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

令和　２ 年　 ５ 月 ２８ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００３９５

研究課題名（和文）タンパク質電子構造DBシステムの拡充

研究課題名（英文）Expansion of Protein Electronic Structure DB System

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）　　3,400,000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、電子状態計算の高速化・効率化および計算構造モデリングと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った。GPUによって電子状態計算の高速化し、メモリマップドファイルによって限られた計算機資源でのタンパク質電子状態計算の達成を実現した。QCLO法に基づく自動計算プログラムの機能追加・改良を施し、入力コードの保守性・視認性を向上させた。電子状態計算に基づく相互作用解析ツールを開発し、小規模タンパク質の実証計算によりその有用性を確認した。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。

研究成果の学術的意義や社会的意義
実験的に得られたタンパク質構造をもとに、タンパク質電子状態データベースの構築・拡張をおこなうため、タンパク質電子状態計算をより自動化し、省力化する研究開発を行った。タンパク質電子状態計算を高速化・効率化し、自動計算の簡略化・機能追加を行った。相互作用解析機能を追加し、基質やアミノ酸残基間の相互作用を可視化することができるようになった。
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 protein electronic state calculation, we have improved the speed and efficiency of the electronic state calculation, and the structure of the calculation and the automation of the electronic state calculation based on the QCLO method: the electronic state calculation was accelerated by the GPU, and achieved with the limited resources of the memory-mapped file. We have developed an interaction analysis tool based on electronic state calculations and confirmed its usefulness in small-scale protein experiments.
The developed source code has been available on the Internet.

研究分野：計算科学

キーワード：タンパク質　電子状態　量子化学

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

電子状態計算(量子化学計算)は、シュレーディンガー方程式に基づき第一原理的に物質の電子状態を明らかにすることができるため、物質の物性・反応性を理解・解析する上で極めて強力なツールである。我々は含金属タンパク質をターゲットとした密度汎関数法計算プログラム 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 の特長は、領域分割することなく、タンパク質全体をまるごと量子化学的に扱う、正準(カノニカル)Kohn-Sham 分子軌道計算を行えることである。カノニカル分子軌道計算により、タンパク質の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直接的に求められる。大規模分子のカノニカル分子軌道の達成は大変困難であったが、申請者らは分散並列計算機に適した第三世代密度汎関数計算法[T. Hirano, et. al., PCCP, 2014]を考案し、これを ProteinDF に実装することによってこれを克服した。

一方、タンパク質の実験科学的手法によって得られたアミノ酸配列や立体構造、代謝ネットワークなどの知見は、情報科学の強力なデータ処理・アルゴリズムを活用してバイオインフォマティクスとして重要な研究・学術領域を確立しており、その有用性は疑う余地も無い。生命実験科学と情報科学の融合によって大きく発展したバイオインフォマティクスは、これから理論科学・シミュレーションとの融合によってさらに研究分野を広げ、生物学におけるさらなる有用性を確立するものと期待している。

２．研究の目的

本研究を通して、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たいと考えている。タンパク質立体構造データベース(PDB)と同様に、タンパク質の電子状態情報を蓄積・共有化する。タンパク質電子状態情報とバイオインフォマティクス技術とを融合することは、戦略的なタンパク質研究のツールとして必須である。しかし、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。膨大な登録数である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った。

本研究では、タンパク質のモデリング・電子状態計算など電子状態計算に関するプロセスの自動化に関する基盤研究を発展させ、タンパク質電子状態 DB の公開とバイオインフォマティクスに展開する解析用ソフトウェアの基盤研究を行うことを目的とした。

３．研究の方法

(1) タンパク質量子化学計算の自動化に関する研究

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失敗せずにタンパク質分子軌道計算を安定に達成する、いわば自動計算ロボットを用意する必要がある。

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., (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。

我々はこの QCLO 法を発展させ、タンパク質のイオン結合、ジスルフィド結合、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計算手法を開発した[T.Hirano, et.al., J.Chem.Phys., 127, 184106, (2007)]。これまでの研究において、この作業をヘテロ原子非保有のタンパク質について自動化する基本プログラムは既に作成済みである。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、エラーの発生をユーザーにメールなどで通知するとともに、失敗した理由の解明をスムーズにするために、収束しなかった分子のエネルギー変遷図、フロンティア軌道近傍の分子軌道図などを自動作成する。原因が分子構造の場合は、それをモデリングに自動フィードバックする機構も検討する。

(2) タンパク質波動関数 DB の開発とデータ作成

100 残基クラスのタンパク質量子化学計算からは、数 GB の量子化学計算(波動関数)データが得られる。大量の(行列)データのうち、係数行列さえあれば様々な物理量を、時間がかかるものの別途再計算することができる。データ保存量・転送量・再計算に伴う計算量を考慮し、計算データから必要なデータを取捨選択し、アーカイブする機構を作成する。このとき、計算機の機種依存性を排除する。

また、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。大規模な計算は適宜並列計算機を借用する。タンパク質電子状態計算の高速化・効率化を図り、また電子状態データのより利用しやすいデータ解析法を開発する。

４．研究成果

(1) ProteinDF プログラムの高速化

① コレスキーベクトル保存方法の改良

第三世代密度汎関数計算法の計算では、SCF 繰り返し計算前に一度だけ分子積分を行う。分子積分は計算精度を考慮した上で、ランクを最小にしたコレスキーベクトルの形で保存される。コレスキーベクトルは SCF 繰り返し計算中にベクトルの形で切り出され、クーロン項や Fock 交換項算出の行列計算に利用される。このとき、SCF 繰り返し計算前の低ランクコレスキーベクトル計算方向と、SCF 繰り返し計算中の計算方向がそれぞれ行方向、列方向と異なる。CPU 計算が高速化し、メモリ速度との解離が大きな今日の計算機において、キャッシュヒットは高速計算に必須である。

そこで本研究では ProteinDF に改良を加え、コレスキーベクトルをブロック化した帯行列クラスで扱うことにした。はじめは行方向に格納したメモリアクセスで低ランクコレスキーベクトルの計算効率を下げることなく、また後の列方向のコレスキーベクトルのアクセスに向けて、メモリ配置の転置操作をブロック化によって効率化した。コレスキーベクトルは大きなものだと数 TB にも及ぶため、mmap を利用したメモリマップド I/O による転置操作をサポートした。mmap の利用は、SSD の利用によりより高速化される。

② GPU による第三世代密度汎関数計算法の高速化

タンパク質のカノニカル分子軌道計算を行うに際し、計算時間の短縮はデータベースの構築において最も効果的である。分子軌道計算のボトルネックとなる SCF 繰り返し計算において、複雑で均等なタスク分配が困難な分子積分を必要とせず、行列演算のみで計算が達成できる第 3 世代法を開発している。行列計算を GPU 上で行うことで計算時間を短縮することに達成した。

GPU を用いる行列演算ライブラリとして、OpenCL を用いた ViennaCL ライブラリを採用した。同時に SIMD 命令を駆使し、コンパイラによる最適化を活かした C++テンプレートライブラリ Eigen も採用した。C++で記述した ProteinDF への実装にあたり、線形演算インターフェースを共通化し、少ないコードの修正で多様なライブラリへ対応するとともに、電子状態計算に関わるアルゴリズムの変更を最小限に留めることができた。

いくつかの実証計算を行なった結果、GPU を用いた高速化を確認した。これにより、第三世代密度汎関数計算法に基づくカノニカル分子軌道計算において、GPU の有用性を示した。本研究によって開発した ProteinDF プログラムのソースコードは、インターネット上で入手可能である[https://github.com/ProteinDF/ProteinDF]。

(2) タンパク質モデリングの自動化・改良

タンパク質のカノニカル分子軌道計算を達成するためには、水素原子一つの過不足も許さないモデリングが求められる。タンパク質のほとんどは 20 種類のアミノ酸のペプチド結合により構成されているため、ヒスチジンや荷電アミノ酸などの一部アミノ酸残基を除いて、ペプチド鎖のモデリングはそれほど難しくない。一方で多種 多様な物性をもたらすタンパク質には、ペプチド鎖の他にヘテロ分子を含むものが多い。ヘテロ分子は種類も豊富である。低分子の量子化学計算と同様に、タンパク質のモデリングも本来ならば量子化学計算により行われるべきである。しかし、量子化学計算の高い計算 コストゆえに現実的に実施困難である。そこで、もう少し計算コストのかからない古典力場でタンパク質のモデリングを代用した。とはいえ、ヘテロ分子のモデリングにも分子力場が必要である。一般的なアミノ酸残基は、種類も少なく、あらかじめよく研究された分子力場が用意されている一方で、ヘテロ分子は種類も多く、分子力場が用意されていない。本研究では、PDB に登録済みのヘテロ分子に対し、自動的かつ網羅的に分子力場を計算する仕組みを構築する必要がある。

QCLObot は、タンパク質のカノニカル分子軌道計算における初期分子軌道の作製において、擬カノニカル局在化軌道(QCLO)計算法に基づく自動計算プログラムである。入力形式として、テキスト形式で構造化モデルを記述できる YAML 形式を採用することによって、複雑な QCLO 作成をプログラミングすることなく、QCLO 計算に伴う手順とモデルの階層化を表現することができる。また、繰り返し何度同じ操作を行なっても同じ結果が得られる、冪等性を確保するように設計した。冪等性は試行錯誤が必要なタンパク質カノニカル分子軌道計算において、必要不可欠で便利な機能である。本研究ではこの QCLObot プログラムをベースとして、タンパク質実験構造から量子化学計算モデルを作製するモデラープログラム(QCLObot_modeler)を作製した。構造緩和の手段として、分子動力学法プログラムである Amber パッケージを利用した。分子動力学計算における力場、ステップ数、温度などの諸条件・パラメータは、タンパク質によって変更すべきである。したがって、各パラメータはデフォルト値の他、QCLObot_modeler 入力ファイルによって変更可能としている。QCLObot_modeler は分子動力学プログラムのラッパーとして機能するため、入力ファイルのキーワード・値の変更なしに、後に Gromacs など他の分子動力学パッケージを利用することも可能である。QCLObot_modeler プログラムを利用して数十残基相当のタンパク質カノニカル計算を行い、モデリングが確実に行われることを確認した。しかし、複雑なヘテロ分子を含むタンパク質構造データのサポートは、まだ未完成である。結晶を作成するときに必要な界面活

性剤や混入する塩などは、タンパク質電子状態を観察・解析する上では不必要と考えられるが、モデリングや電子状態計算に必要・不要の判断を一括りに対処することは難しい。また、ヘテロ分子の分子力場の多くは分子動力学パッケージに用意されていないケースがほとんどであり、モデリング時に自動作成する必要がある。ヘテロ分子の水素付加等の処理も必要であり、包括的な対処にはまだ時間がかかる。モデリング・カノニカル分子軌道計算へのプログラムの自動化はまだ不完全であり、今後の開発・作業事項である。

(3) QCLObot プログラムの高機能化

QCLO 法に基づく自動計算プログラム QCLObot プログラム本体のアップデート・改良を行った。まず、template 機能の追加・拡充を行い、変数を利用した繰り返し処理・条件分岐処理の実装を行った。これにより、数百を超えるタンパク質の QCLO 法計算においても、長々と計算指示することなく、自動的に繰り返し計算することが可能となった。

また、include 文を用意し、外部ファイルの入力をサポートした。これにより、定型のヘテロ分子などの QCLO シナリオを再利用することが可能となった。アップデートした QCLObot プログラムはインターネット上に公開している [https://github.com/ProteinDF/QCLObot]。

(4) 相互作用解析ツールの開発

タンパク質の量子化学計算に対する期待の一つとして、基質およびアミノ酸残基間の複雑化な相互作用の理解が挙げられる。タンパク質を対象とした大規模分子のカノニカル分子軌道計算プログラム ProteinDF と初期値自動計算プログラム QCLObot により、タンパク質カノニカル計算にかかる労苦はかなり改善されたが、依然としてその計算量は膨大で収束困難であり、多大な労力を要する作業である。タンパク質設計において有用なアミノ酸残基間の相互作用を、実験化学的手法で観察することは難しく、計算化学によって相互作用を見積もることが期待されている。任意の分子領域における相互作用エネルギーを見積もる方法として、Energy decomposition analysis 法 [K. Morokuma, J. Chem. Phys., 55, 1236 (1971)]など様々な手法が知られている。その殆どは、全系のカノニカル分子軌道計算の他に、相互作用計算対象となる領域(サブユニット)のカノニカル分子軌道計算が必要である。タンパク質のカノニカル分子軌道計算の場合、高コストであるタンパク質全体のカノニカル分子軌道計算に加えて、サブユニットのカノニカル分子軌道計算を行うことはさらに骨の折れる作業である。大規模系カノニカル分子軌道計算における相互作用の描像を、なるべく簡便に把握できる解析手法が欲しい。

そこで本研究では、相互作用解析手法として Energy density analysis (EDA)法 [H. Nakai, Chem. Phys. Lett., 36, 73 (2002)]を採用した。EDA 法では Mulliken population と同様の計算方法で、エネルギー行列要素からサブユニットに属するエネルギーとその構成要素(一電子項やクーロン項など)が得られる。密度汎関数法で特徴的な交換相関エネルギーの行列要素計算法として、grid-free 法を採用した。本研究では、EDA 法に基づく相互作用解析法を実装し、小規模タンパク質を始めとする具体的な系における相互作用計算と評価を行った。

複雑な塩橋とジスルフィド結合を持つタンパク質(図 1-a)における電子の全エネルギーに関する相互作用解析の結果を図 1-c に示した。青色は負の相互作用エネルギー、赤色は正の相互作用エネルギーを示し、それぞれ親和、反発の相互作用を意味している。EDA 法の計算式から、全成分の総和は全エネルギーと一致する。対角成分は自己相互作用エネルギーとみなすことができ、全エネルギーの大部分を占めている。副対角成分の青色は、ペプチド結合による安定化と考えられる。非対角成分には、塩橋に関連するアミノ酸残基とジスルフィド結合に関与するアミノ酸残基の相互作用が観察できた。塩橋は、荷電アミノ酸の位置関係で議論することができたが、三次元座標ではそれらを定量化することは難しかった。分子軌道計算結果から Mulliken 電荷を求めて議論することも可能であるが、本解析ではさらに複雑に絡み合ったエネルギー分布を得ることができた。Mulliken 電荷では一見クーロン引力が働くように予想されたアミノ酸残基同士でも、負の相互作用がある場合があることが観察できた。ジスルフィド結合まわりでは、システイン同士のジスルフィド結合による安定化が観察できた一方で、隣接するアミノ酸残基の不安定化が観察できた。主鎖・側鎖の相互作用解析により、この不安定化は主鎖骨格のねじれによるものと考察できた。

特徴的な 2 次構造(アルファヘリックス)を持つタンパク質(図 2-b)の相互作用解析結果を図 2-d に示した。アルファヘリックスを構築する主鎖骨格の水素結合に起因する特徴的な相互作用パターンが非対角項に観察できた。EDA 法を利用することにより、このような解析パターンが 1 度のカノニカル分子軌道計算から得られた。アミノ酸残基間に働く複雑な相互作用を可視化することにより、タンパク質の分子軌道計算結果がタンパク質の機能解析や酵素の性能向上、薬剤設計に寄与するものと期待している。

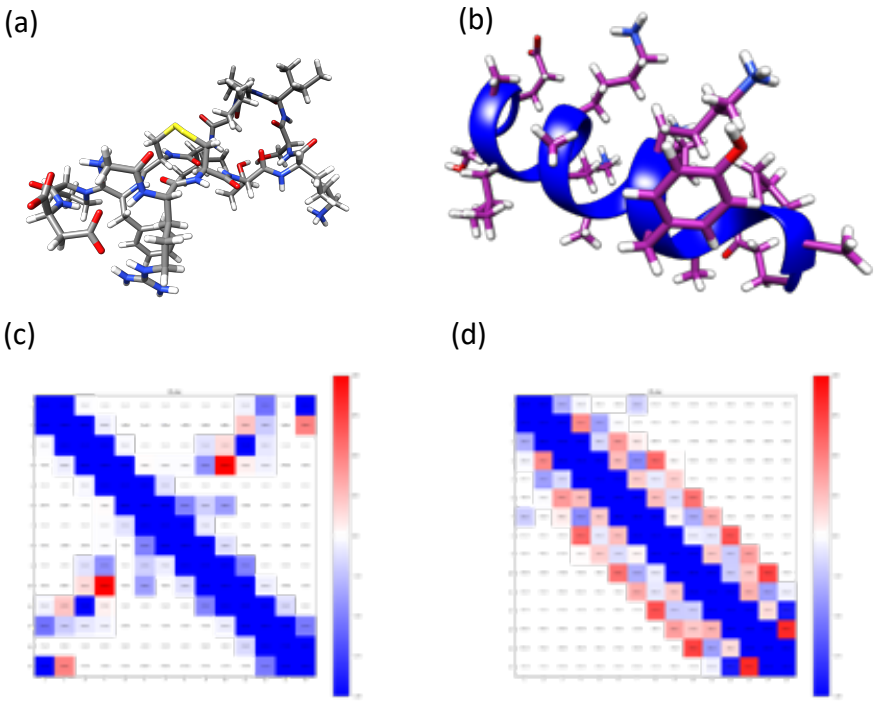

図 1 タンパク質の分子図と相互作用解析結果

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 EGUCHI Haruki、HIRANO Toshiyuki、SATO Fumitoshi	４．巻 17
２．論文標題 Study on the Cloud-like Visualization Model of Protein Molecule Orbitals by the Rejection Method	５．発行年 2018年
３．雑誌名 Journal of Computer Chemistry, Japan	６．最初と最後の頁 189～192
掲載論文のDOI（デジタルオブジェクト識別子） 10.2477/jccj.2018-0062	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 平野敏行, 佐藤文俊	４．巻 61
２．論文標題 大規模分子設計における電子状態計算法と期待	５．発行年 2018年
３．雑誌名 放電研究	６．最初と最後の頁 13-18
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hirano Toshiyuki、Sato Fumitoshi	４．巻 1906
２．論文標題 Study of high-performance canonical molecular orbitals calculation for proteins	５．発行年 2017年
３．雑誌名 AIP Conference Proceedings	６．最初と最後の頁 30019
掲載論文のDOI（デジタルオブジェクト識別子） 10.1063/1.5012298	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　2件／うち国際学会　5件）

１．発表者名 T. Hirano, F. Sato
２．発表標題 GPU acceleration of a canonical molecular orbital calculation program by the third-generation density-functional-theory-based method
３．学会等名 The 59th Sanibel Symposium（国際学会）
４．発表年 2019年

１．発表者名
T. Hirano, F. Sato

２．発表標題
Automated Canonical Molecular Orbital Calculation Engine for Protein: ProteinDF/QCLObot

３．学会等名
16th international congress of quantum chemistry（国際学会）

４．発表年
2018年

１．発表者名
江口晴輝，平野敏行，佐藤文俊

２．発表標題
雲状モデルによるタンパク質分子軌道の新規可視化手法の研究

３．学会等名
日本コンピュータ化学会 2018秋季年会

４．発表年
2018年

１．発表者名
平野敏行，佐藤文俊

２．発表標題
GPUを用いた大規模電子状態計算プログラムProteinDFの高速化

３．学会等名
第12回分子科学討論会

４．発表年
2018年

１．発表者名
平野敏行，佐藤文俊

２．発表標題
カノニカル分子軌道計算によるインフルエンザM2タンパク質の電子構造

３．学会等名
分子科学討論会2017

４．発表年
2017年

１．発表者名
Toshiyuki Hirano, Fumitoshi Sato

２．発表標題
QCLObot: an automation engine of canonical MO calculation in proteins

３．学会等名
CBI学会2017年大会

４．発表年
2017年

１．発表者名
Toshiyuki Hirano, Fumitoshi Sato

２．発表標題
Development of Molecular Orbitals Calculation Applications for Proteins

３．学会等名
International Workshop on Massively Parallel Programming for Quantum Chemistry and Physics 2018（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Toshiyuki Hirano, Fumitoshi Sato

２．発表標題
Electronic structure of the active site on glucose oxidase by using canonical molecular orbital calculation

３．学会等名
The 58th Sanibel Symposium（国際学会）

４．発表年
2018年

１．発表者名
Toshiyuki HIRANO, Fumitoshi SATO

２．発表標題
Study of High-Performance Canonical Molecular Orbitals Calculation for Proteins

３．学会等名
International symposium: Computational Chemistry (CC) in ICCMSE2017（招待講演）（国際学会）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

ProteinDF
https://github.com/ProteinDF/ProteinDF
QCLObot
https://github.com/ProteinDF/QCLObot
ProteinDF software package
https://proteindf.github.io/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考