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　２９　年　　４　月　２６　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2015～2016
課題番号：　１５Ｋ２０９２９

研究課題名（和文）タンパク質電子構造DBシステムの構築

研究課題名（英文）Construction of Electronic Structure Database of Proteins

研究代表者
平野　敏行（Hirano, Toshiyuki）

東京大学・生産技術研究所・助教

研究者番号：６０４５１８８７

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：　タンパク質電子状態計算において、計算構造モデリングの自動化とQCLO法に基づく電子状態計算の自動化に関する研究開発を行い、タンパク質電子状態データベースへの計算事例追加をより加速させるための基盤技術を確立した。YAML形式の入力ファイルを採用することで、プログラミングなしに安全かつ簡便にQCLO法を利用したタンパク質電子状態計算が可能となった。開発したソースコードはインターネット上に公開済みである。

研究成果の概要（英文）：In the calculation of protein electronic structures, we have developed a software related to automation of computational structure modeling and automation of electronic structure calculation based on the QCLO method, and established fundamental technology to further accelerate the addition of calculation examples to the protein electronic structure database. By adopting the YAML format input file, it became possible to calculate protein electronic structure safely and conveniently using the QCLO method without programming. The developed source code has been published on the Internet.

研究分野：　量子化学

キーワード：　蛋白質　ハイパフォーマンス・コンピューティング　電子状態

１．研究開始当初の背景

電子状態計算は、シュレーディンガー方程式に基づき物質の電子状態を明らかにするため、物質の物性・反応性を理解する上で極めて強力なツールである。我々は、タンパク質をターゲットとした量子化学計算プログラム ProteinDF を開発し、インターネット上で GPL ライセンスのもと公開している[http://proteindf.github.io/]。ProteinDF はタンパク質をまるごと扱った、カノニカル分子軌道計算を行えることが特長であり、反応に関与するフロンティア軌道(分子軌道)、エネルギー準位、静電ポテンシャルなど様々な物理量が求められる。また、ガウス型局在化基底関数に基づく密度汎関数法を採用しており、金属タンパク質の計算が可能なほか、物性・反応性に寄与する原子が特定できるため、タンパク質やリガンドの設計に貢献できるものと期待される。

実際に ProteinDF を用いて、Yagi らは数種の甘味タンパク質変異体のフロンティア軌道・電子密度と甘味の強弱の関係を明らかにしている[Y. Yagi, et. al., JSST, 2012]。また、Tokita らは亜鉛置換シトクロム c とシトクロム b の励起電荷移動機構の違いを明らかにし、タンパク質がデバイスとして利用できる可能性を示した[Y. Tokita, et. al., Angew. Chem., 2011]。

タンパク質の機能解析や機能強化設計において、変異体やファミリーなど複数の関連したタンパク質の間で、バイオインフォマティクスを用いて配列情報や立体構造を比較検討するのが常套手段である。さらに電子状態も比較することができれば、タンパク質物性の仕組みをより効果的に理解できると考えられる。

本研究では、物性を左右するタンパク質の電子状態も生物情報の一つとして位置づける。タンパク質立体構造データベース(PDB)と同様に蓄積・共有化し、バイオインフォマティクス技術と融合することは、戦略的なタンパク質研究のツールとして必須であると着想した。実用化には、PDB に登録されているタンパク質の電子状態計算を行い、計算結果を DB に登録する仕組みが求められる。とはいえタンパク質の電子状態計算に必要な生物学的・量子化学的知識・技術は極めて専門性が高く、達成は困難である。膨大なタンパク質構造から電子状態を計算し蓄積するプロセスが、自動的に実行される環境を整備する必要がある。

２．研究の目的

これまでの研究成果をもとに、電子状態計算構造の作成(モデリング)・電子状態計算などタンパク質電子状態シミュレーションに関するプロセスの自動化に関する基盤研究を行い、タンパク質電子状態データベース開発に必要なソフトウェアの研究開発を行う。

① 電子状態計算の前処理であるタンパク質モデリングの自動化

タンパク質モデリングとは、以下の一連の操作である。PDB から取得したタンパク質構造から、欠損原子・水素原子を付加、構造緩和を行い、異常接近原子対のチェックなど、電子状態計算に耐えうる計算構造を算出する。これらの作業を、自動化する。また、分子軌道計算の自動化を行う。分子軌道計算に必要な初期電子密度を用意する必要があるが、すでに予備研究として開発している QCLO 法[T.Hirano, et.al., J.Chem.Phys., 127, 184106, (2007)]を用いることで効率的に求められる。

②トラブルのない安定したタンパク質電子状態計算の自動解法に関する研究

大規模分子の電子状態計算において安定した収束解を得るためには、収束解になるべく近い、高精度の初期値が求められる。特にタンパク質電子状態計算において密度汎関数法を用いて求める場合は、大規模になればなるほど HOMO-LUMO ギャップが狭くなり、収束解を得ることが難しくなる。QCLO 法は、大規模分子をサブユニットに区分し、それぞれの電子状態計算結果から高精度な初期値を作成することができる方法である。QCLO 法は大規模分子の電子状態計算において非常に有用な計算法であるが、大規模になればなるほど試行錯誤も増え、計算手順が複雑となり、人の手には負えなくなる課題がある。本研究では、QCLO 法による電子状態計算をより使いやすく、自動的に行える基盤技術の開発を行う。自動的かつ容易に電子状態計算が得られる環境が整うことで、タンパク質電子状態計算の計算事例を増やし、データベースの増強に繋げる。

３．研究の方法

PDB に登録されているタンパク質構造から、量子化学計算に資する構造を自動的に作成するモデリングと、その構造を元に安全に収束解が得られる電子状態計算法の開発に関する基盤研究を行う。これらを元にタンパク質電子状態データベースに向けた計算事例の蓄積・DB 構築開発と解析ツールに関する研究を行った。

３．１．タンパク質モデリングの自動化

PDB に登録されている構造データの多くは、X 線結晶構造解析と NMR 解析によるものである。水素原子の付加や欠損部位の処理、結晶水やヘテロ原子に対する処置、溶媒の取り扱い方など、量子化学計算に至るまでに必要なモデリングは多段階に及んでいる。厳密なモデリングが必要な理由は、1 原子たりとも矛盾のある分子構造を許さない量子化学計算の要求によるものである。

タンパク質は原子数が多く、異常接近原子対や構造の歪みに対するチェックを、すべて

人の手で行うには限界がある。しかも、DB として有意なデータ数を用意するためには、大量のタンパク質量子化学計算を達成しなければならない。本研究では、PDB 構造から量子化学計算に至るまでのモデリング手順をプログラミングし、インタラクティブ性の無い、自動バッチ処理が可能なモデリング技術の確立を目指した。

モデリングにより付加された原子や PDB 構造に起因して、計算構造が構造的歪みや異常接近原子対を含み、量子化学計算が失敗することがある。本研究では、短時間で効果的なモデリングを達成するため、分子力学法を用いて計算構造の問題を取り除く手法を確立する。この時、各種溶媒の処理も同時に行う。

ヘテロ原子は分子力学法で用いる力場が用意されていない場合が多い。Chemical Component Dictionary に登録されている分子に関しては、あらかじめ GAFF など一般分子用の力場を自動生成しておく。計画通り上手く構造緩和が行かない場合は、半経験的分子軌道法による部分構造最適化を用いて局所的な構造の歪みを取り除くことも検討する。

３．２．タンパク質量子化学計算の自動化

大量の量子化学計算をスムーズに行うために、構造に問題が無ければほぼ 100%失敗せずにタンパク質分子軌道計算を安定に達成する、自動計算ロボットを完成させる。

安定して収束する量子化学計算の達成には、適切な初期電子密度が重要である。低分子の計算で用いられる Hückel 法をタンパク質にそのまま適用した場合、初期電子密度と収束解との差が大きく、用いることができないことを確認している。本研究では、タンパク質の初期値の作成に QCLO 法[H.Kashiwagi, et.al., Mol. Phys., 101, 81 (2003)]を用いる。QCLO 法は、まず 1 残基ごとの量子化学計算を行い、その結果から 3 残基の初期値を作成、3 残基の結果から数残基の初期値を作成・・・と順次計算領域を拡大し、結果として分子全体の良好な初期値を作成する方法である。我々は、イオン結合、ジスルフィド結合、および 2 次構造を考慮に入れることで、更に収束効果が高くなることを確認している[T.Hirano, et.al., J.Chem.Phys., 127, 184106, (2007)]。本研究では複数の計算領域拡大のシナリオのうち、評価点が高いものをリストアップし、自動的に量子化学計算を試行錯誤する仕組みを作成したうえで、ヘテロ原子を含むタンパク質の量子化学計算の自動化に着手し、殆どのタンパク質の自動計算が行えるシステムを構築する。

自動計算がうまくできない場合に備えて、計算・処理のログ出力機能や、エラーの発生をユーザーに通知する機能を実装する。

３．３．タンパク質波動関数 DB の計算事例

PDB に登録されている小さなタンパク質から順に、前述の手法を用いてタンパク質電子状態計算を自動的に行う。大規模なタンパク質の電子状態計算を行う場合は、必要に応じて並列計算機を利用した。

４．研究成果

４．１．計算シナリオ(YAML の採用)

これまで開発した自動計算プログラムは、計算実行者自ら Python スクリプトを記述する必要があった。きめ細やかな処理内容を記述できるメリットがあるものの、Python 言語の他、QCLObot ライブラリの使用法を学ぶ必要があり、使いこなすには難しいものがあった。そこでより簡単に計算・処理内容を QCLObot に指示できる方法として、入力ファイルを YAML 形式のテキストファイルにするよう実装・改良した。YAML 形式ファイルは、リストや辞書などの構造化データに対し、インデントを用いて記述することできる。内容はテキストファイルであるため、可読性があり、テキストエディタで容易に編集可能である。このため、Windows/MacOS/Linux などの利用環境を問わず、簡単に編集することができる。

前述の通り、YAML 形式では構造化データを取り扱うことができるため、QCLO 法における計算単位であるフレーム分子やフラグメントをツリー型で記述できるようにした。フレーム分子やそれを構成するフラグメントには、ユーザーが自由に名前付けできるようにし、計算構造や電子状態を後の計算で再利用できるようにした。QCLO 法ではフレーム分子とフラグメントは親子関係になっており、QCLObot の入力ファイルでは、YAML 形式が持つ構造化データ表現を用いて親子関係を表現している。これにより、QCLO 法の処理を考える上で、プログラミングの問題に悩まされることなく、QCLO 法におけるフレーム分子のスケールアップ計算手順のみに専念できるようになった。

オブジェクト指向言語である Python の特長を活かし、クラスを用いてフレーム分子やフラグメントを記述している。フラグメントは再帰的にグループ化できるように実装し、プログラミングを省力化した。

QCLO 法により初期値を作成する際は、入力したフラグメント情報に基づき自動的に計算される。参照元となるフラグメントからその親となるフレーム分子に局在化軌道計算の指示が渡り、局在化軌道計算の後、各分子軌道のフラグメント分配評価関数に基づき振り分けを行い、それからフラグメントの QCLO 計算を行う。計算すべき対象フレーム分子を構成するフラグメントの QCLO を集めて初期値を作成する。参照元のフラグメントは、計算対象のフレーム分子情報を参照することで、高精度な QCLO が作成できるようになっている。

フレーム分子の電子状態計算において必要な電子数、軌道情報は自動的に計算し、ProteinDF 入力ファイルを作成する。イオン

などの場合は電荷を指定することで、適切に処理できるように実装している。
これらの手続きを全て正しく人の手で行うことは難しいが、QCLObot プログラムはこれらを自動化することに成功した。

４．２．末端モデリング処理機能

アミノ酸残基を計算する際には、N 末端側・C 末端側それぞれに末端処理しなければならない。隣接するアミノ酸残基の主鎖原子座標を元に、自動的に適切な N-メチル基・アセチル基座標を計算・付加する機能を実装した。
ヘテロ分子の場合も同様に QCLO 法におけるフレーム分子電子状態計算が達成できるように、ダングリングボンドに対し、隣接原子座標を参照として、水素またはメチル基の位置を計算、付加する機能を実装した。

４．３．変数・テンプレート機能

YAML 形式の入力ファイルを採用した QCLObot では、タンパク質座標データ(PDB ファイル、または ProteinDF オリジナルデータ形式)以外に、一般的な量子化学計算パッケージで用いられているような、原子座標を羅列した入力形式もサポートしている。YAML のリスト表現を用いて原子種と xyz 座標を記述できるようになっている。これにより、特殊なヘテロ分子や途中計算に必要なダミー原子、カウンターイオンなどの入力が可能になった。xyz 座標などの数値は、変数として入力ファイル中に指定することができる。したがって、原子間距離や角度を徐々に変更する計算など、幾通りのパターンの電子状態計算が一つの入力ファイルで自動的に作成・計算できるようになった。
YAML 形式入力ファイルにおける、変数のサポートは、タンパク質を構成する一つ一つのアミノ酸残基の計算においても効果を発揮する。各アミノ酸残基に対して N-メチル基・アセチル基を付加する処理は、どのアミノ酸にも共通で、かつ繰り返し行う処理である。そのため、YAML 入力ファイルにテンプレートをサポートし、残基番号を変数として繰り返し処理できるように実装した。テンプレートを使うことで、ヒューマンエラーの原因となる入力すべき事項が少なくなり、省力化され、見通しが良くなった。

４．４．モデリング用 QCLObot の作成

QCLO 法における YAML 形式入力ファイルの採用は、計算手順を記録しておくだけでなく、再現性の確保の視点からも利点がある。QCLO 法で利用した YAML 形式入力ファイルを量子化学計算に適したモデリング作業の処理に適用した。
水素付加には reduce プログラムを採用した。X 線結晶構造解析結果をはじめとする PDB データに対し、ヘテロ分子も含め水素付加することができる。
特に結晶構造にはパッキングフォースなどの力が加わり、構造が歪む・異常接近原子が見られるなどの、電子状態計算が失敗する要因となるような異常が見られることがある。電子状態計算構造を作成するために、構造緩和を行う。構造緩和作業をすべて量子化学計算に基づいて行うことは難しい。そこで、分子力学法・分子動力学法を用いて構造緩和を行った。構造緩和に使用したプログラムとして Amber を使用した。プログラム化に際し、計算指示内容をカプセル化し、Amber プログラム以外の GROMACS 等のオープンソース分子動力学法計算プログラムも後日使用できるように設計した。

４．５．タンパク質電子状態 DB の事例追加

本研究で開発したモデリング・電子状態計算自動解法プログラムのテストとタンパク質電子状態 DB への事例追加を目的として、具体的な事例計算を行った。限られた計算時間・計算資源により、多くのサンプルの計算成功事例を上げることはできなかったが、少ないサンプル数ながらも、モデリングの自動化・電子状態計算の自動化を概ね達成できた。ヘテロ分子の対応など、まだ人手を要する部分はあるものの、基盤となる計算手法は確立できた。タンパク質電子状態 DF として、より多くの事例を追加するためには、大規模計算機の活用の他、電子状態計算エンジンのさらなる高速化が必要である。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① 金泰煥, 平野敏行, 佐藤文俊, “カノニカル分子軌道計算に基づく線形回帰法を用いたタンパク質原子電荷の開発”, 生産研究, 68, 213-217 (2016).
② 紀平昌吾, 平野敏行, 佐藤文俊, “量子化学計算によるオキシトシンの安定構造に関する研究”, 生産研究, 68, 219-223 (2016).

〔学会発表〕（計 14 件）

① 平野敏行, 佐藤文俊, “自由なフラグメント分割可能な QCLO 法プログラムの開発”, 理論化学討論会 (2015)
② 金泰煥, 平野敏行, 佐藤文俊, “機械学習を用いたカノニカル分子軌道計算に基づく新規タンパク質形式電荷に関する研究”, 理論化学討論会 (2015)
③ 平野敏行, 王笛申, 佐藤文俊, “第三世代密度汎関数計算法の進展”, 分子科学討論会 (2015)
④ 金泰煥, 平野敏行, 佐藤文俊, “線形回帰法を用いたタンパク質原子電荷に関する研究”, 分子科学討論会 (2015)
⑤ 紀平昌吾, 平野敏行, 佐藤文俊, “大規模分子の構造最適化の収束法に関する研

究”，分子科学討論会（2015）
⑥ 平野敏行, “量子化学シミュレーションで観察するタンパク質”, 東海コンファレンス, 長野, (2015)
⑦ 平野敏行, “分散メモリ型超並列計算機に向けた大規模カノニカル分子軌道計算法の開発”, HPCC2015, 東京（2015）
⑧ 平野敏行, 佐藤文俊, “効率的なタンパク質カノニカル分子軌道計算を目指して”, ProteinDF/ABINIT-MP 研究会, 東京 (2015)
⑨ T. Hirano, F. Sato, “A Third-generation Density Functional Calculation Program: ProteinDF”, Pacifichem 2015, 米国（2015）
⑩ T. Hirano, F. Sato, “An effective grid-free approach based on the third-generation density-functional-theory calculation method”, Sanibel Symposium, 米国 (2016)
⑪ 平野敏行, 佐藤文俊, “グリッドフリー密度汎関数計算におけるエネルギー勾配計算”, 分子科学討論会, (2017)
⑫ 平野敏行, 佐藤文俊, “罰則付き回帰法に基づくタンパク質原子電荷の計算と性質”, 第 39 回ケモインフォマティクス討論会（2017）
⑬ T. Hirano, F. Sato, “A theoretical study of glucose oxidase using canonical molecular orbital calculation”, Sanibel Symposium, 米国（2017）
⑭ 佐藤文俊, 平野敏行, “タンパク質の美しき電子の世界”, 日本農芸化学会 2017 年度大会, 京都（2017）

〔その他〕
ホームページ等
①https://proteindf.github.io/proteindf/proteindf/
②https://proteindf.github.io/proteindf/qclobot
③http://satolab.iis.u-tokyo.ac.jp/

６．研究組織
(1)研究代表者
平野 敏行（HIRANO Toshiyuki）
東京大学生産技術研究所 助教
研究者番号：60451887