様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１５　日現在

機関番号：　１１３０１
研究種目：　若手研究(B)
研究期間：　2016～2017
課題番号：　１６Ｋ１６１４２

研究課題名（和文）大規模シミュレーションとデータベース解析に基づく蛋白質立体構造予測改良法の開発

研究課題名（英文）Refinement of protein structure prediction by using simulation and database analyses

研究代表者

城田　松之（Shirota, Matsuyuki）

東北大学・医学系研究科・助教

研究者番号：００５４９４６２

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：アミノ酸配列からの蛋白質立体構造の予測において構造モデルの精度を向上させることはそのモデルの分子機能の解明や創薬などへの応用可能性を広げる重要な課題である。本研究では既知の蛋白質構造を網羅的に活用することで構造予測改良につなげるデータベースと手法の作成を行った。合わせて、ヒトの蛋白質をターゲットとして、既存の構造とのアラインメントを作成し、構造予測の改良に役立てられるように整備した。さらに、蛋白質立体構造へのマイクロエクソンによるアミノ酸残基挿入やG蛋白質共役型受容体における膜貫通ドメインのアラインメントと変異の統計解析を行って、構造予測の改良を適用する基盤を構築した。

研究成果の概要（英文）：Improving the quality of protein models generated by protein structure prediction is a challenging task that can extend the applicability of the predicted models to function prediction and drug design. In this study, known protein structures are utilized for refinement of protein structure models by comprehensively processing all the available structures. The information from known structures are aligned to the human protein sequences by sequence alignment for improving the structure models of human proteins. As applications of the model refinement methods, two genomic analyses were performed. First, amino acid insertions of human protein structures by splicing-in microexons were analyzed by using public RNA-seq data. Second, amino acid variations within the transmembrane domains of G-protein coupled receptors were summarized. The results of these analyses provide g resource for appying model refinement methods.

研究分野：　バイオインフォマティクス

キーワード：　蛋白質構造　構造予測　RNA-seq

１．研究開始当初の背景

アミノ酸配列からの蛋白質の立体構造予測は遺伝子機能や分子間相互作用の予測や創薬などに広く応用されている。予測された構造モデルの応用可能性を広げるためには、モデルを改良し高精度化することが重要である。構造予測モデルの改良のために、既知の蛋白質立体構造情報に基づく経験的手法と分子シミュレーションに基づく物理的手法が用いられている。しかし、現在の構造予測改良の手法の精度は十分ではなくさらなる研究開発が必要である。

モデルの高精度化には物理化学に基づく力場を用いた分子動力学シミュレーションによる構造サンプリングが有効である場合があるが、この方法ではモデルの質が悪化する場合も多いことが知られている。高精度化においては改良すべき領域がモデル構造全体ではなく、精度の低い部分に限られる場合があり、この場合は高精度化したい領域について蛋白質立体構造データベースから鋳型となる構造情報を検索し利用することが有効となる。

蛋白質の立体構造データベースである Protein Data Bank (PDB)には現在約 14 万のエントリが登録されている。これらは多様な生物種の様々な蛋白質の構造を含み、構造予測の改良においても重要な役割を果たしている。一方で PDB は配列類似性の高い構造を多数含んでおり、構造予測においてはこのような類似した構造をまとめて冗長性を排除した代表構造のみが利用されることも多い。しかし、配列と全体構造が類似した構造でも部分構造においては違いが見られるケースもあり、このような部分構造の多様性は構造予測モデルの改良において精度の低い局所構造をサンプリングする上で重要な情報となる。また、PDB の構造は X 線結晶解析、NMR、電子顕微鏡（電顕）などの手法で解析されているが、蛋白質の統計解析などにおいては大多数を占める X 線構造が主に使われることが多かった。しかし、近年電子顕微鏡の解析技術が進んでおり高分解能で巨大な構造が報告されているが、このような構造は従来の PDB のフォーマットでは取り扱えないケースもある。このような状況から、PDB の情報をより網羅的かつ有効に活用することが既知の構造情報を用いた構造予測改良手法の開発において重要であるという発想に至った。

２．研究の目的

本研究では PDB の情報を網羅的に活用し、シミュレーションを合わせて、蛋白質構造予測モデルを改良する手法の開発を行う。さらに、構造予測の重要なターゲットであるヒトの蛋白質についてこの手法を適用できるようにゲノムから PDB の構造情報への対応付けを行い、構造予測改良の指針とできるように情報を整理する。さらに、ヒトの蛋白質のうち、構造予測および予測モデル改良のターゲットとなる例として、マイクロエクソン (ME) を含む蛋白質、および G 蛋白質共役型受容体 (GPCR)について構造予測改良の適用について検討を行った。

３．研究の方法

蛋白質立体構造予測の改良のために PDB における立体構造情報を用いた。これらの構造についてヒトの蛋白質配列との配列アラインメントを作成し、構造予測・モデル改良に活用できる形にデータベース化した。

ヒトの蛋白質の構造予測改良法の活用対象として ME を含む蛋白質と GPCR について検討した。ME については Short Read Archive (SRA)より RNA-seq のデータをダウンロードし、VAST_TOOLS (Irimia et al. Cell,2014) を用いて評価した。GPCR については既報の配列アラインメント(Cvicek et al. PLoS CB, 2016) をもとに、Exome Aggregation Consortium (ExAC, Lek et al. Nature 2016) の変異の出現頻度を評価した。

４．研究成果

蛋白質の構造予測の改良のために蛋白質構造データベースを網羅的に活用するためのパイプライン作りを行った。一般に既知の構造を利用する際は PDB に存在する蛋白質チェイン（サブユニット）構造のアミノ酸配列を配列一致度によってクラスタリングして立体構造の分解能などからその代表構造を用いるという手法が用いられる。しかし、このやり方では配列は類似しているが微妙に異なる蛋白質の部分構造の情報を利用することができない。そこで、PDB のすべての構造を扱い立体構造予測に活用できるデータベース化することを目指した。

(1).mmCIF への対応

この上で、1970 年代から使われている PDB のレガシー（フラットファイル）フォーマットの 2 つの問題点が明らかになってきた。1 つは固定長フォーマットのために表現できる原子数・チェイン数に制限があり超巨大分子構造が扱えないこと、もう 1 つはメタデータの不足のために構造解析した分子のアミノ酸配列と原子座標情報のあるアミノ酸の間の対応付けがうまくできないケースがあることである。特に後者は UniProt などの外部データベースからリン酸化などの修飾情報を活用する上で障壁となる。これらの問題を解決するために新しい（2014 年～）PDB 構造の標準である macromolecular CIF (mmCIF) ファイルを利用することとした。これにより、2018 年 4 月現在で 139,717 エントリ、うち mmCIF ファイルしかない（レガシーフォーマットにない）ものが 651 エントリあった。これらのエントリの多くはチェイン数が多い（50 以上）ものであり、これらの 651 エントリの総チェイン数は 55,283 にのぼり、レガシーフォーマットのみでは見逃されるデータが多いことを示す。また、この中にはチェイン数が 10 以下のものが 44 エントリ、50 以下のものが 148 エントリ存在する。これらは比較的新しい構造であり、レガシーフォーマ

ットでも記述できるが、PDB が mmCIF（または pdbml）を推奨フォーマットとしたためにレガシーフォーマットのファイルを作成していないものである。従って、新しいフォーマットへの移行は今後の PDB 構造のデータベース解析において大規模構造に限らず、一般的な大きさの構造においてもますます重要となってくることが予想される。

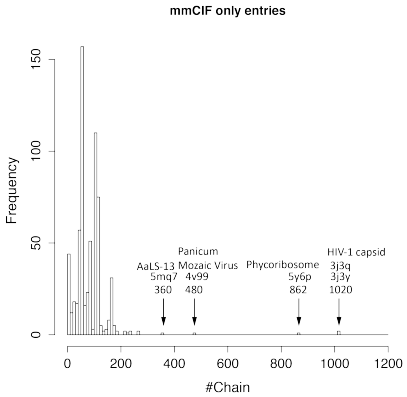

図 1 mmCIF のみのファイルとチェイン数

(2). BioUnit の取り扱い

蛋白質には他の蛋白質や低分子などと複合体を形成して生物学的機能を発揮するものも多い。一方、PDB エントリに含まれる原子座標は結晶の非対称な構造単位であることがあり生物学的に機能する複合体となっていないことがある。たとえば、ホモ 4 量体で機能するイオンチャネルにおいて、通常の PDB ファイルにはサブユニット 1 つしか座標がなく、それを対称操作して残りの 3 つを生成することで生物学的に機能する分子となるケースがある。蛋白質構造予測の改良においては他の分子との相互作用が重要である場合があるため、これらの生物学的な複合体構造を取り扱う必要がある。

PDB において X 線構造や電顕構造の生物学的な複合体構造は BioUnit として提供されている。しかし、BioUnit の大部分はレガシーフォーマットで提供されており、一部の超巨大分子構造のみが mmCIF フォーマットで提供されているという問題があった。そこで、レガシーフォーマットで提供されている BioUnit については、元のエントリの mmCIF とレガシーフォーマットの BioUnit の情報を統合して mmCIF フォーマットの BioUnit を作成した。これにより元の PDB エントリを対称操作することでできる蛋白質間相互作用面やリガンド結合部位の情報も活用することが可能となった。

(3). 構造予測改良のための既知構造の DB 化

これらの立体構造については主鎖の二次構造、側鎖の水素結合・Cation-π相互作用・van der Waals 相互作用などの原子間コンタクト、残基の埋もれ度、リガンドとの距離によって部分構造を分類しデータベース化した。これらの PDB の構造についてはヒトタンパク質のアミノ酸配列と網羅的な配列相同性検索と配列アラインメントを行い、立体構造が未知のヒトのタンパク質の構造予測とその改良に利用できるようにした。

(4). マイクロエクソン(ME)の機能解析への応用

ME は 3 から 15 塩基の（多くはインフレームの）短いエクソンであり、特に選択スプライシングによる調節を受けるものは神経系の発生段階などで重要な働きを持つことが知られている。ME の挿入は主に蛋白質表面や相互作用面の数残基の挿入となり、これにより生じる構造変化と機能への影響を予測することは神経変性疾患や自閉症などの解明に役立つと考えられる。蛋白質のヒトの RNA-seq データを用いて選択的 ME の発現パターンを確認した。

ME の発現調節を調べるために公共の RNA-seq データを用いて、VAST_TOOLS を用いて解析を行った。ME の発現は RNA-seq のリードのうち、ME に隣接するエクソンの間に ME 配列を含むリード（spliced-in）と含まないリード（spliced-out）を比較して Percent spliced-in (PSI)として定量する。SRA のヒト脳の一細胞 RNA-seq データ(Darmanis et al. PNAS 2015)を用いて ME の細胞ごとの発現状況を解析したところ、24 の ME は細胞によって発現（PSI）が変わることがわかった。一般に細胞集団で見ると PSI は 0 と 100%の間の平均値を中心に正規分布に近い分布を示すが、一細胞での発現は 0%と 100%の 2 峰性を示すことがわかった。これは細胞ごとに ME に対応するアミノ酸残基を含んだ蛋白質と含んでいない蛋白質のどちらかが発現していることを意味する。

次に、SRA の国際 1000 人ゲノム計画(1KG)のヨーロッパ系 466 人のリンパ芽球の RNA-seq データ(Lappalainen et al. Nature 2013)を用いて ME の PSI を計算し、全ゲノム関連解析を行った。その結果、3 つの ME についてゲノム上の近傍に PSI の発現量に影響する変異（*cis*-eQTL）が見つかった。このうち、蛋白質 RPB11b1 をコードする *POLR2J2* の 3 塩基の ME は近傍に PSI に強く関連する変異（rs6966982）があり、参照型と変異型のホモでは約 20%PSI が低下した。この ME の挿入により RPB11b1 の 17 番目の Lys の前に Asn が挿入される。Lys17 は RPB3 の E223 と塩橋を形成しており、この ME は複合体の安定性に影響することが推測される。

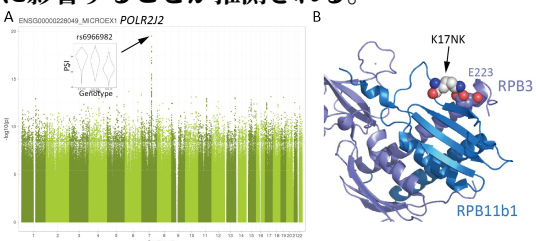

図 2 マイクロエクソンの eQTL 解析。A *POLR2J2* の ME の PSI に関連する変異、B RPB11b1 の ME によるアミノ酸挿入

(5). GPCR への応用

立体構造情報を利用したタンパク質構造予測の改良の具体例として G タンパク質共役型受容体（GPCR）の構造情報の活用についての検討を行った。GPCR はヒトゲノムにおいて約 800 種類存在する重要な創薬ターゲットであるが、立体構造解析がされているものはごく少数であり、構造解析の難しさからヒト以外のホモログで解析されたり様々な変異が導入されたりする。ヒトの GPCR において実験的に解析された立体構造から天然の配列を持った立体構造をモデリング・改良することで GPCR の創薬研究等に応用可能なモデルを作成することを目指した。

GPCR は 7 回膜貫通ドメイン(TMD)と細胞外の N 端ドメイン(NTD)、細胞内の C 端ドメイン(CTD)からなり、TMD はさらに 7 本の膜貫通ヘリックス(TMH)とそれらの間のそれぞれ 3 本の細胞内・細胞外ループ（ICL1-3, ECL1-3）から構成される。NTD, CTD とループは GPCR によって大きく異なるが、7 本の TMH の構造は特に同じ GPCR のサブクラス内ではよく似ている。現在ヒト集団のゲノム解析により GPCR の TMH の変異が多く報告されており、これらの変異による構造と機能への変化を推定し、効果のある変異を抽出することは重要である。

GPCR の TM ドメインについて各クラス(A, B, C, F, Taste2, Olfactory）の TMH のアミノ酸配列をアラインメントし、Sequence Logo とともに Exome Aggregation Consortium の約 6 万人において見つかる変異をプロットした。クラス A の TMH3 の例を図に示す。クラス A には 276 種の蛋白質が含まれるが、その約半数において最も種間の保存度が高い Ballosteros-Weinstein 番号 3.50 のアルギニン（R）残基において変異が見られ、変異が見られた蛋白質数としてはクラス A の TMH の残基の中では最大であった。これはアルギニンは CpG で始まるコドンが多く CpG transition による変異が起こりやすいためと考えられる。一方で、R3.50 は種間でもっとも保存される残基であり、前後の残基と合わせて(D/E)RY モチーフを形成し、GPCR の活性化に関与しているため、この残基の変異はそのシグナル伝達機能に影響を与える可能性がある。一方、これらの R3.50 の変異の変異アレルの頻度は極めて小さく 0.1%以下であり、この残基の変異が抑制されていることを示す。TMH では他の残基にも変異は見られるが変異アレル頻度は極めて低く、変異に対して負の選択圧がかかっていることが考えられる。健康な個人の中に TMH の変異が見られることは、GPCR によっては許容される変異であるか、ヘテロ接合では代償によって問題が起こらないことが考えられる。近年 GPCR の構造解析も進んでおり、構造モデリングによって解析できる天然および変異蛋白質の種類が増加している。GPCR は重要な薬剤のターゲットでもあり、これらの変異の効果を推定することは医療上でも重要な課題となると考えられる。

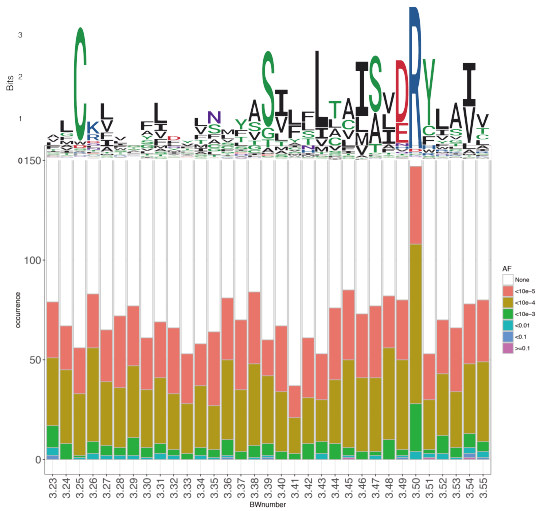

図 3 GPCR クラス A の TMH3 におけるアミノ酸ロゴと ExAC による変異数

５．主な発表論文等

（研究代表者は下線）

〔学会発表〕（計 5 件）

1. 城田松之. NGS で見つかる選択的マイクロエクソンとタンパク質構造・機能への影響. NGS 現場の会第五回研究会. 2017 年 5 月 24 日. 仙台国際センター（仙台市）
2. Matsuyuki Shirota. Population genomics and structural bioinformatics: an analysis of missense variants in a Japanese cohort based on protein structures. 2nd Karolinska-Tohoku Joint Symposium on Medical Sciences. 2017 年 10 月 2 日. 東北大学（仙台市）
3. 城田松之、木下賢吾. 低頻度一塩基多型のタンパク質立体構造を利用した機能アノテーションに向けて. 2017 年度生命科学系学会合同年次大会、2017 年 12 月 7 日、神戸国際会議場（神戸市）
4. 城田松之、木下賢吾「PDBj を利用したデータベース構築の事例紹介」第 6 回生命医薬情報学連合大会、2017 年 9 月 27 日、北海道大学（札幌市）
5. 城田松之「PDB におけるレアバリアント構造の探索」第 16 回日本蛋白質科学会年会、2016 年 6 月 9 日、福岡国際会議場（福岡市）

〔産業財産権〕

なし

６．研究組織

(1)研究代表者

城田松之（Shirota Matsuyuki）
東北大学・医学系研究科・助教
研究者番号： 00549462