様式 C-19

科学研究費補助金研究成果報告書

平成２２年６月１４日現在

研究種目：基盤研究（B）
研究期間：2007～2009
課題番号：１９３１０１３３
研究課題名（和文）タンパク質の不規則領域と選択的スプライシングに関する
ヒトプロテオーム情報解析
研究課題名（英文）Computational analysis of human proteins addressing the
relationship between intrinsic disorder and alternative splicing

研究代表者
西川　建（NISHIKAWA　KEN）
前橋工科大学・工学部・生命情報学科・教授
研究者番号：１００９３２８８

研究成果の概要（和文）：ヒトゲノムにコードされたすべてのタンパク質（２万種余り）を対象として、本研究で開発した方法を適用することにより、それぞれのタンパク質を球状ドメインと構造を作らない不規則領域に分割することに成功した。これによりヒト・タンパク質を構成する構造部分／非構造部分の割合を初めて明らかにした。また、選択的スプライシングの発生部位の解析から､選択的スプライシングは不規則領域の割合に比例して生じるとの結果を得た。

研究成果の概要（英文）：For all human proteins encoded on the genome, we have succeeded to classify protein molecules into structured or intrinsically disordered (ID) regions using the method we developed in this project. As a result, we revealed the total fractions of ordered/disordered regions of human proteins at a first time. Analyzing locations of alternative splicing (AS) data along the protein sequence, we have also revealed that AS events occur in proportion to the fraction of ID regions of proteins.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	5,800,000	1,740,000	7,540,000
2008 年度	4,900,000	1,470,000	6,370,000
2009 年度	4,900,000	1,470,000	6,370,000
年度			
年度			
総　計	15,600,000	4,680,000	20,280,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・応用ゲノム科学
キーワード：情報解析、天然変性タンパク質、スプライシング、真核生物、ゲノム

１．研究開始当初の背景

（１）これまでタンパク質は特異的な立体構造をつくり機能すると考えられてきたが、およそ１０年前から、天然状態で球状構造を形成しない長大な不規則（disorder）領域を含むタンパク質、いわゆる天然変性タンパク質が、真核生物に多く存在することが知られるようになった。このような天然変性タンパク質はとくに細胞内シグナル伝達系や、転写・翻訳制御または細胞周期の制御に関与するものが多いといわれ、それらタンパク質の機能的重要性からも注目されるようになった。一般に不規則領域はアミノ酸組成に顕著な偏り（親水性残基に富み疎水性残基は少な

い）を示すため、それを利用してアミノ酸配列情報から高い精度で予測することができる。一方、タンパク質中の構造ドメインは高性能のホモロジー検索法で同定することができるので、不規則領域予測と併用することにより、個々のタンパク質の（ドメインと不規則領域からなる）分子構成を明らかにすることができる。我々はそのような情報解析の方法を、401 個のヒト転写因子に適用し、それぞれの分子構成を明らかにした。その結果、転写因子は不規則領域の割合が非常に多く、平均として全長配列のほぼ半分（49%）に達することを明らかにした。なお、ドメインの割合は全長の 31%を占め､残りの 20%はどちらとも判定できない未知領域であった。

（２）一方、本研究課題申請の直前（2006 年）に、真核生物の選択的スプライシング（AS）はタンパク質の不規則領域において高頻度で起きている、という論文が発表された（K. Dunker et al., 2006）。不規則領域は本来的に構造的制約を欠くため、構造ドメイン部分と比べてアミノ酸配列は極端に変異しやすい（保存性が悪い）ことは以前より知られていた。同じことは、AS によるエキソン単位の配列の変更に対しても当てはまるため、AS と不規則領域の平行関係は大筋としては理解しやすい。しかし、前者は RNA レベルで生じる現象であり、後者はタンパク質レベルの問題である。両者が単純な対応関係にあるとは思われない。個別タンパク質の分子構成を明らかにした上で、AS によって変化する配列部分がどの領域に対応するかを調べて、不規則領域と AS の関係を具体的に知ることが重要だと考えた。

２．研究の目的

（１）これまでの成果を踏まえ、さらに研究を拡大するために、ヒト転写因子の情報解析で用いた方法を応用して、天然変性タンパク質が細胞内局在の違いにどの程度依存するか、という問題を追求する。具体的には膜タンパク質を対象とし、細胞膜を貫通する膜タンパク質の細胞内部分と細胞外部分に区分したとき、どちら側に不規則領域が多く存在するかを調べる。予想としては、天然変性タンパク質の典型ともいうべき転写因子が核内に存在する細胞内タンパク質であるように、膜タンパク質の不規則領域は細胞内側に多く、細胞外側では少ないだろうと考えられるが、その点を確かめたい。

（２）天然変性タンパク質の情報解析を行うための新しい方法論を開発する。転写因子の解析について述べたように、我々の解析方法を含めて従来の方法はすべて、タンパク質の構造ドメインと不規則領域をそれぞれ別々に同定あるいは予測するものだった。その結果、タンパク質分子は構造ドメインと不規則領域のほかに、どちらとも判別できない未知領域が残ってしまった。しかし、天然変性タンパク質の本来の姿からするとこのような結果は不自然であり、天然変性タンパク質においてはドメインでない部分は不規則領域であり、不規則領域でない部分はドメインと言えるはずである。すなわち、タンパク質は一般にドメインと不規則領域に２分されるはずである。したがって、従来の「不規則領域予測」ではなく、全体を構造部分（ドメイン）と非構造部分（不規則領域）に２分する方法論を開発することが重要だと考えた。このような方法論の開発を第２の目的とした。

（３）上記の「２分法」をヒト全タンパク質に適用し、構造部分／非構造部分の割合を知ることにより、ヒト天然変性タンパク質の全容を明らかにする。また、２分法によってタンパク質の分子構成を知ることは、選択的スプライシング（AS）と不規則領域の関係を解析する上で決定的に重要である。実験によって得られる AS データより、タンパク質分子のどの位置で生じるかという情報が与えられるので、２分法の結果と突き合せることにより、AS と不規則領域の間に相関性があるかどうかを調べることができる。既述の Dunker らの論文は従来法の不規則領域予測に基づいているため、２分法に基づく本研究によって、より一層正確で新規な成果が得られるものと期待できる。

３．研究の方法

（１）研究代表者（西川）の研究室では以前より、ゲノム規模のタンパク質情報解析を行い、その解析結果を GTOP データベース（http://spock.genes.nig.ac.jp/~genome/）として公開してきた。GTOP ではとくに構造情報を重視し、PDB、SCOP（構造ドメイン分類 DB）、Pfam（機能モチーフ DB）に対する高性能の配列ホモロジー検索（PSI-BLAST、HMM を使用）をすべてのタンパク質について行い、その結果は Web 画面上でカラーバー表示することにより、タンパク質のドメイン構成が一目でみられるようにした。現在、ゲノム既知の真正細菌、古細菌、真核生物の総計 700 種以上について公開中である。また、本研究課題で取り扱うタンパク質の不規則領域に関しては、D. Jones（英国 UCL）らの開発した予測プログラム DISOPRED2 を、開発者の許可を得て GTOP に組み込み、その予測結果はすでに GTOP で公開している。こうして現在では、上記の構造ドメインの解析と合せて、ドメインと不規則からなるタンパク質のおおよその分子構成を GTOP 上で知ることができる。

（２）すでに「研究の目的」で述べたように、従来法の不規則領域予測の欠点を克服するために、タンパク質の全長を構造ドメインと不規則領域に２分する予測法を、以下のようにして開発した。まず第１に、GTOP で用いている構造ドメインの同定法および不規則領域予測プログラム（DISOPRED2）はそのまま活用する。しかし、これら２つの方法だけでは全体を２分することはできないので、さらに第３の方法として、新たに CLADIST を開発した。CLADIST は上記の２つの方法とは独立に、タンパク質の全長を構造部分／非構造部分に分割することを狙った方法であり、その基本的なアイデアはタンパク質の構造部分はアミノ酸配列が進化的に変化しにくく、逆に非構造部分の配列は極端に変化しやすい（図１を参照）、という特徴に基づいている。最終的な２分法は、以上の３つの方法を組み合せた複雑なパイプラインからなり、DICHOT システムと命名した。

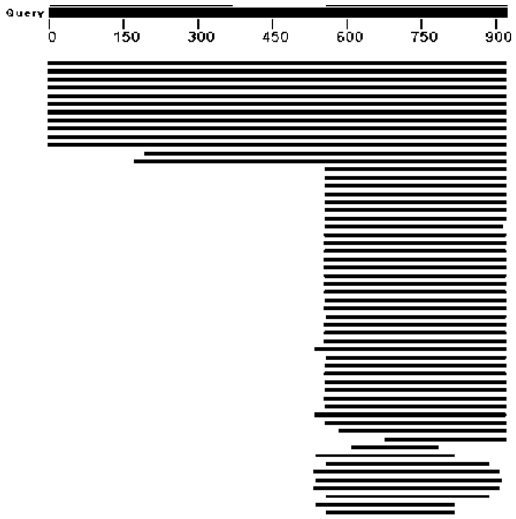

図１．アンドロゲン受容体（AR、920 残基）をクエリとして BLAST によるホモロジー検索を行ったときの出力結果。AR は前半部分が不規則領域、後半は構造ドメインからなることが実験的に知られている。配列ホモロジーは後半部分に偏って検出されることに注意。

４．研究成果

（１）最初に、膜タンパク質における不規則領域の情報解析を行い、以下のような成果を得た。解析対象のデータセットとして、865 種類のヒト膜タンパク質を UniProt データベースから取得した。また、比較のために大腸菌由来の膜タンパク質（594 種類）も同様にして用意した。膜タンパク質は膜貫通ヘリックスの本数と向き（トポロジー）によってファミリー分類することができる。上記の膜タンパク質をそれぞれのファミリーに分類し、各ファミリーの代表的なタンパク質を確認した。それぞれの膜タンパク質について UniProt アノテーションを参照して膜貫通ヘリックスの本数と位置を確認し、構造ドメインと不規則領域に関する情報は GTOP を参照して取得した。その結果、膜タンパク質は膜貫通領域（TM）、構造ドメイン（SD）、不規則領域（ID）とその他、空白（未同定領域、UNA）の４つに分割された。このうち、空白領域には構造未知のドメインが含まれる可能性がある。ヒトと大腸菌の膜タンパク質を相互に比較すると、空白を除く他の３領域を合計した平均の割合はほぼ同じ（約 70%）であったが、ID 領域の割合はヒト（17%）に対して大腸菌（4%）となり、ヒトの方が顕著に多かった。これは、天然変性タンパク質がヒトを含む真核生物には多く存在するが、大腸菌などの原核生物ではほとんど存在しない、という一般的傾向を反映したものといえる。さらに、ヒト膜タンパク質を細胞内ドメインと細胞外ドメインに分割し、ID 領域の割合を調べたところ、その割合は細胞内で多く（13%）、細胞外で少ない（4%）という結果を得た（図２）。両者の差は統計的に有意であり、一般的に天然変性タンパク質は細胞内（細胞質や核内）に多く存在し、細胞外（分泌性など）では少ないという傾向と一致する。以上の成果は論文にまとめて発表した。

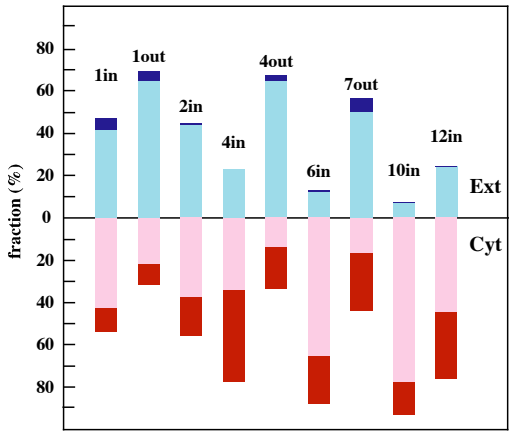

図２．ヒト膜タンパク質の細胞内（Cyt）、細胞外（Ext）ドメインにおける ID 割合の比較。各棒グラフは１回膜貫通型（1in,1out）などタイプ別の平均割合を表し、膜の内外での SD 割合（薄い赤または青）および ID 割合（濃い赤または青）を示す。各棒グラフの長さは一定（100%）である。ID 割合は細胞外よりも細胞内ドメインで顕著に大きいことがわかる。

（２）すでに述べたように、我々は独自の発想に基づき、タンパク質を構造部分／非構造部分に２分する、DICHOT システムを開発した。DICHOT の予測制度を調べるために、既知の天然変性タンパク質（58 種類）に対して

テストしたところ、DICHOT の判別精度は97%に達するという良好な結果を得た。また、従来法と比較するために以前と同じヒト転写因子（401 種類）に DICHOT を適用した。その結果、構造ドメイン（SD）と不規則領域（ID）の残基割合はそれぞれ 38%および 62%となり、ヒト転写因子では ID 領域の割合が全長の 6 割以上に及ぶことを明らかにした。また、SD, ID の割合は両者とも以前の我々の結果（31%と 49%）よりも増加しているが、その理由は、従来法では判別できずに残された空白領域（20%）が DICHOT では SD/ID のいずれかに帰属されたからである。さらに、SD と帰属された構造ドメイン（38%）のうち、34%は構造既知のドメインに帰属されたが、残りの 4%は構造未知ドメインであった（図 3 を参照）。一般にタンパク質の構造ドメインはすべてが実験的な意味で既知ではなく、未知ドメインも含まれるが、そこへ 2 分法の操作を適用すると、既知構造 SD および ID に加えて、おのずから未知ドメインがあぶり出されてくると考えられる。このように、未知ドメインの配列上の位置（図 3）や全長に対する割合が具体的に予測された例はこれまでになく、このことからも我々の 2 分法（DICHOT）が従来法とは異なる新規性を有している、といえるのである。以上の成果を論文にまとめて発表した。

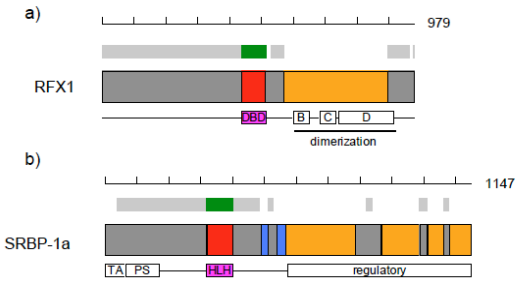

図 3．DICHOT による未知ドメインの予測例。2 つの転写因子（RFX1, SRBP-1a）について、予測結果を上から順に次の 4 段で示す：配列長、従来法による予測（緑は SD、薄いグレーは ID）、DICHOT の解析結果（赤は SD、黄は構造未知 SD、グレーは ID、青は膜貫通ドメイン）、実験的に知られた機能部位や SD など。

（3）最終年度になって、いよいよ DICHOT をヒト全タンパク質に適用した。データセットは UniProt データベースに収録されているヒト・タンパク質の全配列（20,333 本）を用いた。DICHOT のアルゴリズムは当初のものと基本的に同じだが、転写因子のときとは異なり、全タンパク質を対象とすると膜タンパク質や分泌性タンパク質なども取り扱う必要があるので、新たな判定ルールをいくつか加えた。DICHOT 解析の結果、ヒト全タンパク質のアミノ酸残基あたりの内訳は、ID 領域、構造既知 SD、構造未知 SD についてそれぞれ 34%、53%、13%となった（図 4）。こうして、ヒト・タンパク質の全配列のおよそ 1/3 は不規則領域からなること、また 13%は構造未知ドメインによって占められることが初めて明らかになった。ただし、後者の値は現時点のものであり、今後タンパク質の構造決定実験の進展とともに減少する（その分だけ構造既知 SD は増える）ことになる。また、構造未知 SD の中には新規フォールドと見なされる構造がかなり多く含まれると期待されるので、構造ゲノミックス研究に対して格好のターゲットを提供することができる。

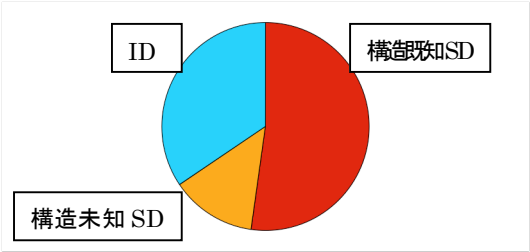

図 4．ヒト全タンパク質における構造／非構造の内訳。構造部分（全体の 66%）はさらに構造既知 SD と構造未知 SD に分かれる。

（4）以上により、ヒト・タンパク質を構造部分／非構造部分に分割することができたので、不規則領域（ID）と選択的スプライシング（AS）の関係を調べるための準備が整ったことになる。AS に関する実験データはすべて UniProt アノテーションを参照して取得した。すでに述べたように、タンパク質の ID 割合は細胞内局在性に依存する（細胞内で多く、細胞外で少ない）が、さらに AS との関係を調べるために、タンパク質を細胞内局在に従ってカテゴリー分類した。この場合も、UniProt アノテーションで与えられているカテゴリー分類に従った。まず、DICHOT 解析によって得られた ID 割合と細胞内局在性を比較したところ、ID 割合は核タンパク質でもっとも高く（47%）、次いで細胞質、膜タンパク質、分泌性タンパク質の順で、もっとも低いのはミトコンドリアのタンパク質（13%）であった（図 5）。これらの順位は先行研究の結果と一致しているが、各カテゴリーにおける明確な ID 割合を示したのは初めてである。次に、AS データより AS がタンパク質分子のどの位置で発生するかを読み取り、その位置が構造部分（SD）／非構造部分（ID）のどちらに属するかを調べて、次のように定義

される割合を算出した：
（ID 領域で発生する AS 数）／（全 AS 数）
細胞内局在のカテゴリー別にタンパク質を分類した上で、各カテゴリーごとに AS 割合を算出したのが図５である。この結果より次のことがいえる。Dunker らの論文と一致して AS は ID 領域でより頻度高く発生すること、また、AS 発生件数はタンパク質の種類（細胞内局在の違いなど）には依存せず、ID 割合に正比例する（一次関数）関係にあることが判明した。現在、これらの成果をまとめた論文を作成中である。

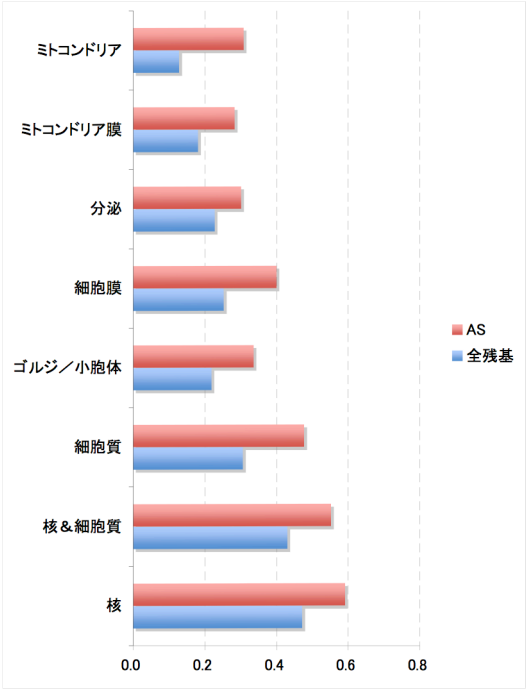

図５．選択的スプライシング(AS)と不規則領域(ID)の関係を細胞内局在性にしたがって調べた。青い棒は各局在カテゴリーにおける平均の ID 割合を、赤い棒は全 AS のうち ID 領域で発生した AS の割合を示す。赤棒が青棒より長いということは、AS が SD よりも ID 側に偏って発生することを意味する。

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 5 件）

① 西川建、天然変性タンパク質とは何か？ 生物物理、査読有、Vol. 49, 2009, pp. 4-10

② Minezaki, Y., Homma, K., Nishikawa, K., Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. *J. Mol. Biol.*, **368**, 902-913 (2007).

③ Fukuchi, S., Homma, K., Sakamoto, S., Sugawara, H., Tateno, Y., Gojobori, T., Nishikawa, K. The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions. *Nucl. Acids Res.* **37**, D333-D337 (2009)

④ Fukuchi, S., Homma, K., Minezaki, Y., Gojobori, T., Nishikawa, K., Development of an accurate classification system of proteins into structured and unstructured regions that uncovers novel structural domains: Its application to human transcription factors. *BMC Struct. Biology*, **9:** 26, pp.1-13 (2009)

⑤ Nishikawa, K., Natively unfolded proteins: an overview. *Biophysics*, **5**, 53-58 (2009)

〔学会発表〕（計 4 件）

① 西川建、天然変性蛋白質の特性：情報解析からわかったこと、第 8 回日本蛋白質科学会、東京、2008 年 6 月。

② 福地佐斗志、細田和男、西川建、ヒト・タンパク質の構造ドメイン/天然変性領域への区分、第 31 回日本分子生物学会、神戸、2008 年 12 月。

③ 西川建、天然変性タンパク質のゲノム情報解析、第 9 回日本蛋白質科学会、熊本、2009 年 5 月。

④ K. Nishikawa, Development of an order/disorder assignment method for protein molecules and its application to the human proteome, PepCon-2010, Beijing, China, 2010 年 3 月。

〔その他〕
研究成果データベース DICHOT:
http://spock.genes.nig.ac.jp/~genome/DICHOT/

６．研究組織
(1)研究代表者
西川　建（NISHIKAWA KEN）
前橋工科大学・工学部・教授
研究者番号：10093288

(2)研究分担者
福地　佐斗志（FUKUCHI SATOSHI）
国立遺伝学研究所・助教
研究者番号：70360336