SACLA でのシリアル結晶学のためのデータ処理パイプライン

中根 崇智

東京大学 理学系研究科生物科学専攻

X 線自由電子レーザ XFEL は、フェムト秒オーダの短時間に、10 の 12 乗個ほどの光子を発生させることができる。この短時間では原子は事実上動かないため、放射線損傷の影響を受けない常温構造を観測できる。励起レーザと組み合わせるポンププローブ法や二液混合系を用いることで、蛋白質の動的な構造変化を捉えることもできる。

蛋白質微結晶は一発の XFEL パルスによって破壊されるので、ランダムに配向した数万から数十万個の結晶から独立した回折像を取得する連続フェムト秒結晶学 SFX によってデータセットを構築する。結晶は脂質キュービック相やグリースなどの高粘度流体に埋め込まれたり、液滴として、X 線との相互作用点に供給されるが、全ての X 線パルスが結晶に当たるわけではないため、生データは回折像のさらに数倍の枚数になる。これは従来の結晶学の百から千倍の規模である。そのため、目視による回折像の評価は非現実的である。また、XFEL のビームタイムは放射光以上に競争率が高く、貴重であるから、データ測定に値する蛋白質の選定や時間配分といった実験戦略を臨機応変に立てて効率的に実験しなければならない。したがって、自動化された、リアルタイム性の高いデータ処理システムが不可欠である。

本発表では、日本の XFEL 施設である SACLA における SFX 実験のために開発したデータ処理パイプラインを紹介する。欧米で用いられているオープンソース・プログラム Cheetah と CrystFEL を元にして、SACLA の実験系や計算機システムにあわせた修正と、高速化・自動化のための改良を行った。 SACLA のデータ処理系はオンラインとオフラインの二系統からなるため、パイプラインも二段階で動作する。オンライン処理系は、画像に含まれる回折点を検出し、ヒット率や検出器飽和度を数秒未満の遅延で表示する。これは、結晶密度の評価、X 線の軸合わせ、減衰率の設定に役立つ。オフライン処理系は、測定が完了した run から十分な回折点を含む画像のみを選別し、補正や圧縮を行ったうえで、波長などのメタデータと合わせて HDF5 形式で出力し、CrystFEL による指数付けと積分を行う。パイプラインの制御と結果の可視化は、wxPython を用いて開発したグラフィカル・ユーザ・インターフェース GUI から可能である。遅延時間を最小化するため、パイプラインにはスレッドとノード間並列化を施した。

本システムは、SACLA における SFX 実験の 9 割以上で使用されている。例えば、ある 4 日間のビームタイムでは、約 20 種類の蛋白質結晶から約 510 万枚の画像を取得した。これは 41 TB に及ぶが、パイプラインによる選別と圧縮を経て、出力サイズは 3.2 TB に抑えられた。蛋白質あたりでは数十から数百 GB であり、放射光実験と同等の規模である。パイプラインの出力を元に分子置換や実験的位相決定を行い、ビームタイム中に初期構造が得られている。このように、放射光施設における回折実験に迫る、効率的な実験が可能となっている。

また、パイプラインは、お仕着せのブラックボックスとせず、ユーザの習熟度や実験の性質に応じてカスタマイズ可能であることが重要である。そのため、GPL ライセンスでソースコードを 公開している。

文献: Nakane et al. "Data processing pipeline for serial femtosecond crystallography at

SACLA" *Journal of Applied Crystallography* 49.3 (2016)