2020 年度　利用報告書

課題名（タイトル）： 深層学習などによる NMR データおよび生体高分子の解析

利用者氏名： 小林直宏

理研における所属研究室名： 放射光科学研究センター NMR 研究開発部門 NMR 応用・利用グループ

1.　本課題の研究の背景、目的、関係するプロジェクトとの関係

本課題においては大規模計算機 HOKUSAI の核磁気共鳴法（NMR）におけるデータ解析の自動化、深層学習などへの応用を目的とした。NMR 法は信号解析の自動化が他の構造解析法、例えば X 線結晶解析やクライオ電子顕微鏡解析などに比べて大きく遅れを取っている。研究代表者は HOKUSAI に搭載されている 40～200 Intel-Core を利用することで大規模な NMR 信号の自動解析の高速化を試みた。またデスクトップ・ワークステーションで独自に開発した深層学習による NMR 信号解析ツールの HOKUSAI 上でのインストールや試験的計算を実行することを目的とした。

2.　具体的な利用内容、計算方法

具体的な計算では NMR 信号の自動帰属プログラムである代表者が独自に開発した MagRO に加え、外部プログラムである FLYA を用いた NMR 信号自動帰属を比較的大型のタンパク質である p120GAP の特殊同位体ラベル体について計算を多数行った。各計算においては 40-Core を利用し、多くのベンチマーク計算を実行できた。また、NMR 化学シフトからタンパク質の2次構造を予測しつつモデリングプログラムである ROSETTA を組み合わせた CS-ROSETTA と呼ばれるツールによる構造解析計算を試験的に実行した。さらにはタンパク質の立体構造から化学シフトを予想させる深層学習ツールを独自開発し、その学習パラメータを用いた試験的な計算を実行できた。深層学習による訓練は Microsoft CNTK を用いて GPU 搭載のデスクトップ・ワークステーションで行い、その訓練データをもとに HOKUSAI 上での 40-Core OpenMP での計算性能などを比較した。

3.　結果

NMR 信号の自動帰属は FLYA を用いた数種のタンパク質において 40-Core による計算により高速かつ高精度な結果が得られることが示された。それらのうち 40kDa の NMR 研究としては比較的大型のタンパク質について自動的な解析に十分な解析精度が実現可能であることを異なる条件で多数示すことができた。また CS-ROSETTA による NMR データとして比較的容易に得られる化学シフト情報を元にしたモデル構造構築計算を小型のタンパク質である CGI38（20kDa）およびやや大きめのタンパク質である p120GAP（40kDa）について実行した。最新のワークステーションでは 1 週間ほどかかる計算を HOKUSAI 上による 200～400-Core での大規模化によってわずか 12～24h で十分な精度の計算を実行できることを示すことができた。さらには深層学習（24 層程度）による化学シフト予測計算も HOKUSAI 上で 40-Core 程度の計算負荷により GPU 搭載のワークステーションでの学習と同程度の計算速度であることを示すことができた。これにより複数ノードを用いることで異なる学習パラメータ最適化を容易に行えることが可能となった。

4.　まとめ

以上、主に NMR データの信号解析や自動帰属、タンパク質構造モデリング計算を高速かつ高精度に実行することに HOKUSAI の大規模計算機が有効であることを多数示すことができた。最新のワークステーション(20-Core) レベルの計算機性能では通常、数日あるいは 1 週間程度の時間を要する自動計算も HOKUSAI による 1-ノード内での計算でかなりの効率化が図れることは簡易利用として極めて有用であった。

5.　今後の計画・展望

上記の成果は来年度における国内外での学会あるいは国際誌への発表を予定している。HOKUSAI による大規模計算機は CPU 数ばかりではなく大規模なメモリ（100～200GB）を必要とするような計算にも有効であり、ラボスケールでの計算機環境では実現し得ないような解析工程の戦略を変え、更なる高効率化が図れるものと考えられる。