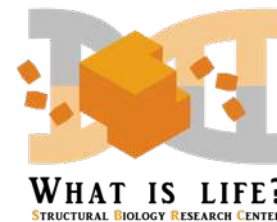


KEK構造生物学研究センター におけるクラウド利用

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

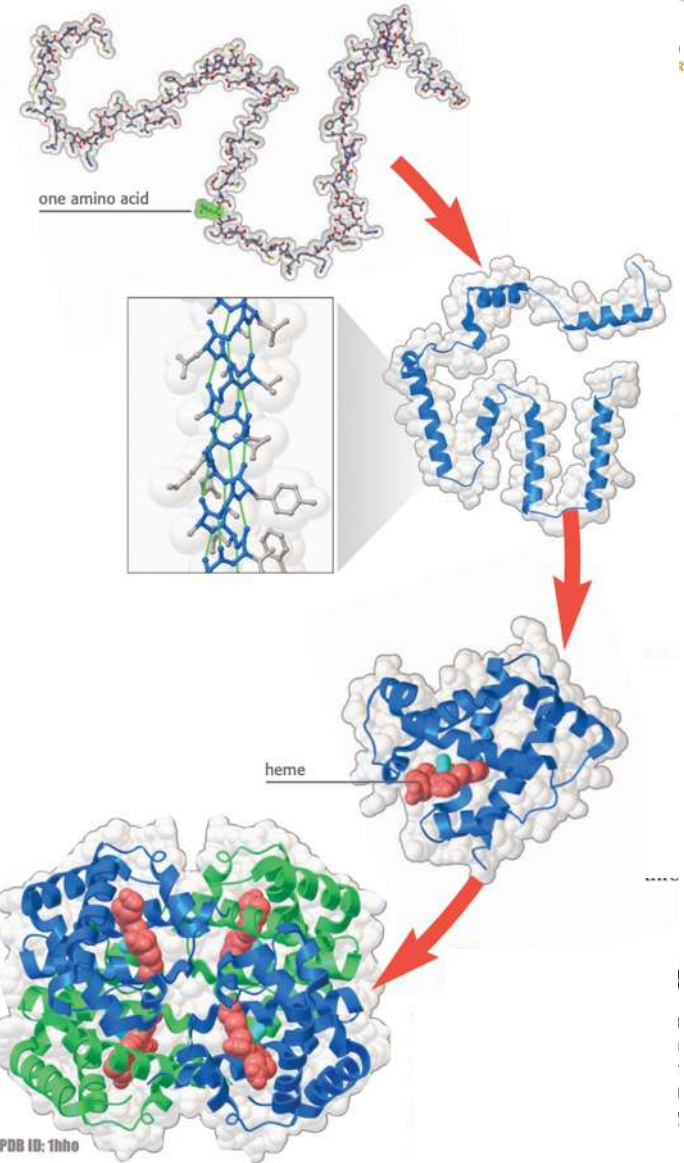
遺伝情報

機能

配列情報

設計図

ポリペプチド鎖(20種類のアミノ酸)



MOLECULAR MACHINERY: A Tour of the Protein Data Bank

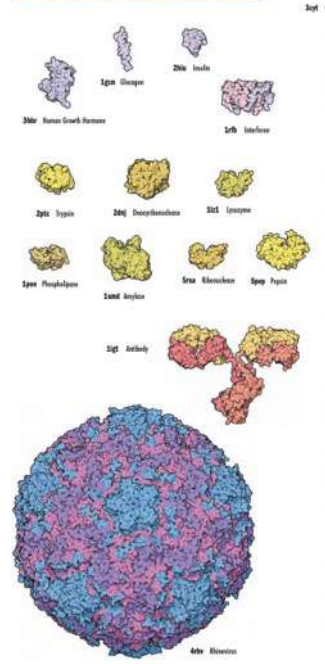


IS LIFE?
OGY RESEARCH CENTER

Living cells are filled with complex molecular machinery, a million times smaller than familiar machines like computers or automobiles. Cells use these tiny molecular machines to perform all of the jobs needed for life. Some are molecular scissors that cut food into cell-sized pieces. Some build new molecules when cells grow or when damaged tissues are repaired. Some are molecular bones and muscles that support cells and help them move and crawl. Some fight off attackers, defending against infection.

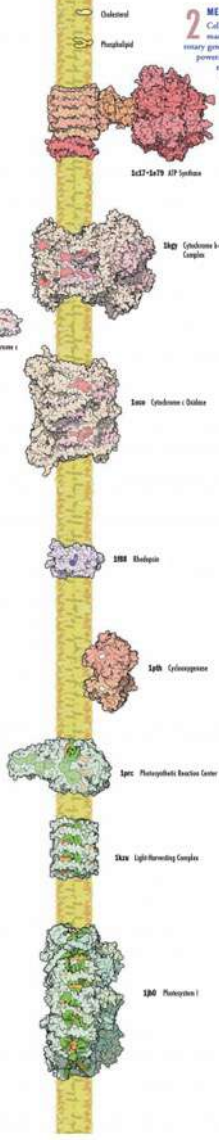
Researchers around the world are studying these molecules and determining their precise atomic structures. These structures are available on the Internet through the Protein Data Bank (<http://www.pdb.org>), the central storehouse of biomolecular structures. A few of the thousands of structures held in the Protein Data Bank are shown here. In these pictures, the molecules are all drawn at a magnification of 3,000,000 times, and each atom is shown as a small sphere. Many of these structures are composed of several subunits, which are indicated by different colors. An enormous range of sizes is shown here: the water molecule at the left has only three atoms and the rhinovirus shown below has hundreds of thousands.

By David S. Goodsell, The Scripps Research Institute, La Jolla, California, USA
Graphic design by Gal W. Bambar, San Diego Supercomputer Center

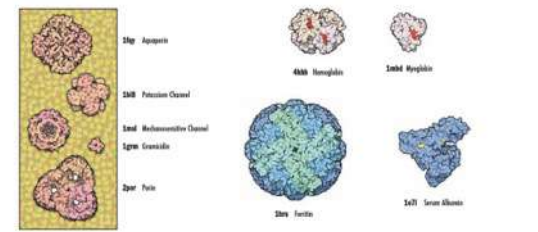


OUTSIDE THE CELL
Some molecular machines perform their jobs outside of cells. Many are compact, so that they can diffuse quickly to their site of action. This is true of the four hormones shown at the top: insulin and glucagon, which together regulate blood sugar levels; interferon, which carries signals in the immune system; and human growth hormone. The seven digestive enzymes (in yellow) are also small and very stable, so that they can survive the hostile environment of the digestive tract. Each of these enzymes has a small groove (oriented towards the top in each) that binds to a different target molecule and digests it. At the bottom is ribonuclease, the virus that causes the common cold, and an antibody, one major defense against viruses. Antibodies bind to viruses and prevent them from binding to cell surfaces, thus blocking infection.

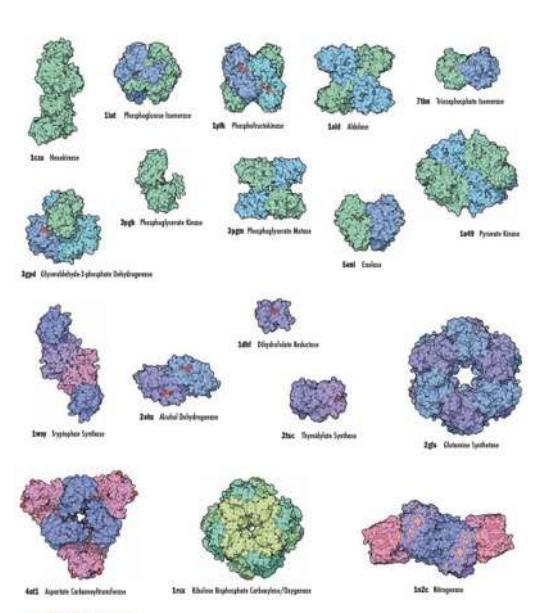
PROTEIN DATA BANK
<http://www.pdb.org/> • info@rcsb.org
RESEARCH COLLABORATOR FOR
STRUCTURAL BIOINFORMATICS
NATIONAL INSTITUTE OF HEALTH
NATIONAL CENTER FOR HUMAN GENOMICS
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY



2 MEMBRANES
Cells are surrounded by a membrane made of lipids, like the phospholipid and cholesterol molecules shown at the top. Membranes keep the cellular machinery inside and unwanted material out. Many proteins are embedded in this membrane, performing a variety of essential tasks. ATP synthase is a rotary generator that produces ATP (adenosine triphosphate), the small molecule used for powering cells. The two large complexes below it charge a battery that powers ATP synthase, and the three protein cytochromes shuttle electrons between them. Rhodopsin is found in membranes in the retina. The small integral molecule inside it changes shape when illuminated, causing the surrounding proteins to send a signal to the brain. Cytochrome *b₆* binds one of the molecules used to signal pain—this cytochrome molecule has, however, is blocked by two molecules of capsaicin, shown inside in white. At the bottom are three molecules involved in photosynthesis, which capture energy from light and use it to power the synthesis of sugar in plant cells.

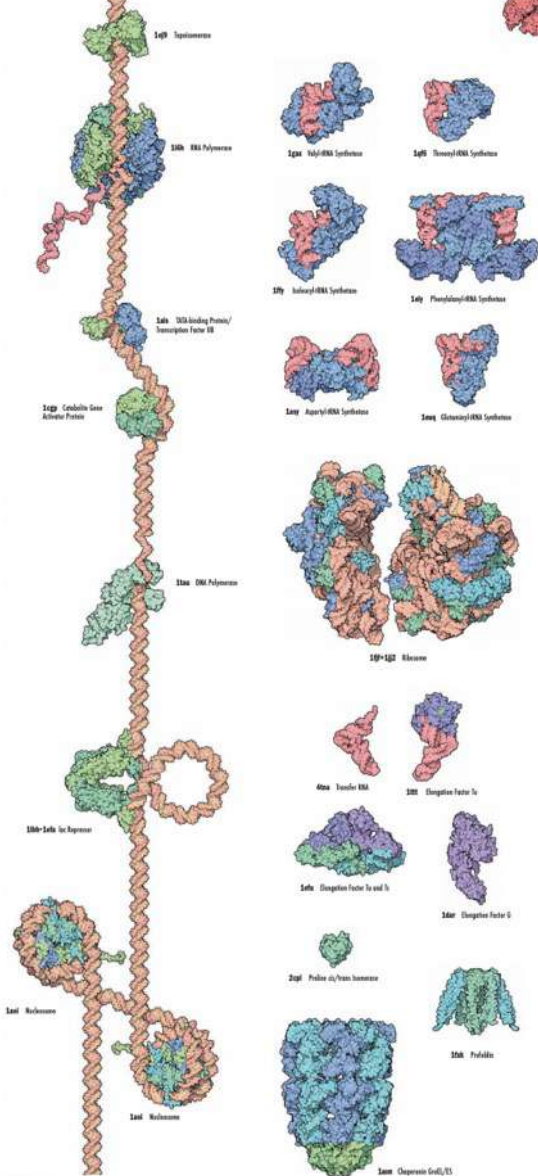


3 TRANSPORT AND STORAGE
Of course, a perfectly sealed membrane would be of little use to cells, because nutrients could not get in and wastes could not get out. The box shows a membrane looking face-on. Five proteins that form channels through the membrane are shown. To the right of the box are several soluble proteins involved in transport and storage of molecules. Hemoglobin and myoglobin carry oxygen. Ferritin forms a hollow shell that stores iron ions. Serum albumin carries many different molecules in the blood.

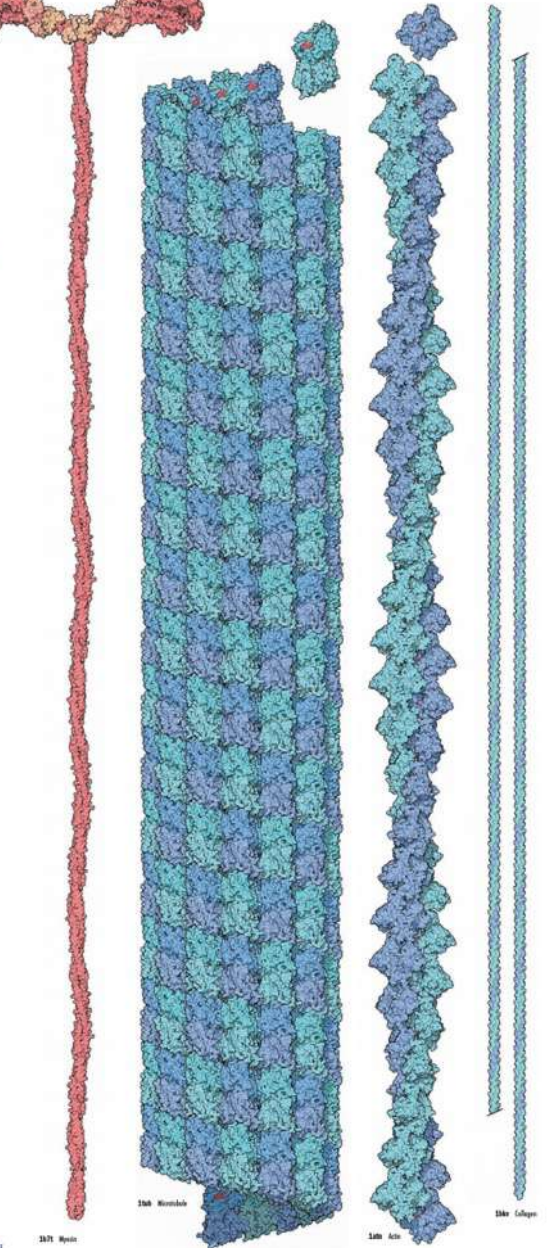


4 CHEMICAL FACTORIES
Cells build a bewildering variety of enzymes—proteins that perform chemical reactions. At the top are the two enzymes that perform glycolysis, the breakdown of sugar to form ATP. Below that are several enzymes that perform different biosynthesizing reactions. Dihydrofolate reductase activates a key cofactor molecule and alcohol dehydrogenase breaks down alcohol. Ribulose biphosphate carboxylase/oxygenase is the most common enzyme on the Earth, and performs a key step in the capture of carbon dioxide by plants to form sugar. The three synthases and the transaminase make different building blocks for creating new molecules. Nitrogenase performs an essential role in the conversion of nitrogen gas into a form that living cells can use.

5 DNA
Genetic information is stored in the DNA double helix, seen running from top to bottom here. Many proteins are used to copy, read, and create this information. RNA polymerase copies the information into a strand of RNA that will be used to direct the construction of new proteins. It is assisted by topoisomerase, which releases tension when the helix is wound and unwound, and guides to appropriate starting points by the protein complex below it. DNA polymerase replicates DNA strands—here, the polymerase is filling a gap in the double helix. Some proteins, like the lac repressor, grab DNA and bend it sharply, or even wrap it all the way around themselves, like the two nucleosomes at the bottom.



6 BUILDING NEW PROTEINS
New proteins are built by ribosomes—complex molecular factories that read the genetic code and use it to direct construction. Many molecular machines are needed to assist the process. Twenty different aminoacyl-tRNA synthetases (as are shown here) load the building blocks onto tRNA, ready to be added to a growing protein chain. Several protein factors, shown below the ribosomes, guide each tRNA into the proper spot. The three chaperone proteins shown at the bottom help each new protein fold into its proper shape.



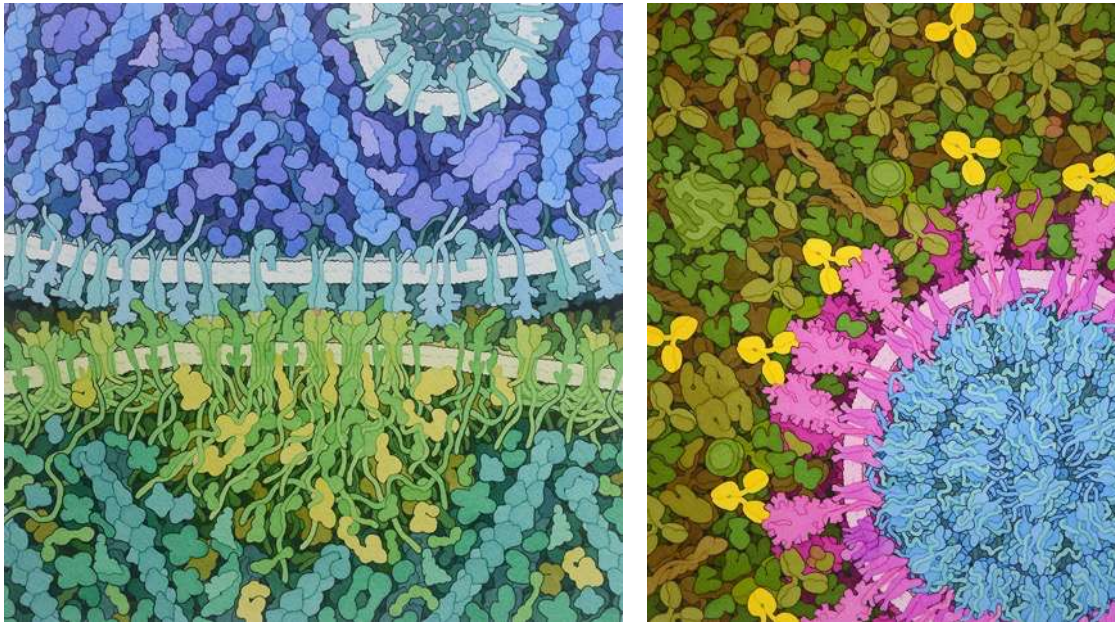
7 BEAMS AND GIRDERS
Cells are bonded and supported by a complex infrastructure composed of many substances stacked like bricks. Myosin moves. Collagen, broken into two pieces here, is actually found

<https://pdb101.rcsb.org/>

構造生物学

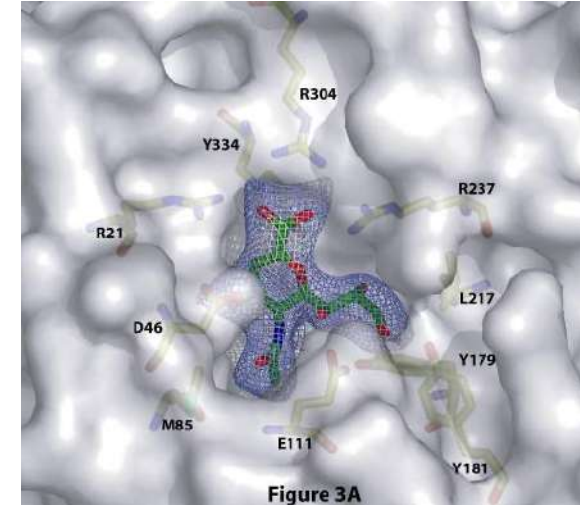
タンパク質の**立体構造を解明**することで、
その**機能を理解**する。

生命現象の理解



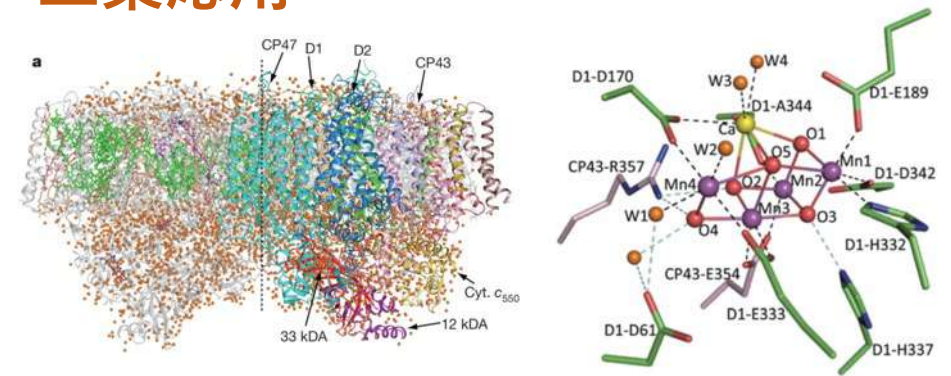
David S. Goodsell, RCSB Protein Data Bank.
doi: 10.2210/rcsb_pdb/goodsell-gallery-022
doi: 10.2210/rcsb_pdb/goodsell-gallery-025

薬剤設計



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

工業応用

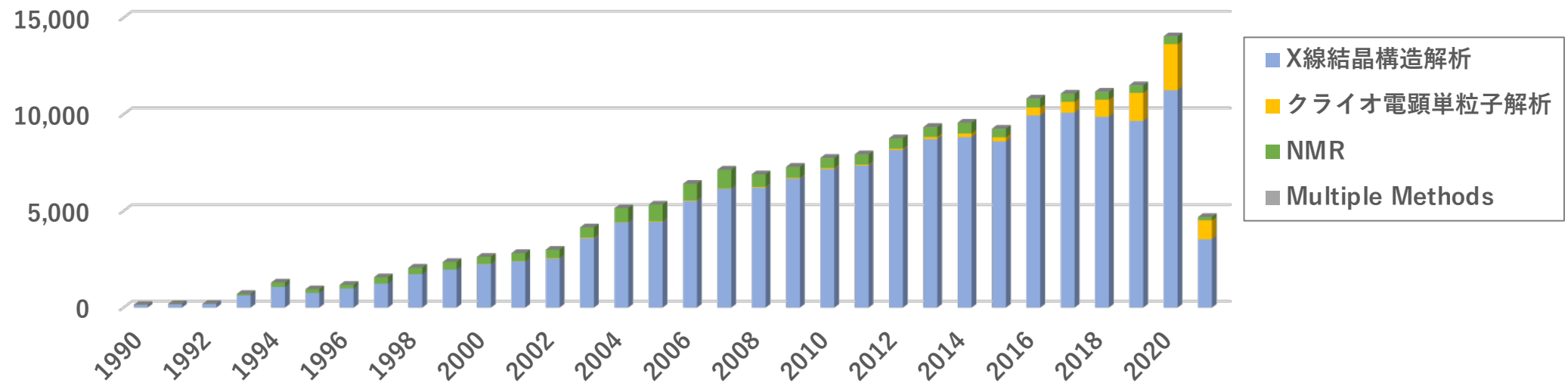


Nature 473, 55-60 (2011).

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

年間PDB登録数



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

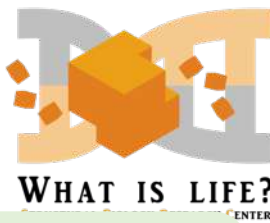
共通基盤研究施設

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)



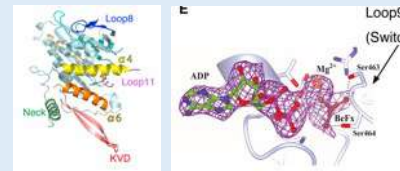
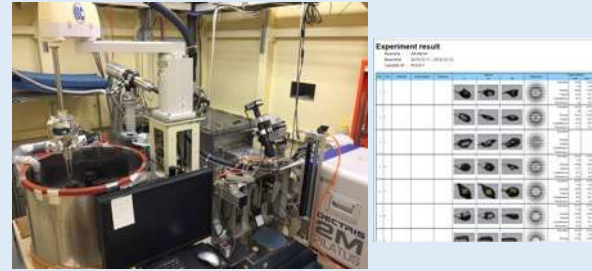
最先端の測定設備、測定手法の開発

共同利用、施設利用

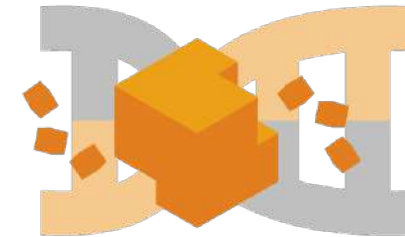
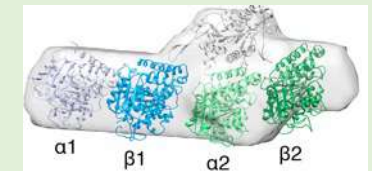
大学・公的研究機関

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

創薬等先端技術支援基盤プラットフォーム事業 (BINDS, AMED)



「知って、使って、進むあなたの研究」

構造解析ユニット (構造解析領域)
最先端ファシリティを駆使して、タンパク質やタンパク質複合体の静的・動的な構造解析をお手伝い致します。

クライオ電顕ネットワーク
最新鋭クライオ電子顕微鏡で、構造解析をお手伝いします。

構造解析ユニット (タンパク質生産領域)
最先端技術を結集してタンパク質生産や結晶化をお手伝い致します。

ケミカルシーズ・リード探索ユニット (ライブラリー・スクリーニング領域)
各機関が保有するユニークな低分子・天然物・ペプチドライブラリーを提供し、スクリーニングをお手伝い致します。

ケミカルシーズ・リード探索ユニット (構造展開領域)
デザイン⇒合成⇒薬理評価⇒ADMET/物性評価のサイクルを回しながら合成展開を行い、効率的なリード化合物の創出をお手伝い致します。

バイオロジカルシーズ探索ユニット
ゲノミクス解析やゲノム改変生物材料の提供、探索的ADMET試験をお手伝い致します。

プラットフォーム機能最適化ユニット
研究成果の最大化に役立つようデータベースクラウドを提供し、利用をお手伝い致します。ワンストップ窓口

インシリコユニット
計算科学を駆使して構造ダイナミクス研究をお手伝い致します。バイオインフォマティクス、ケモインフォマティクス研究もおまかせください。

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！

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)

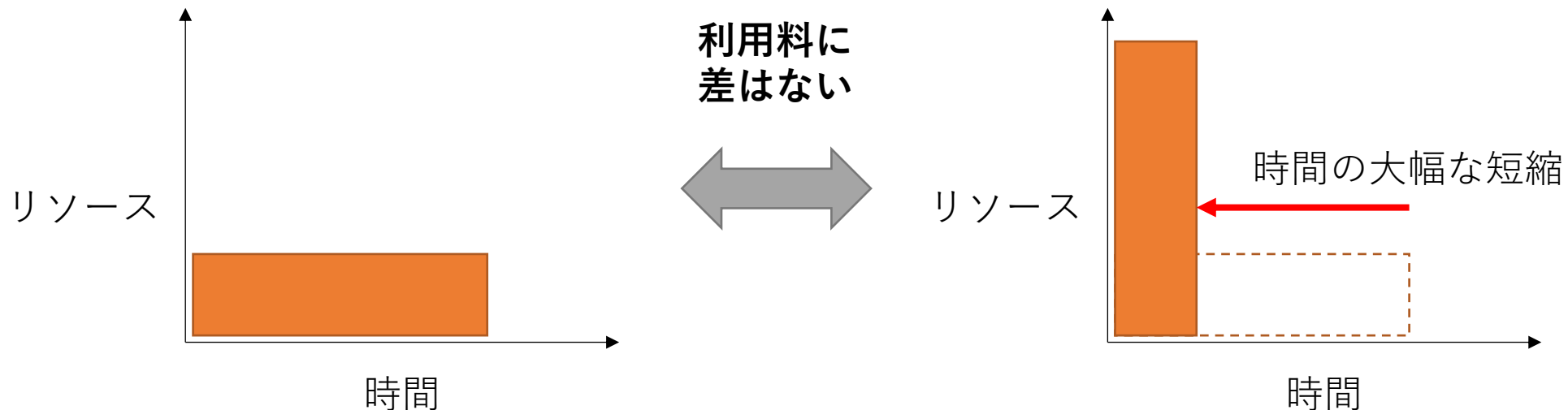
計算リソース

- MX: 解析クラスター 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

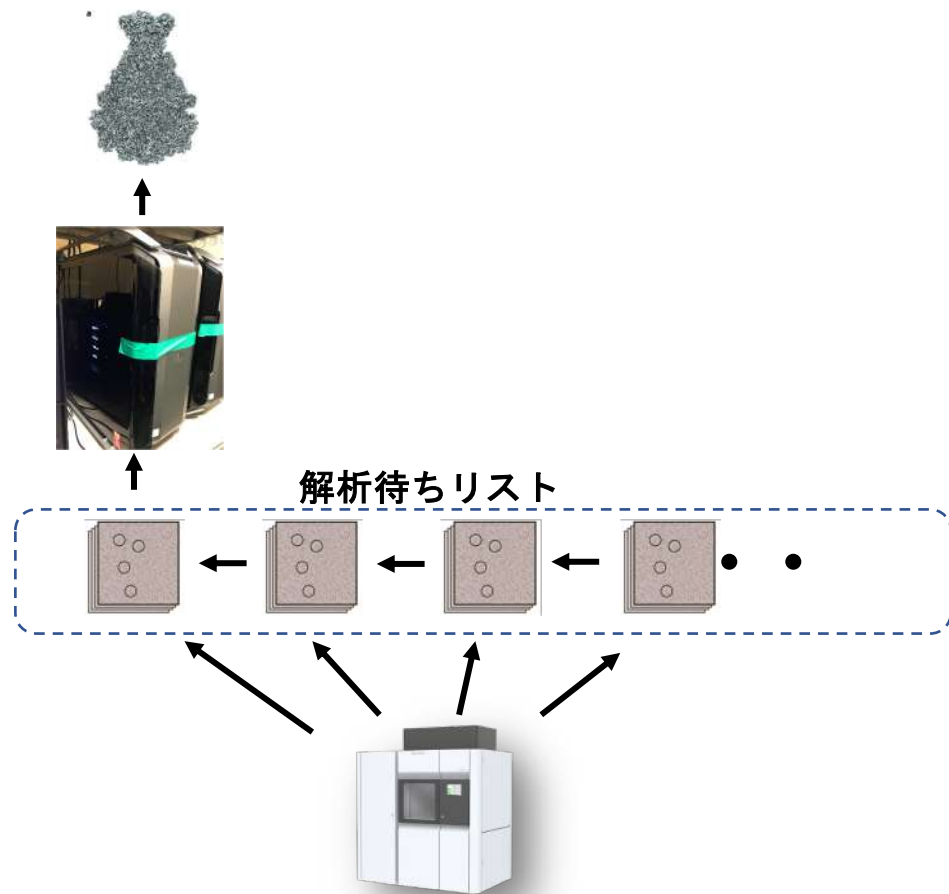
- クラウドコンピューティングを広く一般に提供 (AWS, GCP, Azure, Oracle, ...)
- 多量(ほぼ無尽蔵)のリソース
- マネージド (利用者側での運用管理がほぼ不要)
- 新しいテクノロジーにも迅速に対応
- 従量課金



パブリッククラウドを用いた実験データドリブンの仮想解析環境

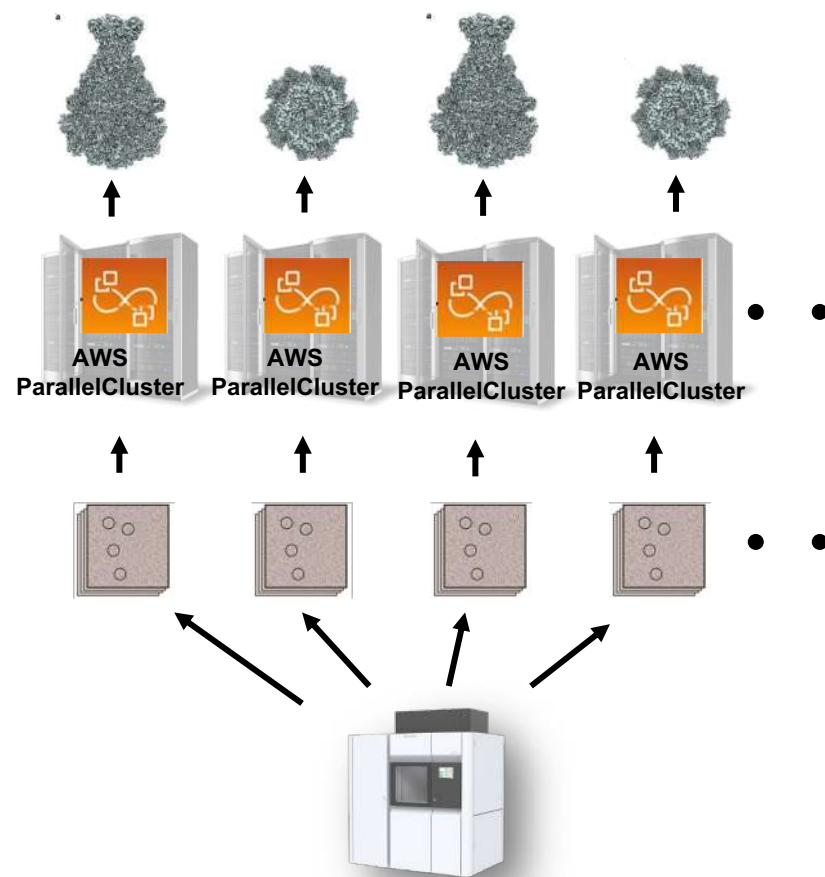
現状

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

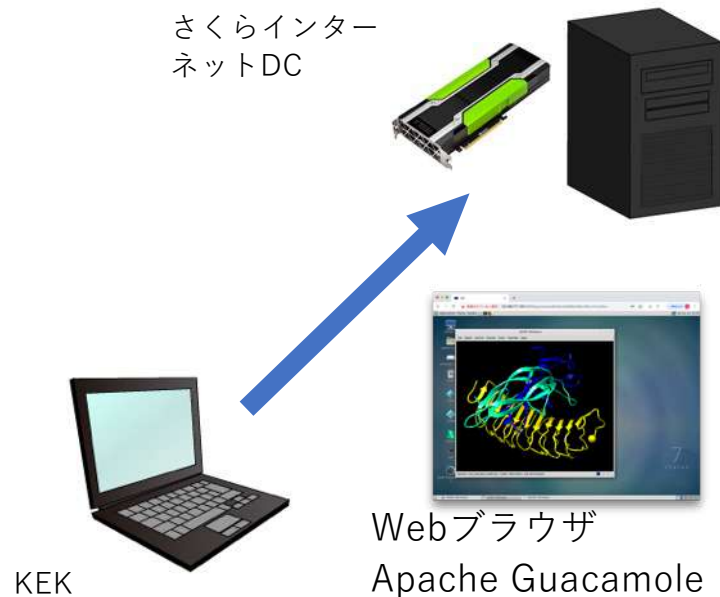
実習インストラクター

重松秀樹：理化学研究所/SPring-8

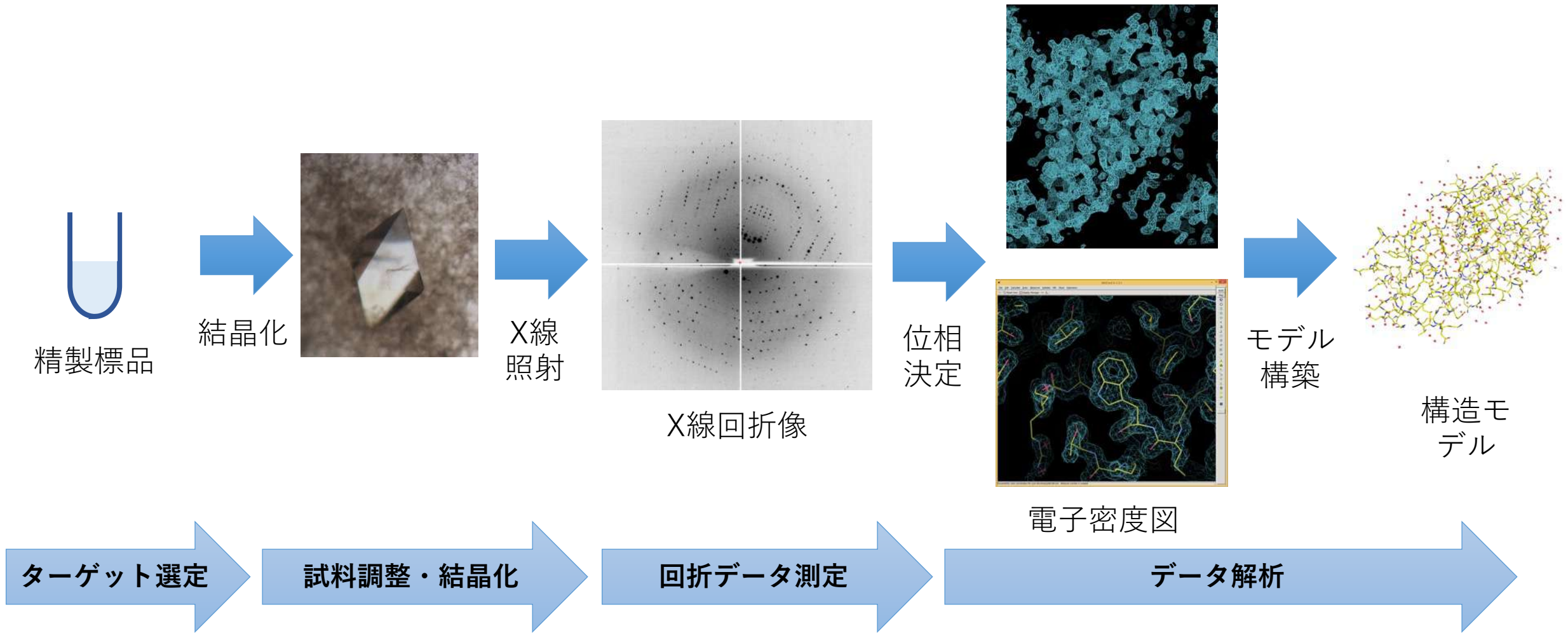
守屋俊夫：KEK・物質構造科学研究所・構造生物学研究センター

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



全自動結晶化スクリーニングシステム(PXS)



加藤龍一



Dispensing and storage



Monitoring drops

Dispenser	Mosquito LCP (0.1 μ L) ~ 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。 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

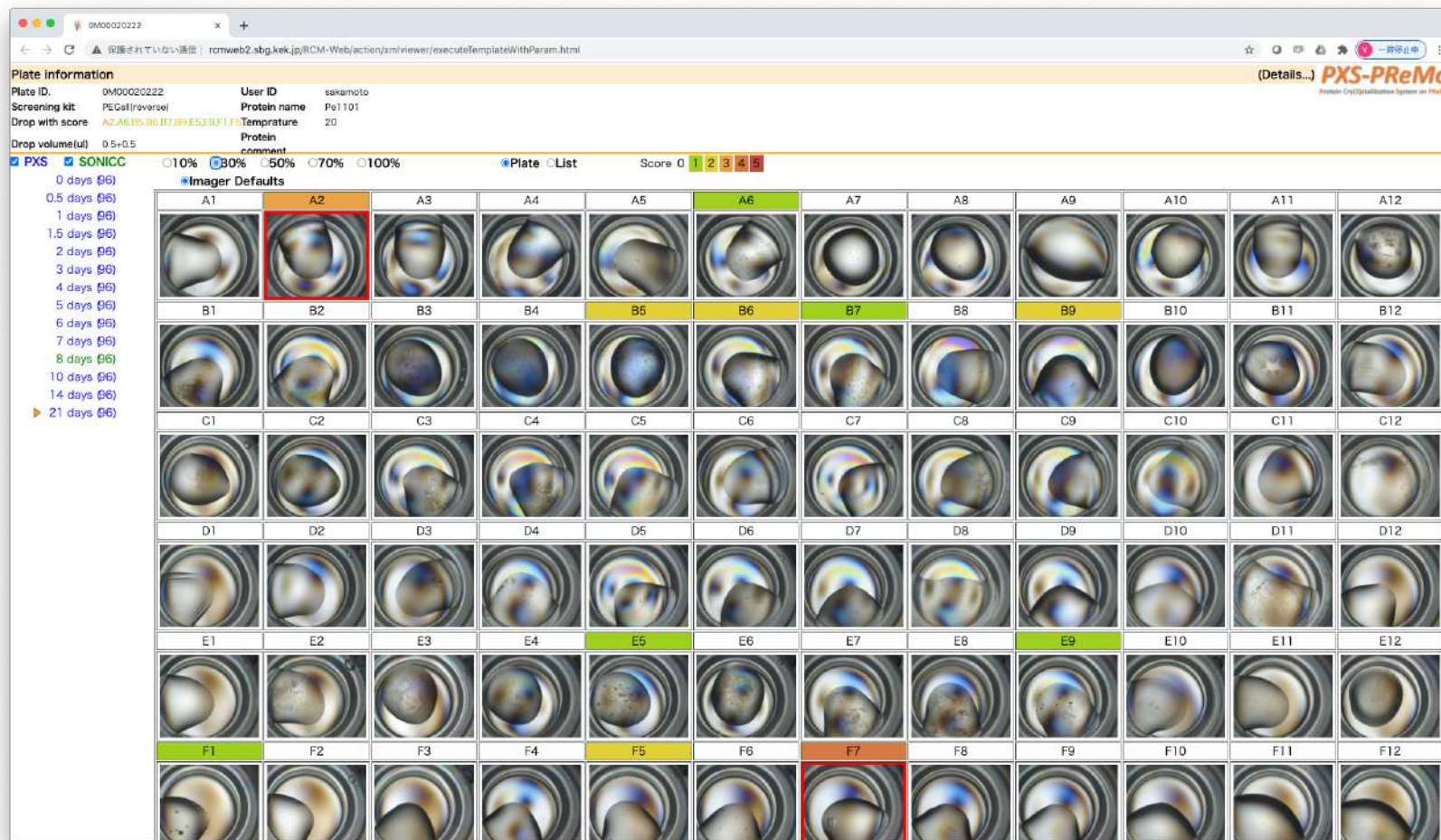


Plate information
Plate ID: DM00020222 User ID: sakamoto
Screening kit: PEGall(reverse) Protein name: Pe1101
Drop with score: A2,A6,B5,B6,B7,B9,E5,E9,F1,F5 Temperature: 20
Drop volume(ul): 0.5+0.5 Protein comment:

Imager Defaults
Score 0 1 2 3 4 5

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12

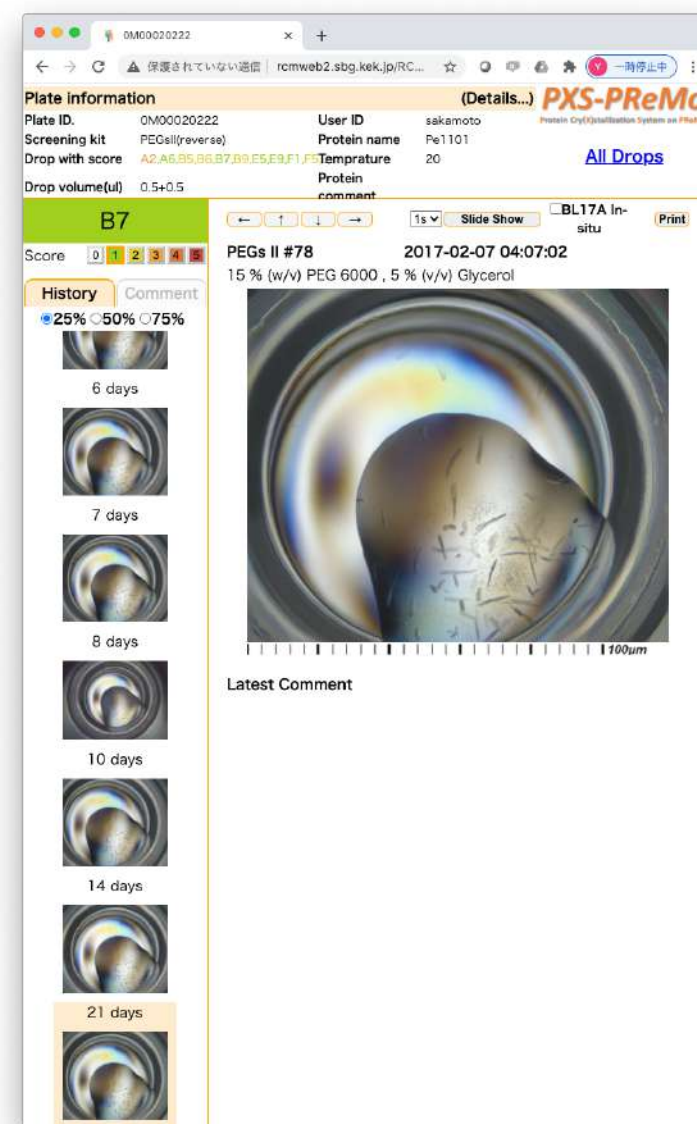


Plate information
Plate ID: DM00020222 User ID: sakamoto
Screening kit: PEGall(reverse) Protein name: Pe1101
Drop with score: A2,A6,B5,B6,B7,B9,E5,E9,F1,F5 Temperature: 20
Drop volume(ul): 0.5+0.5 Protein comment:

B7
Score 0 1 2 3 4 5

History Comment
25% 50% 75%

6 days
7 days
8 days
10 days
14 days
21 days

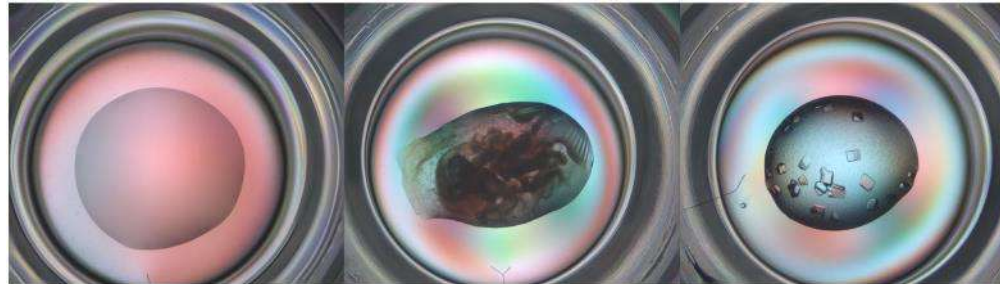
Latest Comment

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)



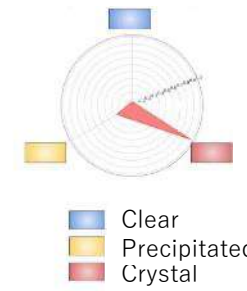
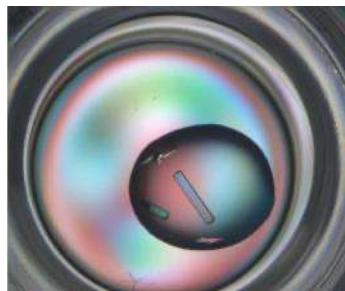
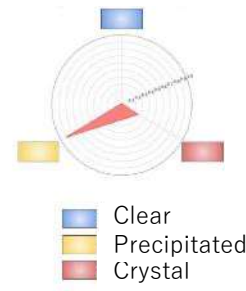
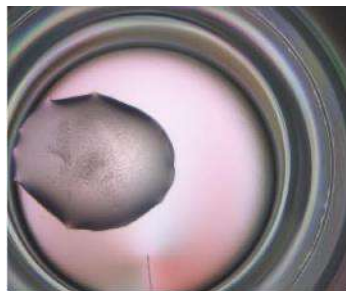
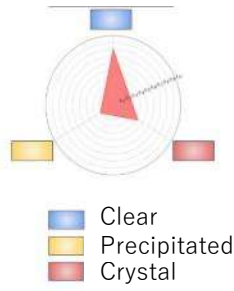
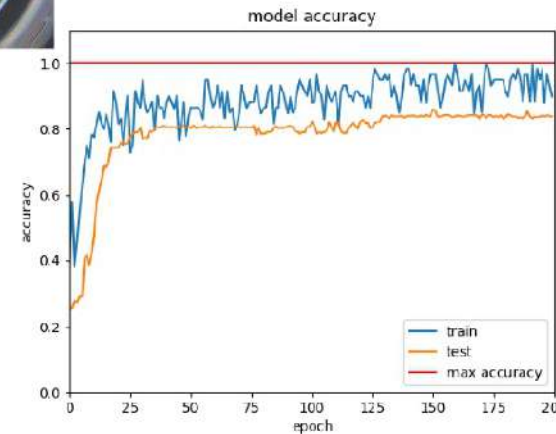
Clear

Precipitated

Crystal



Deep
learning



RESEARCH ARTICLE

Classification of crystallization outcomes using deep convolutional neural networks

Andrew E. Bruno¹, Patrick Charbonneau^{2,3}, Janet Newman⁴, Edward H. Snell^{5,6}, David R. So⁷, Vincent Vanhoucke^{7*}, Christopher J. Watkins⁸, Shawn Williams⁹, Julie Wilson¹⁰

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

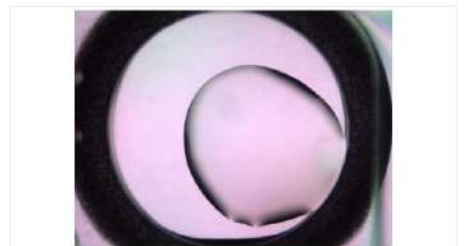
The Machine Recognition of Crystallization Outcomes (MARCO) initiative has assembled roughly half a million annotated images of macromolecular crystallization experiments from various sources and setups. Here, state-of-the-art machine learning algorithms are trained and tested on different parts of this 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)



Images

Clear Crystals Other Precipitate

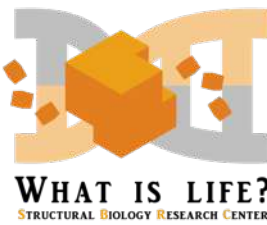


Organization: GSK

Clear

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

深層学習を用いた結晶の認識

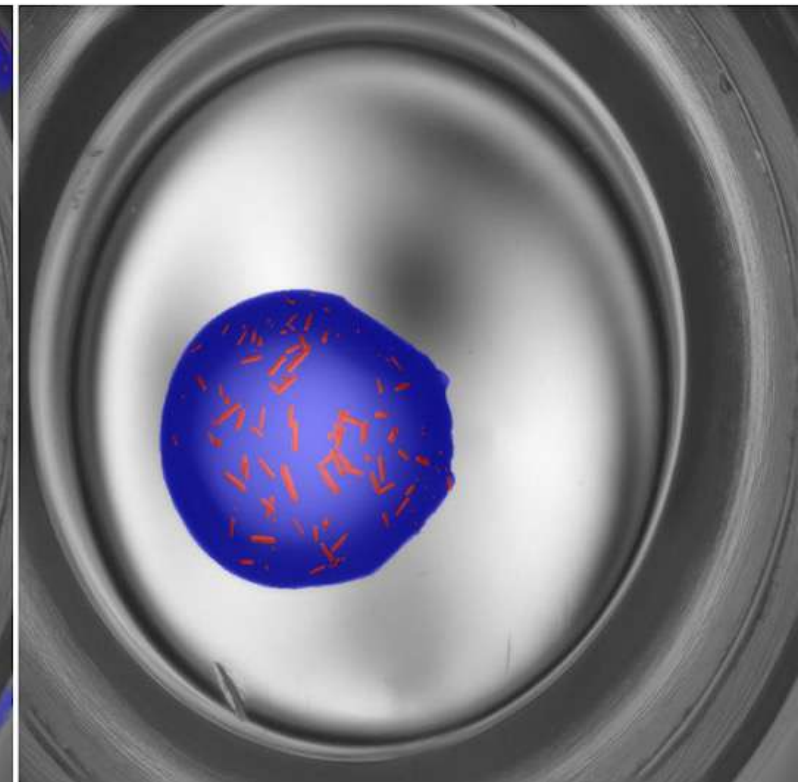
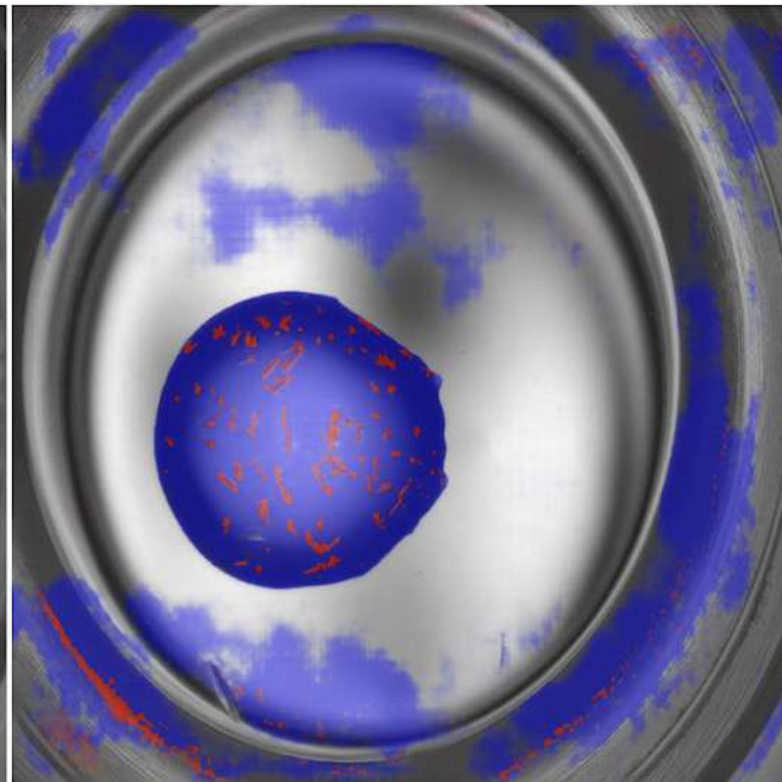
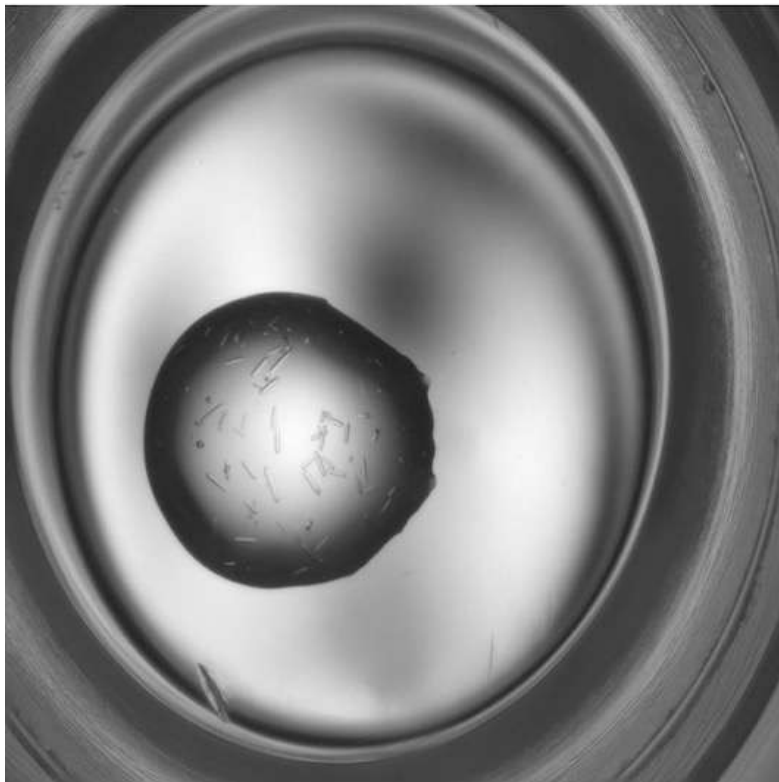


篠田晃

元の画像

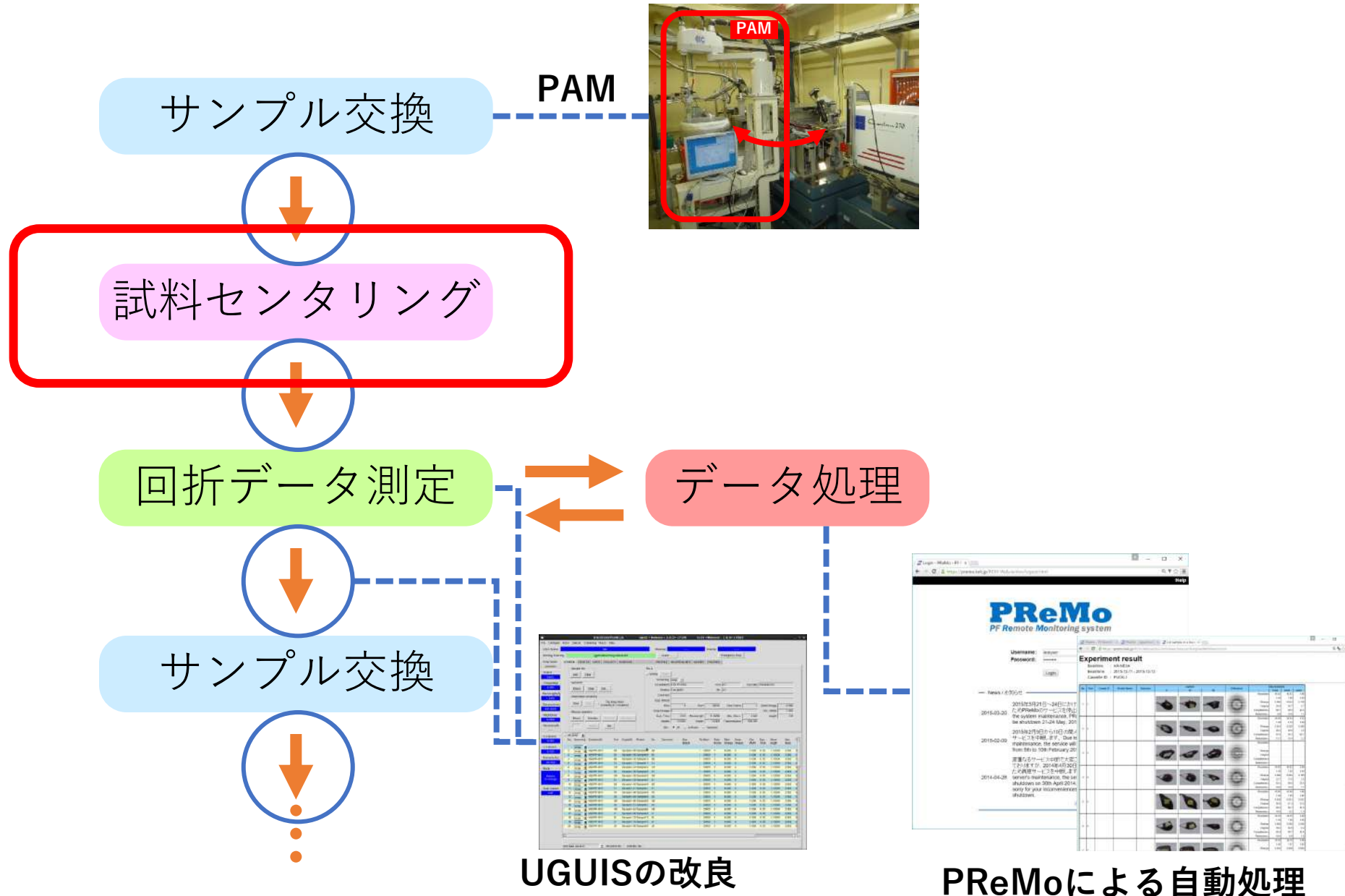
機械学習で推測した画像

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



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

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

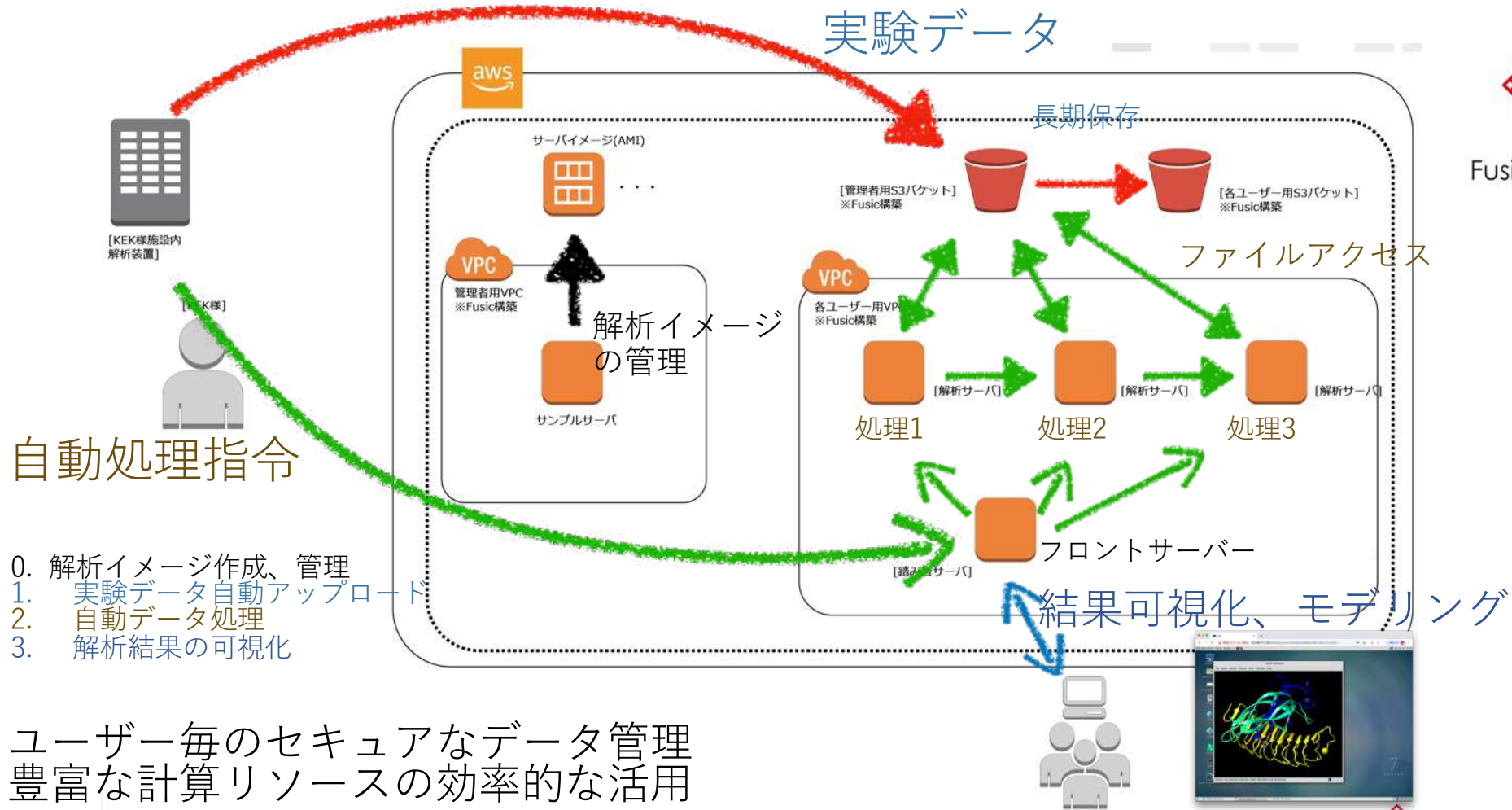
- 回折像 -> 回折強度データ
- 位相決定
- モデル構築
- モデル精密化
-
-



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

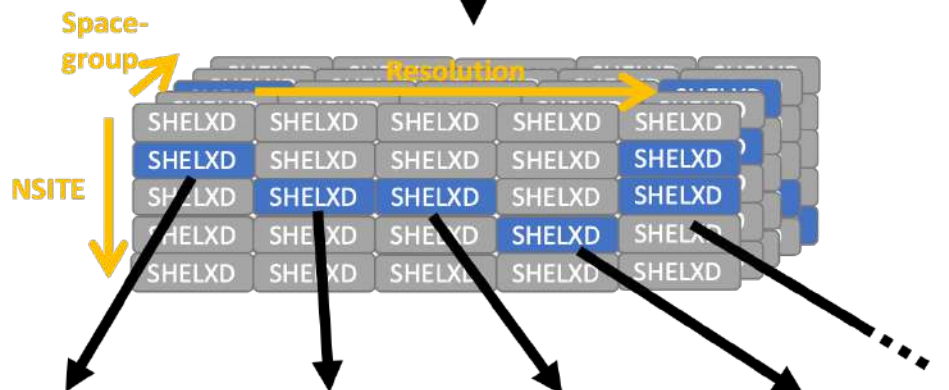
AWS(Amazon Web Service)を用いて構築



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

SHELXC



Solvent content

Original	Inverted	Original	Inverted	Original	Inverted	Original	Inverted
SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE
SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE
SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE
SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE
SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE
SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE	SHELXE

回折強度データ前処理

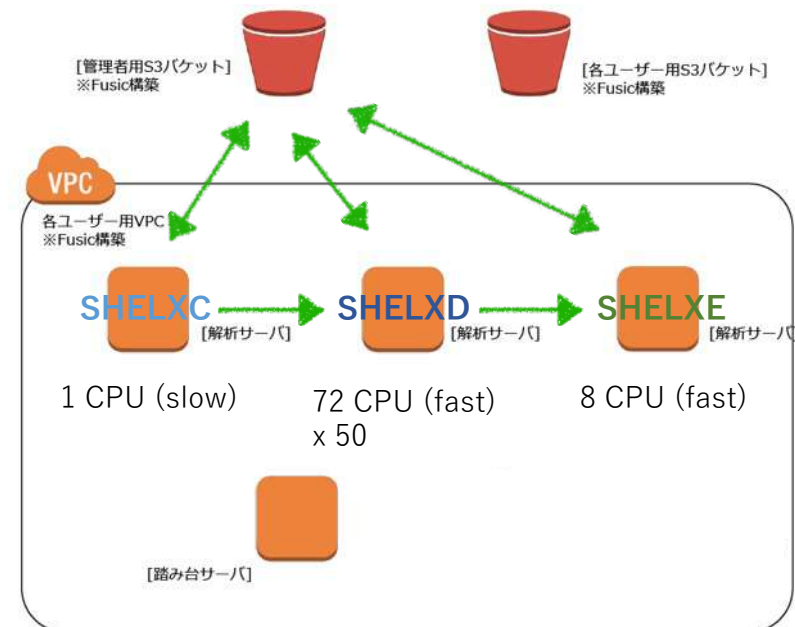
シングルスレッド
シングルパラメータ

異常散乱体(硫黄原子等)の位置同定

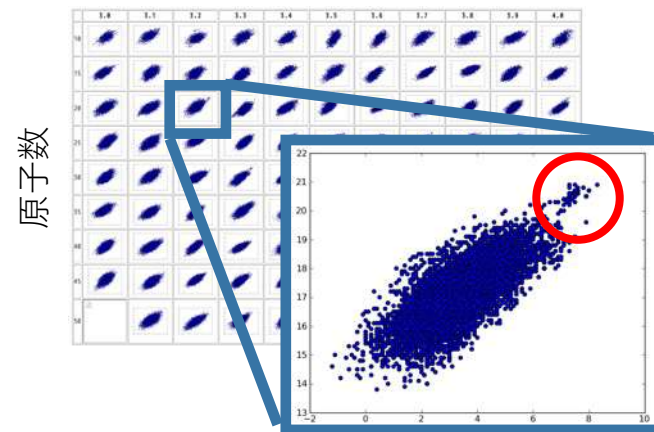
マルチスレッド
マルチパラメータ

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

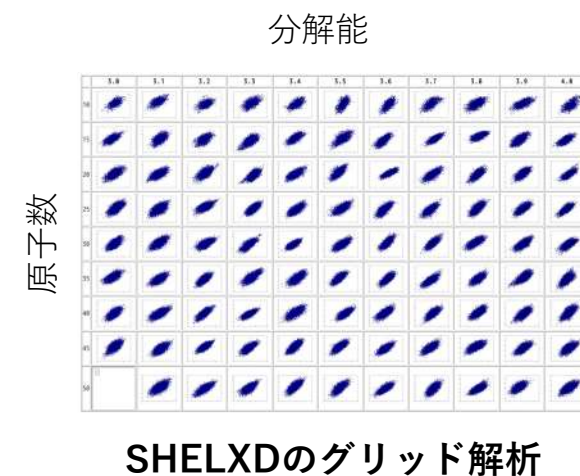
1 TBのデータ: 3時間強

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)



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



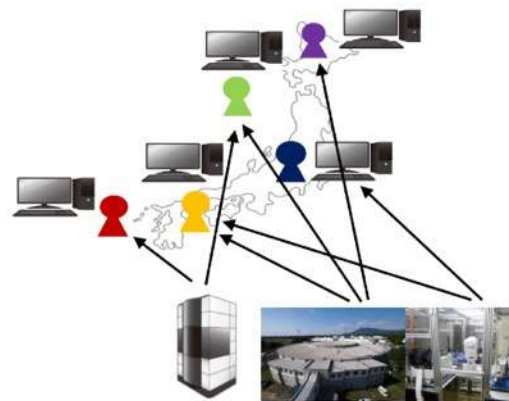
パブリッククラウドの利用

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



小角散乱、結晶化、結晶構造解析、単粒子解析 (CryoEM) には、様々な困難が存在し専門家の助けが未だに必要とされている。共通の実験施設を使いながら知識と経験は分断されている。

構造情報の利用の生物学への浸透が阻害され、機能メカニズムの解明が滞る。創薬の現場でも困難の原因になっている。

大量のデータと解析結果の統合により、原理に基づいた問題解決のみならず、AIによる問題解決も可能にする。その結果、高度な専門知識を持たないユーザーでも、最高レベルの問題解決が可能になり、生物学における構造情報利用の促進につながる。

