様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　3 月　31 日現在

機関番号：63801
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：20510195
研究課題名（和文）　全ヒトタンパク質の構造・変性ドメインのアノテーション
研究課題名（英文）　Annotation of structural and un-structural regions in the human proteome.

研究代表者
福地　佐斗志（FUKUCHI SATOSHI）
国立遺伝学研究所・DDBJ・生命情報研究センター・助教
研究者番号：70360336

研究成果の概要（和文）：本研究の目的はヒトプロテームにおける、構造・変性ドメインの見積を行う事である。変性ドメイン同定は実験的方法論が確立していないため、生命情報学的にこの問に答えることは重要である。一般の変性領域予測プログラムは変性領域のみを出力、構造予測法では既知構造に類似性のある領域のみが出力されるが、本研究で開発したシステム DICHOT は、構造予測法と新たに開発した配列の保存度を用いる変性領域予測法を組み合わせることで、アミノ酸配列の全長にわたり完全に構造領域または変性領域かを判別するという特徴を持っている。DICHOT システムを uniprot データベースに収録された全ヒトタンパク質に適用することで、プロテーム全体で 35%の領域が変性、52%が既知構造と類似性を示す構造領域、13%が既知構造と類似性のない構造領域であると見積もられた。このように全域にわたり構造・変性領域の見積もりを行ったのは世界的に初出である。また、最後の 13%は構造未決定の領域であり、構造ゲノミクスの有望なターゲット領域となる。また、uniport に記述されている多くの機能サイト、たとえば転写因子の活性化部位、リン酸化サイト、o 結合糖鎖修飾部位等が変性領域に多く存在することが示唆された。また、変性領域は核タンパク質に顕著に多く、小胞体、分泌タンパク質の順で少なくなり、ミトコンドリアのタンパク質が最も少なかった。

研究成果の概要（英文）：We estimated the fraction of intrinsically disordered (ID) regions in the human proteome by using bioinforamtics technique. Because it is still difficult to conduct an experimental verification of ID regions in a genome wide scale, bioinformatics is expected to make an answer to this question. Although ID prediction programs generally output only potentail ID regions, our system, DICHOT, can divid an amino acid sequence into two categories, structural domains (SDs) and ID regions. With this unique feature, we can firstly estimate SD/ID fractions in the human proteome to obtain 35% of ID region, 52% of SD with similarity to PDB structures, and 13% of SD without similarity to PDB structures in residue base. The last 13% is the regions, which have not known in 3D structure, thus, can be targets of the structural genomics. Several functional sites such as trans-activation, phpsphorylation, and O-linked glycosylation were estimated on ID regions. The ID fractions differ between protein's sub-cellular locations, where nuclear proteins have the highest and mitochondrial ones do the lowest. Interestingly, phosphorylation and O-linked glycosylation occur in ID regions in secreted proteins, which have less ID regions. Comparison of ID fractions between the proteomes from several model organisms suggest that high fractions of ID regions in the human proteome is common in eukaryotes, but in bacteria, which is agree with the ID distributions by the cellular locations.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
20 年度	1,800,000	540,000	2,340,000
21 年度	500,000	150,000	650,000
22 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・ゲノム情報学
キーワード：生体生命情報学・蛋白質

１．研究開始当初の背景

これまでタンパク質は特異的な立体構造を作り機能すると考えられてきたが、近年になって天然状態で球状構造を形成しない長大な変性領域(disorder 領域)を含む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。このような変性領域タンパク質は特に細胞内シグナル伝達系や、転写・翻訳制御又は細胞周期制御に関与すると言われ、その機能的重要性からも注目を集めている。天然変性領域はアミノ酸組成に顕著な偏りがあり、アミノ酸配列情報から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感度相同性検索法で同定することができるので、変性領域予測と併用することにより個々のタンパク質の構造ドメイン・変性領域構成を明らかにすることができる。しかし、この方法にも問題点があることがわかってきた。申請者の所属する研究室でヒト転写因子に関するドメイン・変性領域構成を解析したところ、３６％が構造ドメイン、４９％が変性領域という結果を得たが、残りの１５％の領域が構造ドメインなのか変性領域なのかは不明であった。また、同様の解析をヒト全タンパク質で行うと４割程度の領域が不明領域となった。空白領域が生じる理由は、変性領域予測とドメイン検索法が完全ではないことに起因する。

２．研究の目的

本研究では、ヒト全タンパク質のドメイン・変性領域のアノテーションを行うことにより、ヒトタンパク質中の構造ドメインの数、未知ドメインの数、天然変性領域の割合、等を生命情報学的に明らかにすることを第一の目標とする。このため、既存の高感度相同性検索による構造ドメインの推定法、天然変成領域予測法の問題点を補完する方法の開発が必要である。申請者は現在ヒト転写因子で配列保存性を利用したドメイン予測を組み合わせた方法をテスト中であり、完成に近づいている。しかし、ここに至るまでに転写因子においてかなりの試行錯誤を経たチューニングを行ってきたことを考えると、このままの設定でヒト全タンパク質に適用しても様々な問題点が出てくると思われる。このため、さらに一般的なタンパク質への適用が可能なチューニングを行い方法を確立することが、本研究の目的を達するために肝要である。

３．研究の方法

高感度相同性検索・変性構造予測を組み合わせれば、タンパク質のドメイン構成知ることが可能である。しかし、先にも述べたとおりアノテーションを行った際、空白部分が生じる。相同性検索による構造ドメインのアノテーションでは、相同性検索の対象として立体構造既知のアミノ酸配列データベースを使用するのが通例だが、すべての立体構造が構造決定されているわけではないので、構造未知のドメインが存在する場合その領域はアノテーションされずに残ってしまう。一般に変性領域予測プログラムの学習データは、変性領域として PDB(Protein Data Bank)で構造決定されたタンパク質の中で動的で座標の決定できない比較的短いループ領域を使っており、長大な変性領域を学習に使用して

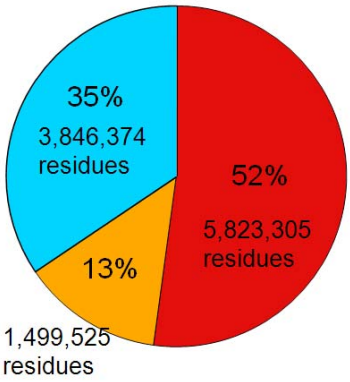
いない。また、天然変性領域は配列保存性が低いことが知られているが、配列保存性を考慮した予測プログラムの例はない。また、変性領域予測プログラムでは、長大な変性領域部分が断片化され、予測変性領域と未知領域の縞模様が見られる。申請者はこれらの予測法を補完するために、配列の保存性を利用した構造・変性ドメイン判別システムを開発した。この方法は現在ヒト転写因子でテスト中であり、このシステムをヒト全タンパク質への適用のためにチューニング行うことで、一つのプロテオーム中の構造・変性領域の区分が可能である。

４．研究成果

まず、サーバを遺伝研に導入し、それまでに開発していた変性領域予測プログラム等の移植を行い、研究環境を整備した。この予測プログラムは、GTOP データベース（発表論文２）で用いられているサイ・ブラスト、隠れマルコフモデルを用いた構造予測システムと統合し、DICHOT システムを完成させた。このシステムは、典型的な変性タンパク質であるヒトの転写因子でテストが行われ、予測エラーは 3%程度であると見積もられた。しかし、実際にヒトの全タンパク質に適用してみると、繊維タンパク質の扱い、サイ・ブラストによる構造予測で配列一致性が低い場合、などに扱いに注意が必要であることがわかった。このため、前者では pfam モチーフ検索の結果を用いることにより、既知の繊維タンパク質領域を同定し構造領域と判定させた。二番目の問題では、該当する予測結果を抜き出し扱いに注意が必要な一致度の閾値を定め、閾値より低い場合は、他の予測法（ブラスト、隠れマルコフ）による予測がある場合はそちらを優先、ない場合は採用しない等のルールを定めた。

このようにしてチューニングされたシステムを、uniprot に収められた全ヒトタンパク質に適用した。その結果、構造領域で既知構造に類似性を示すもの 52%（図中、赤）、構造領域で既知構造に類似性を示さないもの 13%（オレンジ）、変性領域 35%（水色）という結果となった。このように、全プロテームに関して判定されない領域を残さず全域にわたり構造・変性領域の割合を見積もったのは、この仕事が初出であり現在論文投稿の準備中である。また、論文の受理にあわせて、この結果を公開できるようにホームページの開発も既に終えている(http://spock.genes.nig.ac.jp/~genome/DICHOT)。この解析結果をもとに、以下のような知見が得られた。

構造既知領域に関しては、構造単位として構造ドメインの数を見積もる事が出来る。図中の赤い領域に関しては、PDB におさめられた既知構造に類似性を示し、PDB の構造は SCOP データベースにより構造が分類分けされているので、構造ドメインの種類を見積もる事が出来る。また、オレンジの部分に関しては、既知構造に類似性が見られず直接構造ドメインが何種類含まれているかを見積もる事は出来ないが、赤い領域の構造の種類数及び残基数からおおよその種類の数が見積もる事が出来る。結果として、構造既知部分（赤の部分）に 943 種類、構造未知部分に 202 種類の構造ドメイン（SCOP のスーパーファミリー数）が有ると推定された。この事は、オレンジ部分に含まれている 202 種類の新規構造ドメインの構造を決定すれば、ヒトに関してはタンパク質の基本構造単位の種類を尽すことができることを示唆している。



真核生物には様々な細胞内小器官が存在し、各小器官に局在するタンパク質には違いが有る。そこで、構造・変性の割合を細胞内の局在に関して比較してみた。その結果、変性領域の多い順に、核、核および細胞質、細胞質、細胞膜、分泌、小胞体及びゴルジ体、ミトコンドリア膜、ミトコンドリア、となった。もっとも変性の割合の多い核では 53%が変性領域で占められていたのに対し、ミトコンドリアでは 12%にすぎなかった。ミトコンドリアは進化の過程で原核生物が共生する事で生まれたとされており、原核生物型のタンパク質で構成されていると推定される事と一致する。実際、DICHOT を大腸菌等の原核生物に適用すると、10%弱の変性領域しか含んでいない事とつじつまが合っている。

変性領域には様々な機能部位の存在が示唆されてきたが、本研究はこの結果を指示するものである。特に、リン酸化部位はこれまでも変性領域に多いと言われてきたが、今回細胞内局在ごとにタンパク質をクラス分けし検討した結果、変性領域の占める割合の低い分泌タンパク質においても、リン酸化は変性領域に多く起きている事が示唆された。また

O 結合タイプの糖鎖修飾(O-linked glycosyration)も変性領域に特異的に起きている事が示唆された。糖鎖修飾は主に分泌タンパク質に見られる翻訳後修飾であり、多くは構造ドメインに見られると予想された。実際、他のタイプの糖鎖修飾 N 結合型の多くは構造ドメインにおこる事が示唆された。O 結合型の糖鎖修飾が予想に反し変性領域で起きていることが示唆されたことは、O 結合型糖鎖修飾の背後に変性領域の機能等との関係が示唆され大変興味深い。この点は、今後さらに研究を進めてゆきたい。

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 7 件）
1. Fukuchi, S., Hosoda, K., Homma, K., Gojobori, T. and Nishikawa, K. Binary classification of protein molecules into intrinsically disordered and ordered segments. BMC Structural Biology, in press 査読あり
2. Fukuchi, S., Homma, K., Minezaki, Y., Gojobori, T. and 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 Structural Biology 9 26 (2009) 査読あり
3. Fukuchi, S., Homma, K., Sakamoto, S., Sugawara, H., Tateno, Y., Gojobori, T. and Nishikawa, K. The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions. Nucleic Acids Res. 37 D333-D337 (2009) 査読あり

〔学会発表〕（計 12 件）
1. Fukuchi, S. -An annotation system for ordered and disordered regions of proteins. Intrinsically disordered regions in proteins and related topics, Kusatsu, Shiga 2011 年 1 月
2. Fukuchi, S. A database of intrinsically disordered protein. Ths 1st international symposium on intrinsically disordered proteins. Yokohama 2011 年 1 月
3. 福地佐斗志、太田元規「天然変性タンパク質のプロテオーム情報解析」第 32 回日本分子生物学会年会、シンポジウム「天然変性タンパク質による転写調節機構」2009 年 12 月、横浜
4. Fukcuhi, S. Classification of proteins into structured and un-structured regions. The 2nd Annual Protein and Peptide Conference, Seoul, Korea 2009 年 4 月、韓国

〔その他〕
ホームページ等
DICHOT ホームページ
http://spock.genes.nig.ac.jp/~genome/DICHOT

６．研究組織
(1)研究代表者
福地　佐斗志（FUKUCHI SATOSHI）
国立遺伝学研究所・DDBJ・生命情報研究センター・助教
研究者番号：70360336