つことによる．このリピートは RNA の転写プロセスにおいて開始から伸長へと切り替えるために重要な役割を果たす[4].

3.3 アミノ酸ペア間の出現順序の非対称性

次に，異なる 2 つのアミノ酸からなるペアについて，2 つの残基の順番が頻度に与える影響を調べた．ここでは，構造をとるタンパク質として SCOP データベースの構造ドメインを，天然変性タンパク質として DisProt データベースのアミノ酸配列を用いた． その結果，反対荷電を持つアミノ酸ペア（E，D と K，R など）の組み合わせでは，構造をとるタンパク質の配列において，陰性荷電のものが陽性荷電のものの前に来るペアの頻度がその逆よりも高いことが分かった．この傾向は，残基間隔が 4 残基以下の時に加えて，7，10 残基という，α ヘリックスの同じ面に二つの残基が来るような場合に有意であった．このような荷電の順序は二次構造の形成において有利に働くのかもしれない．これに対して，天然変性タンパク質においては，このような荷電の順序による有意な出現頻度の変化は見られなかった．

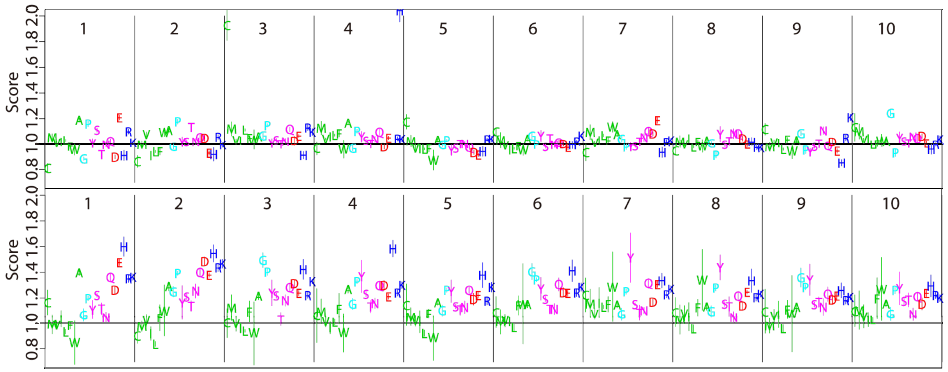

図 1 構造をとるタンパク質（上段）と天然変性タンパク質（下段）における同一アミノ酸残基ペアのスコア．アミノ酸残基は疎水性（緑），G と P（シアン），極性（マゼンタ），陰性荷電（赤）および陽性荷電（青）について色分けをした．

4. おわりに

タンパク質のアミノ酸配列は構造をとる領域においても，天然変成領域においても，アミノ酸ペアの起こりやすさという観点から見ると，ランダムな配列とは大きな違いがあることが明らかになった．前者においては荷電性ペアの順序が，後者においては同一アミノ酸の反復が顕著な性質を示すことがわかった．このような天然のアミノ酸配列についての知見はタンパク質の構造と機能についての示唆を与えるだけでなく，機能を持つタンパク質のデザインにおいても重要であると考えられる．

謝辞 This study was supported by Grant-in-Aid for Scientific Research on Innovative Areas (22136005). Computational time was provided by the Super Computer System, Human Genome Center, The Institute of Medical Science, University of Tokyo.

参考文献

1) Dunker, A.K.: Intrinsically disordered protein. *J. Mol. Graphics Modell.*, Vol.19, pp.26-59 (2001).
2) Suzek, B.E.: UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, Vol.23, No.10, pp.1282-1288 (2007).
3) Ishida, T. and Kinoshita K.: PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.*, Vol.35, pp.W460-W464 (2007).
4) Orphanides, G. and Reinberg, D.: A unified theory of gene expression. *Cell*, Vol.108, pp.439-451. (2002).
5) Chandonia, J.M.: The ASTRAL compendium in 2004. *Nucleic Acids Res.*, Vol.32, pp.D189-D192 (2004).
6) Sickmeier, M. *et al.*: DisProt: the database of disordered proteins. *Nucleic Acids Res.*, Vol.35, pp.D786-D793 (2006).