# 2022 年度名古屋大学 HPC 計算科学連携研究プロジェクト成果報告書

2023 年 3 月 31 日

名古屋大学細胞生理学研究センター　田中　康太郎

## 1.　得られた知見の概要

本課題「クライオ電子顕微鏡法によるタンパク質構造決定におけるスーパーコンピュータ不老の活用検討」では、タンパク質の実験的構造決定で現在最もよく使われるクライオ電子顕微鏡法(単粒子解析)の計算プラットフォームとして不老を利用する検討を行った。

まず単粒子解析ソフトウェアのデファクトスタンダードである cryoSPARC および RELION の環境を不老向けに構築した。３つのデータセットを対象に構造解析を行ったところ、研究室所有のワークステーションを利用するのと近い感覚で、標準的なワークフローでの構造解析を完遂することができた。大量の計算リソースを利用して構造解析の条件検討を十分に行えることと、コストパフォーマンスも大変良いことから、不老を単粒子解析の計算プラットフォームとして利用する利点があると確認できた。

一方で残念なことに、cryoSPARC の利用はログインノードで複数の常駐プロセスを稼働させる必要があり負荷をかけてしまうこと、さらにセキュリティ上の懸念もあること等から、不老に限らず汎用の大規模共有計算機での利用に適さないことが分かった。今回は検討目的で利用したが、同様の手順での利用を一般に公開して推奨することは避けることにした。cryoSPARC 以外のソフト(RELION 等)のみで解析をするならば十分活用できるが、cryoSPARC に関しては AWS 等のクラウド仮想サーバーを利用する手順が整いつつ有るので、オンプレミスの計算機を持たない単粒子解析ユーザーは割高にはなるがそちらを検討するほうが良さそうである。

## 2.　環境構築

### 2.1.　RELION ( https://relion.readthedocs.io/en/release-4.0 )

不老の module で用意されているコンパイラやツールキットを用いて、GPU 版および CPU 版を標準的な手順でビルドした。不老用のジョブスクリプトのテンプレートを作成した。GPU を利用するジョブ、ローカルディスクへの高速アクセスが必要なジョブは TypeⅡサブシステム、それ以外のジョブはクラウドサブシステムで利用した。

### 2.2.　cryoSPARC( https://cryosparc.com )

公式手順を参考に、ログインノードでマスタープロセスを稼働させ、計算ノードでワーカープロセスによる計算を行うクラスタ構成でのインストールを行った。マスタープロセスは複数のプロセスから構成されており、そのうちユーザーインターフェイスとなるウェブサーバーが TCP ポート 39000 番(デフォルト、変更可能)で listen する。ラウンドロビンによるログインノード自動割り当てを回避するため、ssh ポートフォワードはマスタープロセスが稼働しているログインノードの IP アドレスを直接使用した。ジョブはすべて TypeⅡサブシステムで実行した。ログインノードではなく計算ノードでマスタープロセスを動かすことも検討したが、計算ノードからはジョブ発行ができないためマスタープロセスを動かせず、断念した。

## 3. 単粒子解析計算の実践

### 3.1. cryoSPARC T20S extensive workflow

cryoSPARC に付属しているベンチマークデータセット及びワークフローでのテスト(1GPU 使用)を行い、研究室所有の計算機(AMD EPYC 7502P 32core@2.5GHz, 256 GB RAM, GeForce RTX 3090 24GB x 4)と性能比較した。総計算時間は研究室計算機と cx-single ともに 57 分となった。GPU を利用したジョブでは概ね RTX3090 を搭載した研究室計算機が速かったが、ファイル I/O 速度など GPU 以外の点では不老が優れているため同じ計算時間になったようである。

### 3.2. EMPIAR-10248 アポフェリチン

高分解能構造解析のベンチマークとして使われているアポフェリチンの公開データセットを利用して、RELION と cryoSPARC を併用しての高分解能構造解析を試みた。解析条件の検討も含め、総計算時間は 60 時間（cryoSPARC 41 時間、RELION 19 時間）となり、3D 構造再構成の分解能は 1.55 Å に到達した(同データセット公開者らの解析では 1.53 Å)。2022 年度の負担金(10,000 円で 6,500 ポイント)規定では 2,400 円分の計算に相当する。cryoSPARC のジョブはほとんどが 1GPU のみ対応のためほぼ cx-share で計算しており、RELION のジョブは GPU 非対応の CPU ジョブのみが必要だったことから cl-single または cl-small を利用したため、コストパフォーマンス良く計算できた。

### 3.3. EMPIAR-10288 カンナビノイド受容体-G タンパク質複合体

本公開データセットは、cryoSPARC の開発元である Structura Biotechnology 社により、AWS ParallelCluster での cryoSPARC ベンチマークで利用された(2021 年 5 月 10 日 https://guide.cryosparc.com/setup-configuration-and-management/cryosparc-on-aws/performance-benchmarks )。Structura によるベンチマークでは、条件検討のための計算時間は含めず、また高分解能化もそれほど追求しない(> 3Å 分解能)場合で、総計算時間 4.29 時間、利用料金 $47.12(US)と報告している。cx-share と cx-single を利用したところ、条件検討や高分解能化(2.90 Å)のためのジョブも含め総計算時間 19 時間で 1,125 円の利用料金となり、コストパフォーマンスに優れることが確認できた。

## 4. 総括

今回の検討で、不老は単粒子解析計算に関し、高いコストパフォーマンスで標準的な計算を完遂できると確認できた。昨今の現実のデータセットは 3.3 節で使用したものの 2~3 倍程度のスケールになることが多い。難しいデータセットでは実行ジョブ数も数倍になり得る。その場合でも 1 データセット当りの利用料金が 10 万円を超えることはほぼ無さそうであり、興味をもつユーザーもいると思われる。

しかしながら、1 節および 2.2 節で言及した通り、cryoSPARC の利用に適さないという問題があった。昨今は私も含めほとんどの単粒子解析ユーザーが cryoSPARC を利用するため残念であったが、RELION のような HPC フレンドリなソフトウェアを活用した計算や、cryoSPARC や RELION の解析結果を入力としてタンパク質の構造解析を行う機械学習( https://github.com/zhonge/cryodrgn 等)モデルの利用で活用して行きたく、技術ブログなどでその方法の紹介もしていくことにしたい。

以上