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　５ 年　 ６ 月　 ５ 日現在

機関番号：８２６２６
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ０５４３１

研究課題名（和文）国際規格策定にむけた有機フラグメント構造設計プログラムの開発

研究課題名（英文）Development of Structural Design Program Using Organic Fragments for International Standardization

研究代表者
和泉　博（Izumi, Hiroshi）

国立研究開発法人産業技術総合研究所・エネルギー・環境領域・主任研究員

研究者番号：２０３５６４５５

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）：有機分子立体構造(立体配座)ビッグデータで共通部分構造識別に利用する、IUPAC命名法P-94.2順位則修正案の必要性を認めてもらうため、産業応用上の有用性を検討した。提案している立体配座の記述法を用いて、最新のディープニューラルネットワークに応用可能か示すため、タンパク質立体配座可変性予測システムSSSCPredsを開発した。完成当時構造未知で大きな社会問題となっていたSARS-CoV-2を用いて予測精度の検証を行い、レセプター結合ドメインの立体配座可変性予測マップが実験で得られたアミノ酸変異による配列－発現量マップ及び配列－結合マップと極めて類似しており、高い相関を示すことを明らかにした。

研究成果の学術的意義や社会的意義
産業界のニーズとして開発失敗有機分子から的確な改良候補を提案する技術の要請がある。そこで、有機分子立体構造(立体配座)ビッグデータで共通部分構造を識別するのに必要となる順位則を国際科学会議に提案し、国際基準として認められるための取り組みを行っている。そのコンセプトをディープラーニング解析と組み合わせて得られたタンパク質立体配座可変性予測システムは、特定のウィルス表現型との相関を議論可能な程度の精度を有していた。このシステムを用いて、SARS-CoV-2のD614G変異で感染性が増大した原因や多重変異株の中和抗体回避能と構造可変性パターンとの相関を明らかにすることで実社会に役立つことを示した。

研究成果の概要（英文）：The revised sequence rule of P-94.2 of the IUPAC Rules for Nomenclature of Organic Chemistry for an identification of the 3D maximal common substructures (MCSs) has been proposed. For the approval of this rule, it is necessary to show usefulness in industrial applications. Big data of properly selected MCSs may avoid the heavy loads of theoretical calculations and may be available for the design of new functional molecules. For this purpose, the codification techniques of conformations for a comparison of the MCSs were developed, and a deep neural network-based program for the prediction of protein conformational variability (SSSCPreds) was constructed. The predicted conformational variability of the mutation sites for SARS-CoV-2 spike proteins correlated with the neutralization escape ability well. Further, the analysis of D614G mutation demonstrated that the left-handed α-helix-type conformation of G614 contributes to the increase in infectivity because glycine lacks chirality.

研究分野：構造有機化学

キーワード：立体配座　ディープラーニング　SARS-CoV-2　タンパク質　IUPAC命名法　機能性有機分子　分子構造コード化構造検索　超二次構造

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

IUPAC が定める有機化学命名法の P-94.2 には、立体配座の命名に関わる基準として次のように定められている。「(a) すべての原子あるいは置換基の組が異なる場合には sequence rule に従い、最優先のそのもの、(b) 一つの組のみ異なっている場合にはそのもの、(c) すべての組が同一の場合には最も二面角が小さくなるものを選択する。」。この P-94.2 の基準は同一の分子にのみ適用可能である。ところが、創薬をはじめとする分子設計の分野においては、置換基の最適化のために立体配座を比較するニーズが高まっている。その際、異なる分子間で立体配座を比較する必要があるが、現状では、分子モデルを重ね合わせるくらいしか術がなかった。さらに、生体分子構造解析に用いられているクライオ電子顕微鏡法の単粒子解析は構造の違う粒子（タンパク質分子）を分別しながら大量の画像処理を行う工程がある。生体分子からポリマー材料などのソフトマテリアルへ拡張するには一般的に分子の持つ単結合の回転の自由度がより大きくなることから、より詳細な立体配座の分類が必要となっていた。

また、AI の活用にはおよそ 100 万を超える精度の高い正解データが必要である。創薬への応用において、囲碁や将棋のようにこの手を選べば勝率が上がるというような、この手に当たる入力のツールが整備されておらず、さらに全体的に精度の高い正解データが不足していた。これまで、二次元の化学構造式を記号化し、構造活性相関や新規有機分子設計を行う研究は行われてきたが、機能性分子設計を実現した例は限られていた。このことは有機分子の機能性に立体配座が深く関わっており、その動的に変化する立体配座情報が抜け落ちていることが一因として考えられた。

２．研究の目的

有機分子の立体配座について、International Union of Pure and Applied Chemistry（IUPAC）の命名法（P－94．２順位則）が定められている。ところが、異なる分子間での比較が難しい等、課題が存在した。申請者は、異分子間での比較を可能とする Conformational Code for Organic Molecules （CCOM） プログラム（図１）を世界で初めて開発した。さらに、P－94．２順位則修正案を IUPAC に提出し、有機分子全体の立体配座の記述の国際規格に向けた研究を行っている。

一方で、タンパク質分子の先行研究において、アミノ酸残基ごとに立体配座がαヘリックス型(H)かβシート型(S)か分類してコード化を行った。このコード化により、通常の免疫グロブリンと自己免疫疾患に関係する自己抗体とで明確に区別可能な特徴的フラグメントが存在することを見出している（図２）。この成果が認められて、Springer “Methods in Molecular Biology” の chapter を執筆し、「タンパク質超二次構造コード」と名付けた。一般の有機分子を対象とする CCOM プログラムにこの技術を適用することで、適用範囲が著しく拡大し、機能性有機分子設計への活用が期待できる。

申請者が P-94.2 順位則修正の提案を行ったことが示しているように、これまで異なる分子の共通するフラグメントの立体配座を記号として比較することは行われていなかった。すなわち、動的有機フラグメントのデータベースを用いて機能性有機分子設計を行う研究は世界初の試みである。また、正確な動的有機フラグメントの立体配座が検証されている例は少なく、ラパマイシン分子の新規水素結合様式のようにこれまでに知られていない事例が数多く存在しているものと考えられる。そこで、AI を用いた機能性有機分子設計に適用可能な正確な動的立体配座情報の蓄積を進めている。

本研究では、CCOM を基盤技術として用い、近年技術革新の著しいクライオ電子顕微鏡法によるタンパク質構造解析に適用可能な X 線結晶構造及び密度汎関数法計算フラグメントデータベースを構築する。薬剤耐性で問題となっているマクロライドをはじめとする有機分子設計に向けた有機フラグメント構造設計プログラムを開発する。併行して、人工知能（AI）等の機械学習プログラムと統合するため、タンパク質を記述するために考案したタンパク質超二次構造コードを 10 万件以上の Protein Data Bank（PDB） 結晶構造データに適用し、機能やアミノ酸配列と相関する特徴的構造パターンを学習させ、有機分子設計プログラムに活用する。

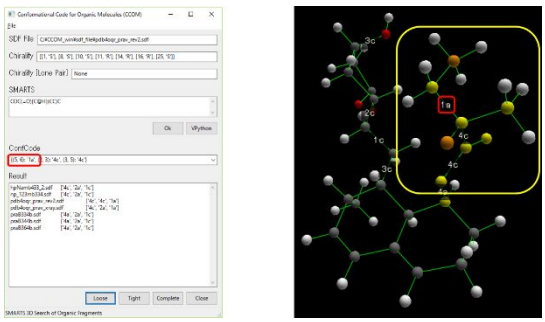

図１．CCOM プログラム

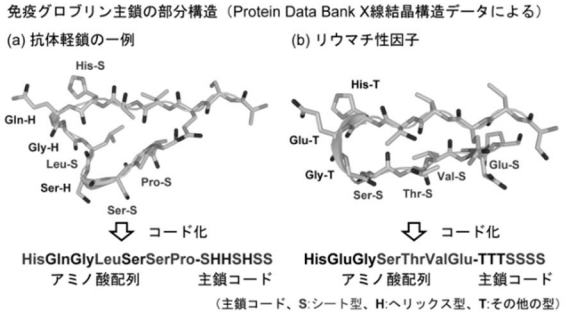

図２．免疫グロブリンと自己抗体フラグメント

３．研究の方法

(1)タンパク質超二次構造コードを用いた構造パターン解析

アミノ酸配列に対応するタンパク質超二次構造コード(SSSC)に変換するプログラムを含むコンフォメーション文字列変換解析プログラム SSSCview を用いて、139,932 個の PDB ファイルから 582,813 個のアミノ酸配列と SSSC 配列が対になったサブユニットの FASTA 形式ファイルを作成した。その中で 379,334 個のファイルが 100 個以上の連続するアミノ酸残基を有していた。その中から、150,000 個のトレーニングデータ、10,000 個のテストデータ、10,000 個、３組の推論システムでの評価データをランダムに選択した。次に、SSSC の H，S，T，D のコードをそれぞれ[1,0,0,0]，[0,1,0,0]，[0,0,1,0]，[0,0,0,1]に変換した(100，4)の行列からなるファイルを作成した。アミノ酸配列についても同様の変換を行い、ディープラーニング用のデータセットを作成した。

得られたデータセットを用いて、様々なネットワークを試し、最終的に Neural Network Console 1.40 の“12_residual_learning.sdcproj”ネットワークテンプレートを改良してディープラーニングを行った。テストデータでいい精度が出たことから、100 個以上のアミノ酸残基を有するサブユニット全般に適用可能な推論システム(SSSCPred)を構築し、性能評価を行った。

低い一致率のサンプルと実際のサブユニットの相関を解析した結果から、低い一致率は同じアミノ酸配列でも異なるコンフォメーションをとりうることと相関していることが判明した。そこで、立体配座可変性を予測するシステムが構築できるのではと考え、新たに 2 個のディープニューラルネットワーク予測システムを構築した。582,666 個のアミノ酸配列と SSSC 配列が対になったファイルの中で 207,738 個のファイルが 200 個以上の連続するアミノ酸残基を有していた。その中から、150,000 個のトレーニングデータ、10,000 個のテストデータ、10,000 個の推論システムでの評価データをランダムに選択した。SSSCPred と同様の手法でディープラーニングを行い、200 個以上のアミノ酸残基を有するサブユニット全般に適用可能な推論システム(SSSCPred200)を構築した。また、100 個以上の連続するアミノ酸残基を有するアミノ酸配列と SSSC 配列が対になったファイルの中から、350,000 個のトレーニングデータ、10,000 個のテストデータ、10,000 個の推論システムでの評価データをランダムに選択し、SSSCPred と同様の 100 個以上のアミノ酸残基を有するサブユニット全般に適用可能な推論システム(SSSCPred100)を構築した。最終的に、1 個のアミノ酸配列に対して、これら 3 個のディープニューラルネットワークを用いて予測を同時に行い、SSSC が 3 個とも一致するか、しないかでコンフォメーションのフレキシビリティーを評価する、立体配座可変性予測システム(SSSCPreds)を構築した（図３）。

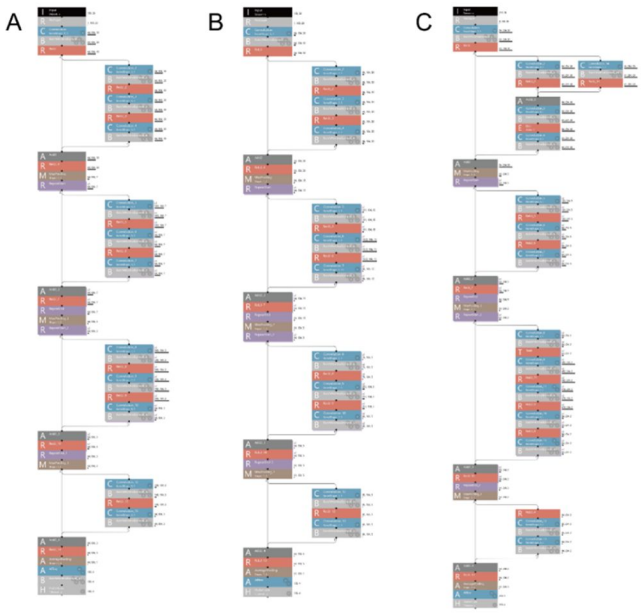

図３．SSSCPreds のディープニューラルネットワークアーキテクチャ（A: SSSCPred、B: SSSCPred100、C: SSSCPred200）

SARS-CoV-2 の出現が大きな社会問題となっていること、deep mutational scanning という、20 種類の 1 アミノ酸変異をアミノ酸配列に対して網羅的に行い、配列－表現型マップを作成する手法を用いたデータが報告されたことから、SARS-CoV-2 の野生型及び多重変異株との相関解析を行うことで、SSSCPreds の予測性能を評価した。

(2)有機フラグメント構造設計プログラムの開発

DFT 計算で得られた Gibbs 自由エネルギー値が小さく、存在比の大きいフラグメント分子の立体配座データのみ格納したデータベースから分子量の大きな分子の DFT 計算の初期構造を作成する RingFragGeneration プログラムを python3.7 に対応させるため、オープンソースである Pybel 及び OpenBabel のライブラリーの他に RDKit、wxPython、VPython、NumPy、GaussView のライブラリー等を利用しプログラムを作成した。また、化学構造式の cdx ファイルから mol ファイルへ変換する cdx_to_mol.py も作成したが、ラパマイシンには対応できず一般に使用されているソフトウェアを使用することとした。

さらに、ケンブリッジ結晶構造データベースの構造データを、二面角分類コード、選択された４個の原子及び詳細な二面角の情報を含む sdf 形式ファイルに変換する python3.7 対応 CCOM2 プログラムを作成し、機械学習用ワークフロー型データ分析プラットフォームに取り込めるようにバージョンアップした。

４．研究成果

(1)タンパク質超二次構造コードを用いた構造パターン解析

SSSCPred の性能評価を行ったところ、従来のタンパク質二次構造予測では不可能であったループ構造のコンフォメーションの予測が可能であることがわかった。10,000 個、３組の推論システムでの評価データから得られた平均一致率は 0.9、また、ディープラーニング用外部評価データセット等の CB513、CullPDB データセットでの平均一致率はそれぞれ 0.88(612 サブユニット)、0.86(17,169 サブユニット)と従来の二次構造予測と遜色のない予測性能が得られた。一方で、一致率が 0.6 未満のサンプルも存在することから、サブユニット名と強く相関があるか統計的に解析した。するとサブユニット名で常に低い一致率を取っているわけではなく、同じサブユニット名で 0.9 を超える高い一致率を取るケースもみられた。すなわち、同じアミノ酸配列でも全く異なるコンフォメーションをとりうるケースが見受けられ、コンフォメーションのフレキシビリティーと相関していることがわかった。

3 個のディープニューラルネットワークを用いて予測を同時に行う SSSCPreds の性能評価を行ったところ、SSSCPred200 は 0.905(CullPDB)、0.911 (CB513)、SSSCPred100 は 0.896 (CullPDB)、0.907 (CB513)という平均一致率の結果が得られた。また、実測値の SSSC との一致ではなく、SSSCPred200、SSSCPred100、SSSCPred 間の予測値の SSSC の一致を検証したところ、それぞれの一致率の値の大きさがサブユニットのフレキシビリティーのよい指標となることがわかった。

大きな社会問題となっている SARS-CoV-2 について、立体配座可変性予測システム SSSCPreds にどの程度の精度があり、CryoEM で報告されていない知見がどう得られるか性能評価を行った。deep mutational scanning により得られた配列－表現型マップにおいて、発現量の低下と結合の低下に明らかな相関がみられるとの報告があり(T. N. Starr, et al., *Cell* **182**, 1295-1310 (2020))、野生型の立体配座可変性予測マップとの比較を行ったところ、予測された rigid なサイトとも高い相関性を示すことがわかった（図４）。このことから、SSSCPreds の立体配座可変性予測性能は特定のウィルス表現型との相関を議論可能な程度といえる。

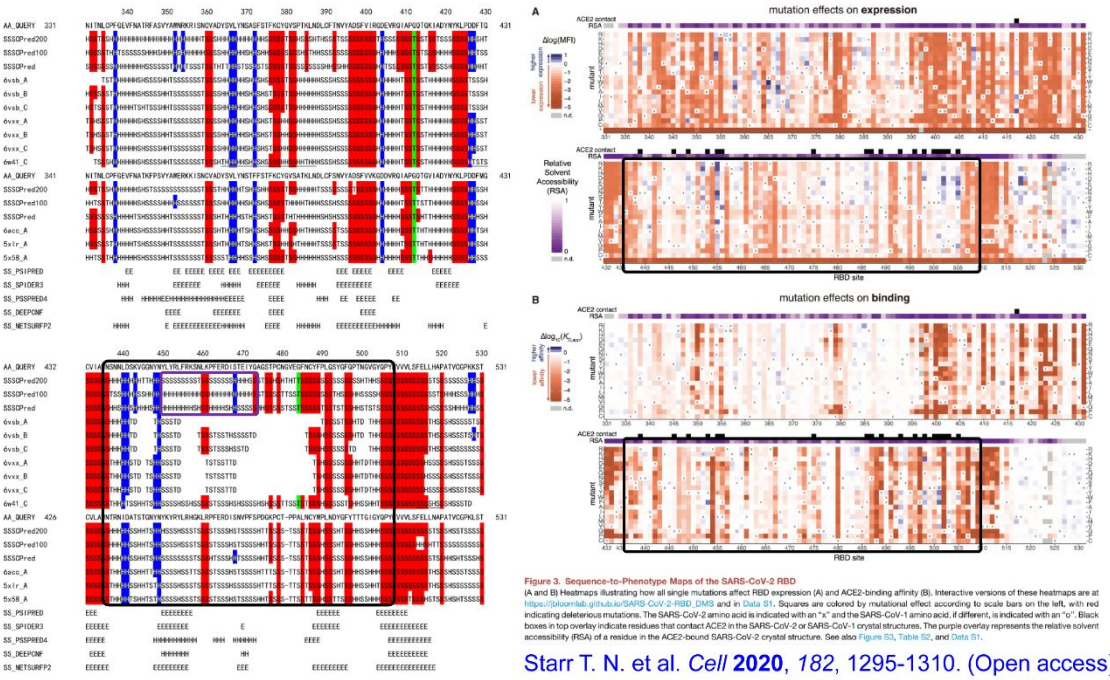

図４．SSSCPreds 立体配座可変性予測マップと deep mutational scanning 結果との比較

D614G 変異は他の多重変異株と異なり、スパイクタンパク質上のこの変異のみで野生型から大幅に感染性が高まった。実験的には、フーリン配列で切断された S1 ユニットの排出を大幅に抑え、完全な形を保ったスパイクタンパク質の発現量が増加することが報告されていたが、他のアミノ酸でなく、なぜ可動性の大きいグリシンのみが発現量を増加させたのかは明らかにされていなかった。

立体配座可変性予測では、G614 のサイトは rigid で、CryoEM の結果と一致する「その他の型」をとるのに対し、D614 のサイトは flexible になるという結果が得られた(図５)。そこで、CryoEM の構造について SSSCview を使って詳細に解析したところ、「その他の型」は左巻きαヘリックスにみられるコンフォメーションと共通する構造であることがわかった（図５）。通常のタンパク質は安定な右巻きαヘリックスをとっている。グリシンのみキラリティーをもたないことから、可動性が大きいものの左巻きαヘリックスを安定化することができる。このことによる、G614 のサイトの rigidity が、S1 ユニットの排出を大幅に抑制し、高い発現量、結果として高い感染性の要因になっていることが明らかになった。多重変異株であるα、β、γ株について、立体配座可変性予測マップはα株のみマップパターンが野生型とよく似ており、α株のみ野生型と免疫に関係する中和抗体回避能があまり変わらないこととよく相関していた。

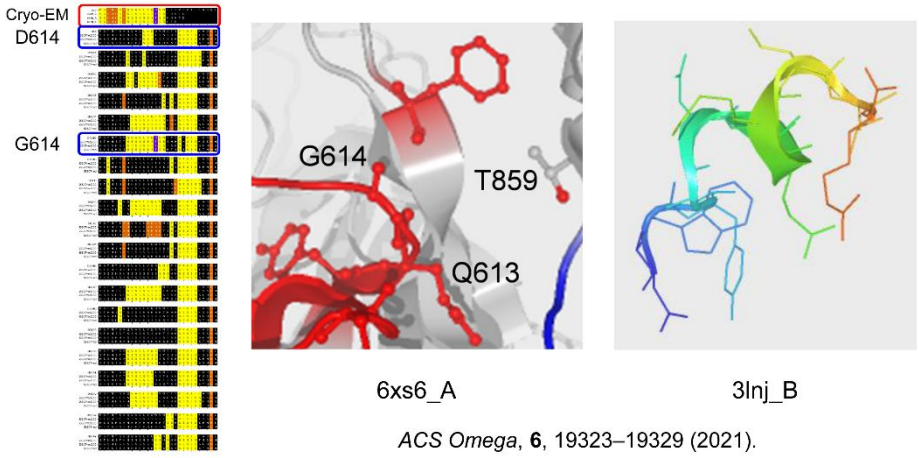

図５．D614G 変異の立体配座可変性予測とその左巻きαヘリックス型コンフォメーション

また、β株の K417N 変異は flexible に、γ株の K417T 変異は rigid にするとの予測結果が得られたが、β株と比べたγ株の弱い回避能とよい相関性を示した。δ株については、L452R 変異により YRYRLFR というアミノ酸配列の rigity が増すという予測結果が得られており、レセプター結合モチーフのフレキシブル部位のエッジの安定化による、発現量の増加に寄与しているものと考えられる。さらに、立体配座可変性予測マップパターンの大きな変化は高い中和抗体回避能ともよく相関した。

野生型に対するδ株の一致比(RBD)0.89、剛性比(RBD)1.10、α株の一致比(RBD)0.97、剛性比(RBD)0.98、β株の一致比(RBD)0.92、剛性比(RBD)0.90 という結果が得られたが、定量的にも、中和抗体力価の野生型に対するδ株の 5.8 倍の低下、α株の 2.6 倍の低下、β株の 4.9 倍の低下の実測値と一致比、さらには 1.0 からの逸脱を考慮すると実測値と剛性比もよい相関性を示した。

さらに数多くの変異を含むο株についても、野生型だけではなく、他の多重変異株とも立体配座可変性予測マップパターンが大きく異なっており、中和抗体回避能とよい相関性を示した。

(2)有機フラグメント構造設計プログラムの開発

RingFragGeneration プログラムの python3.7 への対応において、python2.7 のライブラリーで可能だったものが出来ない操作も存在したが、図６に示すように、ラパマイシンのような分子量が大きい分子でも DFT 計算のための初期構造を自動作成する基本動作は確認した。

さらに、ケンブリッジ結晶構造データベースの構造データを sdf 形式ファイルに変換する python3.7 対応 CCOM2 プログラムに関して、DFT 計算関連構造データでは見られなかった NaN（Not a Number、非数）の多数の発生、機械学習用ワークフロー型データ分析プラットフォームに取り込めない問題に直面した。これらの問題を解決し、ケンブリッジ結晶構造データベースから検索可能な 150,461 個のデータの SMARTS 3D Pattern を含む sdf 形式ファイルを作成した。ただ、結晶データの場合、SMILES に変換できても SMARTS のクエリとして必ずしも検索できないケースがあるのが課題である。また、検索を繰り返した結果、100 万件を超える構造データが登録されたケンブリッジ結晶構造データベースでも rigid な部分構造のヒット数は多いものの、コンフォメーション変化を伴う部分構造について、共通すると判定されるフラグメントは極めて少ないことがわかった。このことから、機能性分子設計のためのディープラーニングには現時点では DFT 計算から得られた構造データを活用する方がよいものと考えられる。

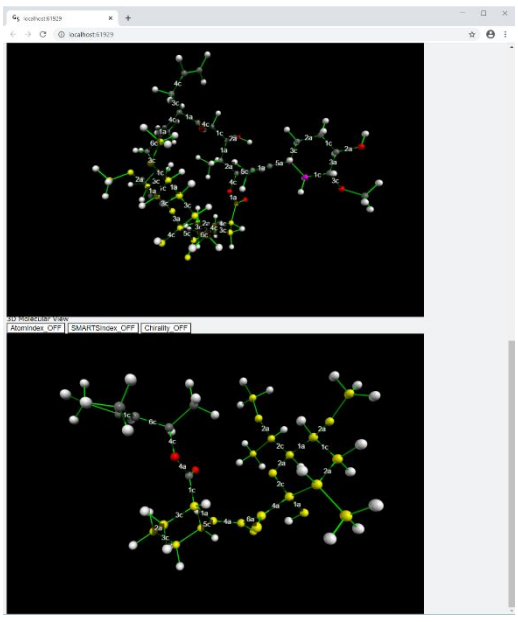

図６．RingFragGeneration プログラム

さらに、25th IUPAC International Conference on Physical Organic Chemistry (ICPOC 25) に参加し、「Requirements for International Standard of Conformational Descriptor for Three-Dimensional Maximal Common Substructure (MCS)」というタイトルで、SARS-CoV-2 の解析を例示し、ディープニューラルネットワークへの展開に IUPAC 規則 94.2 の立体配座命名法の選択則の改良が必要であること、三次元最大共通部分構造記述子の国際基準が必要になってくることについて発表を行った。

５．主な発表論文等

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

１．著者名 Izumi Hiroshi、Aoki Hiroshi、Nafie Laurence A.、Dukor Rina K.	４．巻 8
２．論文標題 Effect of Conformational Variability on Seasonable Thermal Stability and Cell Entry of Omicron Variants	５．発行年 2023年
３．雑誌名 ACS Omega	６．最初と最後の頁 7111～7118
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acsomega.2c08075	査読の有無 有
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 該当する

１．著者名 Izumi Hiroshi、Nafie Laurence A.、Dukor Rina K.	４．巻 6
２．論文標題 Conformational Variability Correlation Prediction of Transmissibility and Neutralization Escape Ability for Multiple Mutation SARS-CoV-2 Strains using SSSCPreds	５．発行年 2021年
３．雑誌名 ACS Omega	６．最初と最後の頁 19323～19329
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acsomega.1c03055	査読の有無 有
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 該当する

１．著者名 Izumi Hiroshi、Nafie Laurence A.、Dukor Rina K.	４．巻 5
２．論文標題 SSSCPreds: Deep Neural Network-Based Software for the Prediction of Conformational Variability and Application to SARS-CoV-2	５．発行年 2020年
３．雑誌名 ACS Omega	６．最初と最後の頁 30556～30567
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acsomega.0c04472	査読の有無 有
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 該当する

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 和泉 博、ローレンス ネィフィー、リナ デュコア
２．発表標題 Requirements for International Standard of Conformational Descriptor for Three-Dimensional Maximal Common Substructure (MCS)
３．学会等名 25th IUPAC International Conference on Physical Organic Chemistry (ICPOC 25)（国際学会）
４．発表年 2022年

１．発表者名
和泉 博

２．発表標題
Deep Learning Analysis of Multiple Mutation Omicron Strains using the Codification Techniques of Conformations

３．学会等名
JCUP XI（国際学会）

４．発表年
2022年

１．発表者名
和泉 博

２．発表標題
立体配座のコード化によるオミクロン変異株のディープラーニング解析

３．学会等名
第32回基礎有機化学討論会

４．発表年
2022年

１．発表者名
和泉　博、ローレンス ネィフィー、リナ デュコア

２．発表標題
立体配座のコード化によるSARS-CoV-2変異株のディープラーニング解析

３．学会等名
第31回基礎有機化学討論会

４．発表年
2021年

１．発表者名
和泉　博

２．発表標題
Development of Conformational Descriptors for Organic Materials Informatics

３．学会等名
日本化学会 第102春季年会（2022）

４．発表年
2022年

１．発表者名
和泉　博

２．発表標題
Deep Learning Analysis of SARS-CoV-2 Using the Codification Techniques of Conformations

３．学会等名
日本化学会 第101春季年会（2021）

４．発表年
2021年

１．発表者名
和泉　博

２．発表標題
Supersecondary Structure Code (SSSC) of Nivolumab, PD-1, and PD-L1

３．学会等名
JCUP X（国際学会）

４．発表年
2019年

１．発表者名
和泉　博、後藤 仁志

２．発表標題
立体配座のコード化によるタンパク質分子のディープラーニング解析

３．学会等名
第30回基礎有機化学討論会

４．発表年
2019年

１．発表者名
和泉　博

２．発表標題
Deep Learning Analysis of Protein Molecules Using the Codification Techniques of Conformations

３．学会等名
日本化学会 第100春季年会（2020）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

ディープラーニング立体配座可変性予測システム公開ホームページ https://staff.aist.go.jp/izumi.h/SSSCPreds/index-e.html プログラム公開ホームページ https://researchmap.jp/read0004613/%E8%B3%87%E6%96%99%E5%85%AC%E9%96%8B 研究者ホームページ https://staff.aist.go.jp/izumi.h/

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
研究協力者	ネイフィー　ローレンス (Nafie Laurence A.)		
研究協力者	デュコア　リナ (Dukor Rina K.)		

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
米国	Syracuse University	BioTools Inc.		