論文内容の要旨

論文題目

Computational Analyses of the Protein Size Influences on Sequence-Structure Relationships

蛋白質の配列構造相関における

サイズがもたらす影響の計算科学的解析

氏名：城田松之

1. はじめに

近年，蛋白質立体構造データベース(PDB)に登録された構造の数は増加しつつある．既知の立体構造の解析により得られる配列構造相関の知識は，蛋白質の立体構造を理解するために重要であり，またアミノ酸配列から蛋白質の立体構造を予測する際にも基礎となる．

一方，蛋白質のサイズ，すなわち残基数は，それが増加すると蛋白質において表面で溶媒と接触できる領域が相対的に減少し，同時に，蛋白質のほとんどの構造要素の数が増加するため、蛋白質の他の特徴に大きく影響を与える要素である．

そこで，本研究では計算科学的手法を用いて，蛋白質のサイズという観点から配列構造相関に関する二つの研究を行なった．第一の研究では，蛋白質のサイズのアミノ酸組成に及ぼす影響を解析した．第二の研究では，立体構造の蛋白質らしさを表す新しい指標の開発を行なった．この指標は，アミノ酸配列異なる蛋白質間では直接比較することが困難であった従来の指標を、配列の情報，特にサイズの影響を考慮して規格化することで、機械学習によって統合したものである．これらの二つの研究を通して，蛋白質のサイズは蛋白質の配列構造相関を考える上で欠かせない特徴であることを明らかにした．

2. 蛋白質のサイズがそのアミノ酸組成へ与える影響

蛋白質中のアミノ酸残基がおかれる環境は蛋白質のサイズによって影響を受ける．図 1 はサイズが異なる蛋白質を表面の残基を黒，内部の残基を灰色で表示したものである．この図が示すように蛋白質のサイズが増加する

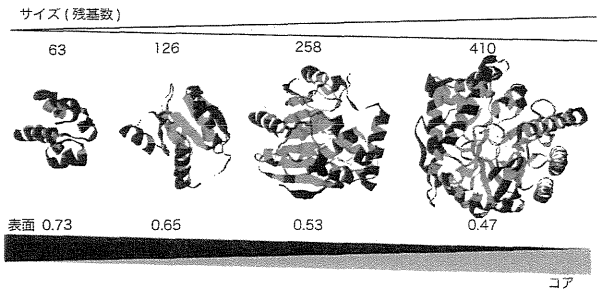

図 1 蛋白質のサイズと表面の比率

と溶媒に露出する表面の残基の比率が減少する．一般に親水性残基は表面に露出する傾向が強いが，蛋白質サイズの増加に伴い、表面の比率が減少すると、これらの残基はその出現頻度を減少させるか、蛋白質コア（内部）に埋もれる必要が生じる．

過去の研究では蛋白質サイズの増加に伴い，コアに埋もれる親水性残基が増加することが報告されている．また，親水性残基全体としては出現頻度に大きな変化は無いが、各アミノ酸種の出現頻度を個別に評価すると，蛋白質サイズの増加とともに親水性残基のうち Lys, Glu, Arg が減少し，Asp が増加するという報告がなされている．これらの知見は興味深いものであるが、その解析手法には、相関を見る際、直接的に影響を与えると考えられる表面の比率ではなく、蛋白質のサイズをそのまま用いているといった問題があり、また、何故親水性残基の中で出現頻度が減少するものとしないものが存在するのかといった点は明らかにされていない．

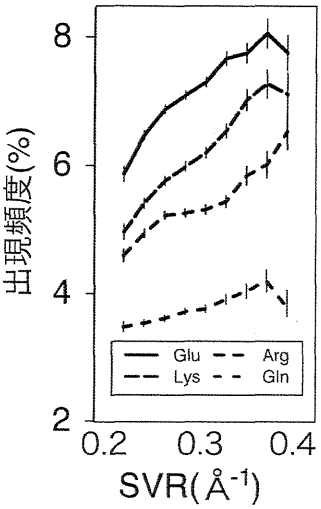

図 2 表面の比率と親水性残基の頻度

そこで本研究では，親水性残基の出現頻度にサイズよりも直接相関すると考えられる指標として，表面積体積比(Surface-to-Volume Ratio, SVR)を蛋白質の溶媒接触表面積と体積の比として定義し，表面の比率の変化がアミノ酸の出現頻度に対してどのような影響を与えるかを解析した．

解析には蛋白質の構造ドメインを定義したデータベース SCOP (Structural Classification of Proteins)より配列冗長性を除いた 7309 の代表構造を用いた．その結果，親水性残基のうちで Lys, Glu, Arg, Gln は，この順番で強く，表面の比率の減少(図 2)およびサイズの増加とともにその出現頻度が減少した．また、蛋白質のサイズと表面の比率のどちらが残基出現頻度と直接の相関を持つか偏相関を用いて解析したところ，表面の比率の方が主な要因であることを見いだした．

一方 Asp, Asn (Asx)は親水性残基でありながら表面の比率の減少に伴い出現頻度が増加した．これらの残基は似たアミノ酸だと考えられる事が多い Glu, Gln (Glx)とは異なり、近傍の主鎖と水素結合を形成しやすいことが知られている。そのため、溶媒に露出していなくとも安定に存在することが可能であり、表面の比率が減少したときに内部に埋もれることができると考えられる．

これらの結果から，蛋白質のサイズという構造的特徴の変化が表面の比率の変化を

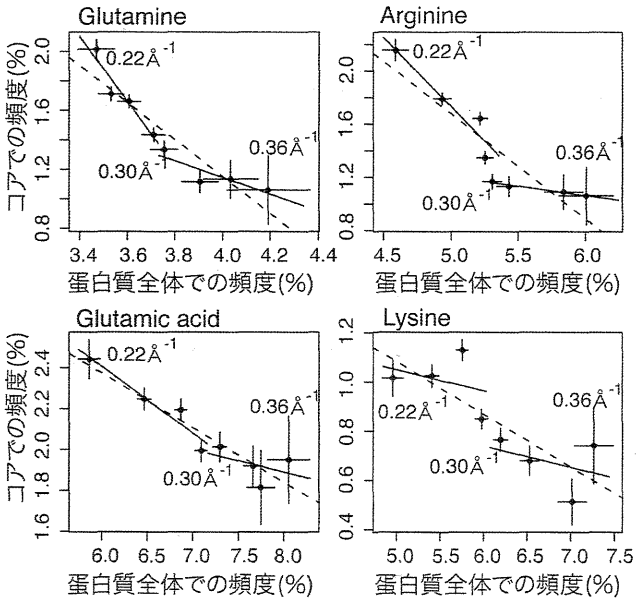

図 3 表面の比率の減少に伴う親水性残基の頻度の減少と埋もれの増加

介して，アミノ酸組成という配列の特徴に影響を与えることが示された．これは通常認識されているアミノ酸配列が構造を決定するのとは逆方向の配列構造相関である．

次に，前述の 4 つの親水性残基について，表面の比率が減少したとき，出現頻度の減少だけでなく蛋白質内部への埋もれの増加が起きているかを調べた．図 3 は SVR 毎の蛋白質全体での出現頻度とコア(蛋白質内部)での出現頻度を示したものである．SVR が減少するとデータ点が左上に移動する傾向を示しており，表面の比率の減少により出現頻度が減少するとともにコアへの埋もれの増加も同時に生じることがわかった．さらに，SVR が 0.3 $Å^{-1}$ 付近に変曲点があり，それを下回ると，埋もれの増加の効果が強く出る(傾きが急になる)ことを観察した．サイズで言うと，親水性残基の蛋白質内部への埋もれは約 130 残基より大きくなってから急激に増加する事を意味している．

3. 立体構造の蛋白質らしさの評価方法の開発

蛋白質の立体構造予測においては生成された立体構造モデルが本当に天然の蛋白質らしい構造をとっているかを評価することが非常に重要である．そのため、様々な立体構造評価手法が提案されてきたが、原子ペアを単位としてその原子間距離も考慮した統計的な手法による構造評価法は精度の点で優れており、立体構造予測のモデル評価などで広く利用されている。しかし，立体構造の予測コンテスト CASP の結果などを見ると，現在の評価手法の精度は未だ不十分なものであり，より高精度の構造評価手法が求められている。

一般に統計的な手法では，構造の蛋白質らしさは，蛋白質天然構造のデータベースにおいて，ある構造要素が見られた観測数と，参照状態と呼ばれる，蛋白質らしさがなくなった仮想的な状態においてその要素が期待される数との比の対数として表される．つまり，期待されるよりもより頻繁に天然構造に現れる構造要素は蛋白質らしく，そうでない構造要素は蛋白質らしくないと判断される．現在では様々な原子ペアを単位とした統計に基づく構造評価手法が提唱されているが，これらの間の違いは参照状態の定義の仕方にある．

本研究ではまず，これらの参照状態の異なる原子ペア単位の構造評価方法の性能を比較した．このために，2006 年に行なわれた CASP7 における 122 ドメインについてのおのおの約 200~400 の予測モデル構造とその天然構造を用いて，「天然構造を予測モデル構造から識別する能力」および「予測モデル構造を天然構造との構造類似性に従って評価付けする能力」について評価を行なった．その結果，これらの方法の中で全ての性能評価項目において優れているものは存在せず、常に良い結果が得られる参照状態の定義がある訳ではないことを明らかにした．この結果から，本研究ではこれらの既存の構造評価方法を組み合わせることで，個々のものより改良された蛋白質らしさの評価手法を作ることを試みた．

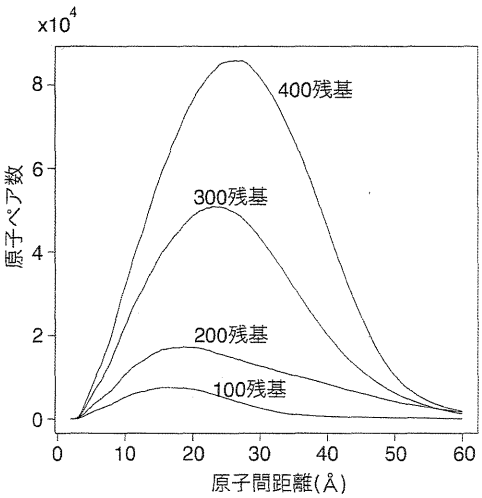

図 4 蛋白質サイズ(残基数)ごとの原子ペアの距離分布

複数の蛋白質らしさのスコアを組み合わせるためには，各々のスコアが同じ値であれば常に同程度の天然構造らしさを意味することが望まれる．しかし，従来の蛋白質らしさのスコアはその絶対値が対象となる蛋白質に含まれる原子ペア数に依存する性質がある．また，図4 が示すように蛋白質のサイズが増加すると原子ペア数も増加するため，これらのスコアの絶対値は対象となる蛋白質サイズの増加に強く影響されるという問題があった．

そこで私は，既存の蛋白質らしさのスコアをサイズとそのアミノ酸配列によって補正する方法を考案した．まず，既知の天然構造のデータベースより対象の蛋白質と同じようなサイズの蛋白質が，どれだけの数の原子ペアを持つかを推定した．これらの原子ペアが天然構造にみられるようなものである時のスコアを 1，参照状態でみられるような原子ペアである時のスコアを 0 として，蛋白質らしさのスコアを規格化することによって補正を行った．

この補正の効果を検証するため，CASP7 における天然構造とモデル構造について，補正後のスコア(横軸)と天然構造との立体構造の類似度 GDT(縦軸)をプロットした（図 5）．ここで，GDT は[0,1]の値をとり，天然構造に近いほど 1 に近づく指標である．補正後のスコアはどのようなアミノ酸配列の構造に対してもほぼ 0 と 1 の間にあり，天然構造との GDT とよく相関することがわかった。また，天然構造との GDT が 1 である点は天然構造自身であるので，これらの点の補正後のスコアは理想的には 1 となるべきであるが，実際に 0.8-1.2 と 1 に近いスコアをとっている事もわかった．また，構造によっては補正後のスコアが[0,1]の範囲を越えているものもあるがこれらは少数であった．

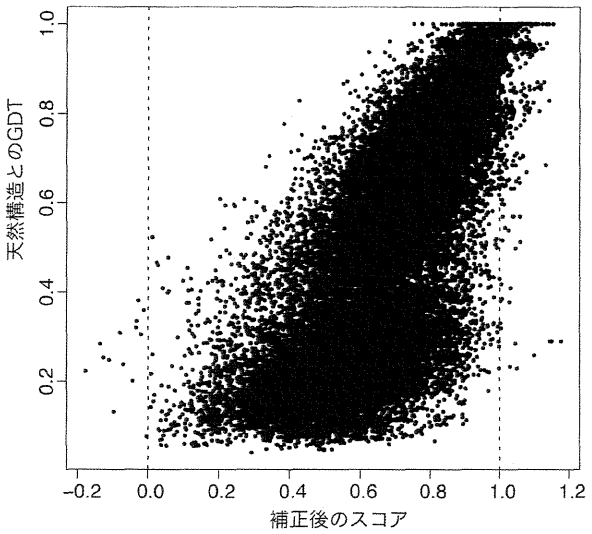

図 5 補正後のスコアと天然構造との GDT

この補正スコアを入力に，構造の天然構造との予測 GDT を出力とするような新たなスコアを機械学習により作成した．このスコアは CASP7 のモデル構造セットをによる評価で，モデル構造の評価付けと天然構造の識別の両者において，個々のスコアからの改善がみられた．

4. まとめ

本研究では，蛋白質のサイズの増加に伴う表面積の減少という構造的変化が親水性残基の出現頻度を減少させることを示した．このことは、アミノ酸配列は蛋白質の構造を決定するが，それと同時に蛋白質のアミノ酸配列はその構造によっても制約を受けるという逆方向の影響もあることを意味し，蛋白質にとってのサイズの重要性を物語る知見である．

また，原子ペアを単位とした蛋白質らしさのスコアをサイズを考慮してアミノ酸配列により補正し，天然構造との近さというアミノ酸配列によらないスコアを考案した．このような規格化は既存のスコアを組み合わせてより改良された構造評価方法を構築することを容易とするため，蛋白質の立体構造予測への応用につなげることが期待できる。