Special Interest Group

X-ray Diffraction Data for Refinement and Deposition

Xinhua Ji

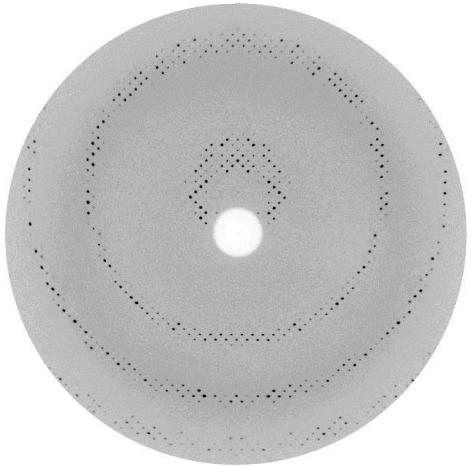
Intensity, Structure Factor, Electron Density, and Structure

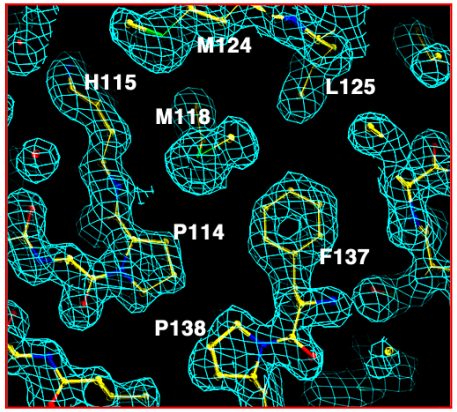
$$I_{hkl} = |F_{hkl}|^2$$

$$F_{hkl} \underset{\int_v dv}{\overset{\sum_{hkl=-\infty}^{+\infty}}{\rightleftharpoons}} \rho_{xyz}$$

$$\| $$

$$|F_{hkl}| e^{i\phi_{hkl}}$$





Data ⟶ Phase Problem ⟶ Results

NIH X-ray Interest Group: Newsletter

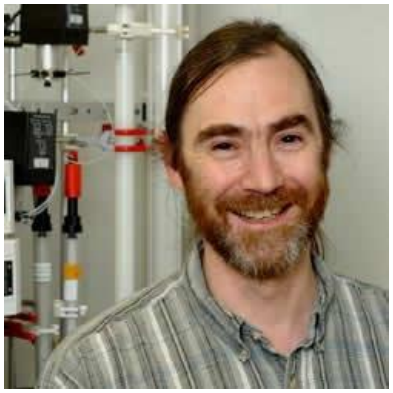
TOPIC DISCUSSION - Data for Refinement and Deposition/Publication

Xinhua Ji (NCI): High-resolution data, even not complete, always helps improve electron density that reveals additional structure features. Therefore, it is beneficial to include more data in the refinement. Claiming a resolution for structure deposition/publication can be done at the final stage of the refinement. A guide line I have been using is shown below. Please comment and/or advise.

	Refinement		Deposition	
	Overall	Last Shell	Overall	Last Shell
Completeness (%)	> 85	> 50	> 93	> 70
$I/\sigma(I)$	> 10	> 1	> 10	> 2
R_{merge}	< 0.10	< 0.50	< 0.10	< 0.50

Mark Mayer, NICHD

TOPIC DISCUSSION - Data for Refinement and Deposition/Publication



Mark Mayer (NICHD): I understand the benefit of using weak and incomplete data in high resolution shells for calculating maps and improving model building, especially with the routine use of rpim, cc and cc* at the stage of scaling supporting use of reflections in shells with with I/sigma < 2, but I don't understand how to proceed to the deposition/publication stage.

After completing model building and refinement using all the data, why would we drop weak and incomplete data in the last round of refinement to achieve > 70% completeness and I/sigma > 2 or some other arbitrary cut off that will satisfy reviewers/PDB annotaters? If maps improve with weak and incomplete data in high resolution shells, then there is useful structural information, so why throw it away?

Mariusz Jaskolski, PAC

TOPIC DISCUSSION - Data for Refinement and Deposition/Publication

Mariusz Jaskolski (Polish Academy of Sciences): Thanks very much for initiating a discussion about the use of high-resolution reflections for refinement and at other stages of structure determination/publication. I have a lot of comments and practical remarks in this area, and I have summarized some of them in a one-page document.

Personally, I am not in favor of using different data for structure modeling and refinement, and different for publication/deposition. Even with the best of intentions, this encourages ghost chasing and complicates reproducibility, even if the reader is scrupulously informed about the procedure. I think an optimal data set should be prepared early on and then used consistently at all stages of structure determination, analysis, validation, and deposition.

Weak Data Do No Harm

How good are my data and what is the resolution?

Philip R. Evans[a*] and Garib N. Murshudov[a]
[a]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, England
Acta D, 69:1204-1214, 2013

At the very least, adding these weak data seems to do no harm for the purposes of either automatic or manual model building.

From our limited tests here, it seems that changing the resolution cutoff over a considerable range (*e.g.* from 2.2 to 1.9 Å) makes only a small difference, so the exact cutoff point is not a question to agonize over, but it seems sensible to set a generous limit so as not to exclude data containing real (if weak) information.

Weak Data Contain Real Information

Better models by discarding data?

K. Diederichs[a,*] and P. A. Karplus[b]
[A] Faculty of Biology, University of Konstanz, M647, 78457 Konstanz, Germany
[B] Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA
Acta D, 69:1215-1222, 2013

Using experimental data sets, the behavior of $CC_{1/2}$ and the more conventional indicators were compared in two situations of practical importance: merging data sets from different crystals and selectively rejecting weak observations or (merged) unique reflections from a data set.

In these situations controlled 'paired-refinement' tests show that even though discarding the weaker data leads to improvements in the merging *R* values, the refined models based on these data are of lower quality. These results show the folly of such data-filtering practices aimed at improving the merging *R* values.

High-Resolution Weak Data Are Important

Inclusion of weak high-resolution X-ray data for improvement of a group II intron structure

Wang J. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. jimin.wang@yale.edu

Acta D 2010, 66:988-1000

Abstract

It is common to report the resolution of a macromolecular structure with the highest resolution shell having an averaged I/sigma(I) > or = 2. Data beyond the resolution thus defined are weak and often poorly measured. The exclusion of these weak data may improve the apparent statistics and also leads to claims of lower resolutions that give some leniency in the acceptable quality of refined models. However, the inclusion of these data can provide additional strong constraints on atomic models during structure refinement and thus help to correct errors in the original models, as has recently been demonstrated for a protein structure…

High-Resolution Weak Data Are Important

A self-spliced group II intron

	Original (3BWP)		Reprocessed (3G78)	
	Overall	Highest res. Shell	Overall	Highest res. Shell
Resolution (Å)	50 – 3.1	3.2 – 3.1	40 – 2.8	2.9 – 2.8
$I/\sigma(I)$	13.9	3.7	20.7	0.4
R_{merge} (%)	14.9	43.9	7.2	> 100
Completeness (%)	99.6	98.7	98.9	92.1
R_{work} (%)	27.6	28.9	19.6	62.7
R_{free} (%)	31.0	28.8	22.6	69.5

High-Resolution Weak Data Are Important

A self-spliced group II intron

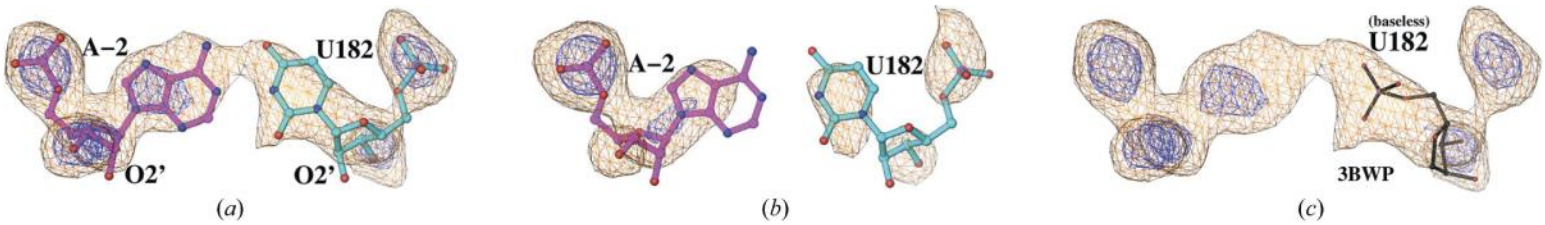


Figure 3. The binding of an RNA product in the catalytic site. (***a***) Our new experimental map superimposed onto our new model for the A2U182Watson–Crick base pair (3g78). (***b***) The original map 3bwp superimposed onto our model 3g78. (***c***) Our experimental map 3g78 superimposed onto the original model 3bwp, which included several 'baseless' nucleotide residues that were built without bases, such as U182. Experimental maps were contoured at 1 (golden) and 3 (blue)

Resolution Cutoff

How good are my data and what is the resolution?

Philip R. Evans[a*] and Garib N. Murshudov[a]
[a]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, England
Acta D, 69:1204-1214, 2013

We cannot set definite rules for this [resolution cutoff], as it depends on what the data are to be used for.

It is therefore a mistake to prematurely apply a harsh cutoff at the data-reduction stage: data can always be excluded later.

Tests carried out here to relate the resolution statistics to final model building and refinement do suggest that extending the data somewhat beyond the traditional limits such as $\langle I/\sigma \rangle = 2$ may improve structure determination, as do the `paired-refinement' tests of Karplus & Diederichs (2012).

Acta D Recommendations

Notes for Authors 2012

11.1. Resolution

The effective resolution should be described clearly. Values of the internal agreement of the data, R_{merge}, together with the multiplicity, the mean value of I/σ and the percentage completeness of the data are required for the overall data set and the highest resolution shell together with the limits of that shell in Å. For high-quality data obtained with synchrotron radiation, completeness > 93% and observable data > 70% should be achievable for the highest resolution shell.

Acta F Requirements

Notes for Authors 2014

5.1. Structural data

Table 3. Data collection and processing

 # If completeness <93% or completeness in outer shell <70%, please provide an explanation [as a footnote here].
† If mean $I/\sigma(I)$ in outer shell is <2.0, please provide an explanation [as a footnote here] and provide resolution at which it falls below 2.0.
‡ Only the redundancy-independent merging *R* factor $R_{r.i.m.}$ or R_{meas} should be reported. If these values are not available, they may be estimated by multiplying the conventional R_{merge} value by the factor $[N/(N - 1)]^{1/2}$, where *N* is the data multiplicity [in such cases, provide a footnote here].

Redundancy-independent R_{rim}

The redundancy-independent merging R factor value Rrim,
also denoted Rmeas, for merging all intensities in this
data set.

$$Rrim = \frac{\text{sum~i~ [N~i~/(N~i~ - 1)]1/2\^{} sum~j~ | I~j~ - <I~i~> |}}{\text{sum~i~ (sum~j~ I~j~)}}$$

I~j~ = the intensity of the jth observation of reflection i
<I~i~> = the mean of the intensities of all observations of
reflection i
N~i~ = the redundancy (the number of times reflection i
has been measured).

sum~i~ is taken over all reflections
sum~j~ is taken over all observations of each reflection.

Ref: Diederichs, K. & Karplus, P. A. (1997). Nature Struct.
Biol. 4, 269-275.
Weiss, M. S. & Hilgenfeld, R. (1997). J. Appl. Cryst.
30, 203-205.
Weiss, M. S. (2001). J. Appl. Cryst. 34, 130-135.

Precision-indicating R_{pim}

The precision-indicating merging R factor value Rpim,
for merging all intensities in this data set.

$$R_{pim} = \frac{\sum_i [1/(N_i - 1)]^{1/2} \sum_j | I_j - \langle I_i \rangle |}{\sum_i (\sum_j I_j)}$$

I_j = the intensity of the jth observation of reflection i
$\langle I_i \rangle$ = the mean of the intensities of all observations of reflection i
N_i = the redundancy (the number of times reflection i has been measured).

$\sum_i$ is taken over all reflections
$\sum_j$ is taken over all observations of each reflection.

Ref: Diederichs, K. & Karplus, P. A. (1997). Nature Struct. Biol. 4, 269-275.
Weiss, M. S. & Hilgenfeld, R. (1997). J. Appl. Cryst. 30, 203-205.
Weiss, M. S. (2001). J. Appl. Cryst. 34, 130-135.

Recommendations

	Scaled Data		Structure (Å)	
	Overall	Last Shell	Overall	Last Shell
Completeness (%)	> 85	> 50	> 93	> 70
$I/\sigma(I)$		> 1		> 2
R_{merge}		< 1		< 1
R_{rim} or R_{pim}				

What about the Pearson correlation coefficient?

Pearson Correlation Coefficient

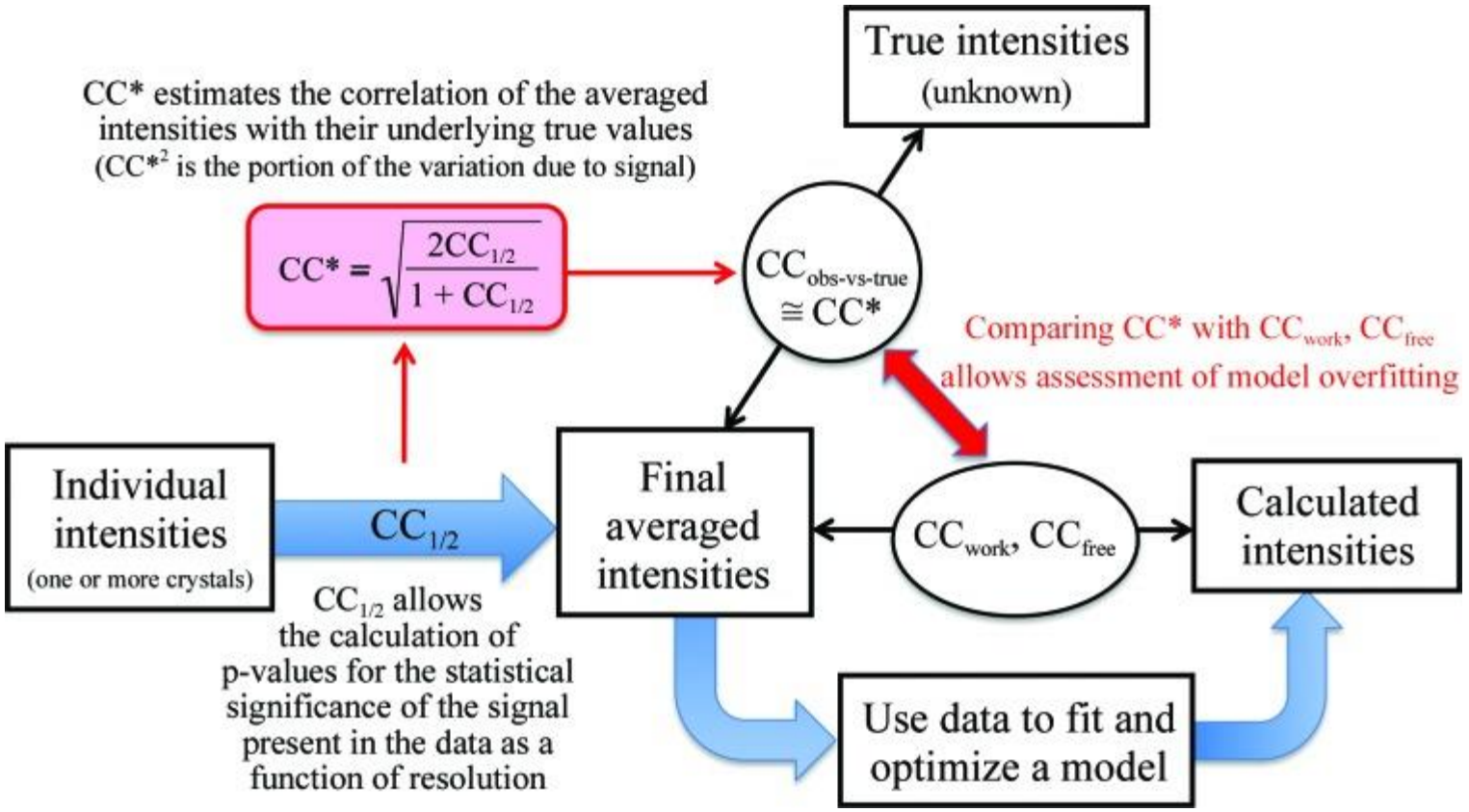
Linking crystallographic model and data quality.

Karplus PA[1], Diederichs K.
[1] Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA.
Science. 2012 May 25;336(6084):1030-3.

Abstract

In macromolecular x-ray crystallography, refinement R values measure the agreement between observed and calculated data. Analogously, R(merge) values reporting on the agreement between multiple measurements of a given reflection are used to assess data quality. Here, we show that despite their widespread use, R(merge) values are poorly suited for determining the high-resolution limit and that current standard protocols discard much useful data. We introduce a statistic that estimates the correlation of an observed data set with the underlying (not measurable) true signal; this quantity, CC*, provides a single statistically valid guide for deciding which data are useful. CC* also can be used to assess model and data quality on the same scale, and this reveals when data quality is limiting model improvement.

$CC_{1/2}$, CC^*, CC_{work}, and CC_{free}



Diederichs, K. & Karplus, P. A. (2013). In *Advancing Methods for Biomolecular Crystallography*, edited by R. Read, A. G. Urzhumtsev & V. Y. Lunin. New York: Springer-Verlag.

Recommendations

	Scaled Data		Structure (Å)	
	Overall	Last Shell	Overall	Last Shell
Completeness (%)	> 85	> 50	> 93	> 70
$I/\sigma(I)$		> 1		> 2
R_{merge}		< 1		< 1
R_{rim} or R_{pim}				
$CC_{1/2}$ or CC*	N/A		N/A	