様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２２　日現在

機関番号：　３４２０４
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2015
課題番号：　２５２８０１０９

研究課題名（和文）生体分子構造データのグラフによる統一の試み

研究課題名（英文）Attempts in unifying biological molecular data via graph methods

研究代表者
白井　剛（Shirai, Tsuyoshi）

長浜バイオ大学・バイオサイエンス学部・教授

研究者番号：００２６２８９０

交付決定額（研究期間全体）：（直接経費）　　12,600,000円

研究成果の概要（和文）：生物学データは階層的であり、これにもっとも適したデータ形式はグラフである。この研究では生物学的グラフを比較する方法を研究開発し、低分子構造グラフによる構造-活性相関、タンパク質超分子複合体モデルのグラフベースの構築方法、系統樹(グラフ)の比較による直系群の同定方法などを開発した。タンパク質複合体モデルについては、ヒト複合体のモデリングにより新たな疾患関連変異をインターフェースに同定した。また、DNA複製系複合体について構築したモデルを、電子顕微鏡単粒子解析などで実験的に検証し、モデルが実験構造によく一致することを示すなど、グラフベースの方法論の有効性を示した。

研究成果の概要（英文）：The biological data are hierarchical, and graph is one of the best data structures for presenting and unifying various biological data from molecules to phylogeny. In this study, multidisciplinary approach was taken to device the methods in comparing biological graphs. The graph-matching method for small molecules had successfully related structures and functions of naturally occurring drugs. The method for comparing phylogenetic trees was used to identify orthologous gene clusters in complex trees. The graph-based supramolecular modeling method had generated 3,197 human protein complex models, and the models revealed disease-related SNVs on the molecular interface for more than 10 disease. The model for EndoMS-PCNA-DNA complex, which specifically cleave mismatched DNA, was verified experimentally by single-particle EM method, and the graph-generated model was shown to be consistent with the experimental structure. The results indicated the usefulness of the graph-based methods.

研究分野：　構造情報生物学

キーワード：　生体生命情報学　分子機械　高分子構造・物性　アルゴリズム　生体超分子構造

１．研究開始当初の背景

オーミクス研究の進展に伴って、生体分子構造生物学の焦点は多様な分子によって構成される複雑なシステムの解析にシフトした。今後は、単分子構造から分子複合体および分子システムまでの、多階層を縦断した解析が必要とされる。しかし現在の生体分子データは、各階層に適当な形式でデータベース化されており、統一性はない(例えば、分子構造は原子座標であり、分子システムはネットワークで表現される(Cloot and Marchal, Curr Opin Microbiol, 14, 599 (2011)等)。今後も膨大に蓄積されるデータに対して、階層を縦断した共通のデータ形式が必要である。

２．研究の目的

グラフは大量データの把握が視覚的に容易であり、ある程度まで定量的な情報を保持できる。グラフによるデータ形式の統一により、共通のアルゴリズムが利用できるので、階層縦断的な解析が促進される。しかし現状では複合体構造などの生体分子構造、X 線解析•電子顕微鏡解析•相互作用解析などの実験データでは、グラフは一般的なデータ形式ではない。このギャップを埋めることが本研究の目的である。

３．研究の方法

以下に示すように、生体分子構造のグラフ表現法とデータベースの開発(生分子複合体構造•分子間相互作用•代謝ネットワーク•分子進化情報•実験データ)から開始し、順次それらのグラフを階層縦断的に比較する方法を開発し、最終的には実験データによる検証を行った。

1) 生理的複合体グラフ推定法の開発
2) 代謝ネットワーク上の生体分子構造のグラフによるデータベース化
3) 複合体構造と電顕•X 線密度マップ(実験データ)のグラフマッチ法の開発
4) 異なる階層のグラフを比較するルールの策定、およびそれらを使って生体分子複合体構造の検索•予測を行う方法の開発
5) 共同研究による X 線結晶解析・電子顕微鏡解析によるシステムの検証

４．研究成果

1)生理的複合体グラフ推定法の開発

低分子化合物構造をグラフマッチにより比較する方法(「高速グラフマッチ検索装置及び方法」, 白井　剛, 特許第 5484946 号)を拡張し、タンパク質など高分子の複合体(4 次構造)に適用可能にした。異なる複合体のグラフ(トポロジー)の類似性を比較するために、タンパク質・低分子の組み合わせを高速に検査する方法をタンパク質モデリングデータベース(SIRD)に実装した。実験的検証を進めていた EndoMS-PCNA-DNA 複合体のモデルをこの方法を使って予測したところ、制限酵素に類似した 4 次構造が示された(図 1)。

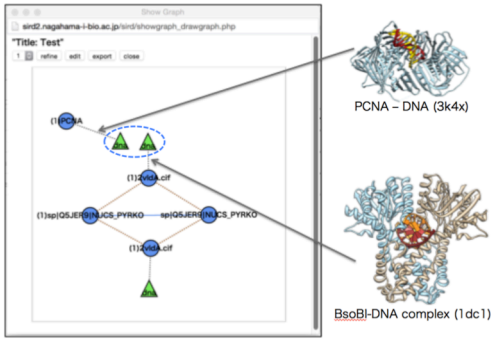

図 1 SIRD による EndoMS-PCNA-DNA のグラフ　ノード(三角や丸など)はタンパク質サブユニット、エッジ(線)は相互作用が存在することを示す。

2)代謝ネットワーク上の生体分子構造のグラフによるデータベース化

上記 1)の方法を PDB 登録の低分子リガン

ドや生薬(Natural drug)の構造分類および構造-機能相関解析に適用したところ、分子構造と薬効の間に有意な相関が示された(主な発表論文 3)。この報告は Molecular Informatics 誌 Best Paper Award 2014 を受賞した。

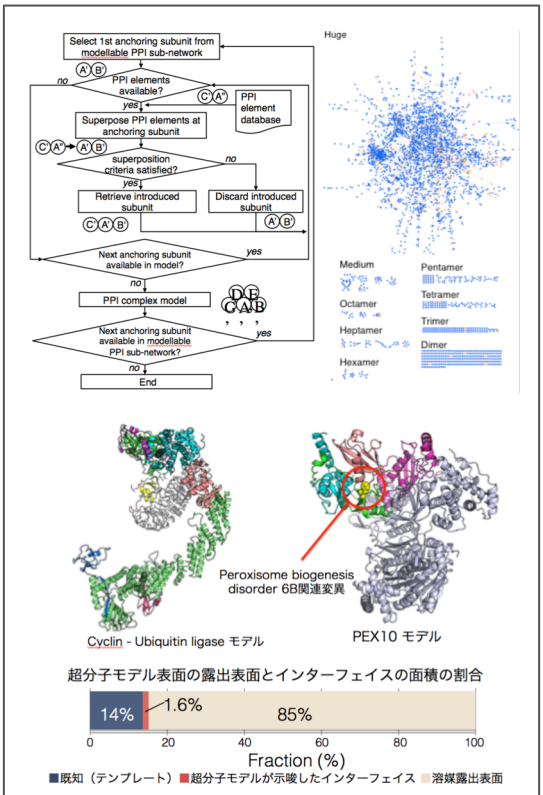

図 2　グラフによる超分子複合体のモデリング
(上左)アルゴリズムの概要, (上右)ヒト PPI ネットワークから抽出された複合体, (下左)PPI ネットワークからモデリングされた cycline-ubiquitin ligase 複合体, (下右)PEX10 モデルとインタフェースの疾患関連変異, (下)モデリングによりインタフェースにマッピングされた分子表面

ヒトゲノムに代表されるように、ゲノムの DNA 配列情報はすでに多数の生物種について完成されている。一方、そこから発現するタンパク質やタンパク質複合体の立体構造情報について完成されている生物種は、現時点で 1 つも存在しない。このゲノムと立体構造の完成度の差は、現在の生物学が抱える重大な情報ギャップの一つである。

このギャップを解決するために、実験データベースからの知識抽出と統合によるグラフベースのタンパク質複合体モデリング手法を開発した。この方法では、タンパク質間相互作用データベース IntAct とタンパク質立体構造データベース PDB を組み合わせて、タンパク質間相互作用ネットワークからタンパク質複合体を網羅的にモデリングする。

複合体モデリングのストラテジーを図 2 に示した。これは、例えばタンパク質 A、B、C からなる相互作用ネットワーク（C-A-B）があった場合に、

a) タンパク質 A と B および A と C の複合体構造がそれぞれ明らかになっている
b) 2 つの複合体において A が大きな構造変化を起こしていない
c) 2 つの複合体構造を A で重ねあわせたときに B と C が衝突しない

これら 3 つの条件を満たせば、2 つの複合体構造をタンパク質 A（重ね合わせに用いるタンパク質をアンカータンパク質と呼ぶ）で重ねあわせることで相互作用ネットワークグラフに基づく 3 量体モデルを構築することができる。

この考え方に基づくアルゴリズムでは、アンカータンパク質による重ね合わせのあと、アンカータンパク質に隣接したタンパク質にアンカータンパク質を移動することで、次々と相互作用するタンパク質をつなぎ合わせて、より大きな複合体構造をモデリングも可能である。この過程は、アンカーにできるタンパク質が無くなる、または新たに導入できるタンパク質候補が無くなるまで繰り返される。

この方法で得られたモデリング可能な相互作用ネットワークには、2,884 個のヒトタンパク質と、それらの間の 5,455 の相互作用が含まれており、上に述べた方法によって 3,197 個の独立したヒトタンパク質複合体モデルが構築された。しかし、IntAct に登録された相互作用のうち 6%は、モデル構造上では直接相互作用せず、1 個以上の別

のタンパク質を介してして相互作用していた。相互作用データは様々な実験手法によるが、例えば免疫沈降法などによる実験では共沈降を指標とするため、間接的に相互作用する例が多数登録されていることが原因と考えられる。

得られた複合体の平均サブユニット数は3.3で、55%（1,756/3,197）の複合体は2量体であった。2量体が多い理由は、全ヒトタンパク質の80%が、それぞれ1つしかインターフェースを持たないためである。

全複合体モデルに存在するインターフェースを解析した結果、3,959個がユニークであり、このうち21%（970/3,959）はIntActに登録されていない新規のモデリングにより予測されたインターフェースであった(図2)。また、複数のタンパク質と相互作用することができるオルタナティブインターフェースは全インターフェースの35%（1,370/3,959）にものぼることが示唆された。構築したモデルの例としてcyclin A2-cyclin dependent kinase-CDK inhibitor1-ubiquitin ligase complexの複合体モデルを図2に示した

さらに、疾患変異データベースMutation@A Glanceを利用して、構築された3,197個のヒトタンパク質複合体モデル上に疾患に関連するSNVによるアミノ酸変異をマッピングしたところ、複合体モデルに含まれるタンパク質領域にマッピングされたもののうち約20%（456/2,319）がインターフェースに存在した。このうち、モデリングによって新規に予測されたインターフェースに存在するものは、12%（54/456）に止まったが、単位面積当たりにマッピングされた数は、既知インターフェースと予測インターフェースでほぼ同等であった(10,000Å^{2}あたり0.13個のSNVが存在する)。

解析した487種類の疾患のうち、26種類の疾患に関連したSNVが複合体モデルのインターフェース領域にマッピングされた。さらにこのうち11種類は、新規に予測したインターフェース（図2）にマッピングされることがわかった。このことから、複合体構造におけるインターフェース破壊は疾患原因として無視できない存在であり、また開発した方法が疾患原因の解析に有効であることが示された。

この成果は(主な発表論文 2)として報告され、生命医薬情報学連合大会 IIBMP2015 Excellent Research Award を受賞した。

3)複合体構造と電顕•X線密度マップ(実験データ)のグラフマッチ法の開発

密度マップを球の集合で近似し、それらの大きさをノード情報、相対配置をエッジ情報とし、同様にグラフ化した分子立体構造と比較する方法を開発し検討した。この方法は予備的評価では十分な性能を示せず、最終年度での実装には至らなかった。

4)異なる階層のグラフを比較するルールの策定、およびそれらを使って生体分子複合体構造の検索•予測を行う方法の開発

1)~2)で示した分子構造をグラフ比較する方法と並行して、より高次のグラフとして分子系統樹(デンドログラム。木構造のグラフである)を相互に比較し「重ね合わせ」する方法を開発した。これは、遺伝子重複により多数のパラログに分化した分子(例としてヘモグロビン)の系統樹1と、基本的にオーソログだけからなる分子(例としてミオグロビン)の系統樹2を比較することで、系統樹1の内部のオーソログ系統を予測することを可能にする(図3)。

5)共同研究によるX線結晶解析・電子顕微鏡解析によるシステムの検証

ゲノムDNA複製においてラギング鎖の合成を行うDNAポリメラーゼ-Fen-リガーゼ-PCNA-DNA複合体の構造をSIRDを使って予測し、電子顕微鏡単粒子解析によるFen-リガーゼ-PCNA-DNA複合体構造と比較した。結果、サブユニットの構造はよく類似していた。また、EndoMS-PCNA-DNA複合体についても同様に予測構造と電子顕微鏡解析の結果を比較したところ、両者が類似していることが示された。これにより開発した方法の有効性が示された（図4)。

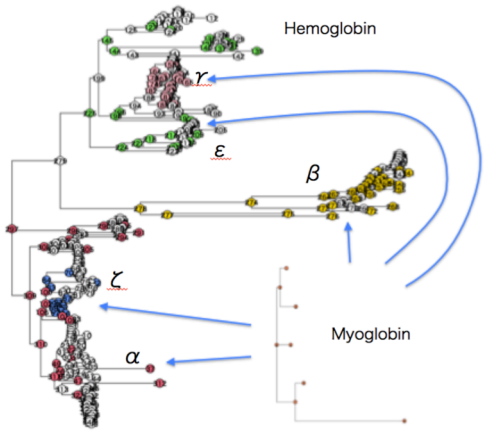

図 3　系統樹グラフの比較によるオーソログ系統の推定　右下のミオグロビンに基づく種系統樹を左上のヘモグロン系統樹とグラフ比較することで、α鎖(赤)，ζ鎖(青)，β鎖(黄)，ε鎖(緑)，γ鎖(桃色)のオーソログ系統が推定できる。

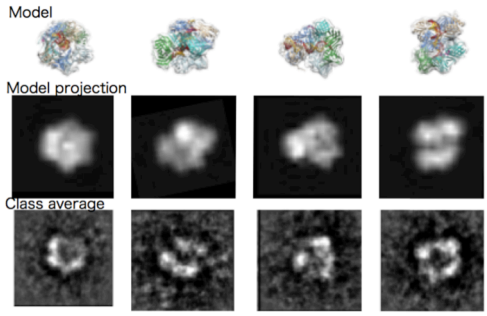

図 4　EndoMS-PCNA-DNA モデルと電子顕微鏡平均化像との比較
Model は SIRD により構築されたモデル、Model projection はモデルから推定される投影像、Class average は実験により得られた平均像である。

後者の結果の一部を報告した論文(主な発表論文 1)　は Nucleic Acid Research 誌の Breakthrough Article に選定された。また EndoMS についての研究は、ミスマッチ DNA を高い特異性で切断するこの酵素の特性から用法特許(産業財産権 PCT/JP2015/07560，平成 27 年 9 月 9 日)の提出につながった。

５．主な発表論文等

〔雑誌論文〕（計 10 件，抜粋 5 件）

1) Ishino S, Oda S, Uemori T, Sagara T, Takatsu N, Yamagami T, Shirai T, Ishino Y, Identification of a mismatch-specific endonuclease in hyperthermophilic Archaea. ***Nucleic Acid Res.*** **44,** 2977-2986 (2016) (査読有 NAR Breakthrough Article) doi: 10.1093/nar/gkw153

2) Tsuji T, Yoda T, Shirai T*, Deciphering Supramolecular Structures with Protein-Protein Interaction Network Modeling. ***Sci. Rep***. **5**, 16341 (2015) (査読有) doi: 10.1038/srep16341

3) Ohtana Y, Azamimi A, Altaf-Ul-Amin M, Huang M, Ono N, Sato T, Sugiura T, Horai H, Nakamura Y, Morita A, Lange KW, Kibinge NK, Katsuragi T, Shirai T*, Kanaya S*, Clustering of 3D structure similarity based network of secondary metabolites to reveal their relationships with biological activities, ***Mol. Inform.***, **33**, 790-801 (2014) (査読有: Molecular Informatics Best Paper Award 2014) http://onlinelibrary.wiley.com/doi/10.1002/minf.201400123/abstract

4) Shirai T*, Saito M, Kobayashi A, Asano M, Hizume M, Ikeda S, Teruya K, Morita M, Kitamoto T, Evaluating prion models on comprehensive mutation data of mouse PrP, ***Structure***, **22**, 560-571 (2014)(査読有) doi: 10.1016/j.str.2013.12.019

5) Nakae S, Ito S, Higa M, Senoura T, Wasaki J, Hijikata A, Shionyu M, Ito S, Shirai T*, Structure of novel enzyme in mannan biodegradation process 4-O-beta-D-mannosyl-D-glucose phosphorylase MGP, ***J. Mol. Biol.*** **425**, 4468-4478 (2013) (査読有) doi: 10.1016/j.jmb.2013.08.002

〔学会発表〕（計 33 件, 抜粋 3 件）

1) タンパク質超分子モデリングによる疾患関連変異の解析, 口頭発表, 辻 敏之, 依田隆夫, 白井 剛, 生命医薬情報学連合大会2015年大会, 2015/10/30, 京都府宇治市 (Excellent Research Award).
2) Development of a protein-protein docking method based on matching of vector-presented amino acid residues, ポスター発表, Hijikata A, Shionyu M, Shirai T, 生命医薬情報学連合大会, 2015/10/29-31, 京都府宇治市 (Poster Award).
3) 白井 剛, Supramolecular modeling pipeline for correlative structural analysis and rational drug-design, ワークショップ 「構造バイオインフォマティクスによる蛋白質機能予測・解析」、第53回日本生物物理学会年会、札幌、2014年9月26日

〔図書〕（計 4 件, 抜粋 2 件）

1) 日本バイオインフォマティクス学会編, バイオインフォマティクス入門, 慶應義塾大学出版会 pp. 1-175 (2015) (編集代表者 白井 剛)
2) 辻 敏之, 白井 剛, ビッグデータからの展開:古代タンパク質解析と超分子モデリング, 生命のビッグデータ利用の最前線 (植田充美 監修), シーエムシー出版, pp. 225-231 (2014) (1 章担当)

〔産業財産権〕

○ 出願状況（計 1 件）

1) 特許：「耐熱性ミスマッチエンドヌクレアーゼの利用方法」, 発明者: 上森隆司, 石野良純, 相良武宏, 石野園子, 山上 健, 白井 剛, 権利者：タカラバイオ株式会社, 国立大学法人九州大学, 学校法人関西文理総合学園長浜バイオ大学, 国内：特願2014-184934 平成26年9月11日, 国外：PCT出願番号: PCT/JP2015/075603 (平成27年9月9日)

○ 取得状況（計 1 件）

1)特許：「高速グラフマッチ検索装置及び方法」, 発明者: 白井 剛, 権利者：学校法人関西文理総合学園長浜バイオ大学, 国内：特許第 5484946 号 平成 26 年 2 月 28 日 (出願番号：特願 2010-31526 平成 22 年 2 月 16 日), 国外：PCT 出願番号：PCT/JP2011/053280(平成 23 年 2 月 16 日)

〔その他〕

1） 日刊工業新聞「遺伝性脳症の原因遺伝子変異を解明 京大」 2014 年 7 月 15 日
2） 近江毎夕新聞「脳症の原因遺伝子新たに発見 長浜バイオ大が共同研究で成果」2014 年 7 月 13 日

６．研究組織

(1)研究代表者 白井 剛（SHIRAI TSUYOSHI） 長浜バイオ大学 ・バイオサイエンス学部 ・教授 研究者番号: 00262890

(2)研究分担者 大山 拓次 （OOYAMA TAKUJI） 山梨大学・医学工学総合研究部 ・准教授 研究者番号: 60423133

(3)研究分担者 真柳 浩太 (MAYANAGI KOUTA) 九州大学・生体防御医学研究所・助教 研究者番号:50418571

(4)連携研究者 石野 良純 (ISHINO YOSHIZUMI) 九州大学・農学研究科・教授 研究者番号: 30346837