様式Ｃ-19

科学研究費補助金研究成果報告書

平成２１年　５月２６日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：19500251
研究課題名（和文）　タンパク質の局所配列・構造情報に着目した機能予測法の開発

研究課題名（英文）　Development of protein function prediction system based on local sequence-structure relationships

研究代表者
中村　周吾（NAKAMURA SHUGO）
東京大学・大学院農学生命科学研究科・准教授
研究者番号：90272442

研究成果の概要：

本研究では、アミノ酸配列情報および構造情報の局所部分の一致の数に着目して、タンパク質間の類似度をネットワークのように定義する方法を、新たに構造未知のタンパク質に適用することで、アミノ酸配列だけからタンパク質の機能予測を行う新しい方法の開発を行った。まず、局所構造の両端のα炭素原子間の距離をその局所構造の「端間距離」と定義し、予測２次構造情報と配列プロファイル情報を入力とし、予測端間距離を出力する、サポートベクタ回帰をベースとしたツールを開発した。これを、さまざまな立体構造を含むタンパク質群に適用したところ、ループ長が短いところから長いところまで、予測端間距離と実際の端間距離がよく相関することが明らかになった。とくにループ領域については、これまでβターンなど、端間距離が短いものについては、アミノ酸配列と立体構造との関係性についてさまざまな研究がなされていたが、端間距離が長いものについては研究例が少なく、本研究によって、端間距離が長くなるようなアミノ酸配列傾向がとらえられたことは、局所配列が局所構造を制限し、結果として、タンパク質のフォールディングにおいて、立体構造全体の構造空間がかなり大きく制限されている可能性を示唆する興味深い結果である。また同様の解析を、複合体形成によってディスオーダーからオーダーへ転移することが知られているタンパク質領域に適用したところ、オーダー領域ほどではないが、ランダムよりもよい予測が可能であることが明らかとなった。この局所構造の端間距離予測ツールと、配列一致検出ツール、および２次構造予測ツールを組み合わせたタンパク質機能予測ツールを開発し、その性能を確認することができた。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,300,000	390,000	1,690,000
2008 年度	900,000	270,000	1,170,000
年度			
年度			
年度			
総　計	2,200,000	660,000	2,860,000

研究分野：総合領域
科研費の分科・細目：情報学・生体生命情報学
キーワード：タンパク質、構造予測、機能予測、局所配列構造相関

１．研究開始当初の背景

さまざまな生物種のゲノム情報が蓄積し、機能未知タンパク質のアミノ酸配列や立体構造情報が指数関数的に増えている現在、ゲノムスケールでタンパク質の機能予測を行うことは、生命科学研究において非常に重要な作業となっている。機能未知のタンパク質に対して遠縁の既知ホモログを発見する方法は、PSI-BLAST の開発以来急速に発展しており、進化情報をとりこんだプロファイル同士を比較するもの、さらにはプロファイルをもとにして構築した Hidden Markov Model 同士を比較するものが高い性能を実現している。しかし、これらの技術を用いてもなお、既知タンパク質と明確なホモロジーがないタンパク質は多く存在する。

ここで、新たな可能性を生みそうなのは、タンパク質の局所部分(あるいは「フラグメント」)に限定した配列・構造の類似に着目する考え方である。立体構造が得られた場合に、タンパク質の立体構造と機能を結びつけるものとして提案されている、SCOP、CATH に代表される立体構造分類のデータベースや手法のほとんどは、タンパク質をドメインに分割した上で、階層的な分類を行っている。これらはタンパク質の構造空間を、フォールドごとの離散的なものとしてとらえる考え方といえる。これと相補的なものとして、従来のドメイン(数十-百数十残基程度)よりもさらに小さい、10 残基程度から少数の２次構造を含む程度の局所部分を単位にして立体構造を記述し、また構造類似度を定義する方法が最近提案されてきている。Friedberg らは、5 から 20 残基の類似の配列かつ類似の構造をもつフラグメントを共有する数を指標にフォールド間の構造類似度を定義し、これが Gene Ontology をもとに定義した機能類似度と明確な相関を示すことを明らかにしている（Friedberg et al., 2005)。また Petrey らは、異なるフォールドをもつタンパク質の一部の構造が類似し、同じ機能を担っている例を報告している（Petrey et al., 2005)。

フラグメントをもとにした考え方の利点は、既知タンパク質ドメインをいくつかのグループに分類して、未知タンパク質をそのいずれかのグループに同定するのではなく、既知タンパク質ドメイン(グループ)を互いに類似度で結びつけることでドメイン群をネットワークとしてとらえ、未知タンパク質をそのネットワークの中に位置付けることができる(必ずしも 1 つのグループに割り当てる必要はない)という点である。また相同性比較を局所部分同士に限定することで、わずかに残る進化類縁の痕跡あるいは収斂進化による配列・構造類似の機能モチーフをとらえやすくなると考えられる。実際に、配列相同性だけでは同定できない、複数フォールドに共通に含まれる金属結合モチーフなどの機能モチーフがいくつか発見されている。

２．研究の目的

そこで本研究は、局所部分に着目する考え方を構造未知のタンパク質に適用することで、アミノ酸配列だけから機能予測を行う新しい方法の開発を行う。この方法の特色の 1 つは、局所的な立体構造情報に関する予測を行い、それらを積極的に取り込んだ予測を行うことである。Friedberg らの報告によれば、単なる配列プロファイル類似スコアだけでは、フラグメントベースで類似度を定義しても機能類似度との間に相関がみられなかった。これは配列類似フラグメントが構造も類似しているという条件が、機能類似を発見するために必要であることを示している。酵素反応や基質結合に特定残基の配置が必須である多くの例を考えると、このことはもっともであるといえる。残基ごとの 2 次構造予測あるいは主鎖二面角予測などを含めて相同性を定義することはスレッディングなど過去にも多くの研究例があるが、本研究でもアミノ酸配列から予測可能な局所構造特徴を用いて、構造既知のフラグメントライブラリ中の多数のフラグメントのそれぞれにどのくらいマッチするかを予測する。15 残基以下の短いフラグメント構造は、アミノ酸配列を考慮しなければ、現在のタンパク質データベースでとりうる構造空間のうちの大部分をすでに満たしているという報告もあり、フラグメント構造が逆に全体構造からの影響を少なからず受けていることを考慮しても、フラグメント単位で構造を予測することは有効であると思われる。もう 1 つの特色は、上述のように局所類似の積み重ねで全体類似を定義することである。スレッディングは既知フォールドのいずれかに未知配列をマップする予測方法であるが、本手法は局所構造の範囲においてのみ既知構造へのマッピングを行うため、未知タンパク質が新規フォールドをもつ場合でも、既知タンパク質との構造情報を加味した類似度が定義できる。また従来手法においてしばしば問題となるギャップペナルティを考慮する必要がなくなり、さらにドメインスワッピングなどが起きている場合にも適用できる可能性が高まると

思われる。

３．研究の方法

手順としては、まずアミノ酸配列や配列プロファイルから予測できる構造に関する情報を用いて、できるだけ正確なフラグメント構造予測手法を確立する。具体的には、フラグメントの両端の Cα原子間の距離をそのフラグメントの「端間距離」と定義し、局所アミノ酸配列から端間距離を予測することができる、サポートベクタ回帰（Support Vector Regression, SVR）をベースとしたツールを開発する。サポートベクタ回帰は、機械学習の一種であるサポートベクタマシン（Support Vector Machine, SVM）を利用して、入力データをもとに 1 つのスカラー価を出力するもので、入力にはアミノ酸配列そのもの、アミノ酸配列から生成した配列プロファイル、および、それらと 2 次構造予測ツール PSIPRED（McGuffin et al., 2000）を利用した予測 2 次構造文字列を使用する。次に、どの程度予測構造情報が一致している場合に「一致」とみなすかを、既知のタンパク質データベースを用いて検証する。詳細な構造情報の予測は構造情報としては有用であるが精度が落ちることが予想されるので、構造情報の「淡さ」と精度の関係を把握する。そして、これらの知見を取り込み、構造・機能未知のタンパク質に対する大規模な探索に適用できるようなシステムの構築を図る。

一方、機能類似に関しては、本システムに適当である定量的な尺度の見極めを行う。2 つのタンパク質の機能類似の指標として、両者の Gene Ontology（GO）のキーワードがどのくらい共通であるかをもとにした指標がいくつか考案されている。本研究では、いくつかの指標を検討したうち、GO のツリー構造をたどり、出現確率がもっとも低い一致 GO キーワードの出現確率の–log をとる、Lord らによって提案されているものを採用する（Lord et al., 2003）。タンパク質間の機能類似は、それぞれのタンパク質に割り当てられている GO キーワードの相互類似度と自己類似度の比から算出されるスコアを用いる（Friedberg et al., 2005）。

以上の構造情報の予測および構造類似と機能類似に関する情報を統合し、予測システムとしての構築を図る。そして、配列および構造に幅広いバリエーションをもたせたテスト用のタンパク質セットを構築しそれに対して適用することで、構造未知のタンパク質の既知タンパク質との機能類似度が予測できるかどうかを検証する。

４．研究成果

まず、アミノ酸配列情報をもとにした、フラグメント端間予測ツールの構築に成功した。予測構造情報としては、構造がある程度決まっているヘリックスやストランドよりも、それ以外のフレキシブルなループ領域の構造情報が有用であると考えられるため、予測精度はとくにループ領域に着目することとした。タンパク質構造全体の類似のために局所的な配列・構造が類似する効果をのぞくために、ASTRAL データベース（Chandonia et al., 2004）から SCOP の Fold 代表ドメイン構造を取得した。SVR への入力として、5, 9, 17, 33 残基幅のアミノ酸配列および位置特異的アミノ酸置換行列（PSSM）を利用し、性能評価は 5 -fold cross validation によって行った。

その結果、PSSM を入力とした場合、ループ領域における端間距離の予測と実際の間の相関係数は、それぞれのウィンドウ幅で 0.486, 0.502, 0.449, 0.377、全フラグメントについては 0.674, 0.666, 0.557, 0.377 であった。これにより、αヘリックスやβストランドなどの 2 次構造を形成しているフラグメントのみならず、ループ部のフラグメントについても、端間距離が短いものから長いものまで全体的に予測可能であることが示された。また PSSM を入力として用いた場合は、アミノ酸配列のみを入力とした場合よりも有意に性能が向上することを明らかにした。

次の図 1 および図 2 に、ウィンドウ幅 9 のときの結果を示す。縦軸は予測された端間距離、横軸は実際の端間距離である。図 1 が、ループ領域のフラグメント（9 残基幅のうち、5 残基以上の 2 次構造が非ヘリックスかつ非ストランドと判定されているフラグメント、43450 本）への適用結果、図 2 がすべてのフラグメント（164703 本）への適用結果である。ともに、ループ長が短いところから長いところまで、予測端間距離と実際の端間距離がよく相関していることがわかる。また、とくにループ領域については、これまでβターンなど、端間距離が短いターンについては、アミノ酸配列と立体構造との関係性についてさまざまな研究がなされているが、端間距離が長いものについては研究例が少なく、本研究によって、端間距離が長くなるようなアミノ酸配列傾向がとらえられたことは、局所配列が局所構造を制限し、結果として、タンパク質のフォールディングにおいて、立体構造全体の構造空間がかなり大きく制限されている可能性を示唆する興味深い結果である。本内容については、日本生物物理学会第 45 回年会にて発表を行った。

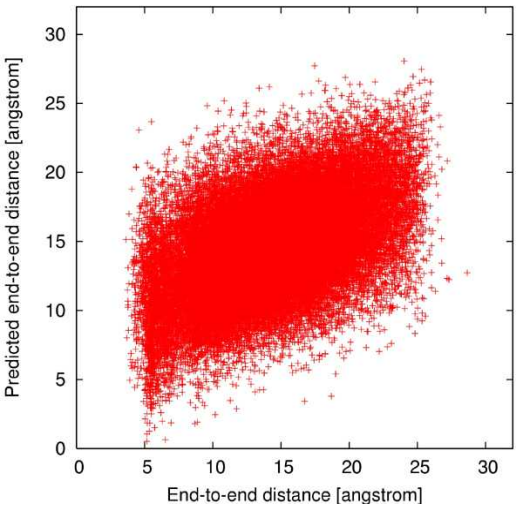

図 1 ループ領域の端間距離予測結果
（9 残基幅、PSSM 入力）

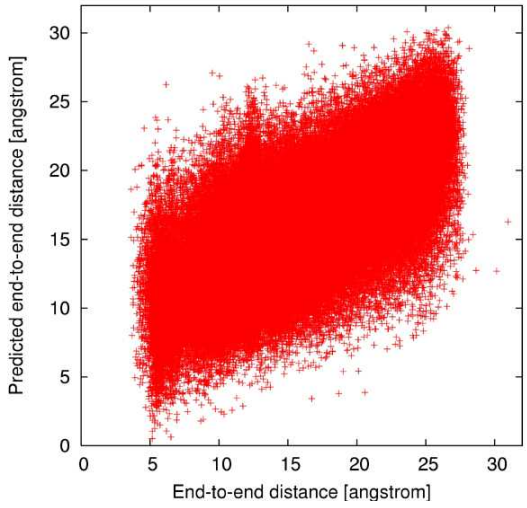

図 2 全フラグメントの端間距離予測結果
（9 残基幅、PSSM 入力）

また同様の解析を、複合体形成によってディスオーダーからオーダーへ転移することが知られているタンパク質領域に適用したところ、9 残基幅において相関係数は 0.484 であり、オーダー領域ほどではないが、明らかにランダムよりもよい予測が可能であることが明らかとなった。本内容については、第８回日本蛋白質科学会年会において発表を行った。

フラグメント端間距離予測ツールの精度のさらなる評価として、アミノ酸配列からのタンパク質立体構造予測の真のブラインドテストである CASP8 のターゲットのうち、配列類似の既知構造テンプレートがないものについて、予測精度を求めたところ、これまでの見積りと同程度の精度が得られていることが確認できた。

次に、予測ツールと２次構造予測情報の組み合わせを試みた。フラグメント端間距離予測ツールへの新たな入力として PSIPRED による２次構造予測結果を PSSM の配列プロファイル情報に加え、端間距離予測精度の変化をみた。その結果、ループ領域においては数%から 10 数%程度、全領域でみたときには 20%以上の性能向上がみられた。この結果を受けて、フラグメント単位で配列プロファイル情報の類似、２次構造予測結果の類似、端間距離予測結果の類似から、構造類似度を定義し、類似フラグメントの数を数えることで、タンパク質間あるいはフォールド間の構造類似度を計算するツールを開発した。配列プロファイルの類似度には、よい性能を示す報告が多い相関係数を用いた。

さらに、構造類似と機能類似の関係を解析するために、２つのタンパク質の間の機能類似度を Gene Ontology(GO)をもとに計算するツールを開発した。そしてこれらのツールを統合し、既存のフォールドの間の機能・立体構造の類似度を求めてネットワーク構造を構築した。最終的に、未知のアミノ酸配列を入力として、その配列の既知フォールドとのフラグメントでの予測構造類似をもとに、求めたネットワーク構造の上に入力配列を置き、機能予測を行うためのシステムの構築に成功した。例として、SCOP の分類における g.41（Rubredoxin-like fold）に適用した結果、類似した金属結合モチーフを含む g.39（Glucocorticoid receptor-like fold），g.37（C2H2 and C2HC zinc fingers fold）などをネットワーク中の近いノードとして検出することができた。以下にその出力の一部を示す。

…

g.41	g.36	29379	10124	4	0.000
g.41	g.37	29379	125	11	0.088
g.41	g.39	29379	10599	13	0.001
g.41	g.41	29379	29379	140	1.000
g.41	g.44	29379	14447	7	0.000

…

立体構造特徴を局所配列から予測するツールは、本研究で開発した端間距離予測のほかにも、二面角予測などいくつか提案されており、その精度も年を追うごとに向上している。本研究で構築したシステムに、これらのツールによる複数の予測構造特徴を組み合わせることで、さらなる性能向上の実現が期待される。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　４件）

1. W. Cao, K. Sumikoshi, T. Terada, S. Nakamura, K. Kitamoto, K. Shimizu, Computational Protocol for Screening GPI-anchored Proteins, Proceedings of the First International Conference on Bioinformatics and Computational Biology (BICoB) 2009, Springer Lecture Notes in Bioinformatics Series, in press.

2. J. Inaba, S. Nakamura, K. Shimizu, T. Asami, Y. Suzuki, Anti-metatype peptides, a molecular tool with high sensitivity and specificity to monitor small ligands, Anal. Biochem., 388(1), 63-70, 2009.

3. M. Morita, S. Nakamura, K. Shimizu, Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures, PROTEINS, 73(2), 468-479, 2008.

4. M. Kakuta, S. Nakamura, K. Shimizu, Prediction of protein-protein interaction sites using only sequence information and using both sequence and structural information, IPSJ Transactions on Bioinformatics, 49, SIG 5(TBIO 4), 25-35, 2008.

〔学会発表〕（計　２件）

1. 中村周吾、タンパク質ループ領域の配列構造相関、第８回日本蛋白質科学会、2008.06.10、タワーホール船堀

2. 中村周吾、タンパク質ループ領域の配列構造相関の解析、日本生物物理学会第45回年会、2007.12.22、パシフィコ横浜

６．研究組織
(1)研究代表者

中村　周吾
東京大学・大学院農学生命科学研究科・准教授
90272442

(2)研究分担者

(3)連携研究者