様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月３１日現在

機関番号：１２６０１
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：２１５００２８９
研究課題名（和文）コンピュータを用いたディスオーダー蛋白質の構造情報の網羅的解析

研究課題名（英文）Comprehensive analysis of the structural information of disorder proteins using computers

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（和文）：
アミノ酸配列を入力とした立体構造予測法および分子動力学シミュレーションによる構造サンプリングを多数のディスオーダー領域・ディスオーダータンパク質に適用した結果、これらのアミノ酸配列は局所的には球状タンパク質のループ部とほぼ同じ程度に立体構造とりやすさの情報を内在していること、いくつかのディスオーダータンパク質については複合体形成時の相手となるタンパク質の情報がなくても、全体構造を予測できることを明らかにした。

研究成果の概要（英文）：
I have applied protein structure prediction method and molecular dynamics simulation to amino acid sequences of various intrinsically disordered regions / proteins (IDRs / IDPs). I found that these amino acid sequences include structural information as the same degree of loop regions of ordered proteins and that tertiary structures of some IDPs after complex formation are even predictable without information of partner proteins.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,200,000	360,000	1,560,000
2010 年度	800,000	240,000	1,040,000
2011 年度	600,000	180,000	780,000
年度			
年度			
総　計	2,600,000	780,000	3,380,000

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：コンピュータシミュレーション、ディスオーダー蛋白質

１．研究開始当初の背景

近年、天然状態において一定の構造をとらないディスオーダータンパク質（Intrinsically Disordered Proteins, IDPs）あるいは球状タンパク質中のディスオーダー領域（Intrinsically Disordered Regions, IDRs）が注目されてきている。IDPs/IDRs の研究の進展によって、これらが転写調節やシグナル伝達などにおいて中心的役割を担っていることが次第に明らかになってきている。とくに真核生物で存在割合が多く、また IDPs/IDRs が原因で起きる疾患も数多く報告されている。このような重要性から、IDPs/IDRs に関する研究報告は年々増

えており、また実験的に確かめられたIDPs/IDRs のデータベースも構築されている。しかし球状蛋白質と比べると研究の歴史が浅く、まだよくわかっていないことは多い。とくに、そのアミノ酸配列には大きな特徴があることが明らかとなっているが、アミノ酸配列と立体構造の関係や立体構造とりやすさに関する情報については、分光学的方法などによる個々のタンパク質の立体構造解析結果は報告されているが、induced folding の原子レベルでのメカニズム、さらにはアミノ酸配列の違いによる影響を調べるためのタンパク質網羅的な解析などは研究例が少ない。そこで IDPs/IDRs について、ディスオーダー状態に残存する構造がどの程度あるのか、それらと複合体形成時の構造との対応の有無などを、多くの IDPs/IDRs について詳細に解析することによって、IDPs/IDRs の機能を理解するための手がかりが得られると思われる。

２．研究の目的

本研究は、以上のような現状を踏まえ、IDPs/IDRs を含む多くのタンパク質に対して、アミノ酸配列に内在する構造とりやすさはどの程度みられるか、またそれらと複合体形成時の構造との対応などを、アミノ酸配列を入力とした立体構造情報の予測および分子動力学シミュレーションを用いて解析することで、IDPs/IDRs の機能発現の理解に原子レベルの立体構造情報の視点からアプローチする。

３．研究の方法

本研究は、アミノ酸配列を入力とし、その情報から可能な限り立体構造情報を引き出す立体構造予測システムを用いてアミノ酸配列に内在する構造情報を知る方法と、分子動力学シミュレーションを用いて、入力としたアミノ酸配列がとりうる立体構造を数多くサンプリングし、それらを解析することによって構造とりやすさの傾向を把握する方法の２種類を用いて、IDPs および IDRs の原子レベルでの解析を行った。

(1) タンパク質のアミノ酸配列情報を入力とし、それをもとに進化情報を加えた局所的な配列プロファイルを作成して、その情報に基づいて大まかな局所の立体構造（具体的には局所フラグメントの両端の Cα原子間距離）を予測するプログラムを開発した。このプログラムは、機械学習の方法の１つである Support Vector Regression (SVR) を用いている。また、Support Vector Regression 単独では予測値として１つの値が出力されるだけであるが、複数の Support Vector Machine (SVM) を組み合わせることで、予測値の連続分布が得られる方法を新たに開発した。予測精度が高ければ、局所的なアミノ酸配列情報に局所的な構造情報がより多く含まれていたということがわかる。

この予測ツールを、まず球状タンパク質のループ部に、次に複合体形成に伴ってタンパク質の立体構造がディスオーダーからオーダーにシフトする領域に適用した。

(2) アミノ酸配列を入力とし、種々のフォールド認識ツール、フラグメントアセンブリ法をベースとした構造構築・改良、モデル品質評価手法などを組み合わせて、全体予測構造を全自動で出力できる立体構造予測サーバを開発した。このサーバは、単独チェインの立体構造のみならず、入力タンパク質のホモマー形成傾向をも予測し、4 次構造を考慮した上で立体構造を出力する機能を備えている。

この全自動サーバを、複合体形成に伴ってタンパク質全体がディスオーダーからオーダーにシフトする IDPs であることが明らかになっている Meszaros2007 リスト（Meszaros et al., JMB, 2007）のタンパク質（33 残基以上の 20 個, 入力配列のホモログは立体構造予測には利用しない）と、完全なブラインドテスト環境における CASP9 ターゲット 116 個のタンパク質の IDRs に対して適用し、解析を行った。

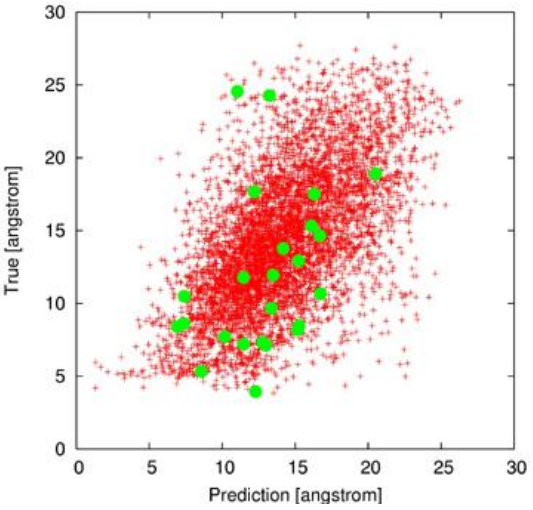
(3) Meszaros2007 リストに挙げられている IDPs のうち、33 残基未満の 17 個について分子動力学シミュレーションを用いた解析を行った。構造サンプリングには、エネルギー曲面の局所最小点へのトラップを回避し、広範な構造サンプリングを可能とするレプリカ交換法（4 レプリカ、シミュレーション温度は 300 K, 330 K, 363 K, 400 K）を用いた。2000 ステップのエネルギー極小化後、300 K, 500 ps の平衡化を行い、その後 2 ps ごとにレプリカ交換判定を行いながら、5 ns のシミュレーションを行って、構造をサンプリングした。スナップショット構造を 2 ps ごとに 1 タンパク質あたり 2500 個サンプリングし、2 次構造および局所フラグメントの両端の Cα原子間距離の分布を求めて、それらを複合体中のものと比較した。分子動力学シミュレーションには、AMBER10 プログラムパッケージを用いた。また力場パラメータには ff03、Generalized Born 陰的溶媒モデル (OBCII 法) を用いた。

総残基数は 282、複合体形成時の 2 次構造は、ヘリックスが 93 (33.0%)、ループが 189 (67.0%) であった。

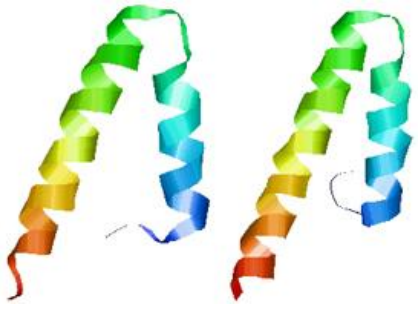
４．研究成果

(1) 開発した SVR/SVM ベースの予測ツールを球状タンパク質 (SCOP1.71 のフォールド代表構造) のループ部配列 43219 本に適用したところ、天然構造における 9 残基幅フラグメント端間距離と予測端間距離との相関係数は 0.551 であり、フレキシブルであるといわれているループ部でも、有意に局所配列情報に局所構造情報が含まれていることが明らかとなった。またディスオーダータンパク質あるいはディスオーダー領域のデータベースである DisProt とタンパク質の立体構造データベース Protein Data Bank のデータを突き合わせ、複合体形成に伴ってタンパク質の立体構造がディスオーダーからオーダーにシフトする領域のデータベースを作成した。これらに上記の予測ツールを適用した結果、そのような領域についても、天然構造における端間距離と予測端間距離の間には相関がみられ、局所配列情報に局所構造情報が有意に含まれていることを明らかにした。

(2) 開発したアミノ酸を入力とするタンパク質全体構造の全自動予測サーバを、CASP9 ターゲット 116 個に対して適用した。これらは、様々な２次構造含有率、フォールド、残基長をもつバラエティに富むテストセットであり、また予測の段階では立体構造が公開されていないため、完全なブラインドテストであるという大きな利点がある。まず、CASP9 ターゲットについて、5 残基以上の連続したループ領域を含む残基長 9 のフラグメント (5765 本) の端間距離を予測し、実際の端間距離と比較した。その結果、両者の相関係数は 0.586 であり、(1) で述べた SCOP フォールド代表構造のループ部配列に対して計算されていた値 0.551 と同等かそれよりも高いといえる相関がみられた。このことから、このツールを用いて配列に含まれる構造情報の解析を行うことの妥当性が確認できた。5765 本のフラグメントのうち、連続する 3 残基以上のディスオーダー領域を含むフラグメントは 130 本、連続する 5 残基以上のディスオーダー領域を含むフラグメントは 24 本であった。これらのフラグメントについての予測端間距離と実際の端間距離の相関係数は、それぞれ 0.581, 0.356 であり、フラグメント全長に対して半分以上がディスオーダー領域であっても、ランダム予測よりも明らかによい予測ができていた。また２次元プロットを描いてみると (次図)、これらのディスオーダー領域を多く含む予測プロット点群 (緑) は、5 残基以上の連続したループ領域に対する予測のプロット点群 (赤) の範囲におさまり、ディスオーダー領域を全長の半分以上含んでいてもループ領域とほぼ同等の精度で端間距離を予測できることが示された。



全体構造予測では、開発した全自動構造予測サーバを用いて、アミノ酸配列のみの情報から立体構造を予測し、予測立体構造と、実際の複合体中の立体構造を比較した。予測立体構造は１タンパク質あたり５つとし、その中でもっとも実際の立体構造に近かったものを解析対象とした。その結果、構造類似を表す TMscore (0 が全く似ていない場合、1 が完全一致) の平均は 0.361 であった。一般に TMscore は 0.4 以上であるとランダムとは異なる意味のある構造類似であるといわれており、今回解析対象とした IDPs 全体の傾向としては、複合体形成時の全体構造をアミノ酸のみから予測することは困難であるとの結論を得た。一方で、1j2jB, 1kilC, 1kilD で、実際の立体構造との TMscore がそれぞれ 0.663, 0.740, 0.785 の予測構造が得られた。例として、1j2jB の天然構造と予測構造を次図に示す。左が天然構造、右が予測構造である。これらのタンパク質については、有意に実際の構造と似た予測構造が得られたといえる。

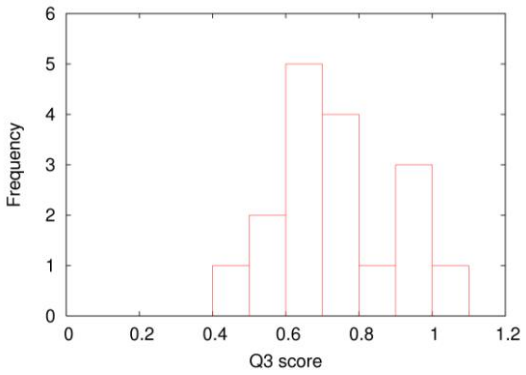


以上より、ディスオーダー領域 (IDRs) においてアミノ酸配列に少なからず構造情報が含まれていることが完全なブラインドテスト環境においても確認できた。また、ディスオーダータンパク質 (IDPs) の複合体形成時

における全体構造はアミノ酸配列情報からだけでは構築が難しいが、配列情報のみから複合体形成時の立体構造を予測できるものも存在するということが確認できた。

(3) Meszaros2007 リストに挙げられているもののうち、33 残基未満の 17 個について、レプリカ交換法を用いた構造サンプリングを行った。1 タンパク質あたり取得した 2500 個のスナップショット構造の 2 次構造および 9 残基幅フラグメントの両端の Cα原子間距離の分布を求めて、それらを複合体中のものと比較した。

2 次構造の解析の結果、全長の 20%以上のヘリックスを含むタンパク質 11 個について、天然構造でヘリックスである領域において、スナップショット構造中でもヘリックスを形成している割合は、平均 53.8%と高いことが明らかとなった。またループ部がスナップショットでもループである割合は、17 個の平均でさらに高い 78.8%であった。2 次構造の 3 状態分類（ヘリックス、ストランド、それ以外）において、天然構造とスナップショット中で一致している割合（Q3 スコア）をタンパク質ごとに求め、ヒストグラムにしたものを次図に示す。



さらに、得られたスナップショットの各構造において、9 残基幅フラグメントの両端の Cα原子間距離に着目し、天然構造において欠損していない座標が 10 残基以上ある 11 個について、スナップショット中における端間距離分布の平均から 1σの範囲に複合体形成時の天然構造における端間距離が含まれるかどうかを解析した。その結果、解析対象とした 155 フラグメント中 1σの範囲に含まれたものが 78 本、1σの範囲から逸脱していたものが 77 本であった。前者と後者の平均端間距離は、それぞれ 14.3 Å および 17.7 Å であった。なお、9 残基幅フラグメントがヘリックスを形成した場合の端間距離はおよそ 12 Å である。前者には多くのヘリックス領域が含まれるが、ヘリックス領域以外の伸展構造でもスナップショット中の端間距離分布が複合体形成時の端間距離に近い例もみられた。

以上により、分子動力学シミュレーションによる構造サンプリングの結果、複合体形成時にヘリックス・ループを形成している領域においては、フリー状態においても同じ 2 次構造をとる傾向が強いということが明らかとなった。

以上バイオインフォマティクスおよびシミュレーションの両面から、多数のディスオーダー領域（IDRs）およびディスオーダータンパク質（IDPs）に対する解析を行った結果、とくに IDRs においては、通常のループ部とほぼ同程度、アミノ酸配列そのものが局所的に構造とりやすさをもっていることが確認できた。また IDPs においても、局所的には複合体形成時に現れる立体構造の特徴がアミノ酸配列に内在していることが明らかとなった。一方全体構造については、解析した過半数の IDPs においては、予測構造と全体構造の類似度は低かった。一方で、すでにフリー状態においても複合体形成時の立体構造を予測できるものも存在することが明らかになった。この結果は、アミノ酸配列情報のみ用い、複合体形成の相手となるタンパク質の情報を入れていないことを考えると、ディスオーダータンパク質であっても複合体形成時の立体構造の情報がアミノ酸配列に含まれている可能性を示唆するものであり、大変興味深い。

立体構造予測および分子シミュレーションの両方を用いて、多数の IDRs/IDPs の解析を行った例はこれまでほとんどなく、これらの解析を、とくに真核生物のゲノム中に多数すると予測されている IDPs について適用し、また複合体形成の相手に関する情報も加えて解析することで、機能面まで含めた新たな知見が得られる可能性がある。本研究は、その端緒を開いたものといえる。

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 11 件）

1. （査読有）M. Morita, M. Kakuta, K. Shimizu, S. Nakamura, Blind prediction of quaternary structures of homo-oligomeric proteins from amino acid sequences based on templates, J. Proteome Sci. Comput. Biol., 1, Article ID:1, 2012. http://www.hoajonline.com/jpscb/1/1/1

2. （査読有）K. Sumikoshi, T. Terada, S. Nakamura, K. Shimizu, Protein-protein docking using multi-layered spherical basis functions, Int. Proc. Chem., Biol. Environ. Eng., 5, 342-347, 2011. http://www.ipcbee.com/vol5/76-X10027.pdf
3. （査読有）M. Morita, T. Terada, S. Nakamura, K. Shimizu, BUDDY-system: A web site for constructing a dataset of protein pairs between ligand-bound and unbound states, BMC Res. Notes, 4, 143, 2011. DOI: 10.1186/1756-0500-4-143
4. （査読有）W. Cao, K. Sumikoshi, S. Nakamura, T. Terada, K. Shimizu, Prediction of N-myristoylation modification of proteins by SVM, Bioinformation, 6(5), 204-206, 2011. http://www.bioinformation.net/006/97320630006204.pdf
5. （査読有）S. Someya, M. Kakuta, M. Morita, K. Sumikoshi, W. Cao, Z. Ge, O. Hirose, S. Nakamura, T. Terada, K. Shimizu, Prediction of carbohydrate-binding proteins from sequences using support vector machines, Adv. Bioinform., 2010, Article ID 289301, 2010. DOI: 10.1155/2010/289301
6. （査読有）S. Nakamura, K. Shimizu, Comprehensive analysis of sequence-structure relationships in the loop regions of proteins, Genome Inform., 23(1), 106-116, 2009. http://giw.hgc.jp/giw2009/pdf/33-Nakamura.pdf

〔学会発表〕（計６件）

1. 中村周吾, アミノ酸配列からのタンパク質立体構造予測と４次構造予測, 日本生化学会大会, 2011 年 9 月 23 日, 国立京都国際会館
2. 中村周吾, 蛋白質のホモマー構造予測, 計算タンパク質科学研究会, 2011 年 8 月 30 日, 伊良湖
3. 中村周吾, Bilab in CASP9 ～モノマーおよび複合体構造予測～, IPAB 公開セミナーCASP9 特集, 2011 年 2 月 24 日, 東京工業大学
4. 中村周吾, 蛋白質モデル構造の品質評価について, 計算タンパク質科学研究会, 2010 年 9 月 4 日, 城崎大会議館
5. 中村周吾, Development of protein local structure prediction method based on combination of multiple support vector machines, 生物物理学会年会, 2009 年 10 月 30 日, アスティとくしま
6. 中村周吾, タンパク質の局所配列構造相関について, 計算タンパク質科学研究会, 2009 年 9 月 3 日, 聴泉閣かめや

６．研究組織

(1)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

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：