高分解能生体単粒子構造解析を目指したX線回折像の敏速分類

理研　計算科学研究機構
計算構造生物学研究ユニット
研究員
徳久　淳師

目的

フェムト秒パルスX線レーザーを用いて生体高分子の新しい構造解析法を開発する

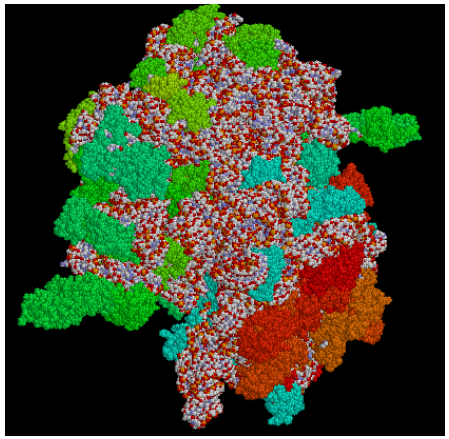
施設
- X線自由電子レーザー
- 検出器
- 光学系
- データー収集系
- データ処理・診断系

実験
- 試料調製
- 試料導入法
- 実験チャンバー
- 実験条件

協力

方法論
- 微結晶構造解析
- 単粒子構造解析

高分解能データセットに対する解析
低分解能データセットに対する解析

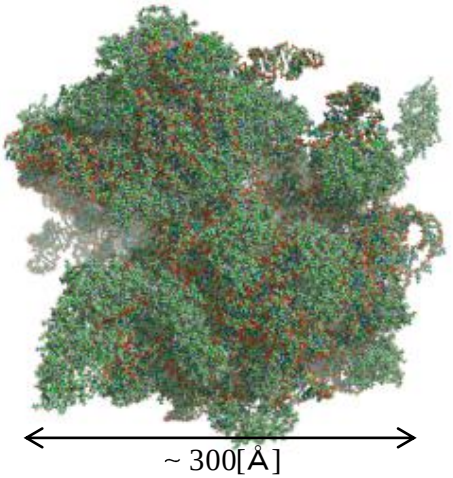


- 開発項目は多岐にわたり施設の方々や実験家と協力して進めていく必要がある

アウトライン

フェムト秒X線パルスレーザーを用いた高分解能構造解析の可能性を追求

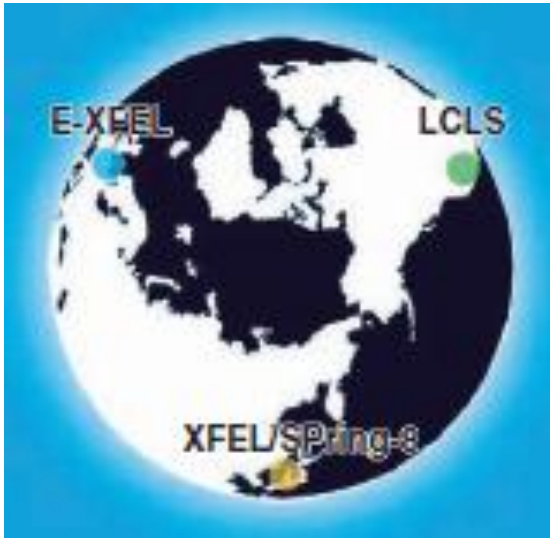
- **高分解能単粒子構造解析の概要**
  →コヒーレントイメージング法の1つ
- **開発した2つのアルゴリズムの説明**
  1. 回折像の分類・平均法
  2. 回折像の相対位置の決定法
- **達成可能な分解能**
  →原子分解能の達成が視野
- **大量の回折像を分類するための2つの取り組み**
  1. 相関線の自動判定法の導入
  2. 京コンピュータへのアルゴリズム実装
- **実験の現状に適した方法**

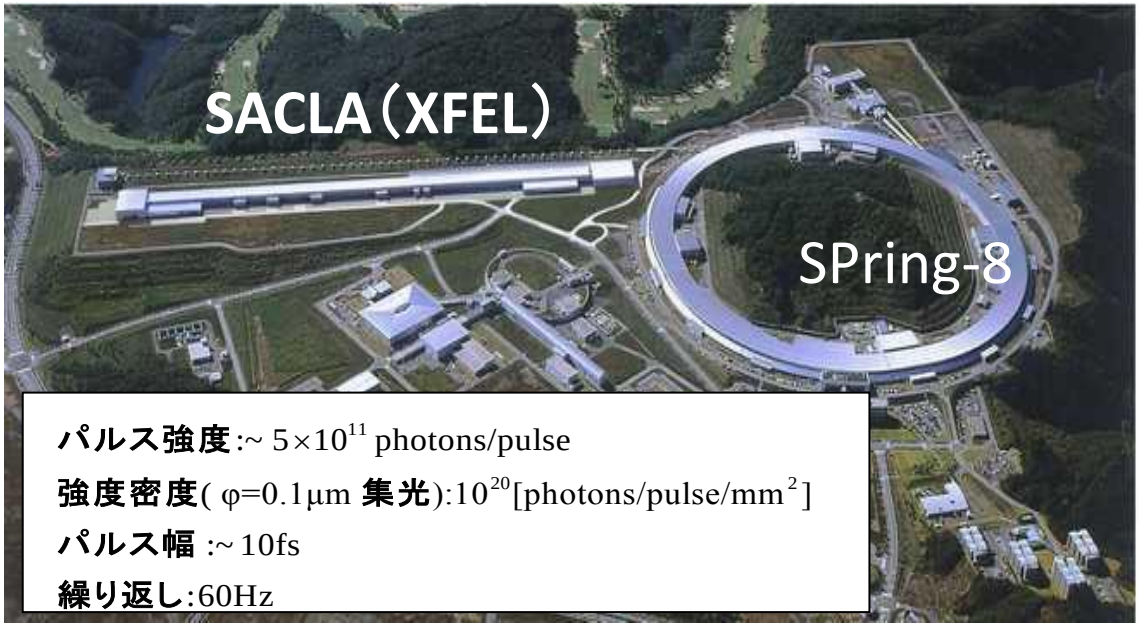


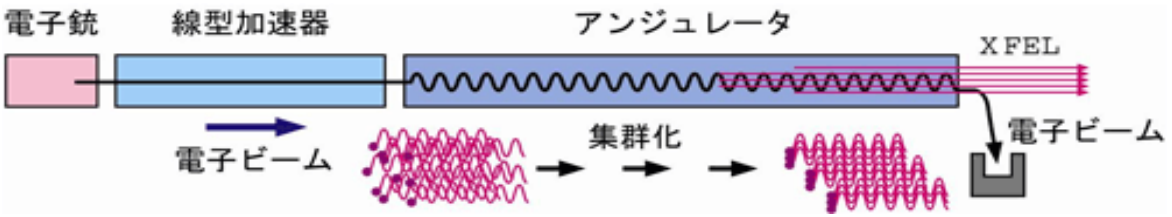
生体分子立体構造
（電子密度分布）

数値シミュレーションによるアプローチ

X線自由電子レーザー（XFEL）







フェムト秒X線パルスレーザーを発振

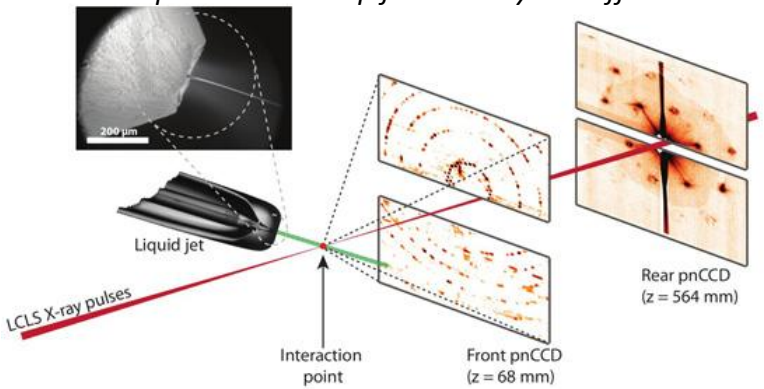
- 高い空間コヒーレンス
- 短パルス幅
- 高ピーク輝度

Ishikawa, T. et al. *Nature Photon*. 6, 540-544 (2012)

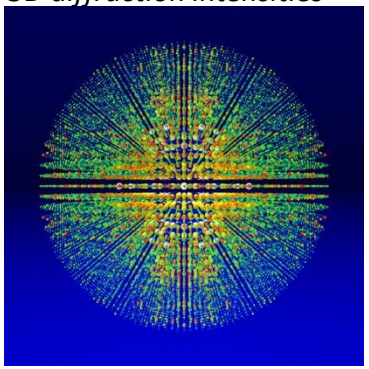
XFELを用いた構造解析の2つの可能性

➢微結晶構造解析

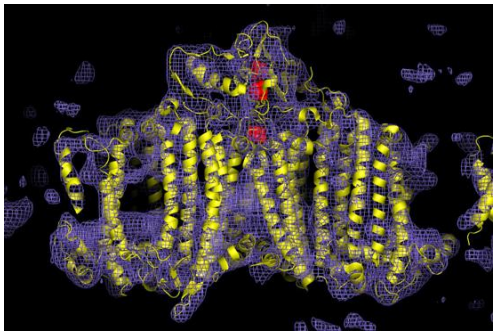
Experimental setup for nanocrystal diffraction



3D diffraction intensities



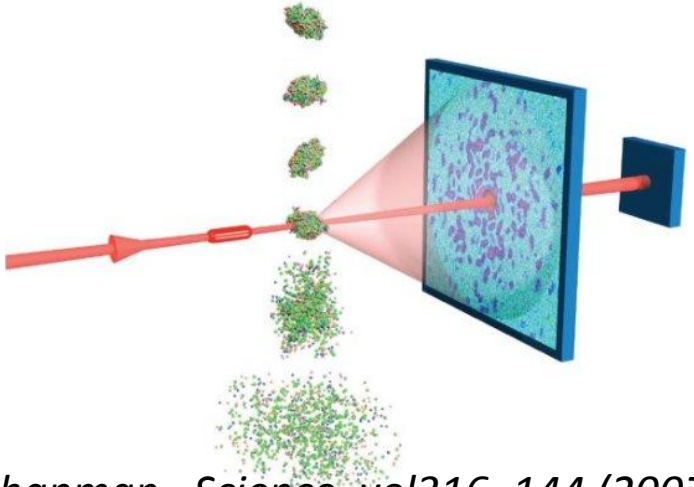
Electron density map of the photosystem I



Nature 470, 73 – 78 (2011).

➢単粒子構造解析

Experimental setup for single particle CDI



Chapman, Science, vol316, 144 (2007)

“挑戦的かつ重要な課題”

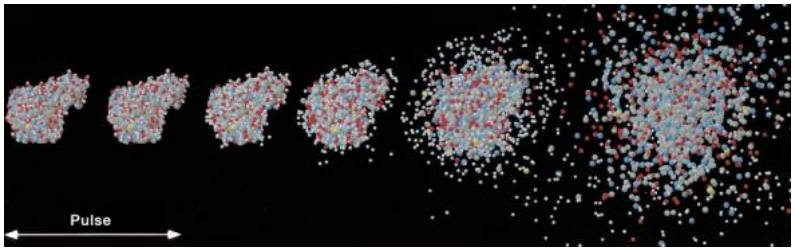
➢この方法が確立されれば構造生物学に与える影響は大きい

- 試料の結晶化が不要
- 病気メカニズムの解明と創薬へつながる

単粒子構造解析に焦点を当てる

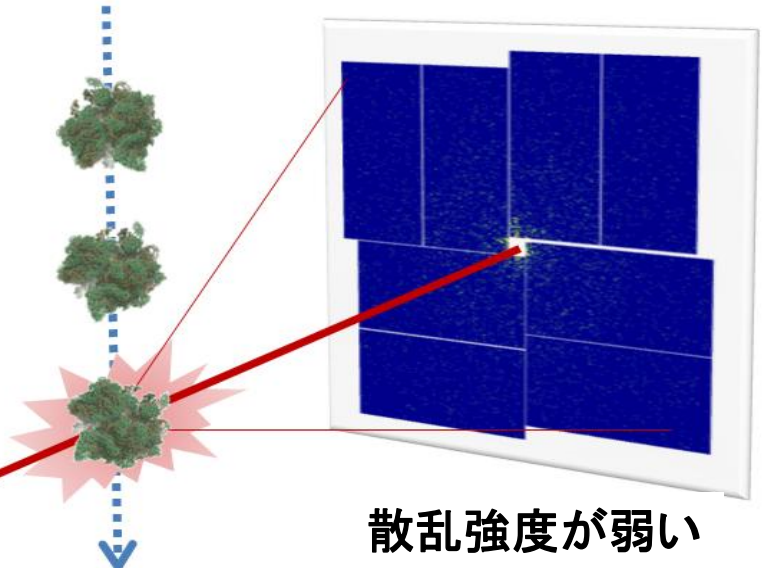
高分解能単粒子構造解析

“probe-before-destroy”



R.Neutze et al.:*Nature* 406(2000)752

散乱強度：$s(\boldsymbol{k}) = I_i r_c^2 \omega \left| F(\boldsymbol{k}) \right|^2$



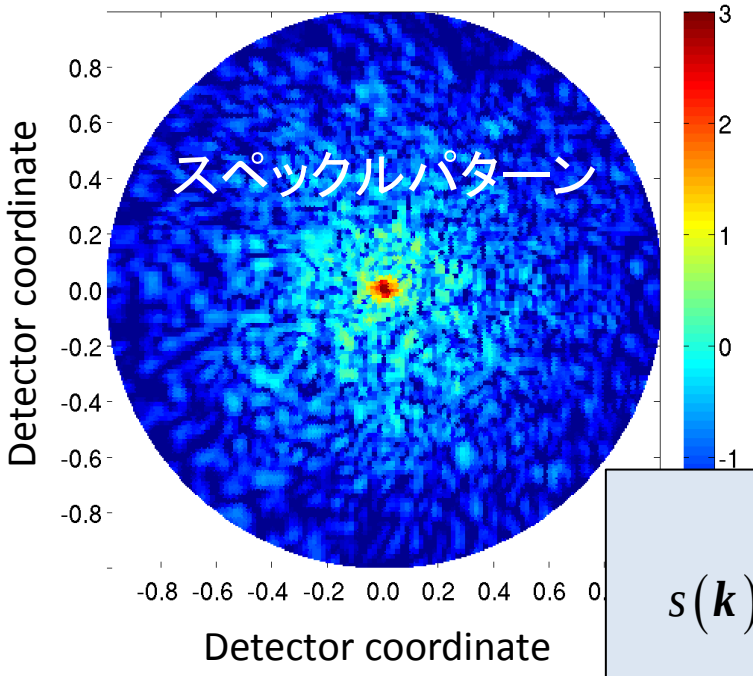
推奨する実験系
分子飛翔法

e.g. Kassemeyer, S., et al.,*Opt. Exp.* Vol. 20, Issue 4, pp. 4149-4158 (2012)

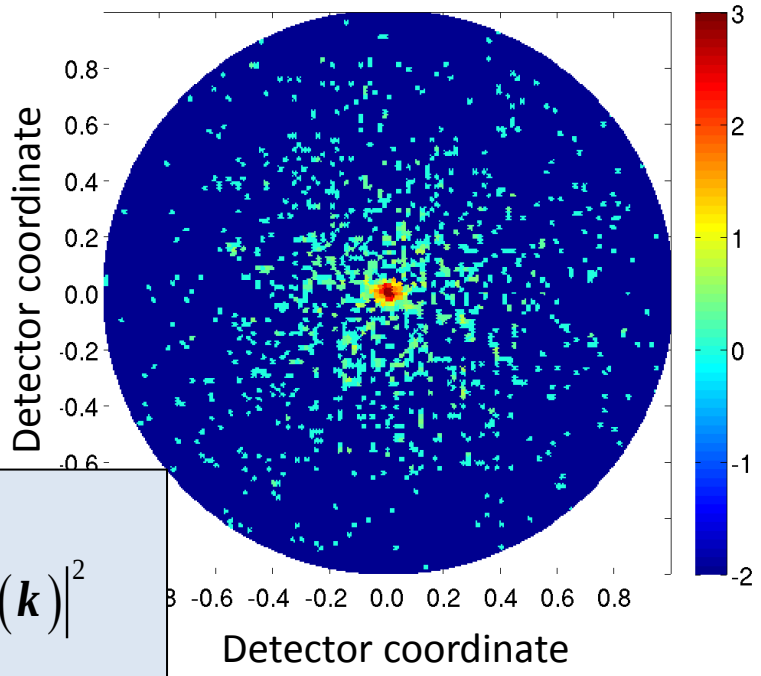
- 分子が壊れる前に測定を終える
- 分子方位は未知
- 多数回の測定が必要

観測される回折像

回折光期待値



量子雑音



回折強度

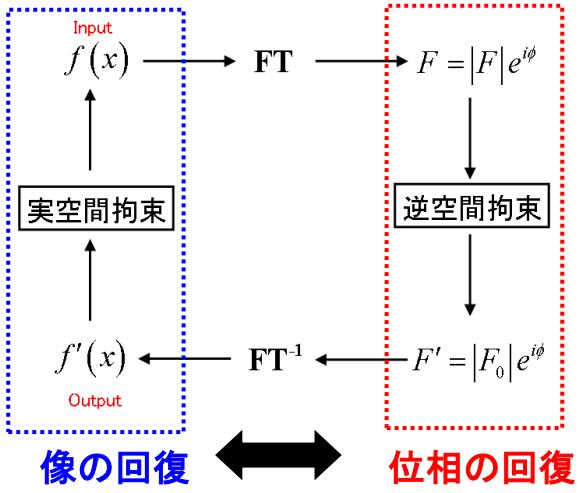
$$s(\boldsymbol{k}) = I_i r_c^2 \omega \left| F(\boldsymbol{k}) \right|^2$$

$$F(\boldsymbol{k}) = \int_{-\infty}^{\infty} d\boldsymbol{r}\, \rho(\boldsymbol{r}) e^{-2\pi i \boldsymbol{k}\boldsymbol{r}}$$

- 軽元素からなる生体分子単粒子試料からの散乱強度は弱い
  →強い量子雑音を伴ったS/N比の悪い像が観測される
- 回折強度は電子密度のフーリエ変換値の大きさに比例（位相情報が欠落）
  →実像を得るには位相情報を回復する必要がある

オーバーサンプリング法と位相回復アルゴリズム

HIO位相回復法

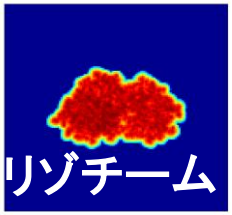


拘束をかけながらフーリエ変換を繰り返す

J.R.Fienup *APPLIED OPTICS* Vol.21, No.15 / 1 August 1982.

2次元のシミュレーション結果

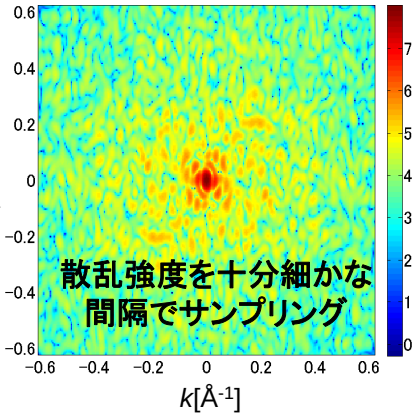
2次元平面射影実像

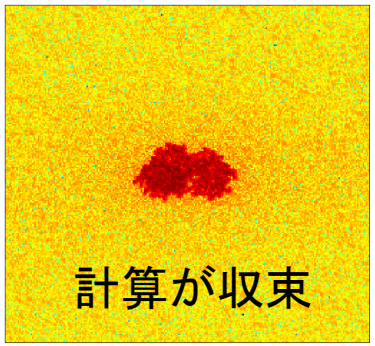


回折像

回復像

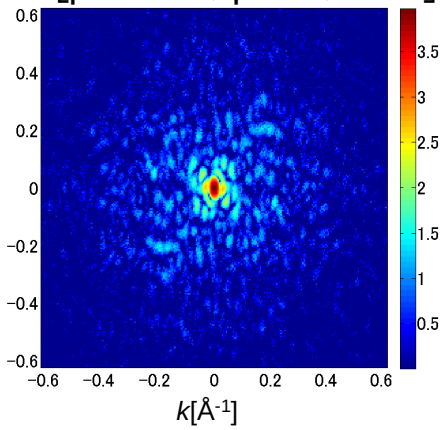
量子雑音なし

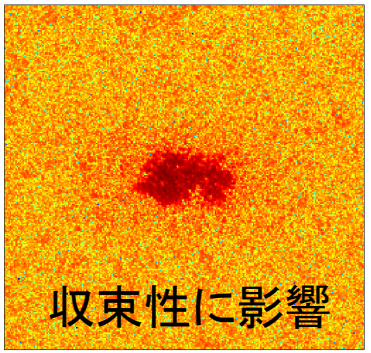




量子雑音あり

10^{22}[photons/pulse/mm^2]





- 散乱強度が十分強い場合欠落した位相情報を回復できる
- 量子雑音は位相回復計算の収束性に影響を与える

高分解能単粒子構造解析
解決すべき問題点

- 散乱強度が弱い（S/N比が悪い）
  →強度不足の１枚の回折像に対しては位相回復が困難
- 入射光に対する分子方位が未知
- ３次元情報を得るには多数の回折像が必要
  →波数空間での相対位置を決める方法が必要
- 位相情報の欠落
- 観測に伴う放射線損傷は無視できない

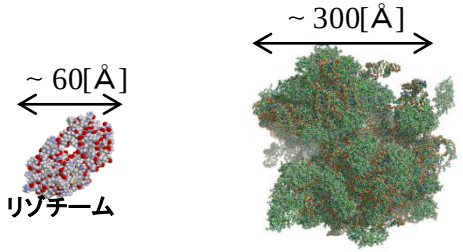
  →弱い散乱強度も有効信号として扱える方法が必要

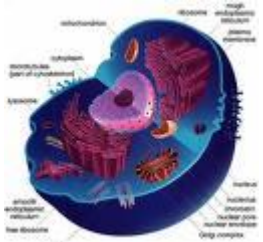
問題点を考慮した構造解析法が必要

アウトライン

- 高分解能単粒子構造解析の概要
  →コヒーレントイメージングの手法の1つ
- **開発した2つのアルゴリズムの説明**
  1. 回折像の分類・平均法
  2. 回折像の相対位置の決定法
- 達成可能な分解能
  →原子分解能の達成が視野
- 大量の回折像を分類するための2つの取り組み
  1. 相関線の自動判定法の導入
  2. 京コンピュータへのアルゴリズム実装
- 実験の現状に適した方法

試料サイズと立体構造構築手順





観測対象

膜タンパク質
球状タンパク質
ウイルス
リボソーム
植物性細胞
動物性細胞
原子

分子サイズ

1Å　10Å　10^2Å　10^3Å　10^4Å (1μm)　10^5Å

弱い　強い

散乱能

手順 A

2次元回折像の分類
同一グループの像を平均して精度向上

⇩

平均化後の像の交円を求めて相互配置
3次元回折像を構築

⇩

3次元回折像に対して位相回復

⇩

フーリエ変換で3次元分子構造を得る

手順 B

個々の2次元回折像にオーバーサンプリング法を用いて位相回復

⇩

フーリエ変換で2次元実像を得る
近似的に射影像

⇩

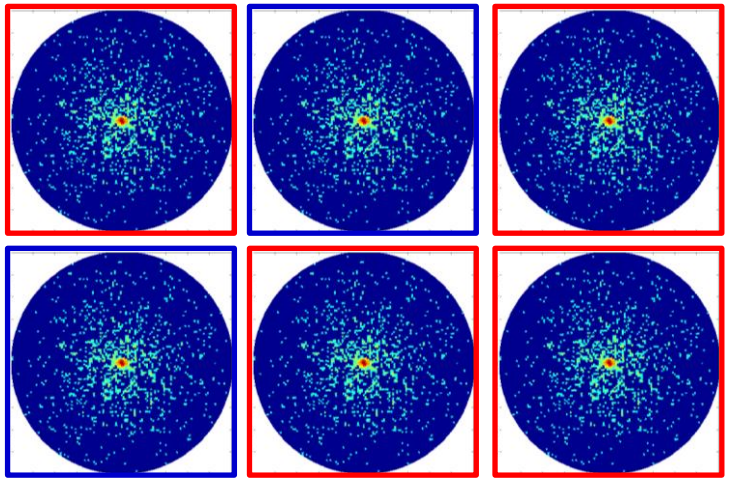
Tomographyにより立体構造をえる

➢実験条件に応じ適した解析法が存在する

高分解能単粒子構造解析の手順

基本手順: **Huldt, G., Szoeke, A., & Hajdu, J. J. Str. Biol. 144, 219-227 (2003)**

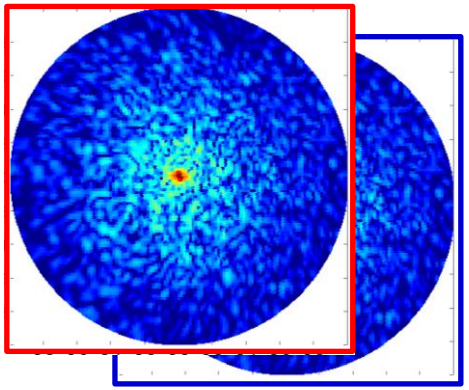
測定された回折像（悪いS/N比）



図柄の類似度により分類

Step1:
分類・平均

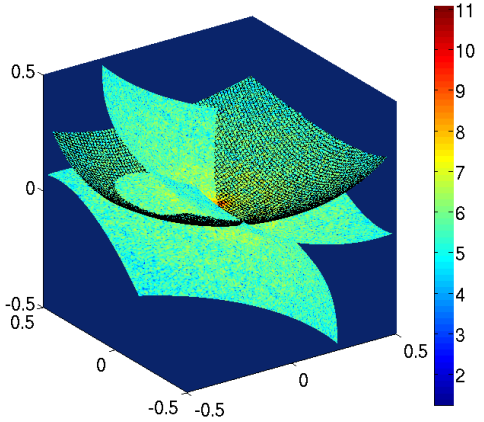
S/N 比の向上した回折像



グループ内で回折像を平均しS/N比向上

Tokuhisa, A., Taka, J., Kono, H., & Go, N. *Acta. Cryst. A* 68, 366 (2012)

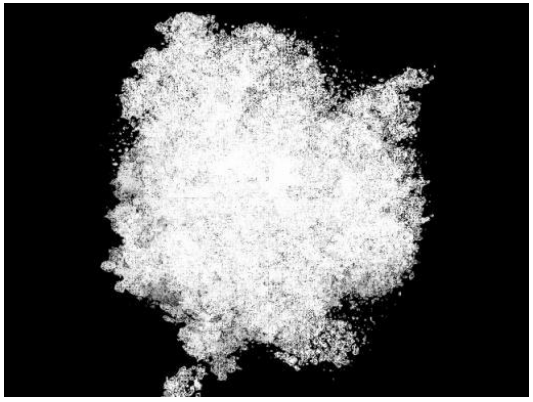
Step2:
相対位置決定



3次元強度関数 (k-space)

Step3:
位相回復

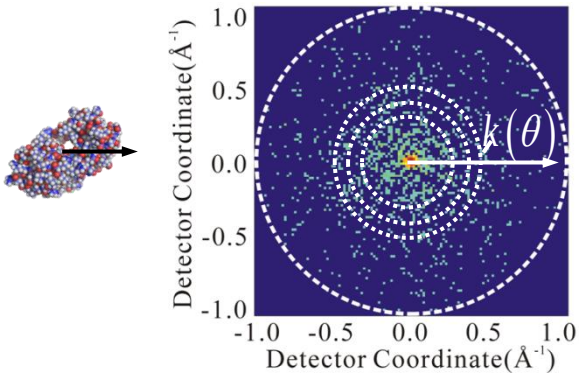
HIO method



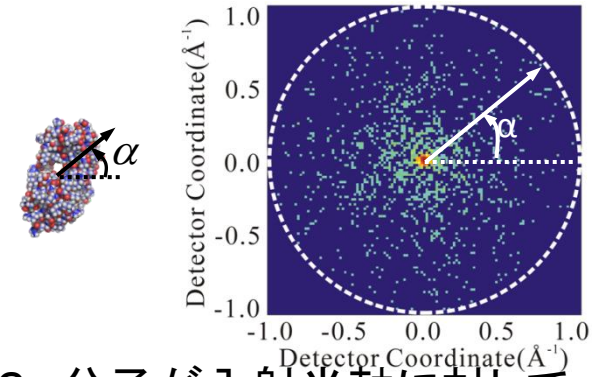
電子密度関数(real-space)

Step1:雑音に強い回折像の分類法

１対の回折像

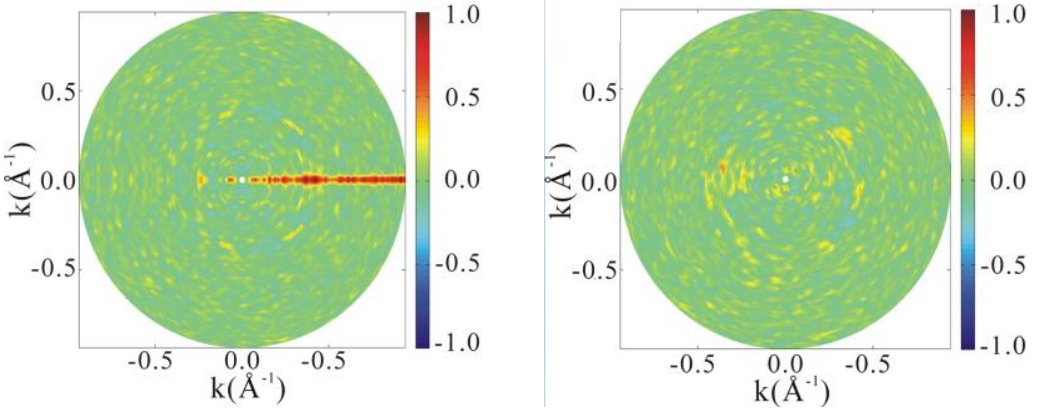


1．大きな波数ほど量子雑音は顕著



2．分子が入射光軸に対してα回転すれば回折像もα回転する

相関図



図柄が似ている場合　　**似ていない場合**

波数kと回転角αの関数として１対の回折像の相関を定義

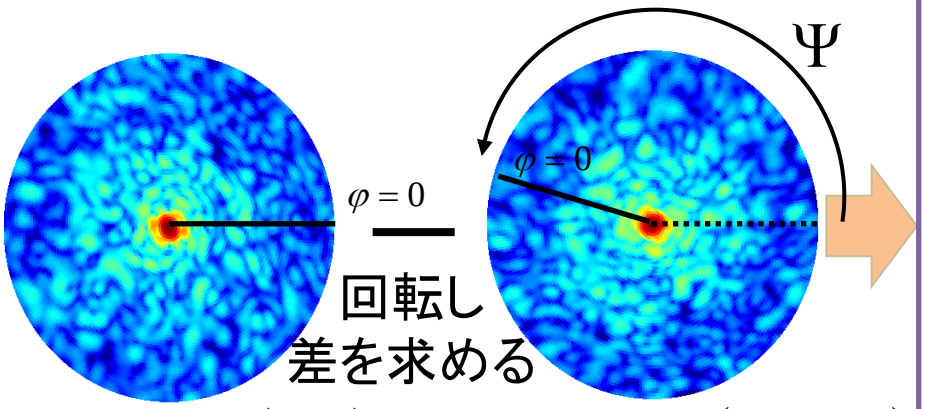
$$c_{ij}(\theta,\alpha)=\frac{1}{N_\theta}\sum_{l=0}^{N_\theta-1}\left(\tilde{s}_Q\left(i;\theta,\frac{2\pi l}{N_\theta}\right)-1\right)\left(\tilde{s}_Q\left(j;\theta,\frac{2\pi l}{N_\theta}+\alpha\right)-1\right)$$

$$\tilde{s}_Q\left(i;\theta,\frac{2\pi l}{N_\theta}\right)=s_Q\left(i;\theta,\frac{2\pi l}{N_\theta}\right)\Big/\bar{s}_Q(i;\theta)$$

$$\bar{s}_Q(i;\theta)=\frac{1}{N_\theta}\sum_{l=0}^{N_\theta-1}s_Q\left(i;\theta,\frac{2\pi l}{N_\theta}\right)$$

- 有効画素あたり1/10光子まで有効信号として扱える
- １対の回折像の図柄がどれくらい似ているか判断するための方法
- 回折像がもつ2つの特徴を考慮
- 図柄が似ている場合相関線が現れる→同じグループに分類→平均によりS/N比向上

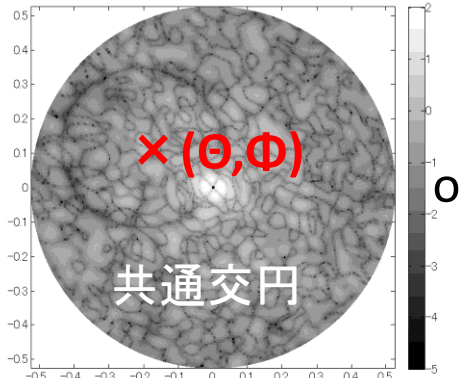
Step2:3次元強度関数の構築



Group1 : $s_{\mathbf{A}(1)}(\theta,\varphi)$　　Group2 : $s_{\mathbf{A}(2)}(\theta,\Psi-\varphi)$

S/N比を向上した2次元回折像

反転した場合

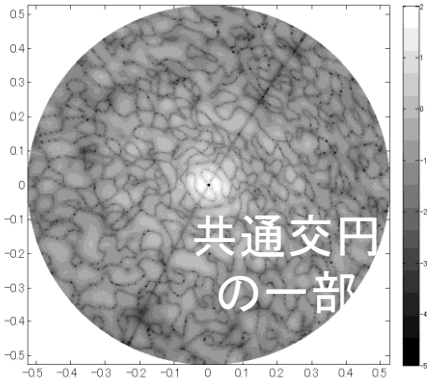


$s_{A(1)}(\theta-\varphi)-s_{A(2)}(\theta,\Psi-\varphi)$

$\Psi=170°$

or

非反転の場合



$s_{A(1)}(\theta-\varphi)-s_{A(2)}(\theta,\varphi-\Psi')$

$\Psi'=130°$

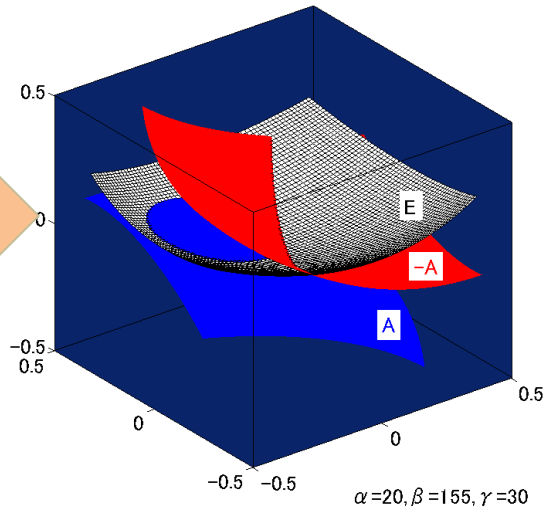
相対的な分子方位(α,β,γ) は Ψ,Θ , Φ から決まる

$$\begin{cases}\alpha=\Psi-\Phi=\Phi'-\Psi'-\pi\\ \beta=\pi-2\Theta=2\Theta'\\ \gamma=\pi-\Phi=-\Phi'\end{cases}$$

(α,β,γ) : 一対の回折像の相対的オイラー角

Ψ : 共通交円が現れる回転角

(Θ, Φ) : 共通交円の極座標



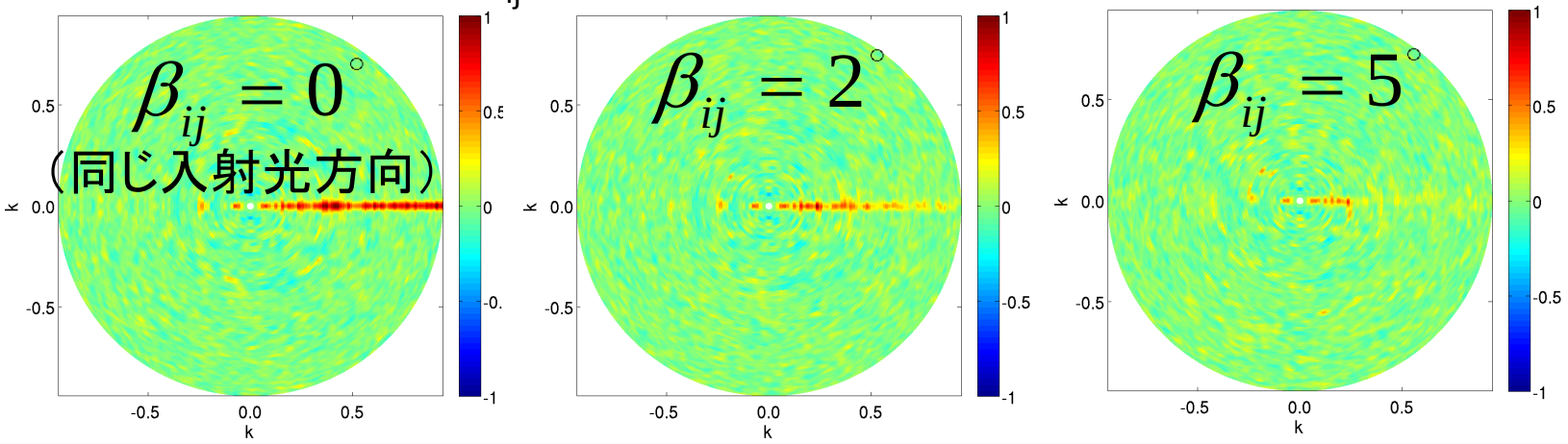
- 1対の回折像は波数空間で交わりを持つ
- この共通部分を差像より求めることで相対位置を決める

アウトライン

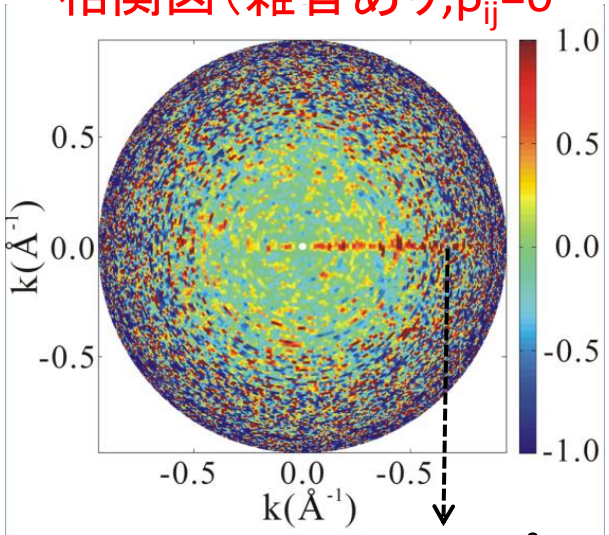
- 高分解能単粒子構造解析の概要
  →コヒーレントイメージングの手法の1つ
- 開発した2つのアルゴリズムの説明
  1. 回折像の分類・平均法
  2. 回折像の相対位置の決定法
- **達成可能な分解能**
  適した実験条件を考えるための指標を与える
  →分子？入射強度？測定回数？
- 大量の回折像を分類するための2つの取り組み
  1. 相関線の自動判定法の導入
  2. 京コンピュータへのアルゴリズム実装
- 実験の現状に適した方法

量子雑音の程度が分解能を決める

相関図（雑音なし）　β_{ij}：1対の回折像(I,j)の分子から見た入射光方向の違い



相関図（雑音あり,$\beta_{ij}=0^\circ$）



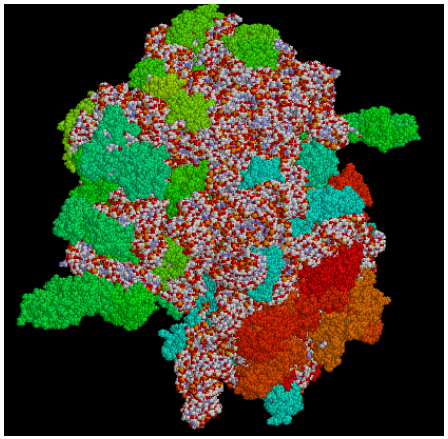
雑音が顕著になる波数：k_N(Å^{-1})

k_N(Å^{-1})の逆数を分解能とした

1. β_{ij}が大きくなるにつれ相関線が見える波数領域が狭くなる
2. $\beta_{ij}=0^\circ$ でも量子雑音により波数k_N[Å^{-1}] で相関線が認識できなくなる
3. k_N[Å^{-1}] より外側は似ているか判断できない
4. 平均によりk_N[Å^{-1}] までS/N比が向上

➢量子雑音の程度を散乱強度動径関数から評価することで雑音が顕著になる波数を見積もることが可能

達成可能な分解能（SACLA設計値）



270 Å

70s-Ribosome
PDB_ID:1YL3,1YL4

- 入射X線強度（SACLA設計値）：$5x10^{11}$ [photons/pulse]
- 集光径：50[nm]
- サンプル:70s-Ribosome(L=270 Å)
- 入射X線強度密度:$2.55x10^{20}$[photons/pulse/mm^2]
- 入射X線波長：1[Å]=12.4[keV]
- 達成可能な分解能：3.0 Å（k_N=0.33 $Å^{-1}$）
- k_Nにおける光子数期待値：0.09
- グループの角：0.64[deg]
- グループ数（波数空間の対称性利用）：16,000
- グループ内で平均に必要な枚数：88枚
- 必要有効回折像：$1.4x10^6$枚
- 分類計算の回数 ：$2.2x10^{10}$回（グループ数x回折像数）
- 総回折像データ量：23.5TB（16.8MB/回折像）

➢播磨設置の10Tフロップス計算機を用いて分類した場合
→100日ほどかかる見積もり（1相関計算0.5秒）

ここまでのまとめ

- 高分解能単粒子構造解析のアルゴリズムを開発
- XFEL光を数十nm集光（～10^{20}[photons/pulse/mm^2]）することで原子分解能の達成が示唆
- 高分解能単粒子構造解析を実現するための課題:
  - ①大量の回折像を分類するための計算機構
  - ②効率よく測定を行うための実験系の構築
  - ③放射線損傷の影響
  - ④試料回りの水の影響
  - ⑤構造多型の影響

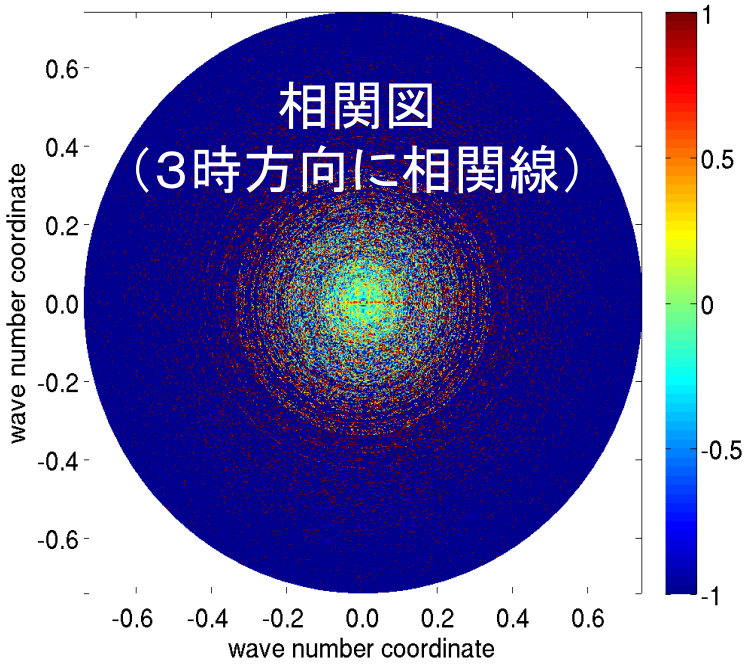
アウトライン

- 高分解能単粒子構造解析の概要
  →コヒーレントイメージングの手法の1つ
- 開発した2つのアルゴリズムの説明
  1. 回折像の分類・平均法
  2. 回折像の相対位置の決定法
- 達成可能な分解能
  →原子分解能の達成が視野
- **大量の回折像を分類するための2つの取り組み**
  1. 相関線の自動判定法の導入→感度も向上
  2. 京コンピュータへのアルゴリズム実装
- 実験の現状に適した方法

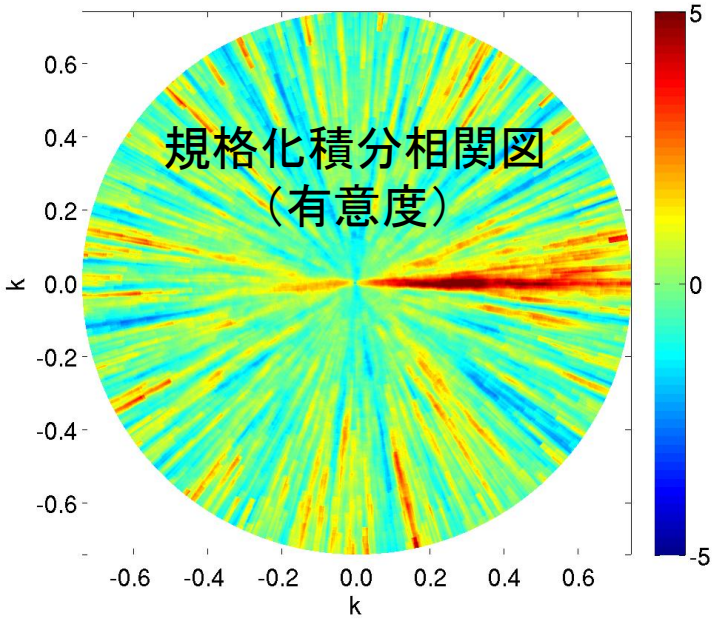
相関線自動判定法の導入

Ribosome,
I_i= 5x10^{19}[photons/pulse/mm^2].
βij=0.59°

導入前



導入後



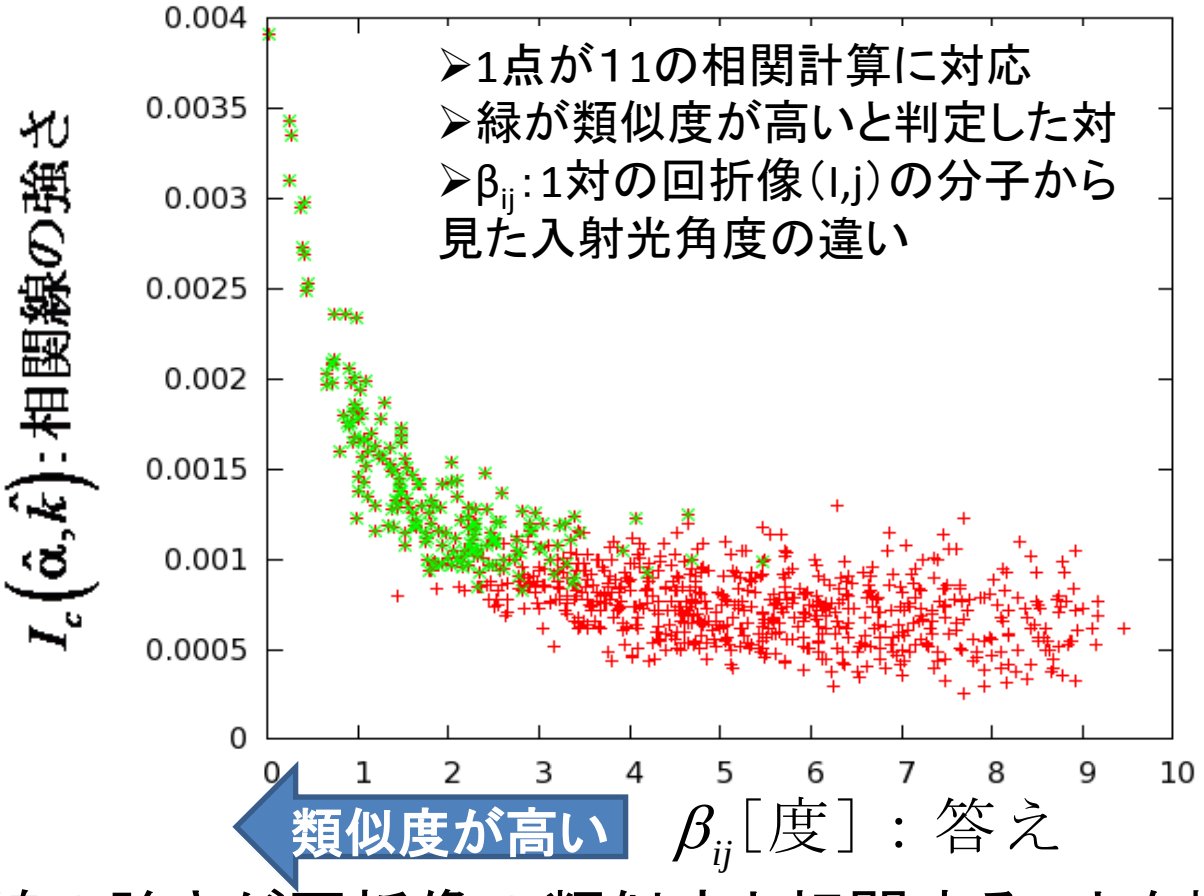
相関図の積分値を導入することで相関線の①方向②長さ③強さを自動判定
→大量の回折像を分類可能

相関図の積分値を考える背景
1. １枚の図の中で雑音を平均化
2. 微弱な信号を集積

→ 感度向上!!

自動判定法の結果（SACLA設計値）

Ribosome,I_i=2.55x10^{20}[photons/pulse/mm^2](設計値を 50nmに集光した強度)

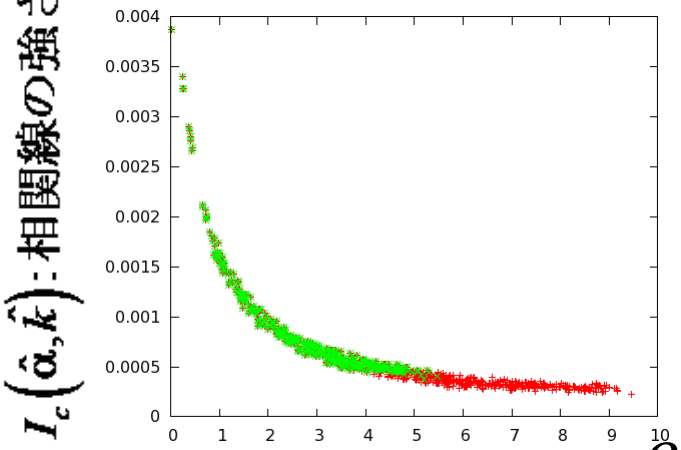


➢相関線の強さが回折像の類似度と相関することを期待
→図柄の類似度が高い対を正しく拾うことができている
➢I_cとβ_{ij}の相関が高いことから狭いβ_{ij}角でグループ化が可能
→前述の見積もりよりも高い分解能を狙える可能性も出てきた

いくつかのパラメータ最適化前の暫定的な結果

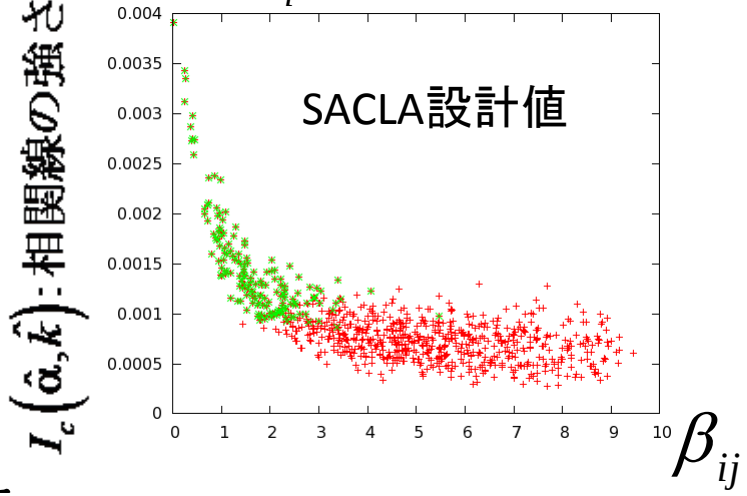
自動判定結果の入射強度依存性

量子雑音無し

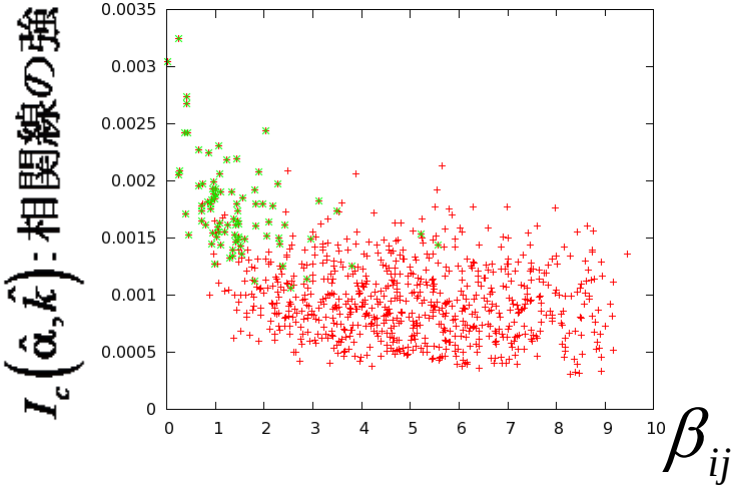


β_{ij}：答え

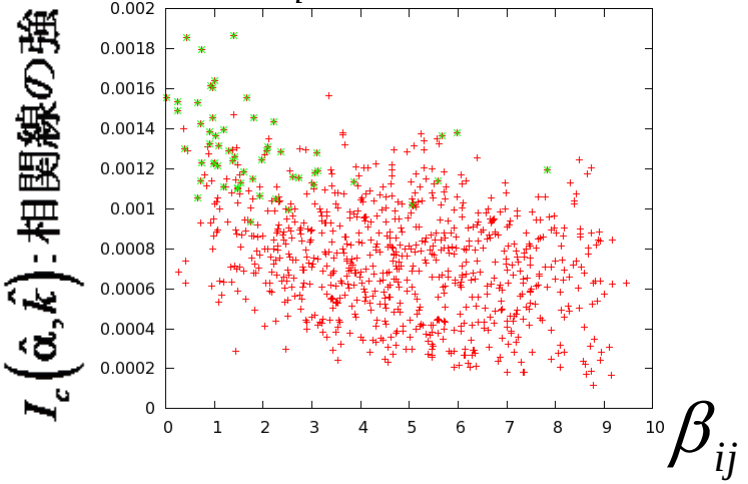
$I_i = 2.55\times10^{20}$



$I_i = 8\times10^{19}$



$I_i = 5\times10^{19}$



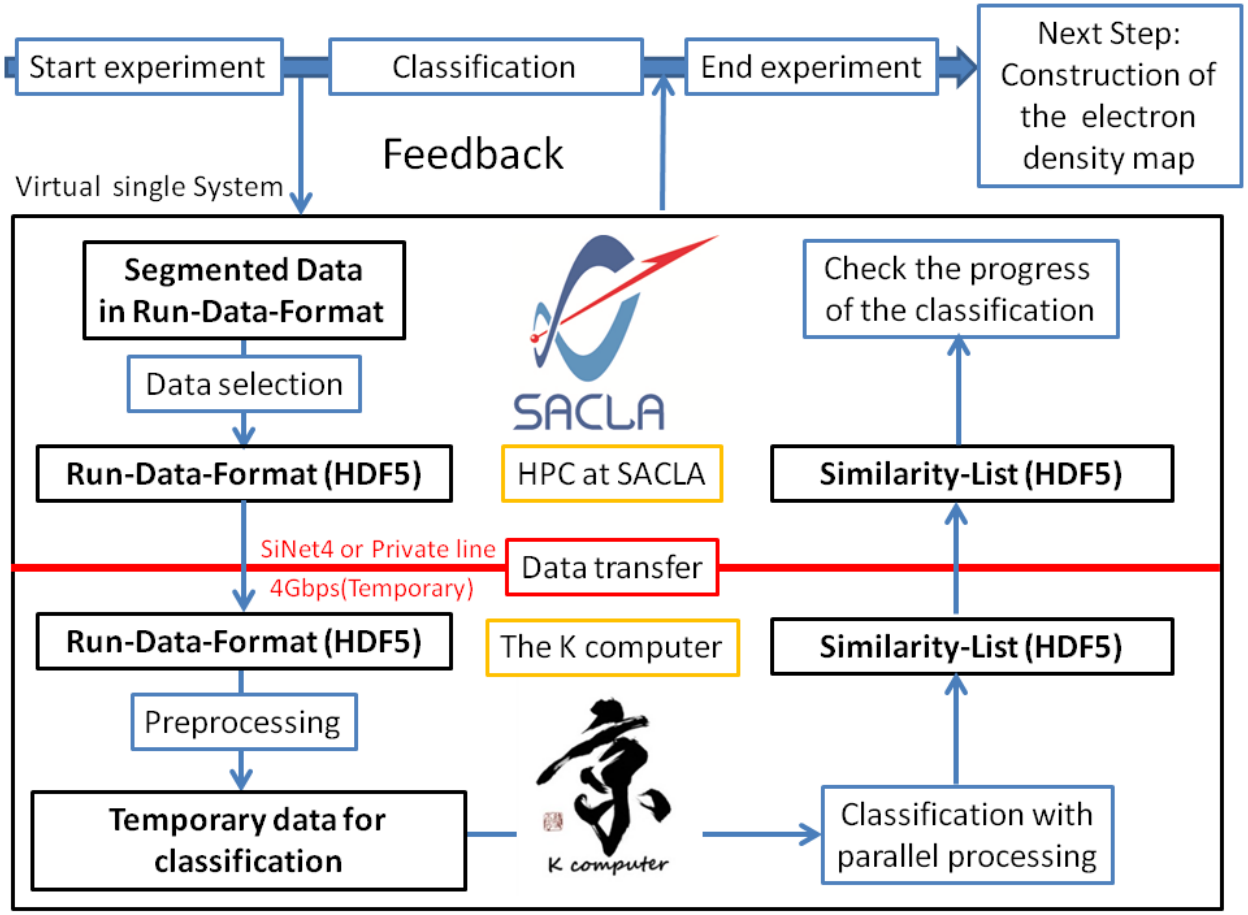
- 入射強度を下げると雑音の影響でI_cとβ_{ij}の関係が崩れる
- 入射X線強度8×10^{19}[photons/pulse/mm^2]以上で回折像を分類可能

開発したアルゴリズムの京への実装

- 初版（2013年～3月）：とにかく動くものを作成
  担当：東京大学　石川研　修士　新井淳也氏
  成果：385ノード1時間のジョブを255回行い、100万枚の回折像を1.4万のグループに分類した
  問題点１：I/O時間が全体の約半分を占めていた
  問題点２：京の全ノードを用いた場合MPIのメモリ消費が無視できない
  問題点３：代表回折像の不足→約3割の回折像が無所属
  （Tokuhisa et.al. , J. Synchrotron Radiation, 2013 in press ）
- 第2版（2013年4月～）：初版の問題点を改良し高度化
  担当：理研　計算科学研究機構　吉永一美博士
  成果：京の全ノードを用いて115分で、160万の回折像を2万のグループへ分類した

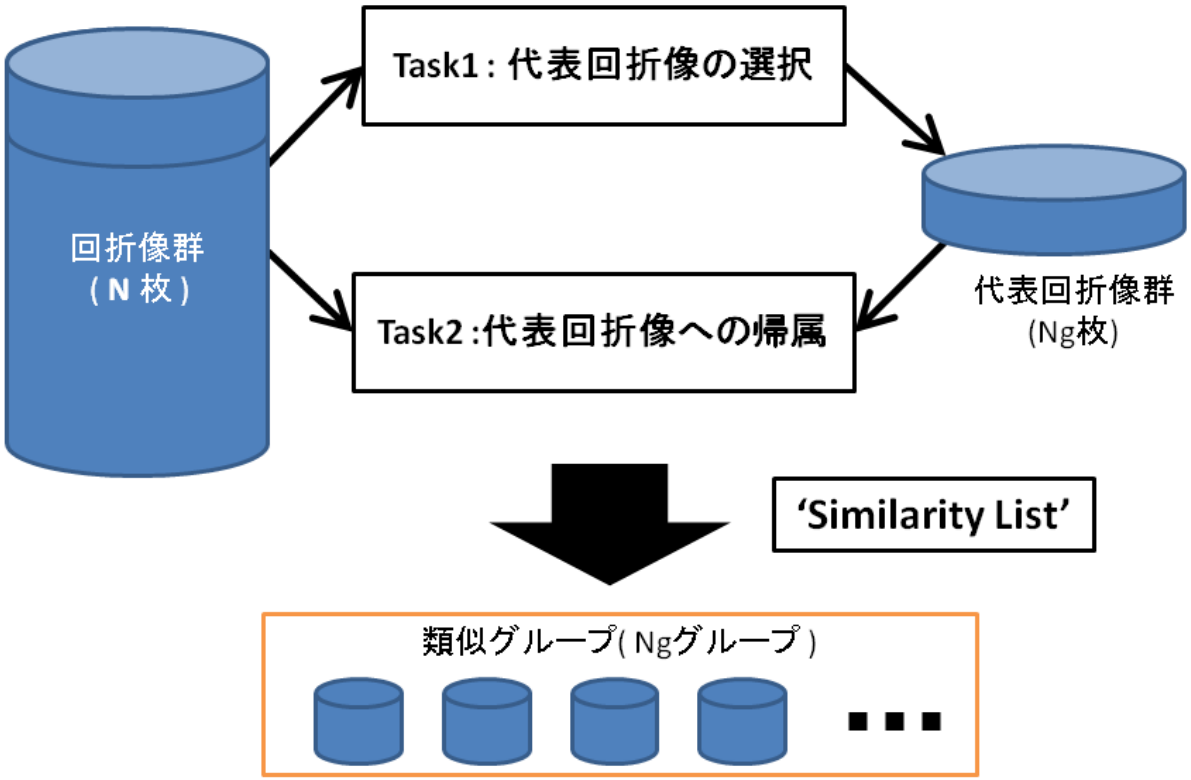
サポート：
- 理研理事長ファンド（FY2012-2013）
  「SACLAと「京」の融合利用による生体超分子の立体構造解析技術の開拓」
- 京利用一般公募枠（課題番号hp120213）
  「SACLA大量回折像データの超並列計算による迅速クラスタリングツールの整備」

京コンピュータが必要な理由



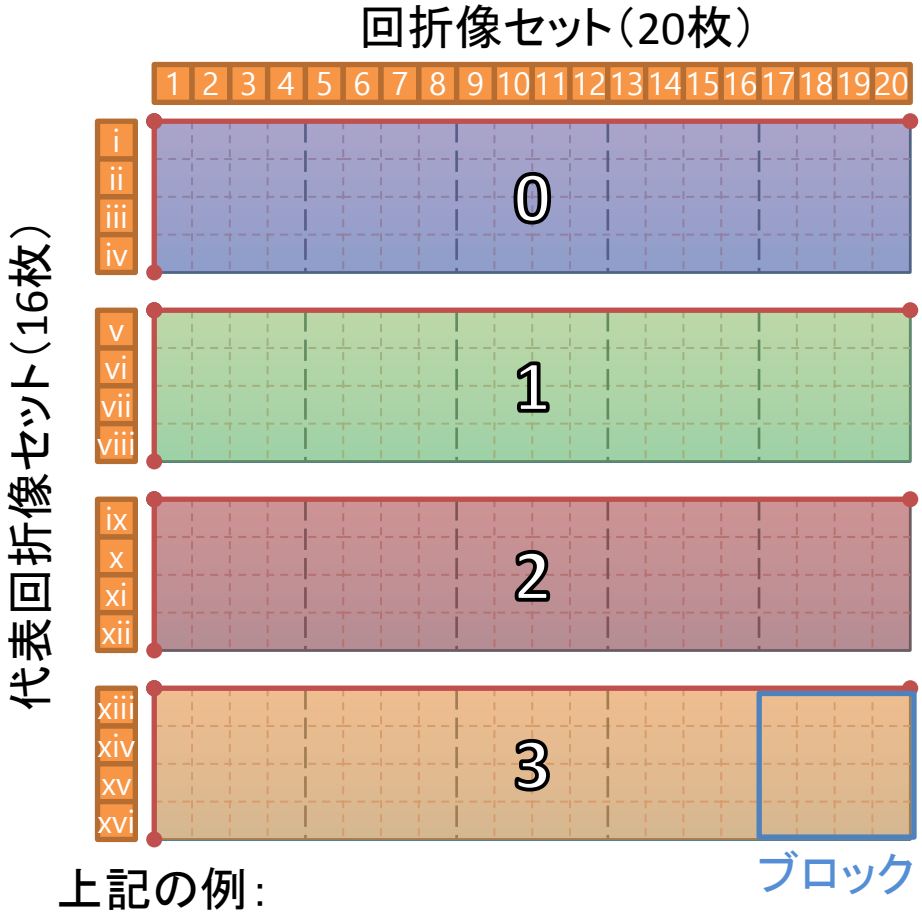
➢データ診断のためのフィードバックシステムを構築したい
→京を用いて回折像の分類を並列処理することで敏速に実行

分類スキーム



- 実験と並行した解析を実現するため、2つの過程に分離
  - Task1:代表回折像選出過程
    - 一部の回折像群に対して総当たりの相関計算
  - Task2:代表回折像に対する帰属過程
    - 代表回折像群とその他回折像群の直積集合要素について相関計算
- 利点 : 相関計算の分割に対して高い柔軟性を持つ

Task2の並列化手法（初版）

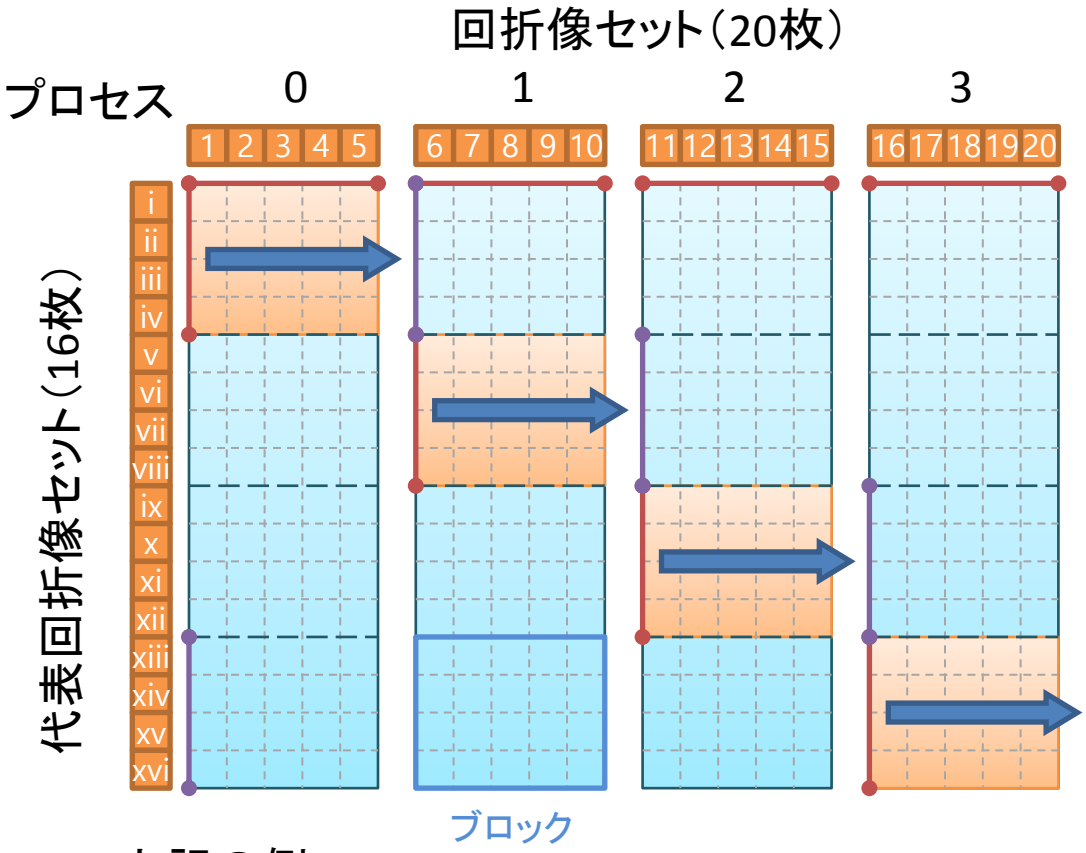


上記の例：
- 20の回折像を16の代表回折像に帰属
- 20のブロックを4つのプロセスが処理

- 小さなマス目が1相関計算に対応
- 処理を複数の「ブロック」に分割し、プロセスに均等に割り振る
- 相関計算に必要な回折像はプロセス毎にファイルから読み込む
  →同じ画像を何度も読み込むことになる
- フラットMPIで実装
- 1ノード8プロセス
  →京全ノードを使うと663,552プロセス
  コミュニケータによるメモリ消費が莫大

- 問題点１：I/O時間が全体の約半分を占めていた
- 問題点２：京の全ノードを用いた場合MPIのメモリ消費が無視できない

Task2の並列化手法（第2版）



- 小さなマス目が1相関計算に対応
- 処理を複数の「ブロック」に分割し、プロセスに均等に割り振る
- I/Oの集中を防止するため

1. 各プロセスが分担して初回ステップに必要な画像のみ読み込む
2. 相関計算の裏で、画像を隣のプロセスへ送る

- ハイブリッドMPIで実装
  →ノード間はMPIで、ノード内はOpenMP
- 1ノード1プロセス

上記の例：
・20の回折像を16の代表回折像に帰属
・16のブロックを4つのプロセスが4ステップで処理

- 画像読み込み量を削減した
- メモリ消費量を削減した

新旧並列化手法の実行時間比較

- 107万枚回折像の分類計算を多数のセットに分割して実行
- 1セット：13,252枚の代表回折像と5,079枚の回折像の相関計算（385ノード使用）

	旧手法（初版）	新手法（第2版）
1セットの処理時間	49分30秒	26分57秒
相関計算の割合	52%	97%
107万枚回折像の分類計算	57,500ノード時間（※実測値）	31,400ノード時間（予測値）

※成果：京コンピュータを用いて255セットの分類計算を行った（大規模計算1回目）

➢ファイル読み込み時間が削減され処理時間が大幅に短縮された

京を用いた大規模計算 (2回目)

大規模計算1回目の問題点

問題点1：I/O時間が全体の約半分を占めていた

問題点2：京の全ノードを用いた場合MPIのメモリ消費が無視できない

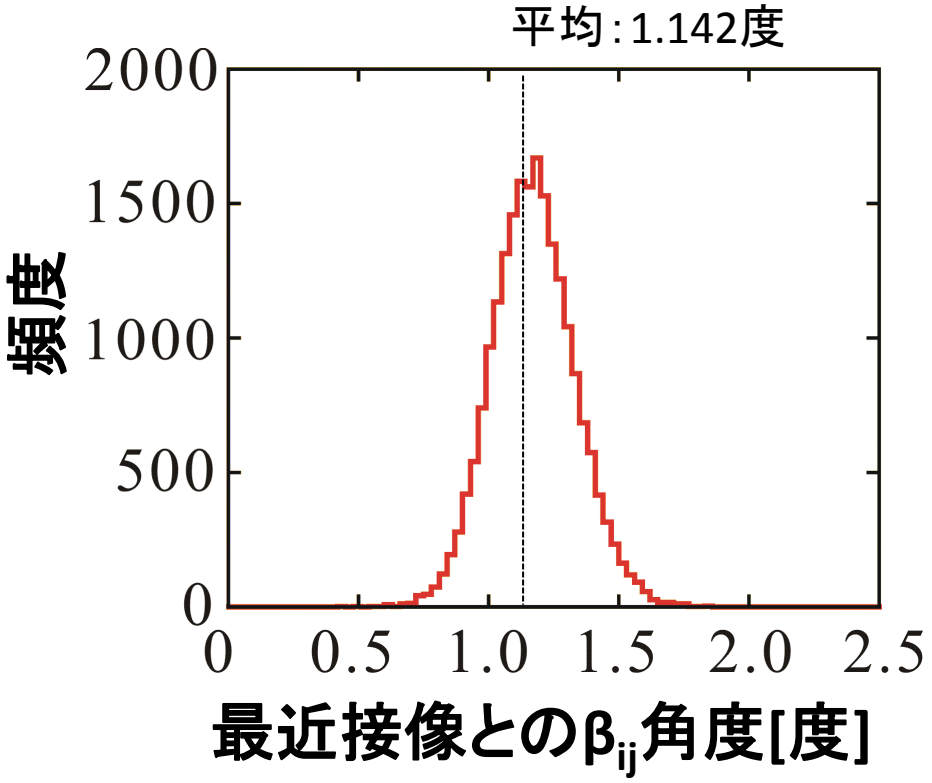
問題点3：代表回折像の不足→約3割の回折像が無所属

設定した計算条件	
目的分解能	5.5 Å
サンプル	70s-Ribosome
入射X線強度密度	2.55×10^{20}[photons/pulse/mm^2]
グループ角	1.167度
グループ数	※20,821
グループでの平均枚数	80枚
必要な回折像数	166,5万枚
回折像1枚のデータ量	15MB

※波数空間の対称性は利用していない。代表間の隙間をなくすための係数を2.16とした。

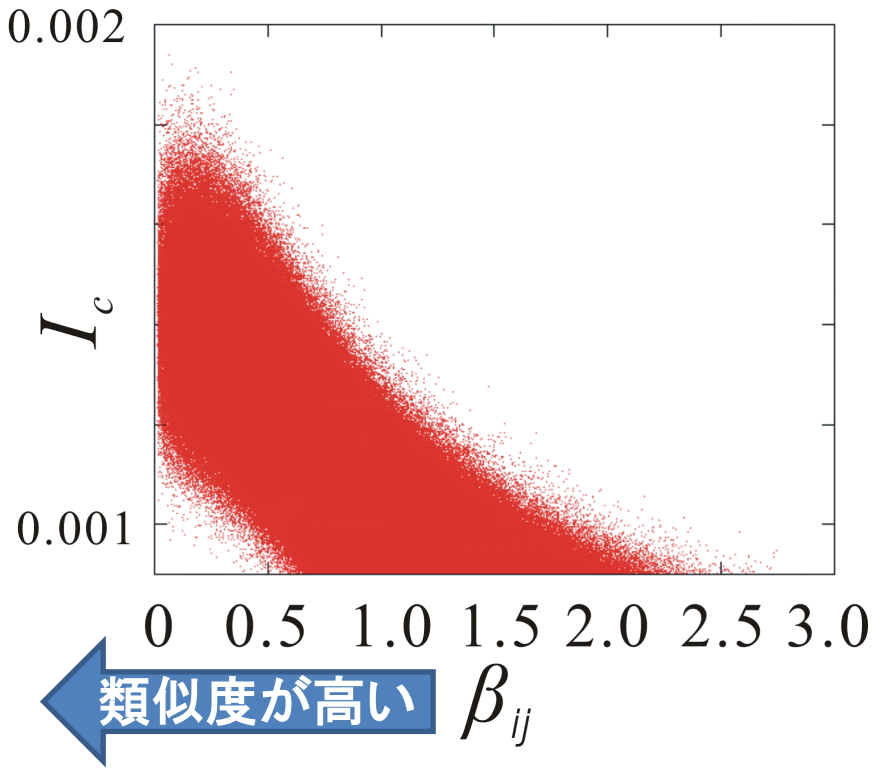
➢京の全ノードを用いて、約167万の回折像セット（25TB）を2万のグループへ分類

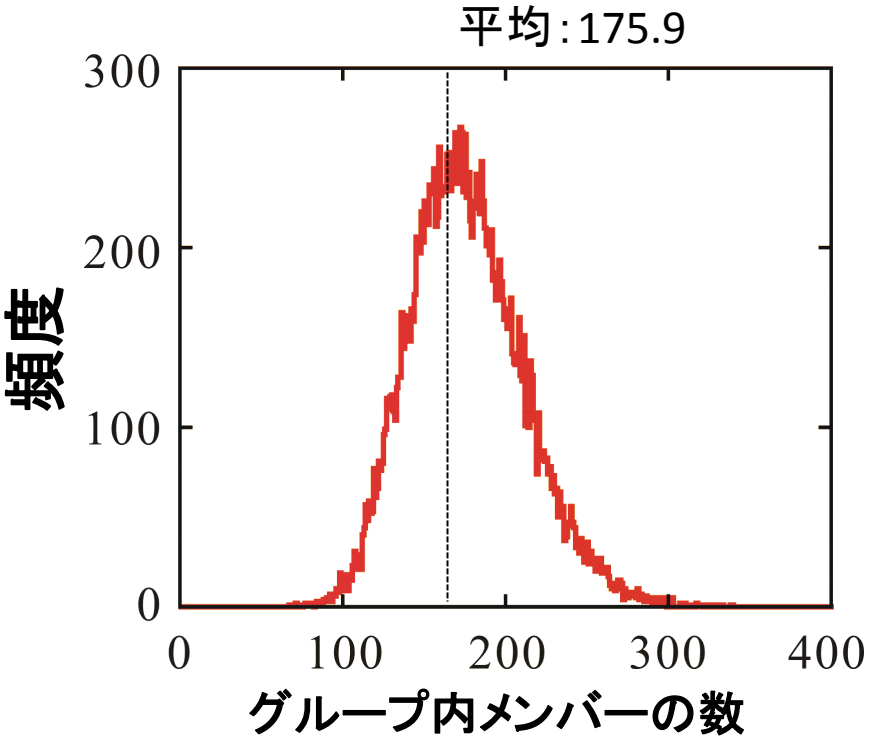
Task1:代表回折像選出過程



- 代表回折像の定義:他のどの代表回折像と閾値よりも似ていない回折像
- 15万枚の回折像群に対して総当たり計算
- 類似度に対する閾値:$I_{c,th}$=0.0009
- 選出された代表回折像数:20,951(目的の数: 20,821)
- 最近接像との平均角度:1.142度(目的の角度: 1.167度)
- 複数回に分割して実行→約2.8万ノード・時間

Task2：代表回折像への帰属





- 代表回折像数：20,951、回折像数:158.6万→332.3億回の相関計算
- 京の全ノード(82,944)を用いて115分→約16万ノード・時間
- 閾値0.0009→368.5万の対が似ていると判定された
- 代表回折像に対して平均175.9枚の回折像が帰属された（目標値は80枚）

分類結果の評価

‘名前’	‘説明’	‘値’
$I_{c,threshold}$	類似度に対する閾値（高いと似ていると判定）	0.0009
$P_{right} = N_{A\cap B\cap C} / N_C$	正解率（補足された集合のうち正しい率）	0.66
$P_{capture} = N_{A\cap B\cap C} / N_A$	補足率（答え集合のうち正しく補足した率）	0.70
N_A	β_{ij}<1.167 度の対の数	344,200pairs
N_B	Δα< 1 .167度の対の数	304,421,232pairs
N_C	$I_c > I_{c,threshold}$の対の数	3,685,344pairs
$N_{A\cap B\cap C}$	A,B ,Cの積集合の対の数	2,418,154pairs
$N_{orphans}$	どの代表にも属さなかった回折像の数	4,885/158.6万
$N_{count=1}$	1つの代表にのみ属した回折像の数	232,172
$N_{count=2}$	2つの代表に属した回折像の数	712,155

- 99.7%の回折像を有効利用
- 正解率・補足率を向上することが今後の課題

大規模解析のまとめ

- 京の全ノードを用いて115分で約160万の回折像セット（24 TB）を2万のグループへ分類した→5.5Å分解能
- 処理時間の内訳（新手法）
  - 相関計算（5.5割、うち約半分がフーリエ変換）
  - 通信（2.5割）
  - I/O（2割）

  →全ノードを用いると画像読み込み速度の低下がみられた
- 3Å分解能の計算時間予測
  - 1相関あたりの計算時間：0.065秒→0.45秒(約7倍)
  - 代表回折像の数： 20,951 →69,984(約3.5倍)
  - 回折像の数：158.6万→615.9万(約4倍)

  →京全ノードで約111時間（921万ノード時間）

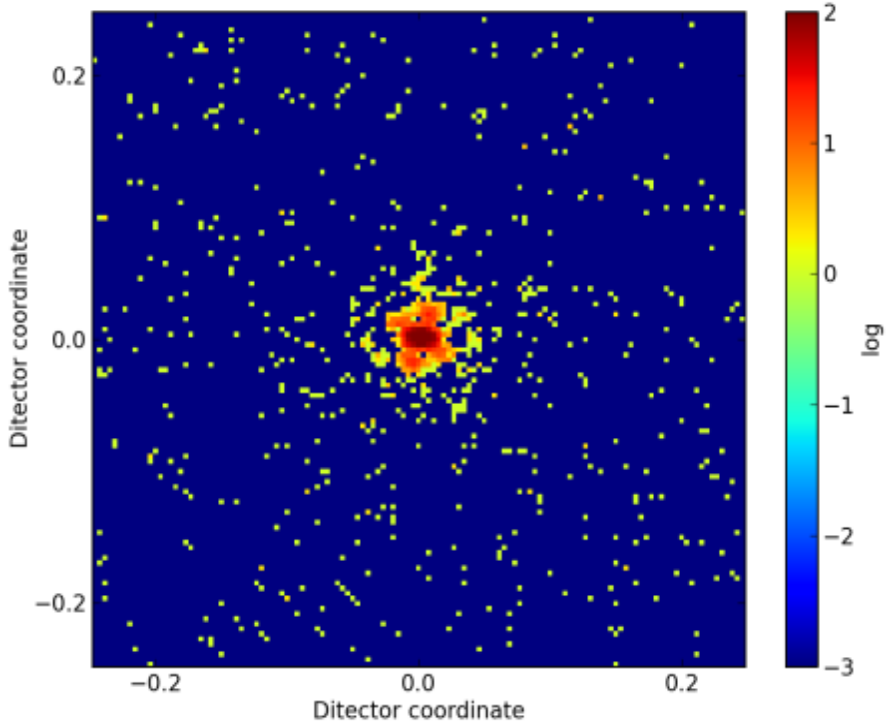
- 通信とI/Oに関しては現在原因究明を行い、改良する予定

アウトライン

- 高分解能単粒子構造解析の概要
  →コヒーレントイメージングの手法の1つ
- 開発した2つのアルゴリズムの説明
  1. 回折像の分類・平均法
  2. 回折像の相対位置の決定法
- 達成可能な分解能
  →原子分解能の達成が視野
- 大量の回折像を分類するための2つの取り組み
  1. 相関線の自動判定法の導入→感度も向上
  2. 京コンピュータへのアルゴリズム実装
- **実験の現状に適した方法**

実験の現状

低分解能データセットを取得できる



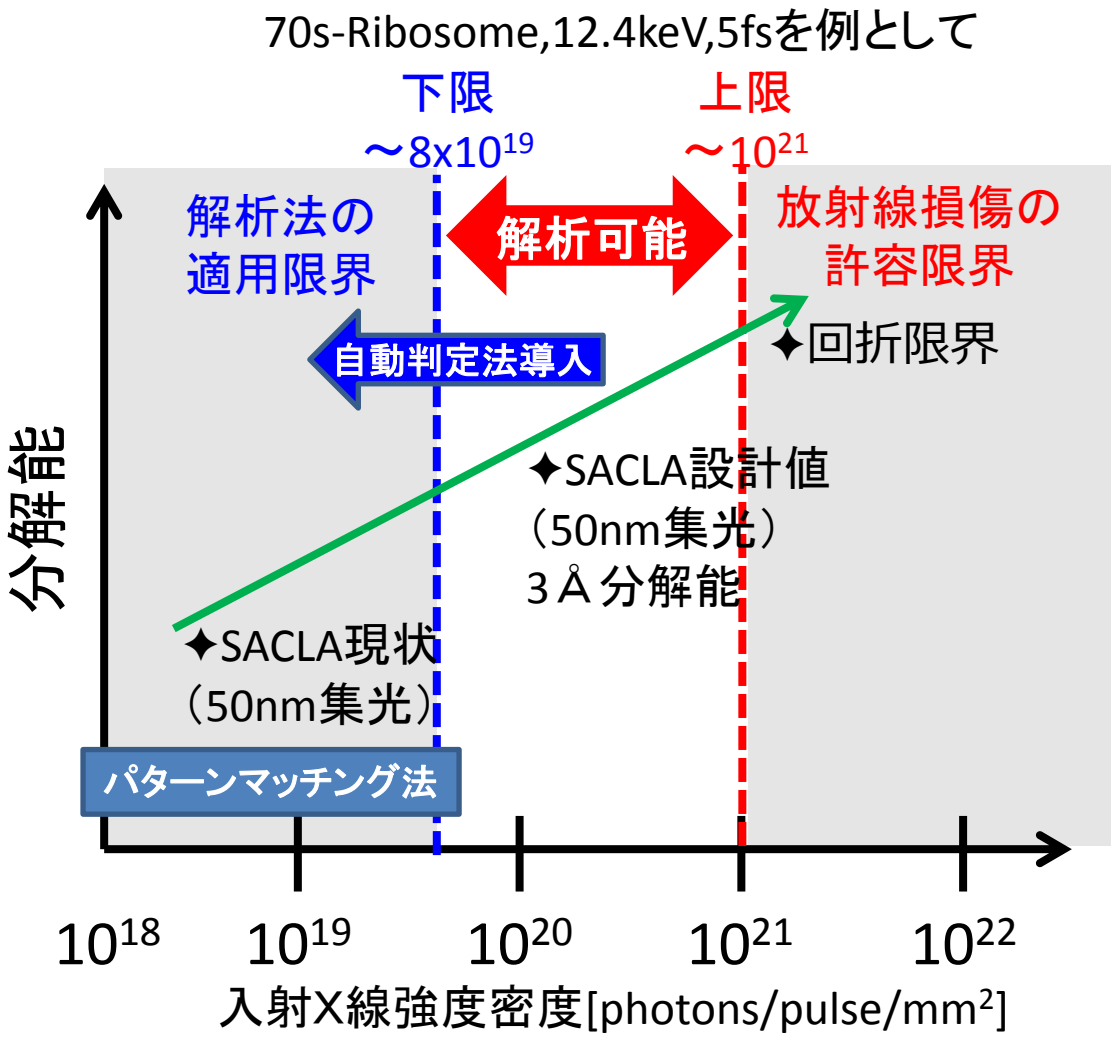
波数～0.2 Å →～5 Å 分解能

- 弱い強度
- 小さな散乱角
- 少数の回折像

（立体構造の部分情報）

- 低分解能データセットに適した解析法とは？

まとめ



解析可能な領域を広げた!!

フェムト秒X線パルスレーザを用いた高分解能構造解析を目指して

- ➢高分解能達成可能なアルゴリズムを開発
- ➢SACLA設計値で原子分解能の達成が視野
- ➢回折像類似度の自動判定法導入により入射光強度の下限を下げた
  - →入射強度の下限
  ～8×10^{19} [photons/pulse/mm^2]
  - →入射強度の上限
  （放射線損傷が決める）
  ～10^{21}[photons/pulse/mm^2]
- ➢データ診断系の構築を目指した大規模データ解析を京により実現した
- ➢構築した大規模解析機構は低分解能データセットの解析に対しても有効

謝辞

- **構造解析アルゴリズムの開発**

  郷信広（京大名誉教授）、河野秀俊（原子力機構）

- **京への実装**

  京コンピュータの方々：

  石川裕（東大）、堀敦史（理研）、吉永一美（理研）、新井淳也（東大）、亀山豊久（理研）、大野善之（理研）、畑中正行（理研）、山本啓二（理研）、ゲローフィ・バリ（理研）、黒川 原佳（理研）、庄司文由（理研）、島田明男（理研）、横川三津夫（理研）

  SACLAの方々：

  初井宇記（理研）、城地保昌（JASRI）、田中良太郎（JASRI）、山鹿光裕（JASRI）、杉本 崇（JASRI）、岡田謙介（JASRI）

- **パターンマッチング法の開発**

  Tama　Florence（理研）、宮下治（理研）

謹んで感謝いたします