

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

名古屋大学細胞生理学研究センター

田中 康太郎

2022年度 名古屋大学HPC計算科学連携研究プロジェクト

JHPCN萌芽型共同研究課題

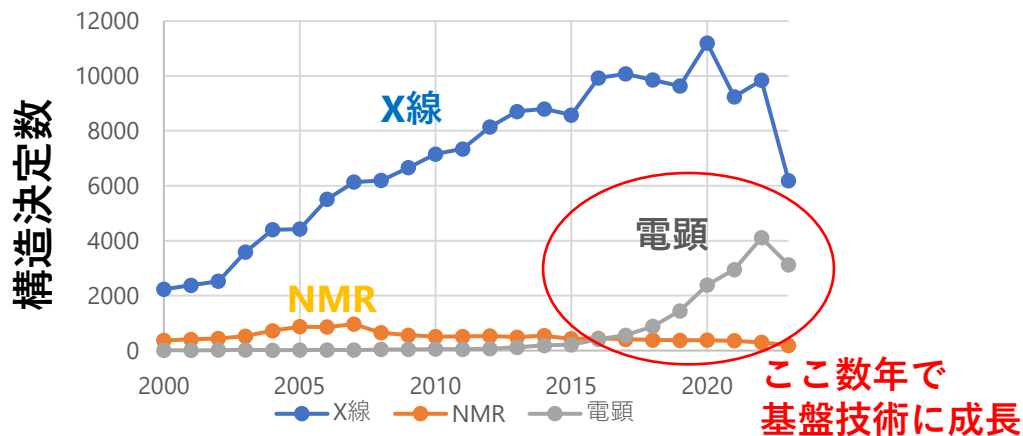
タンパク質の構造を調べる方法

計算による推論

- 機械学習、バイオインフォマティクス (例：AlphaFold)
 - 正確さの問題
- 計算物理学 (例：MD計算)
 - 計算量の問題

実験による構造決定

- X線結晶構造解析
- 核磁気共鳴法(NMR)
- **クライオ電子顕微鏡法(Cryo-EM)**



The Nobel Prize in Chemistry 2017



© Nobel Media AB. Photo: A. Mahmoud
Jacques Dubochet
Prize share: 1/3



© Nobel Media AB. Photo: A. Mahmoud
Joachim Frank
Prize share: 1/3

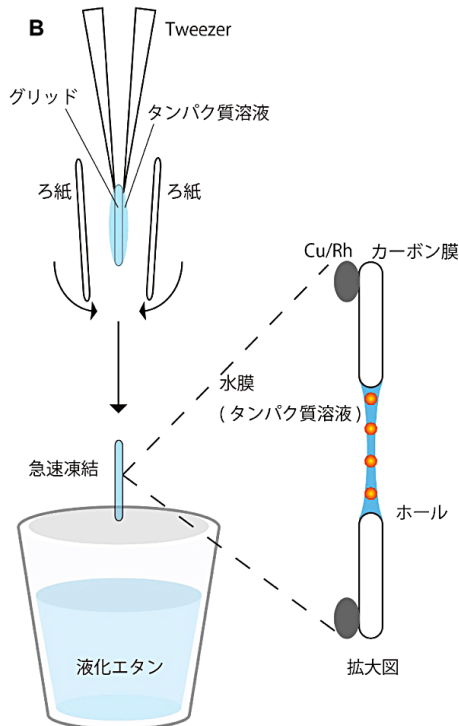


© Nobel Media AB. Photo: A. Mahmoud
Richard Henderson
Prize share: 1/3

The Nobel Prize in Chemistry 2017 was awarded jointly to Jacques Dubochet, Joachim Frank and Richard Henderson "for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution"

クライオ電子顕微鏡法

含水試料を凍結固定



日立ハイテク「構造細胞生物学のための電子顕微鏡技術」

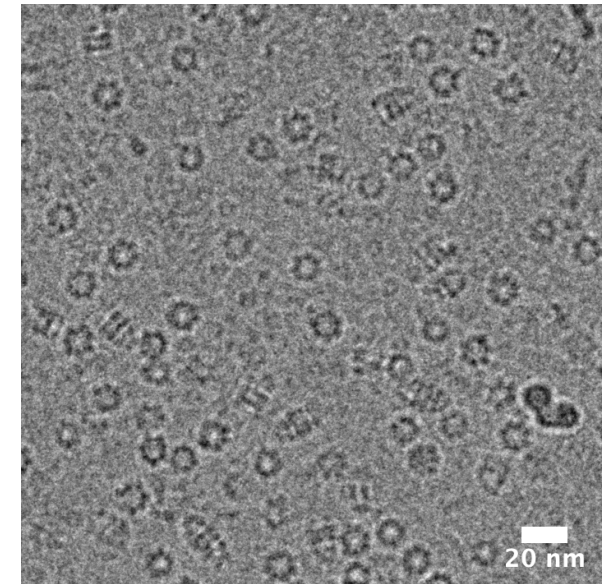
クライオ電子顕微鏡 試料を液体窒素冷却できるTEM



Krios (Thermo Fisher Scientific)



試料の拡大投影像 (2D)



画像解析による立体構造計算

- タンパク質試料を撮影して構造解析する手法を特に**単粒子解析**と呼ぶ

単粒子解析の基本解析フロー例

クラス平均像

入力データ
数千 movies (数百GB ~ 数TB)
~4k x 4k画素, 数十フレーム/movie

試料ドリフト・変形の補正
数秒~数十秒/movie
補正の上で平均化した画像を出力

像ボケ推定
数秒/image

粒子検出
数秒/image

粒子切り出し
数秒/image

粒子画像セット
数十万~数百万 particles
数百 x 数百画素 /particle
数十 x 数十画素にダウンサンプリング

2Dクラス分類
数十分~数時間
Alignment有の教師無しクラス分類

クラス選別
クラス平均像が良いものを目視で選抜

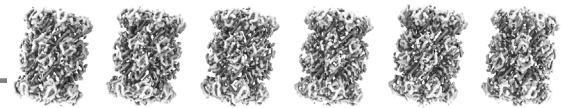
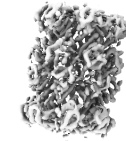
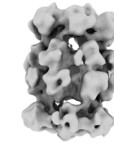
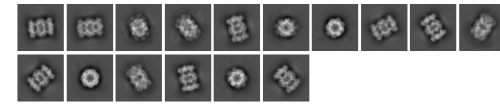
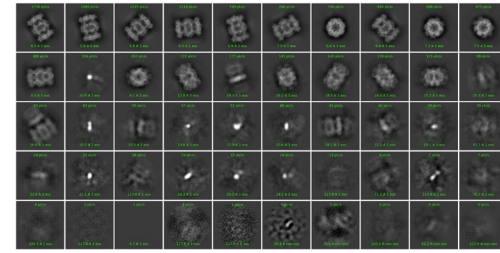
ab initio 3D構造再構成
数十分~数時間

3Dリファインメント(精密化)
数十分~数時間

3Dクラス分類
数十分~数時間

クラス選別
細部まで見えているもの、
構造が異なるものを目視で選抜

3Dリファインメント
数十分~数時間



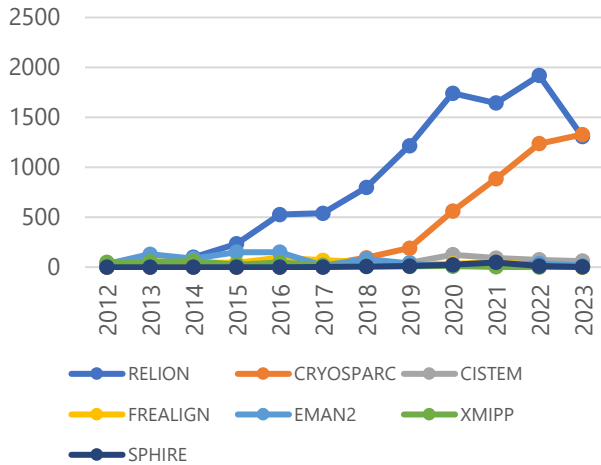
高解像度で粒子再切り出し

- 試行錯誤が必要で、アドホックなワークフロー
- 速ければ全体で数日程度の作業

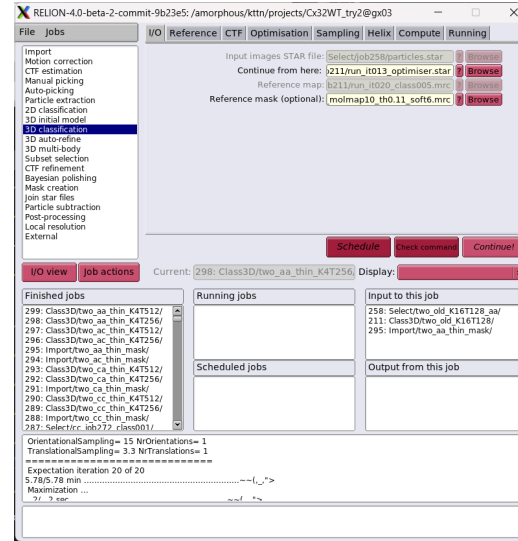
単粒子解析のソフトウェア

RELION

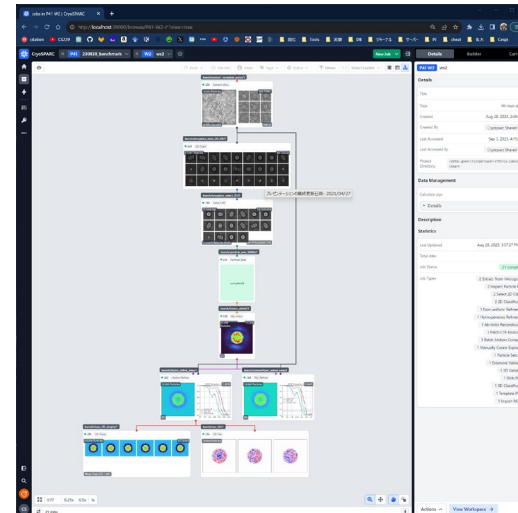
単粒子解析による構造決定で 使用されたソフトウェアの内訳



- RELIONとcryoSPARCがデファクトスタンダード



cryoSPARC



- 単粒子解析普及の立役者
- 英MRC-LMB Sjors Scheresらが開発主体
- C++/CUDA
- フリーソフトウェア (GPLv2)
- Cryo-ET計算にも対応

- Structura Biotechnology社が開発
- クローズドソース
- ライセンス制
 - 商用利用は有償
- 一部アルゴで特許取得
- ウェブアプリとして実装されたUIが優れている
- 処理が高速

単粒子解析における不老の活用検討

所属研究グループの計算環境

- 主力 GPUワークステーション 3 台
 - 単粒子解析向けとしてはミドルレンジ～ハイエンドに相当
 - スペックは十分だが、リソース不足を感じる場合があった
 - 計算プロジェクトが立て込んでいるとき
 - パラメータを振ってチューニングしたいとき
 - かといって高額なGPUワークステーションを増設する程でもない
- 不老で単粒子解析の環境を準備し、追加リソースとして利用したい

不老の活用検討

- RELION・cryoSPARCの環境構築
- 構造解析を実行
 - 性能、使い勝手、コストの評価
- 課題抽出

GPUワークステーション

Real Computing inc.
水冷GPUサーバー nami4 II

- AMD EPYC 7502P 32core
- 256 GB RAM
- GeForce RTX3090 (24 GB) x 2~4
- 1 TB NVMe, 10 TB HDD



Type II サブシステム

クラウドシステム



RELIONの環境構築@不老

ビルド

- 手順は公式ドキュメントに倣った
 - <https://relion.readthedocs.io/en/release-4.0/Installation.html>
- GPUジョブ実行用
 - GCC 11.3.0
 - CUDA 11.8.0
 - OpenMPI 4.1.5
- CPUジョブ実行用
 - Intel CPU用に最適化されたコードパスがあるので、GPU用と別途ビルドした
 - Intelコンパイラ(2020.4.304)
 - 脆弱性があるため新しいoneAPIで要再ビルド
 - CPU専用ノードのクラウドシステムはローカルSSDが無いため、ローカルSSDをスクラッチディレクトリとして使う場合はGPUノード(Type II サブシステム)で実行

不老向け設定

- モジュールファイル作成
- ジョブスクリプトテンプレート作成

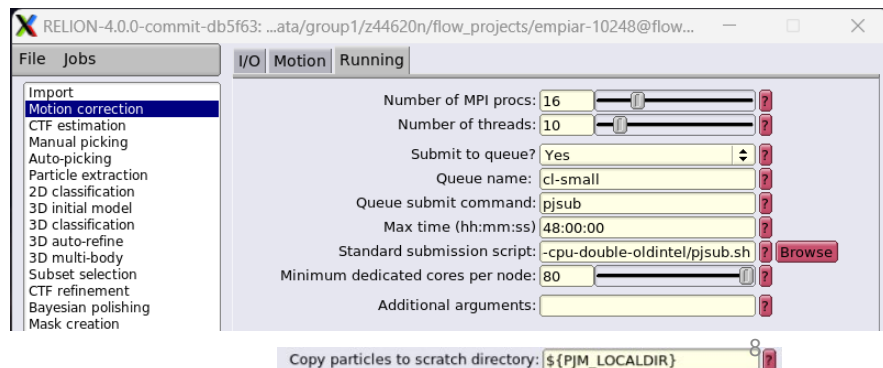
```
#!/bin/sh
```

```
#PJM -L rscgrp=XXXqueueXXX  
#PJM -L node=XXXnodesXXX  
#PJM -L elapse=XXXextra1XXX  
#PJM --mpi proc=XXXmpinodesXXX  
#PJM -o XXXoutfileXXX  
#PJM -e XXXerrfileXXX  
#PJM -S
```

```
module load relion/build-4-db5f63-cpu-double-oldintel
```

```
mpirun -machinefile $PJM_O_NODEINF -np XXXmpinodesXXX XXXcommandXXX
```

- X forwardingでGUIをローカルへ飛ばす
- パラメータ指定してジョブ投下



cryoSPARCの環境構築@不老

- 公式手順のうちcluster configurationを参考に構築

- <https://guide.cryosparc.com/setup-configuration-and-management/how-to-download-install-and-configure>

- ラウンドロビンでたまたま割り当てられたログインノードで構築・稼働
- 10個の連番TCPポートが必要なので、デフォルトの39000~39010を確保

ローカルPC

ウェブUIの操作



マスターノード

マスタープロセス群(常駐)

supervisord
管理プロセス(親プロセス)

HTTP

Web app server

MongoDB

Command Core server

Command Vis server

Command RTP server

Web app API server

pjsub

ジョブスケジューラ

ワーカーノード

cryosparc_worker
各種計算を実行

HTTPS

ライセンスサーバー
<https://get.cryosparc.com>

I/O

ストレージ

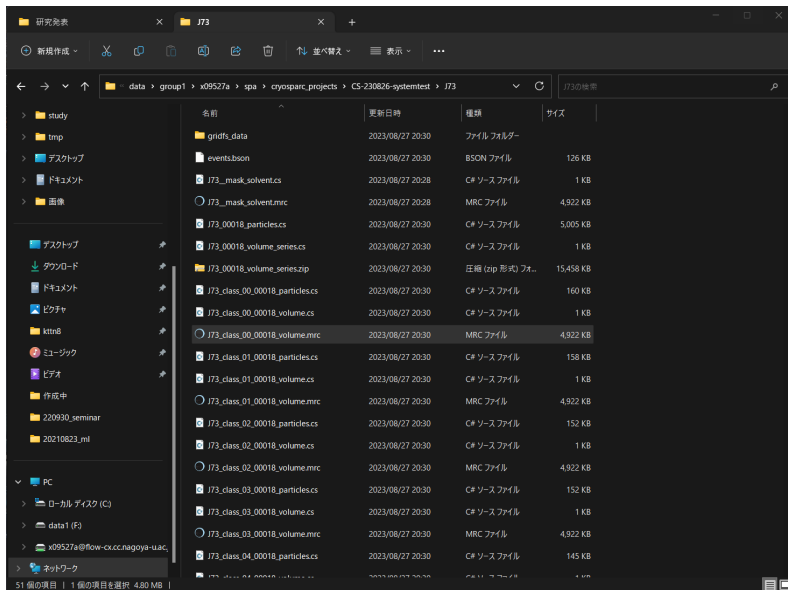
- cx-share, cx-single用のコンフィグおよびジョブスクリプトテンプレート作成

- マスタープロセス稼働中のログインノードのTCP 39000に対しsshポートフォワードで接続

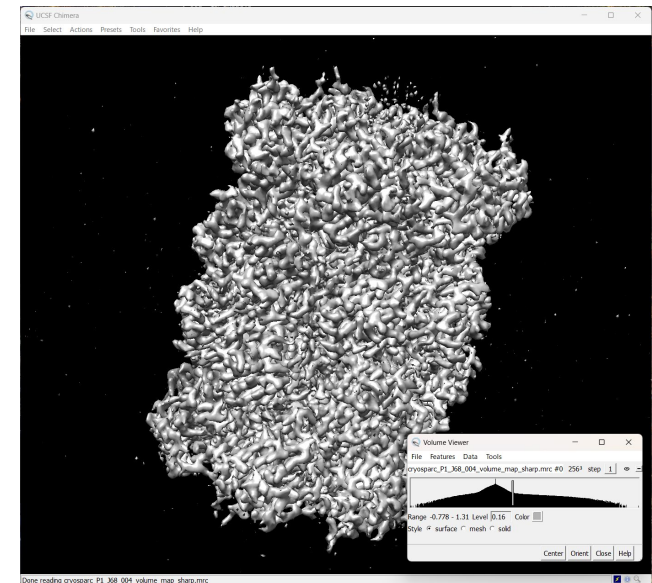
計算結果の確認・利用方法

- 結果の可視化・原子モデル構築など、3D可視化が適時必要
 - ローカルPCのWindowsにインストールしてある可視化ソフトを利用
- WinFSP/SSHFS-Winを利用したファイルブラウザが便利だった
 - 学内回線であればストレス無く実行できた

Windowsのエクスプローラーで不老のストレージが見れる



Windowsに入れてあるソフトで3D可視化



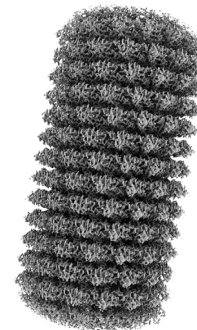
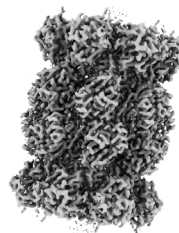
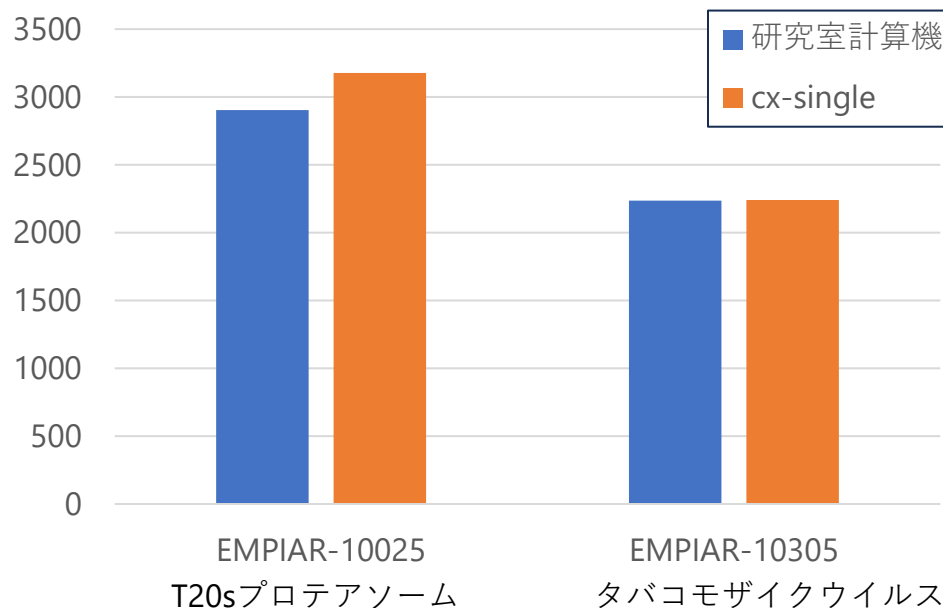
cryoSPARCベンチマーク

- cryoSPARC標準搭載のベンチマーク取得機能を利用
- ラボ所有の計算機とType II サブシステム(cx-single)で比較
 - cryoSPARCのほとんどのジョブはGPU 1基での計算のため、GPU 1基のみで計測

	ラボ計算機	cx-single
CPU	AMD EPYC 7502P 32core 2.5GHz	Intel Xeon Gold 6230 40core 2.1GHz
RAM	256 GB	384 GB
GPU	RTX3090 (24 GB)	V100 (32 GB)
ストレージ	内蔵HDD 10 TB	ホットストレージ
スクラッチ	NVMe 1 TB	NVMe 6 TB

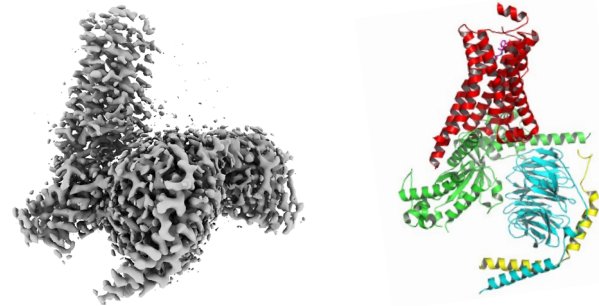
- 単粒子解析WSとしてミドルレンジ～ハイエンドのラボ所有計算機と同等の性能であると確認できた

ベンチマークデータ処理時間 [秒]



実践計算② カンナビノイド受容体 (EMPIAR-10671)

解析フロー



- Structura社がAWSでの構造解析ベンチマークとして利用した中規模データセット (2021年)
- 2,756 movies → 488,490 particle images
- AWS ParallelCluster (2021年 Structura社)
 - 計算時間 4.29 時間
 - 分解能 > 3 Å
 - 計算コスト \$47.12 (US)
- 不老cx-single/cx-share
 - 計算時間 19 時間
 - 分解能 2.90 Å
 - 計算コスト 1,125 円
- 大変コストパフォーマンスが良い

まとめ・課題点

成果

- 単粒子解析ソフトのデファクトスタンダードであるRELIONとcryoSPARCを不老向けに環境構築
- 一通りの構造決定計算を遂行できることを確認
- コストパフォーマンスに大変優れることが確認

課題点 – cryoSPARCの運用

- ログインノードで一定のリソース専有・負荷がかかる
 - 複数の常駐プロセスの連続稼働が必要
 - 10個連番のTCPポート確保 (複数ユーザーが使い始めたときに最初に問題になる)
 - CPU・RAM消費は基本的に小さいが、CPU 数コア100% 消費が続く場合もある
 - 機密性
 - ウェブUIは別ユーザーでもアクセスできてしまう
 - ウェブUIとの通信、cryoSPARC内のプロセス間通信は暗号化されない
- ログインノードの仮想化システムや、cryoSPARC用の専用ログインノードがあればよい？

マスターノード最小要件

Component	Minimum Requirement
CPU	4+ cores
RAM	16GB+
System Storage	250GB+ HDD
Fast Local Storage	Not Required
GPU	Not Required
Network	1Gbps link to storage servers

将来展望

- RELIONを主軸とした計算のプラットフォームとしては即利用可能
- cryoSPARCの課題点を解決できれば、多くの単粒子解析ユーザーにとって魅力的な計算プラットフォームとなる
- 今後の検討事項
 - 大量の並行計算が可能なことを利用した新しい計算ワークフローの研究開発
 - メタヒューリスティクスによる自動パラメータチューニングなど
 - ニューラルネットワークを利用した単粒子解析計算での活用
 - Cryo-ET計算での活用
- 名古屋大学に自動撮影型のクライオ電顕が導入されたらば、こんなことも？
 - 不老と直結してon-the-flyでの前処理を実行
 - そのまま不老で構造解析を完結
- 今後とも活用させていただきます。

以上