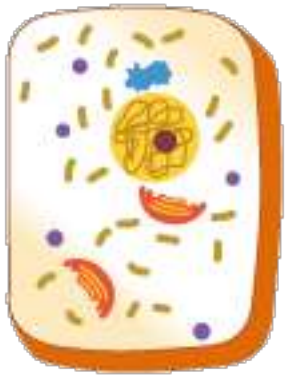
KEK構造生物学研究センター におけるクラウド利用

山田悠介

高エネルギー加速器研究機構 物質構造科学研究所 構造生物学研究センター

総合研究大学院大学 高エネルギー加速器研究科 物質構造科学専攻

構造生物学とは



核酸

タンパク質

炭水化物(糖)

脂質

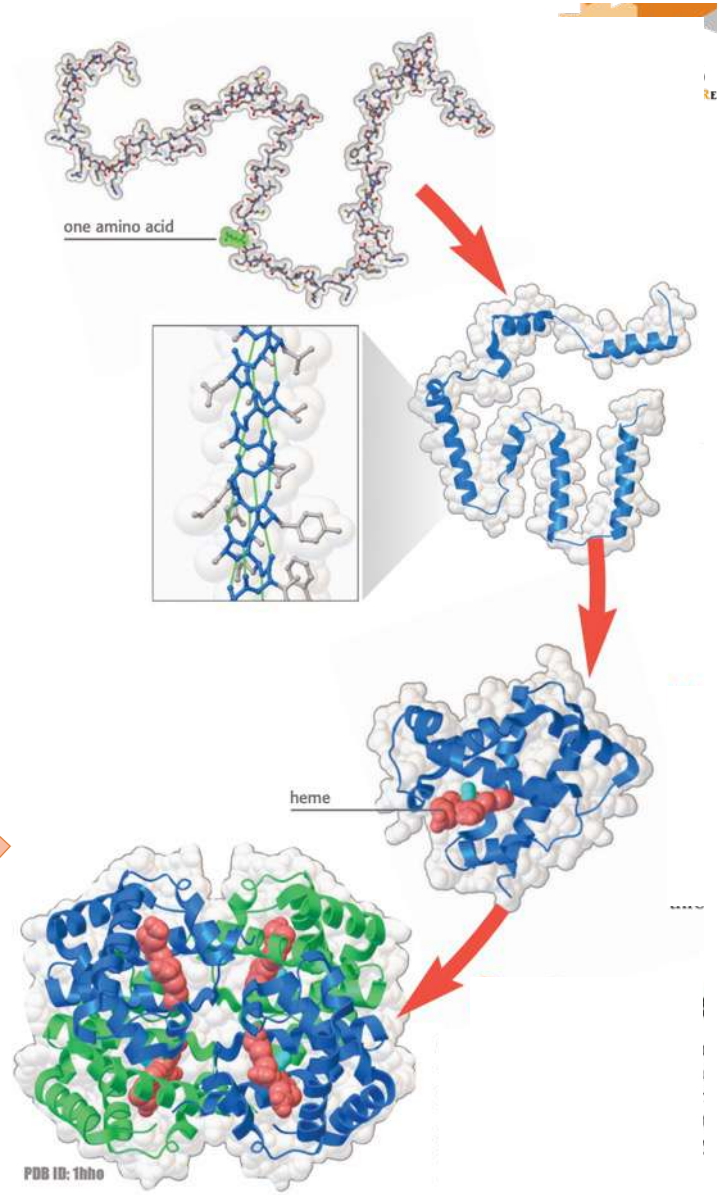
遺伝情報

配列情報

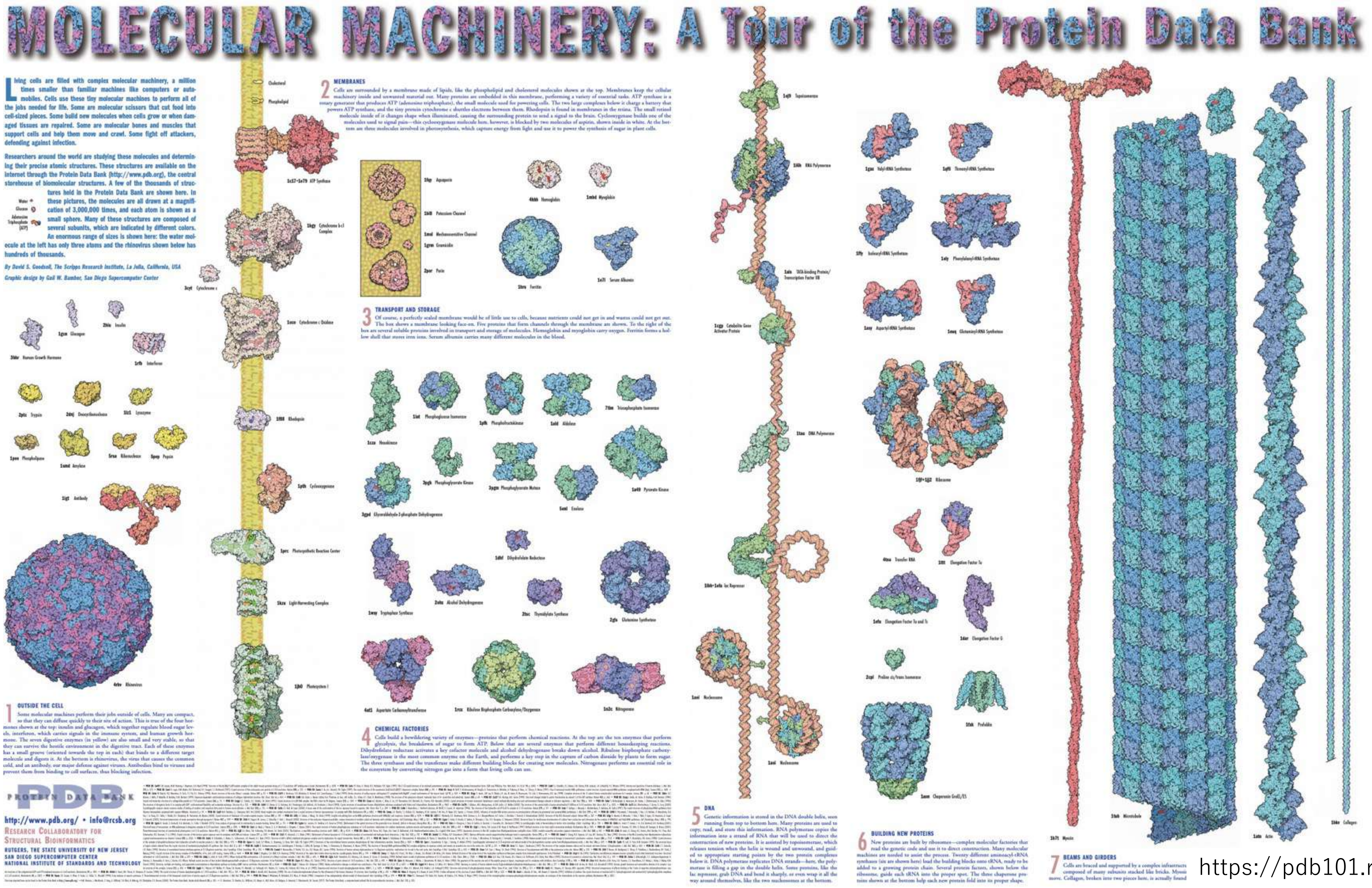
設計図

機能

ポリペプチド鎖(20種類のアミノ酸)



https://pdb101.rcsb.org/



IS LIFE?
.OGY RESEARCH CENTER

https://pdb101.rcsb.org/

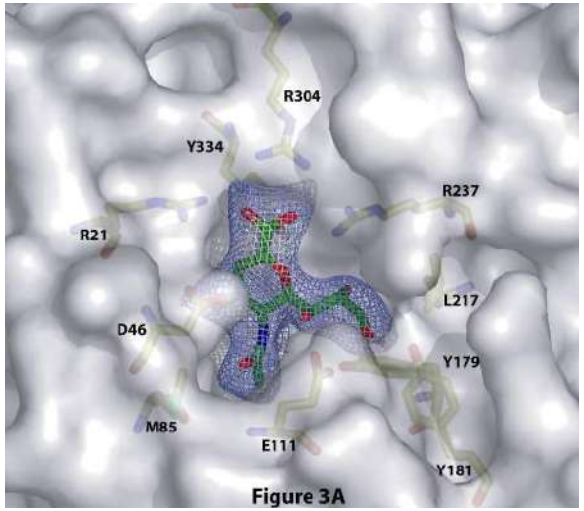
構造生物学

タンパク質の**立体構造を解明**することで、
その**機能を理解**する。

生命現象の理解

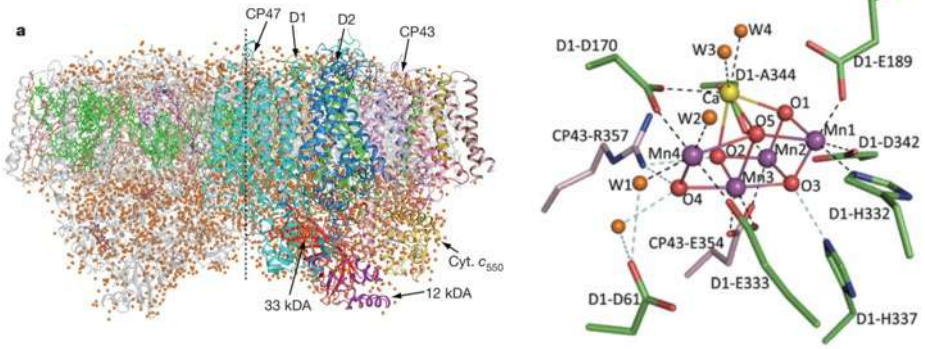
David S. Goodsell, RCSB Protein Data Bank.
doi: 10.2210/rcsb_pdb/goodsell-gallery-022
doi: 10.2210/rcsb_pdb/goodsell-gallery-025

薬剤設計



J. Biol. Chem. 280, 469-475 (2005).

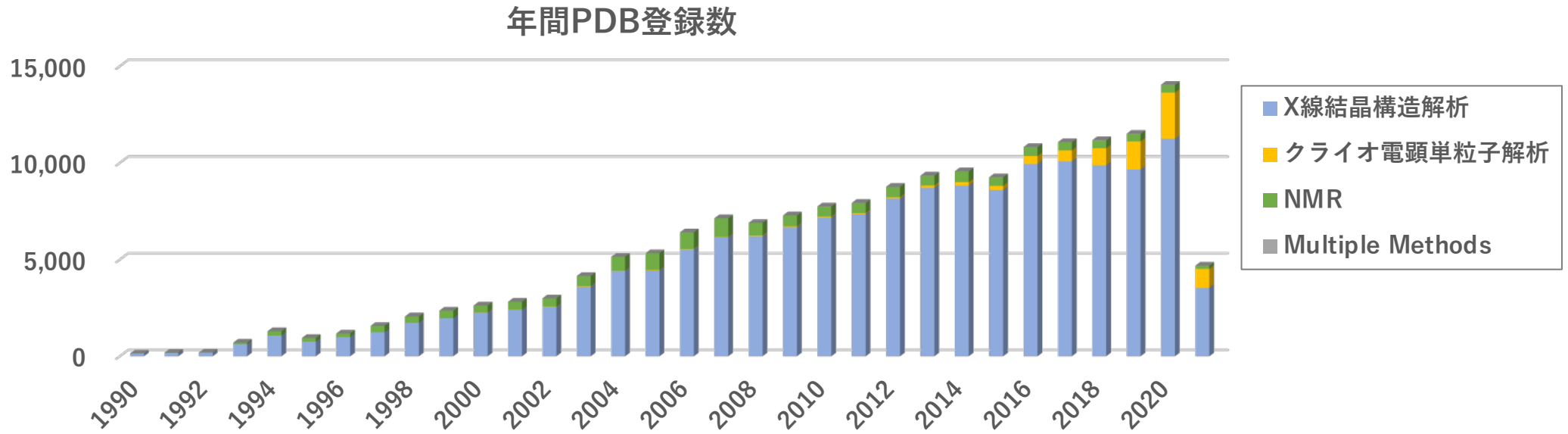
工業応用



Nature 473, 55–60 (2011).

タンパク質の構造解析手法(代表的なもの)

	X線結晶構造解析	クライオ電顕単粒子解析	NMR
測定に用いる試料	結晶	溶液の氷薄膜	溶液
分子の大きさ	問わない	100 kDa以上	50 kDa以下
構造状態	静的	静的・動的	動的
分解能	中～高	低～中・高	中～高
データ測定時間	数秒～数分	数時間～数日	数時間
データ解析時間	数十分から数十時間	数日～数週間	数日～数週間



大学共同利用機関 高エネルギー加速器研究機構(KEK)

- 素粒子原子核研究所
- **物質構造科学研究所**　物質・生命の研究
  - **構造生物学研究センター**
- 加速器研究施設
- 共通基盤研究施設
  - 計算科学センター

KEKつくばキャンパス

KEK 構造生物学研究センター（SBRC）

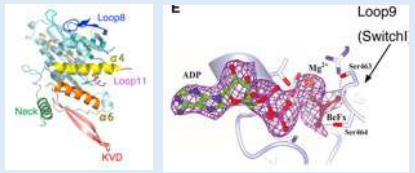
最先端の測定設備、測定手法の開発

共同利用、施設利用

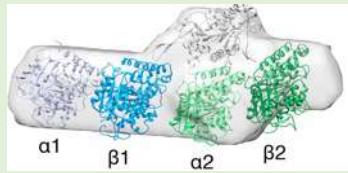
大学・公的研究機関

民間企業(製薬会社)

X線結晶構造解析



X線溶液散乱



WHAT IS LIFE?
STRUCTURAL BIOLOGY RESEARCH CENTER

結晶化スクリーニング

クライオ電子顕微鏡

Ogawa *et al.* Cell Reports, 20, 2626-38 (2017)

創薬等先端技術支援基盤プラットフォーム事業 (BINDS, AMED)

WHAT IS LIFE?
STRUCTURAL BIOLOGY RESEARCH CENTER

初心者から専門家まで全ての生命科学研究者に構造生物学を！

https://www.binds.jp

KEK SBRCで運用するITリソース

ネットワーク

- 各拠点(放射光ビームライン、クライオ電顕、結晶化装置、データセンター)を10 Gbpsで接続

ストレージ

- 並列ファイルシステム(GPFS)　280 TB (2021年度内に1 PBに増強)

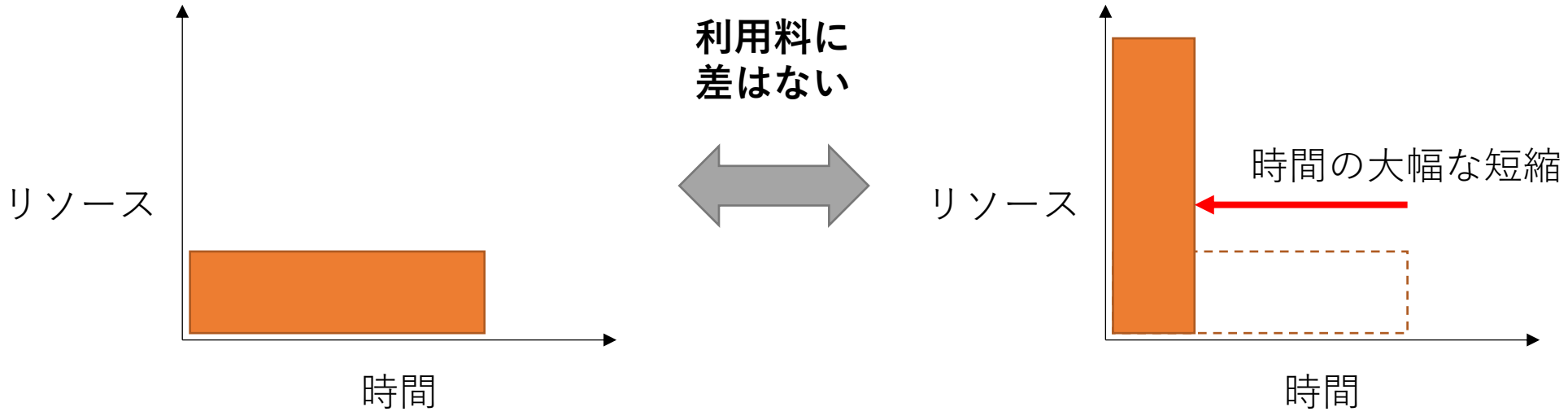
計算リソース

- MX: 解析クラスタ 17ノード (420 CPU、1,896 GB memory)
- CryoEM: 解析ワークステーション(2 ~ 4 GPUs/WS) 9台

パブリッククラウドの利用

パブリッククラウド

- クラウドコンピューティングを広く一般に提供 (AWS, GCP, Azure, Oracle, ...)
- 多量(ほぼ無尽蔵)のリソース
- マネージド (利用者側での運用管理がほぼ不要)
- 新しいテクノロジーにも迅速に対応
- 従量課金

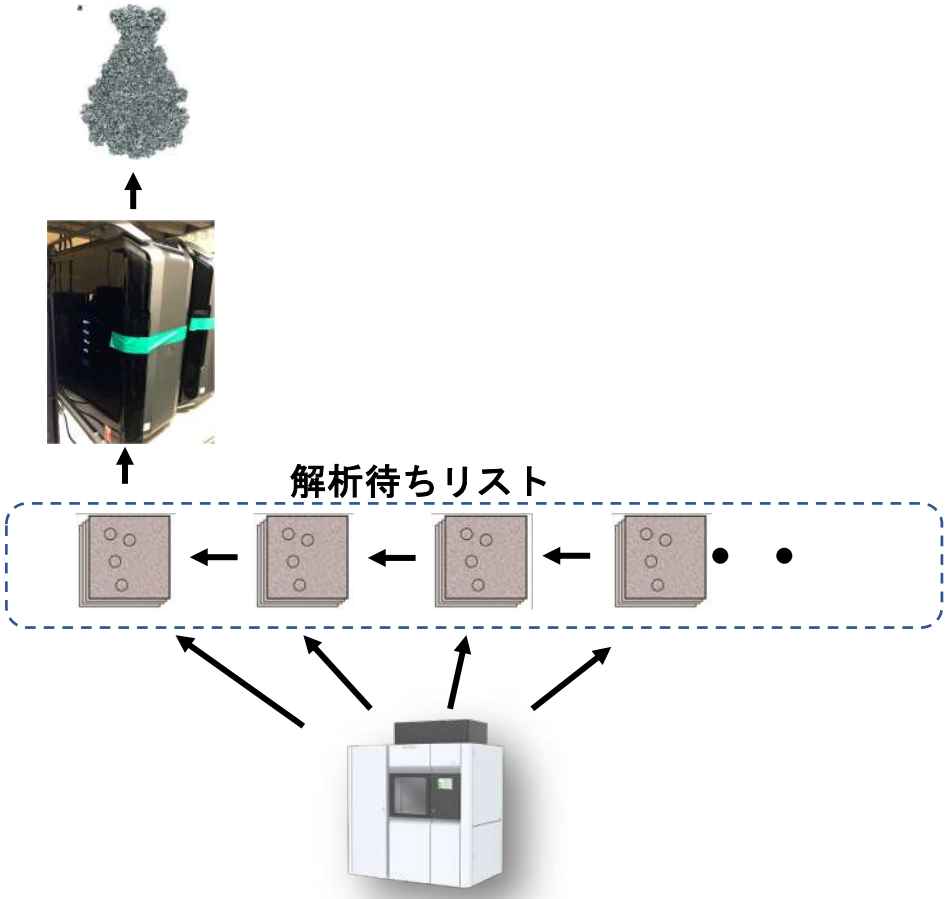


パブリッククラウドを用いた実験データドリブンな仮想解析環境

WHAT IS LIFE?
STRUCTURAL BIOLOGY RESEARCH CENTER

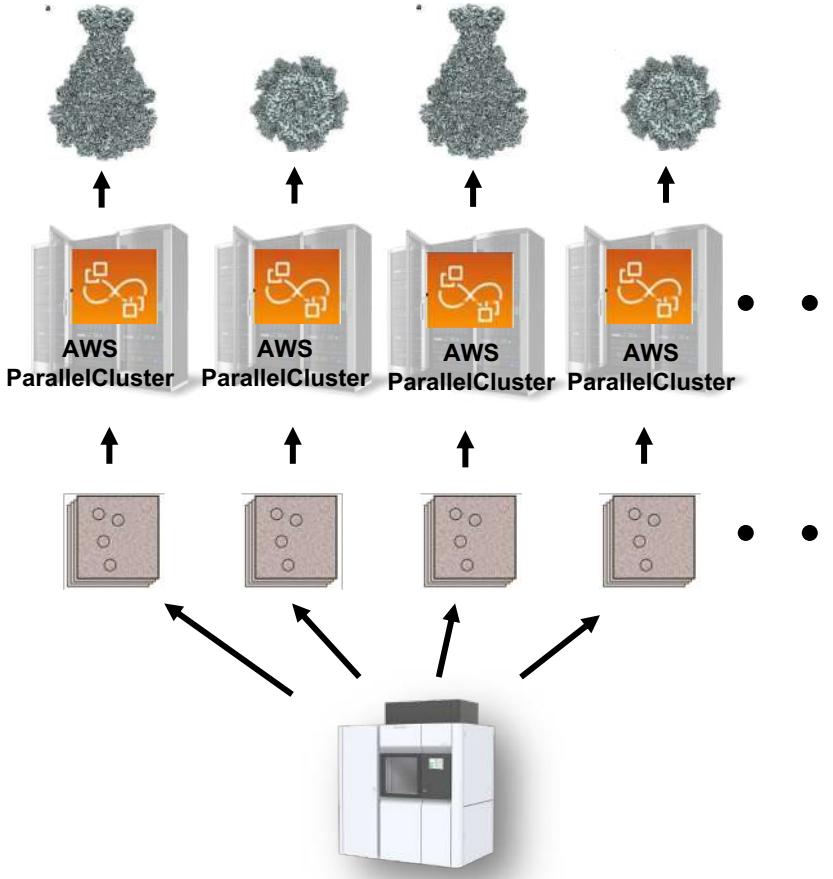
現状

限られた計算リソースを複数プロジェクトで共有
測定完了後に手動で解析
解析待ちリストが発生



これから

実験データ測定時に仮想的な専用解析環境をAWS上に瞬時に構築(*)
測定途中で自動解析
ハイスループット、低コスト



クラウド利用の始まり

第1回　クライオ電顕解析初心者講習会～データ処理～

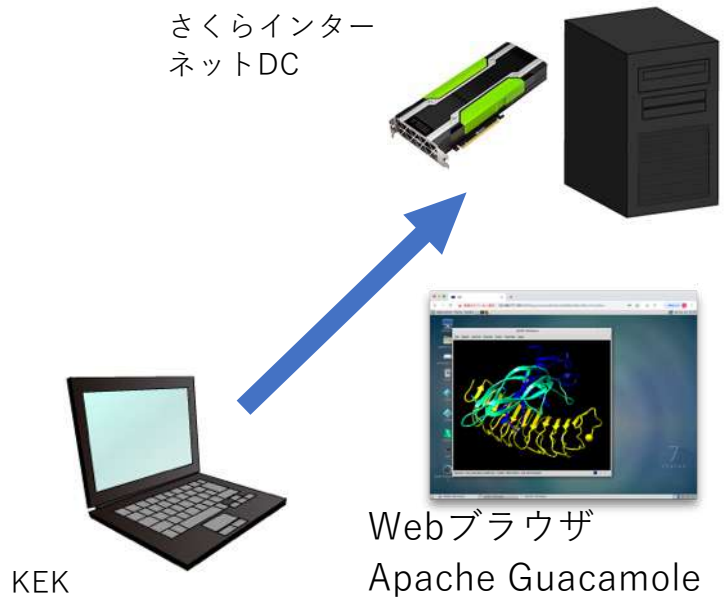
クライオ電子顕微鏡によるタンパク質の単粒子解析に興味はもっているものの、実際にご自身では解析されたことがない初心者の方を対象に、構造解析ソフトウェアRELION-3.0beta（Sjors Scheres, MRC-LMB）を中心としたデータ解析講習会の第1回目を開催します。

2018年12月13日（木）～14日（金）　開催終了

講師
安永卓生：九州工業大学情報工学部

実習インストラクター
重松秀樹：理化学研究所/SPring-8
守屋俊夫：KEK・物質構造科学研究所・構造生物学研究センター
荒牧慎二：Tietz Video & Image Processing Systems GmbH

さくらインターネットのデータセンター(DC)に設置されたGPU搭載マシン25台に参加者それぞれがWebブラウザからアクセスし、RELION-3.0を用いて解析演習を行った。



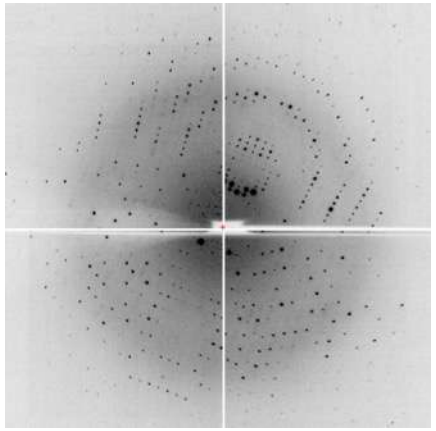
タンパク質X線結晶構造解析の流れ

WHAT IS LIFE?
STRUCTURAL BIOLOGY RESEARCH CENTER

精製標品

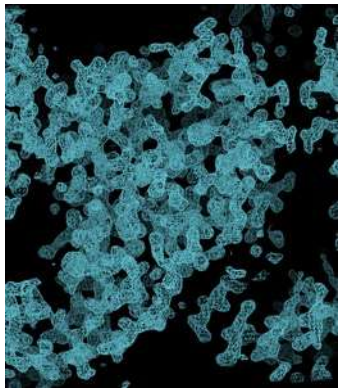
結晶化

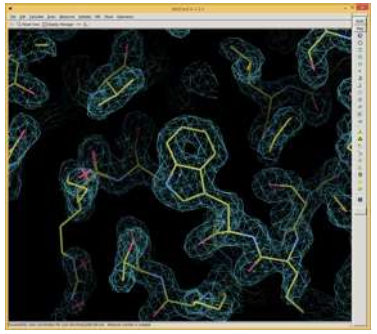
X線
照射



X線回折像

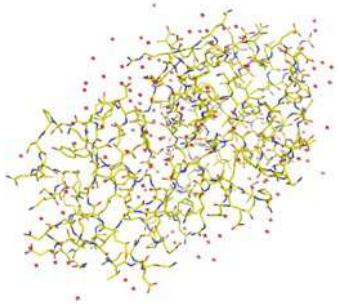
位相
決定





電子密度図

モデル
構築

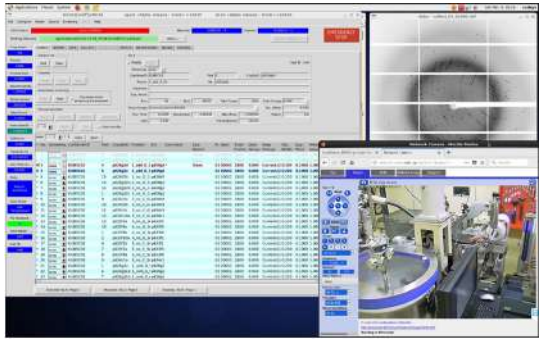


構造モデル

ターゲット選定 → **試料調整・結晶化** → **回折データ測定** → **データ解析**

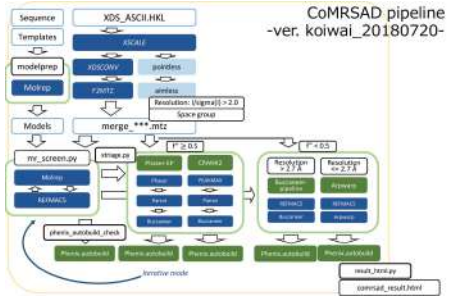
SBRCでのタンパク質結晶構造解析自動化開発

全自動結晶化
スクリーニング
システム



全自動回折データ
収集システム

データベース/
自動処理



全自動構造解析
パイプライン

ターゲット選定 → **試料調整・結晶化** → **回折データ測定** → **データ解析**

タンパク質結晶構造解析の流れ

全自動結晶化スクリーニングシステム(PXS)

Dispensing and storage　　Monitoring drops

Dispenser	Mosquito LCP (0.1 μL) ~ 3 min/plates
Imager	RockImager (20 and 4 deg.), SONICC (UV and SHG)
Incubator capacity	1,920 plates (20 and 4 deg.)

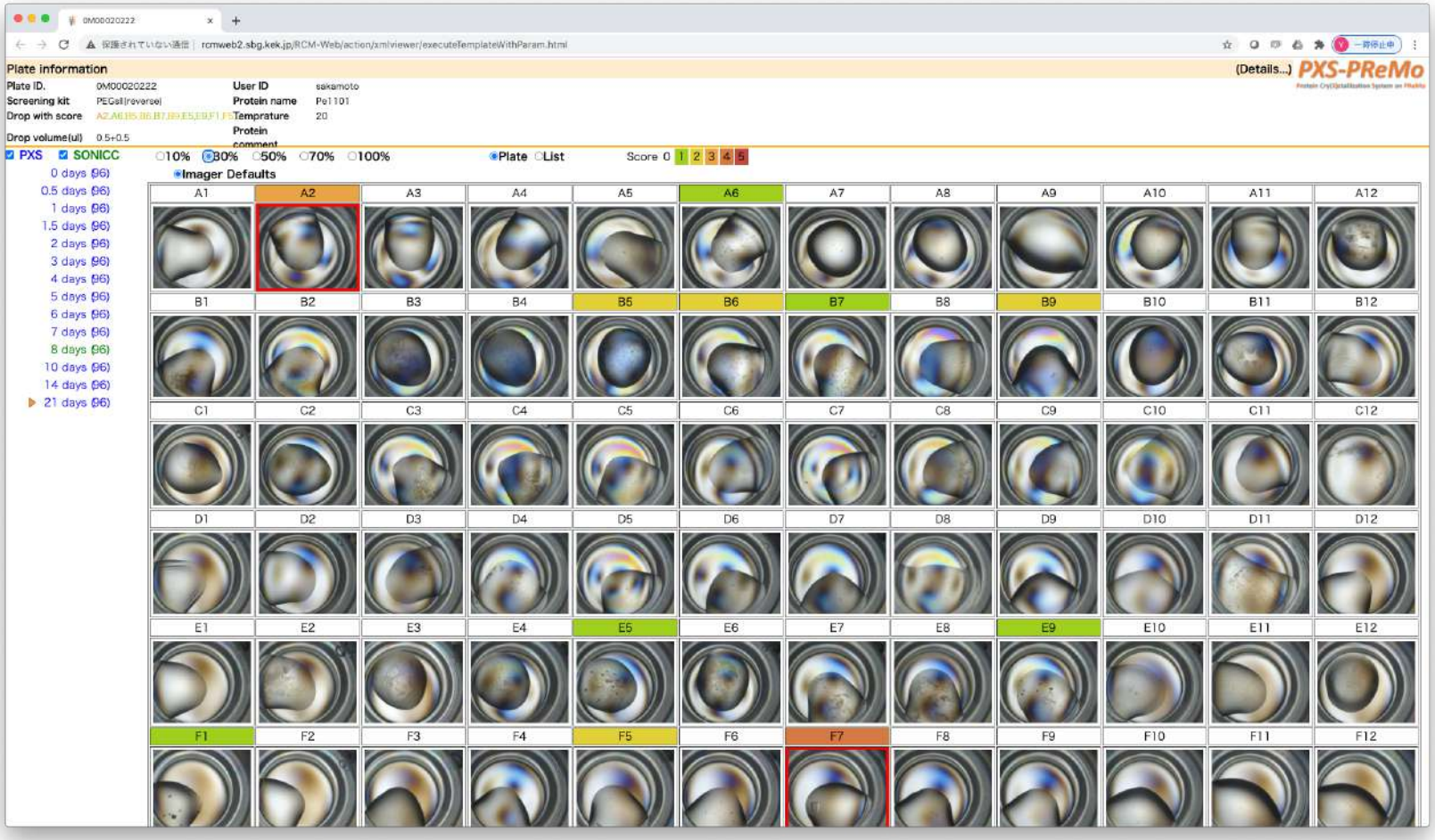
結晶化ドロップ作成と保管、ドロップ観察が全て自動化

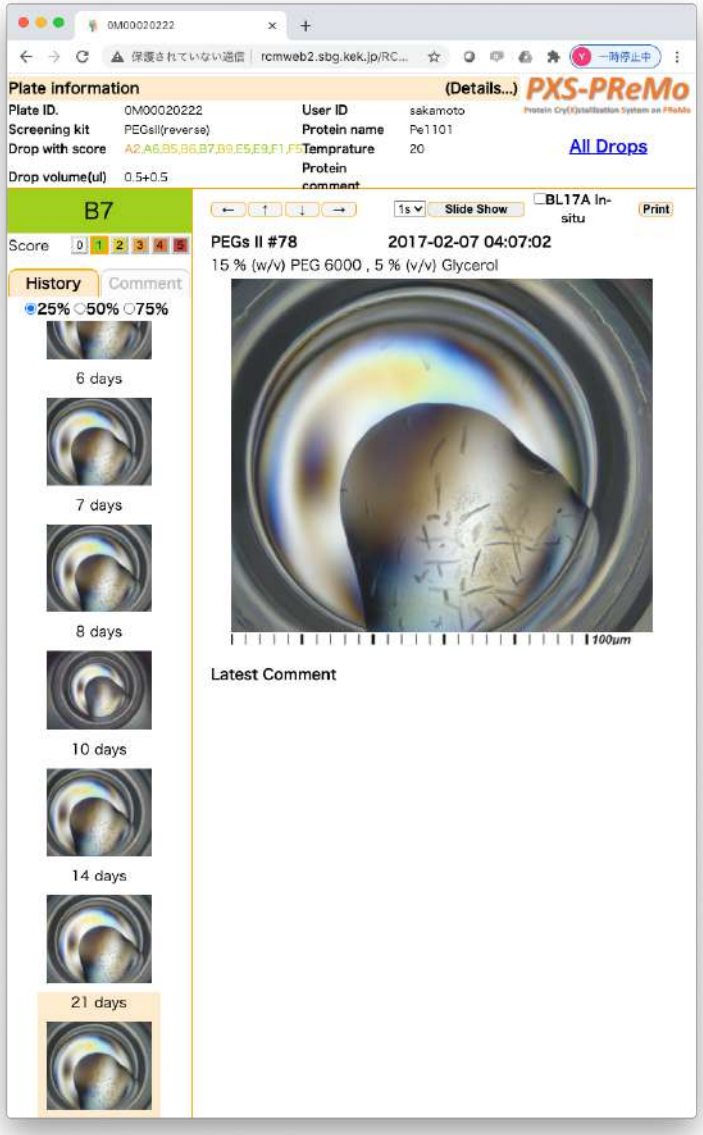
2019年度は906プレートを作成した。

加藤龍一

Kato *et al. Acta Cryst.* **F77** 29-36 (2021).

PXS-PReMo: 結晶化スクリーニングデータベース

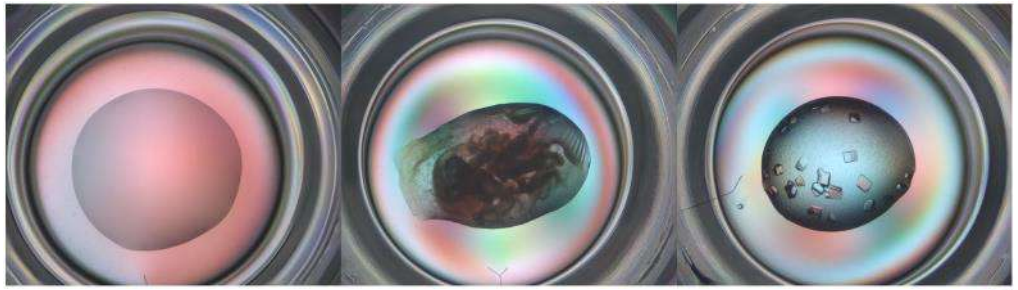




現在は、全ての画像を人が目視で評価

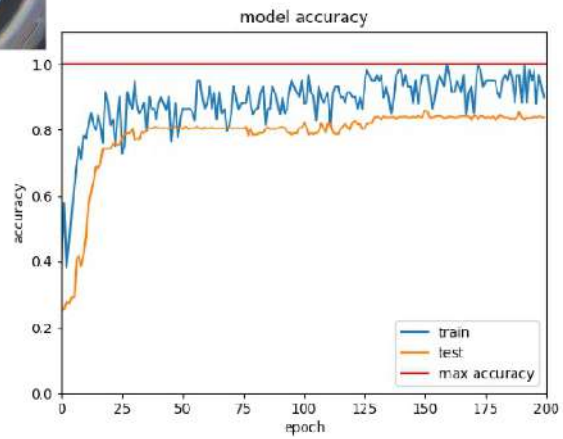
深層学習を利用した結晶化ドロップ評価の自動化

三浦佑晟
櫻井鉄也
(筑波大
AIセンター)



Clear　Precipitated　Crystal

Deep learning



RESEARCH ARTICLE

Classification of crystallization outcomes using deep convolutional neural networks

Andrew E. Bruno[1], Patrick Charbonneau[2,3], Janet Newman[4], Edward H. Snell[5,6], David R. So[7], Vincent Vanhoucke[7]*, Christopher J. Watkins[8], Shawn Williams[9], Julie Wilson[10]

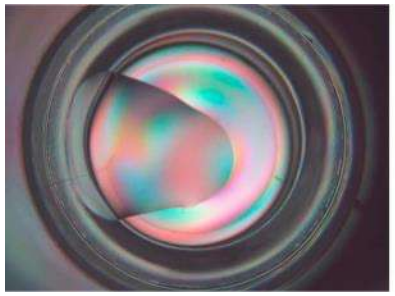
1 Center for Computational Research, University at Buffalo, Buffalo, New York, United States of America, **2** Department of Chemistry, Duke University, Durham, North Carolina, United States of America, **3** Department of Physics, Duke University, Durham, North Carolina, United States of America, **4** Collaborative Crystallisation Centre, CSIRO, Parkville, Victoria, Australia, **5** Hauptman-Woodward Medical Research Institute, Buffalo, New York, United States of America, **6** SUNY Buffalo, Department of Materials, Design, and Innovation, Buffalo, New York, United States of America, **7** Google Brain, Google Inc., Mountain View, California, United States of America, **8** IM&T Scientific Computing, CSIRO, Clayton South, Victoria, Australia, **9** Platform Technology and Sciences, GlaxoSmithKline Inc., Collegeville, Pennsylvania, United States of America, **10** Department of Mathematics, University of York, York, United Kingdom

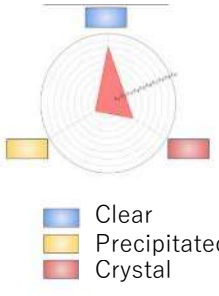
* vanhoucke@google.com

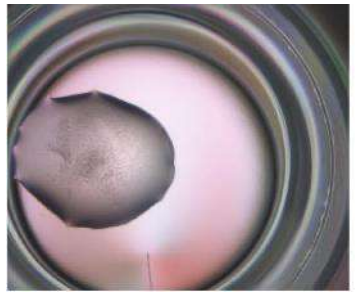
Abstract

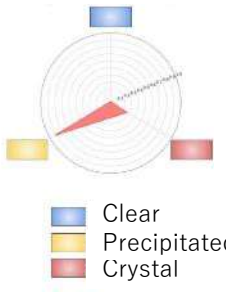
The Machine Recognition of Crystallization Outcomes (MARCO) initiative has assembled roughly half a million annotated images of macromolecular crystallization experiments from various sources and setups. Here, state-of-the-art machine learning algorithms are trained and tested on different parts of this data set. We find that more than 94% of the test images can be correctly labeled, irrespective of their experimental origin. Because crystal recognition is key to high-density screening and the systematic analysis of crystallization experiments, this approach opens the door to both industrial and fundamental research applications.

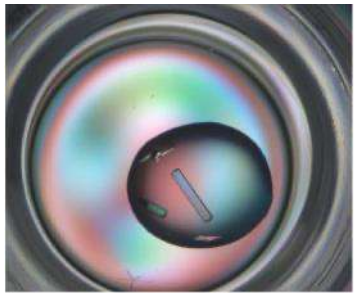
Bruno *et al*. PLoS ONE (2018)

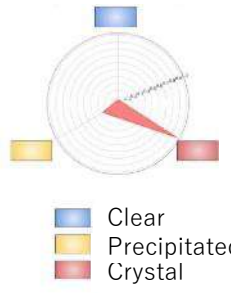


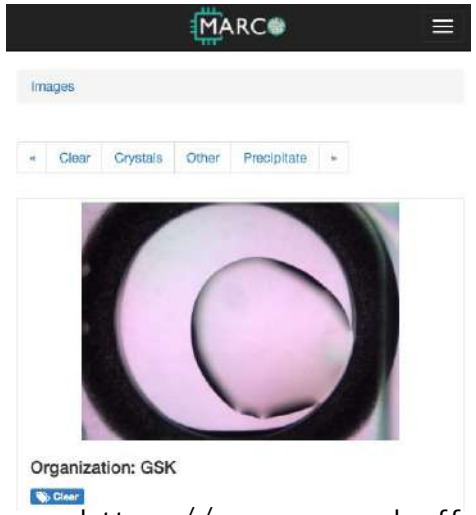










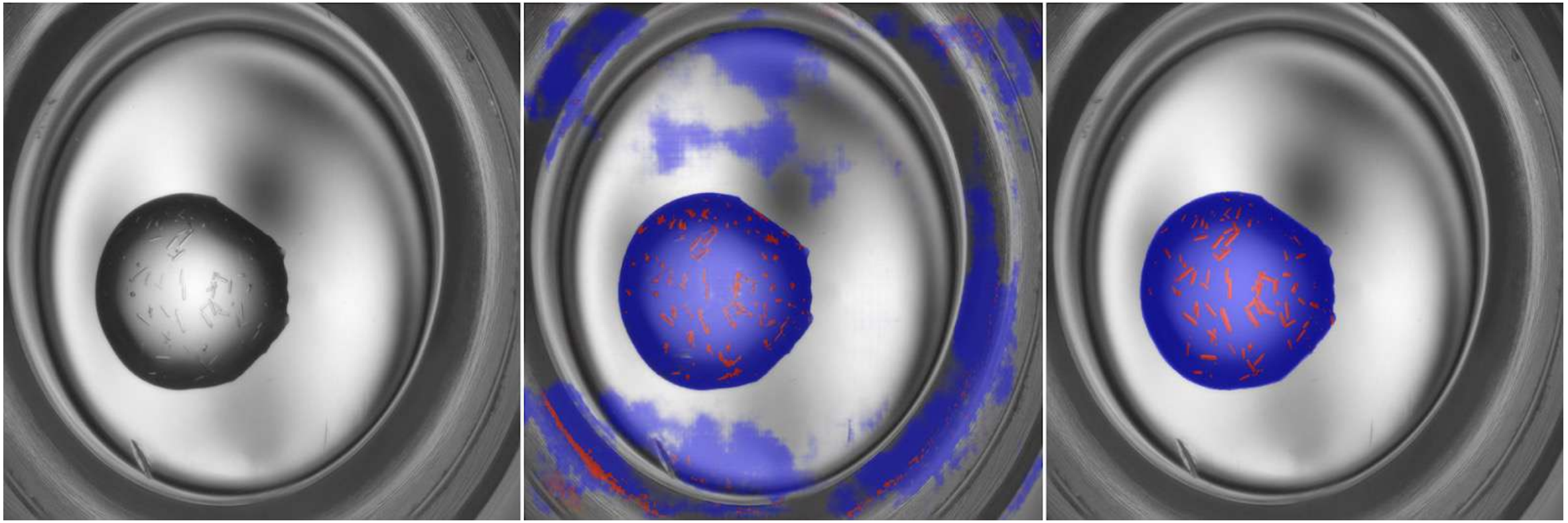


https://marco.ccr.buffalo.edu/

深層学習を用いた結晶の認識

篠田晃

元の画像　機械学習で推測した画像　正解ラベルの画像



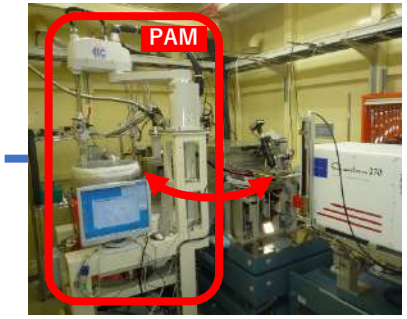
現在ラベル画像が200枚程度。今後をラベル画像増やしていくことで認識精度を向上させていく。

全自動回折データ収集・処理システム

WHAT IS LIFE?
STRUCTURAL BIOLOGY RESEARCH CENTER

サンプル交換

PAM



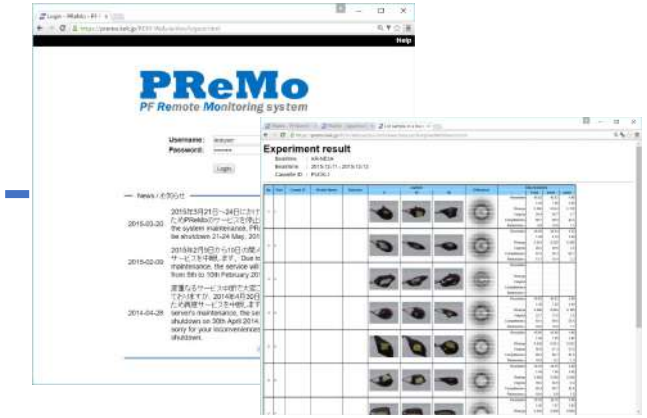
試料センタリング

回折データ測定

データ処理

サンプル交換

UGUISの改良



PReMoによる自動処理

8時間のビームタイムで測定できる試料数の目安

	BL-1A	BL-17A	BL-5A	AR-NW12A	AR-NE3A
センタリングに要する時間(分)	6	6	3	3	3
データ収集に要する時間 (分)	1.2	1.2	6	6	3
1時間あたりの試料数 (個)	8.3	8.3	6.7	6.7	10
8時間あたりの試料数 (個) (30分が準備だとする。)	62	62	50	50	75
1試料あたりのデータ量 (GB)	5	8	5	3	3
8時間あたりのデータ量 (GB)	~300	~500	~250	~150	~250

タンパク質結晶構造解析の全自動化に向けて

WHAT IS LIFE?
STRUCTURAL BIOLOGY RESEARCH CENTER

回折データ収集の迅速化・全自動化

- 200データセット/日

構造解析パイプライン

- XDS, CCP4, Phenixなどのソフトウェアのパイプライン化
- オープンソース、頻繁なアップデート

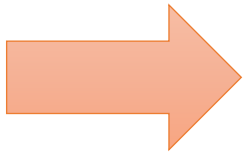
多様な実験スタイルと放射光施設の運転スケジュール

- 実験により必要とされる計算リソース量は異なる
- 年間3000時間(125日)程度の稼働時間

データ解析の工程

- 回折像 -> 回折強度データ
- 位相決定
- モデル構築
- モデル精密化

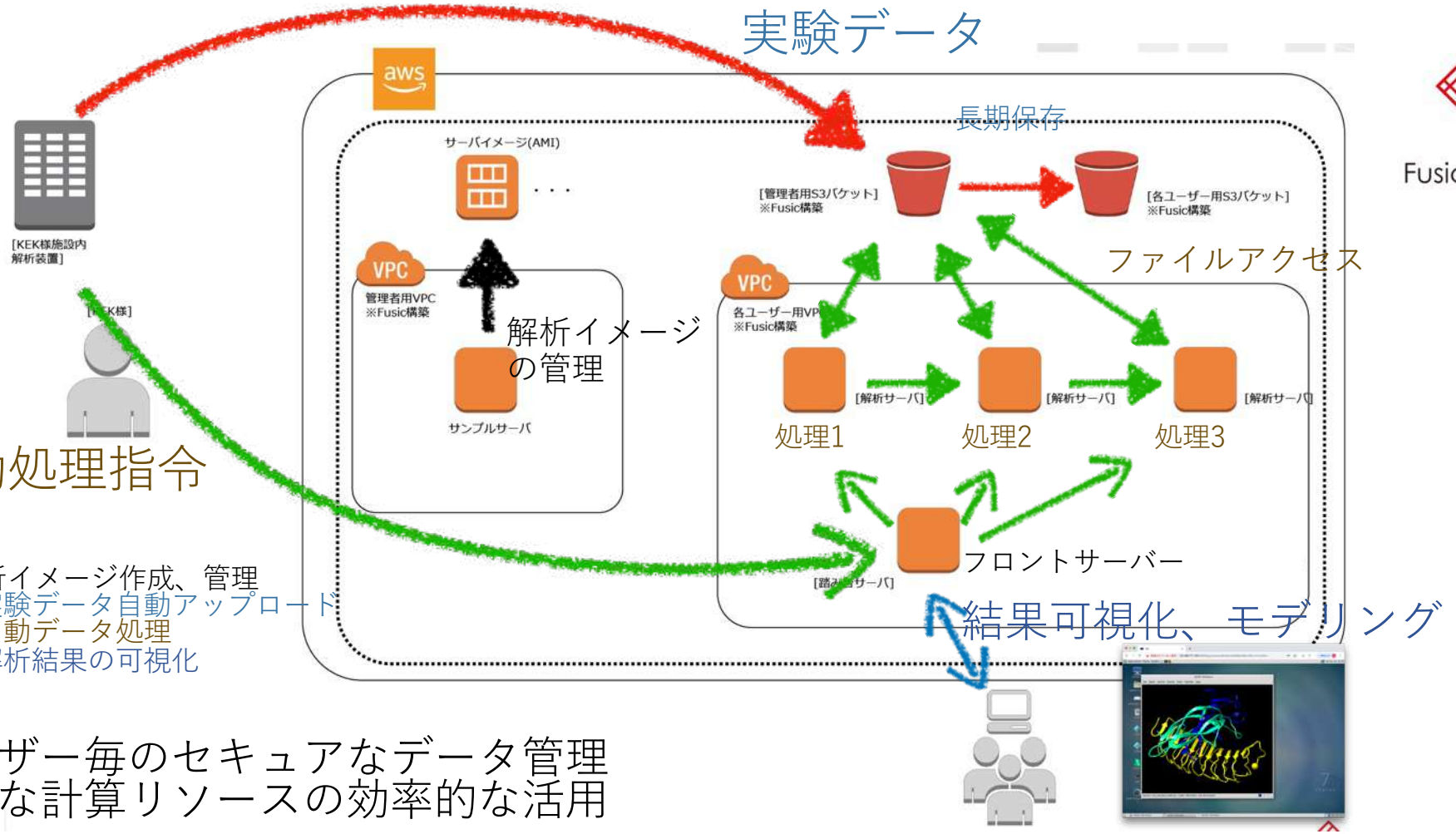
・
・
・



クラウド利用

パブリッククラウドを利用したオンデマンド解析環境構築

AWS(Amazon Web Service)を用いて構築



Fusic Co., Ltd.

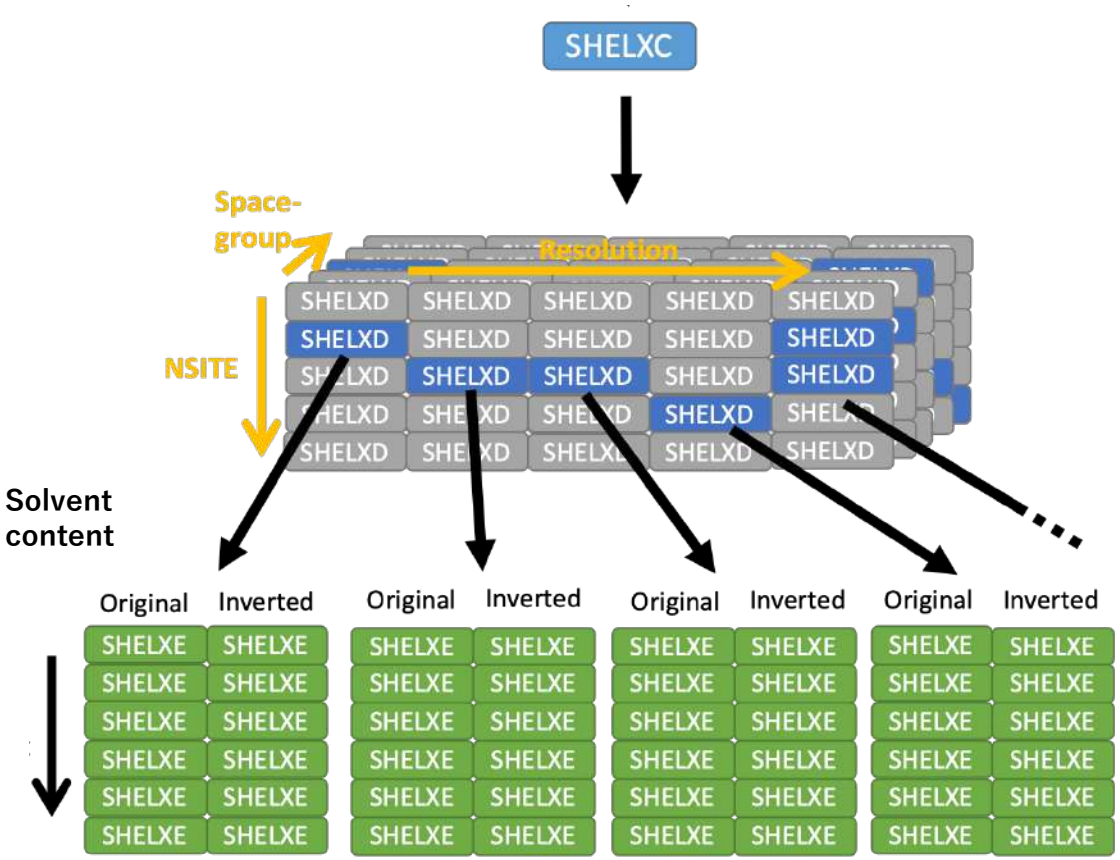
0. 解析イメージ作成、管理
1. 実験データ自動アップロード
2. 自動データ処理
3. 解析結果の可視化

ユーザー毎のセキュアなデータ管理
豊富な計算リソースの効率的な活用

BINDS産学連携高度化課題として実施(2018年度～)

Native SAD法での自動構造解析パイプライン

SHELXC, D, Eの各プロセスで必要なパラメータを網羅的探索



回折強度データ前処理

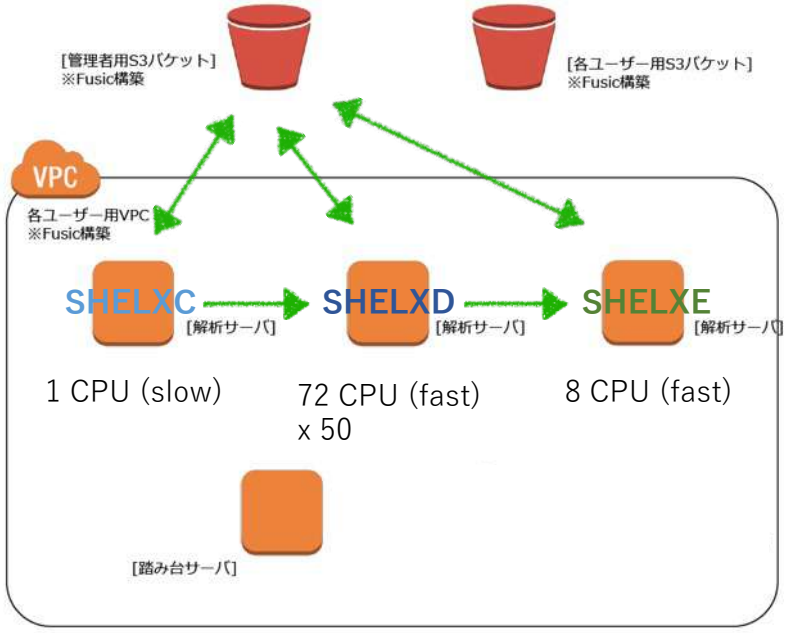
シングルスレッド
シングルパラメータ

異常散乱体(硫黄原子等)の位置同定

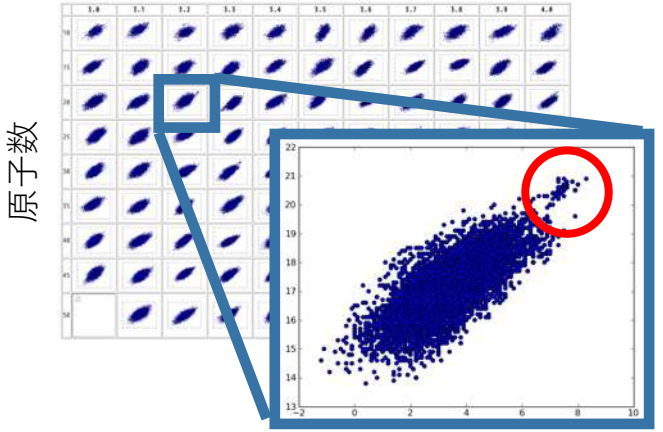
マルチスレッド
マルチパラメータ

位相計算、電子密度修飾、モデルビルディング

シングルスレッド
マルチパラメータ



分解能



KEK
(72 CPU x 1)
120 min

AWS
20 min

パブリッククラウド利用での懸念点

データのアップロード

KEK -> AWS: 80 ~ 100 MB/sec. (ほぼ1 Gbps)

1 GBのデータ:　10秒強

1 TBのデータ:　3時間強

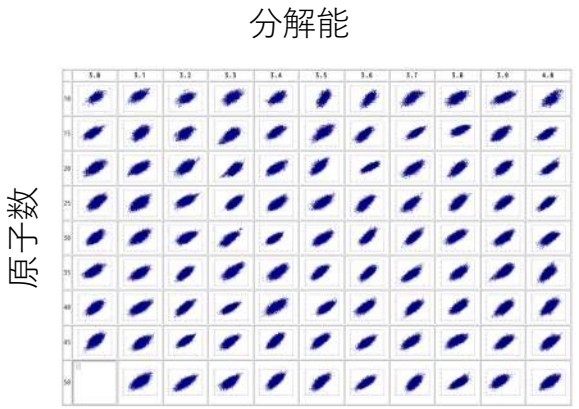
利用料 (SHELXCDEの解析の場合)

SHELXDのグリッド検索 (7 x 15 = 105)

	並列ノード数	時間 (分)	利用料 (円)
KEK 解析WS (32CPU)	1	218	
AWS c5.24xlarge (48 CPU)	10	19	600
	20	9.5	600

解析の出力ファイル(2.6GB)をダウンロードすると+ 34円

(KEK 解析WSは1台60万円)



SHELXDのグリッド解析

まとめと今後の展望

KEK構造生物学研究センター

測定設備・手法の発展に伴う計算リソースの需要増

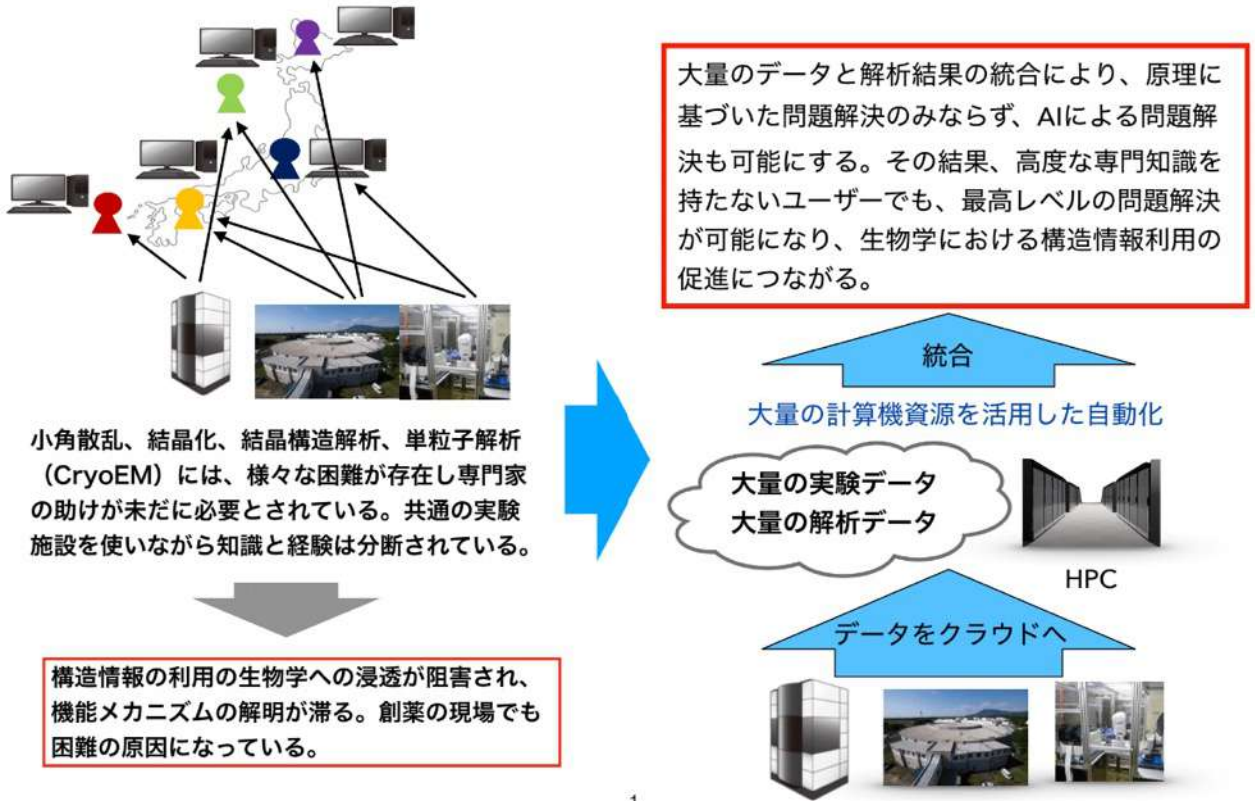
パブリッククラウドの利用

今後の展望

共有の場としてのクラウド利用

- 施設間連携
- 異分野連携

測定データ、解析データに新たな価値を



KEK SBRC