Assessing data quality

Kay Diederichs

Protein Crystallography /
Molecular Bioinformatics
University of Konstanz, Germany

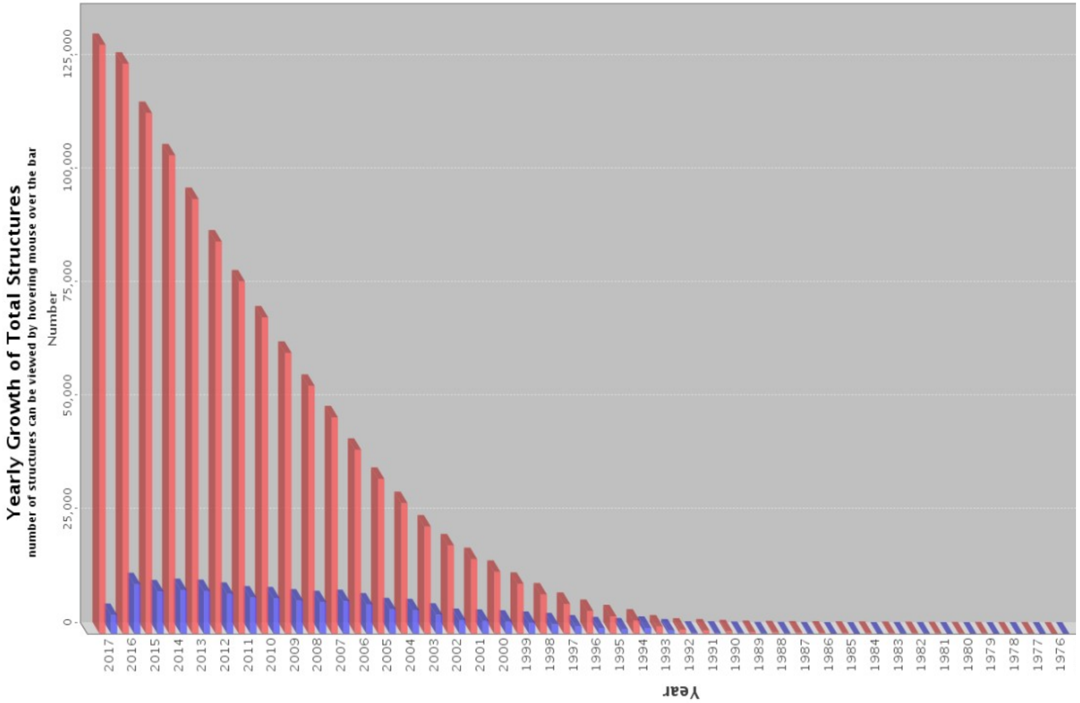
Outline

- 1st example: *meaning* of “quality”
- 2nd example: *measuring* “quality”
- 3rd example: common misunderstandings
- 4th example: resolution

+ practical hint for data processing

Crystallography has been extremely successful

Protein Data Bank : ~135.000 entries



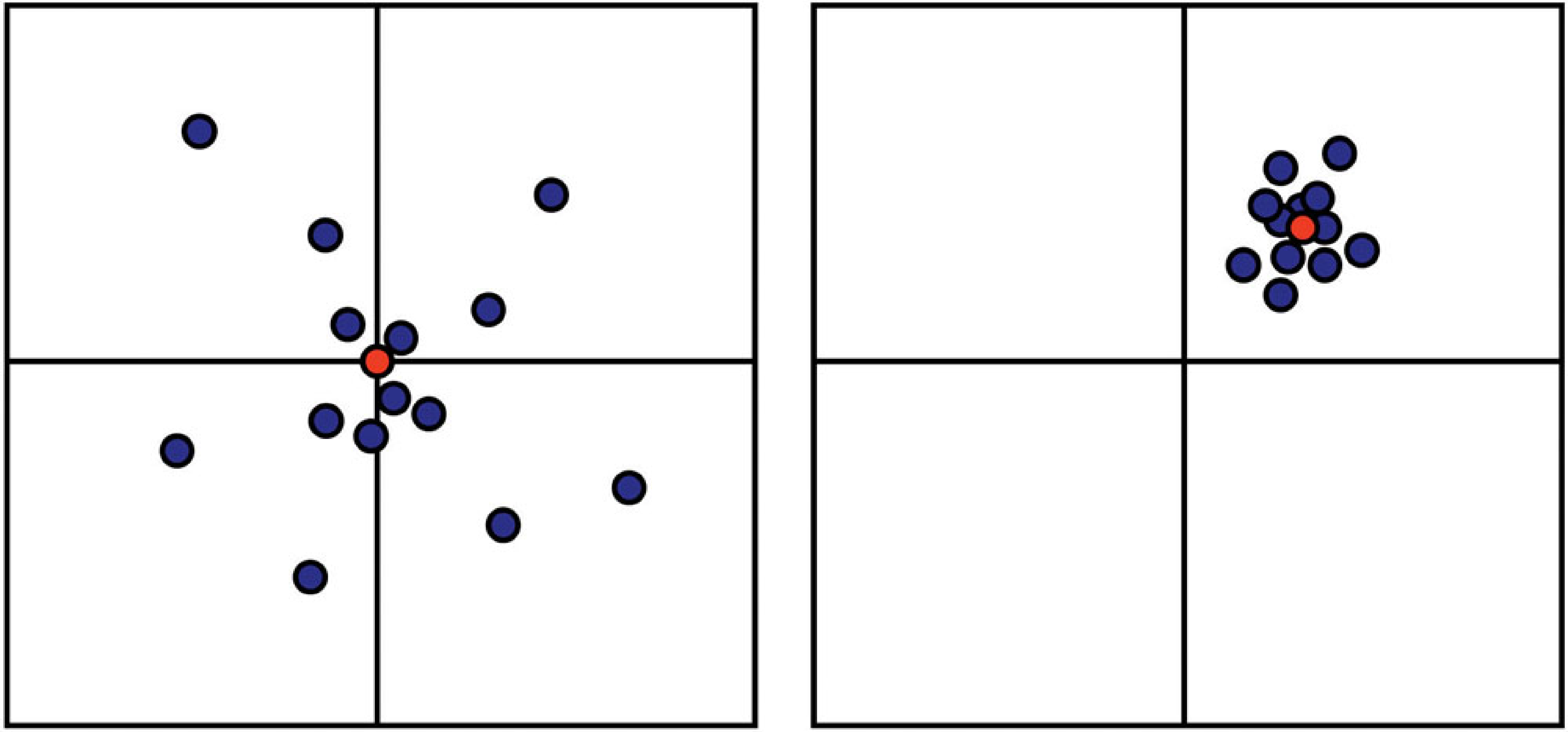
Could it be any better?

Four examples for

- *Rules* that may have been useful in the past under different circumstances, but are still commonly used today and result in wrong decisions
- *Concepts* resulting from first principles that would, if applied, deliver the information to reach the correct decision

1st example: Not understanding the difference between, and the relevance of **precision** and **accuracy**

“Quality”



© Garland Science 2010

B. Rupp, Bio-molecular Crystallography

Accuracy – how different from the *true value*?

Precision – how different are *measurements*?

Numerical example

Repeatedly determine π=3.14... as 3.1, 3.2, 3.0 :
observations have medium precision, medium accuracy

Precision= mean relative absolute deviation from average value=
(0+0.1+0.1)/(3.1+3.2+3.0) = 2.2%

R_{merge} formula!

Accuracy= mean relative absolute deviation from true value:
=(|3.14-3.1| + |3.14-3.2| + |3.14-3.0|)/(3*3.14) = 2.5%

$$R_{merge} = \frac{\sum_{hkl}\sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i(hkl)}$$

Repeatedly determine π=3.14... as 2.70, 2.71, 2.72 :
observations have high precision, low accuracy.

Precision= mean relative absolute deviation from average value=
(0.01+0+0.01)/(2.70+2.71+2.72) = 0.24%

R_{merge} formula!

Accuracy= mean relative absolute deviation from true value=
(|3.14-2.70| + |3.14-2.71| + |3.14-2.72|)/(3*3.14) = 13.7%

Relation of precision, accuracy, and systematic error

- if only **random error** exists, accuracy = precision (on average)
- accuracy and precision differ by the unknown systematic error
- if unknown **systematic error** exists, true value cannot be found from the data themselves
- true values may be known from other approaches (e.g. F_{calc}^2 may be considered an estimate of the true value)
- precision can easily be calculated, but not accuracy

All data quality indicators estimate *precision* (only), but YOU (should) want to know *accuracy*!

Precision *versus* Accuracy

➔ ***Rules***: “The data processing statistics tells me (and the reviewers!) how good my data are.
To satisfy reviewers, the indicators must be good.”

- *Suboptimal result*: these rules encourage
  - overexposure of crystal to lower R_{merge}
  - data collection “strategy” with low multiplicity

➔ ***Concepts***:

- Data processing output reports the *precision* of the data, *not* their accuracy.

- averaging increases accuracy unless the data repeat systematic errors
- rejecting too many data as outliers *increases* the precision, but *decreases* accuracy!

2nd example: confusion by multitude and properties of crystallographic indicators

Confusion – what do these mean?

R_{merge}

$CC_{1/2}$

R_{sym}

Mn(I/sd)

CC_{anom}

I/σ

R_{pim}

R_{meas}

R_{cum}

Calculating the precision of unmerged (individual) intensities

$\langle I_i / \sigma_i \rangle$ σ_i from error propagation & error model

$$R_{merge} = \frac{\sum_{hkl} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

has low-multiplicity bias (U.Arndt 1968)

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

bias-corrected (Diederichs & Karplus 1997)

$R_{meas} \sim 0.8 / \langle I_i / \sigma_i \rangle$ relation between quantities

Averaging ("merging") of intensities from equivalent observations yields improved estimates of intensities of unique reflections

Taking the sigmas as weights,

$$\bar{x} = \frac{\sum_{i=1}^{n} \left(x_i \sigma_i^{-2}\right)}{\sum_{i=1}^{n} \sigma_i^{-2}}, \quad \text{and} \quad \sigma_{\bar{x}} = \sqrt{\frac{1}{\sum_{i=1}^{n} \sigma_i^{-2}}}, \quad \text{with n=multiplicity}$$

(Wikipedia "weighted arithmetic mean")

- If the sigmas are equal, then the merged intensity is just the straight average, and its sigma is reduced by √n from that of the individual observations

These improved "merged" intensities and sigmas are used for
- experimental phasing
- molecular replacement
- refinement

Calculating the precision of merged intensities

a) using the √n law of error propagation (Wikipedia “weighted arithmetic mean”):

<I/σ(I)>

$$R_{pim} = \frac{\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

relation: R_{pim} ~ 0.8 / <I/σ >

b) by comparing averages of two randomly selected half-datasets X,Y:

H,K,L	I_i in order of	Assignment to	Average I of	
	measurement	half-dataset	X	Y
1,2,3	100 110 120 90 80 100	X, X, Y, X, Y, Y	100	100
1,2,4	50 60 45 60	Y X Y X	60	47.5
1,2,5	1000 1050 1100 1200	X Y Y X	1100	1075
...				

- calculate the correlation coefficient $CC_{1/2}$ (Karplus & Diederichs 2012) on X, Y

Measuring the precision of merged data with a correlation coefficient

- Correlation coefficient has clear meaning and well-known statistical properties
- Significance of its value can be assessed by Student's t-test:
  e.g. CC>0.3   is significant at p=0.01 for >100 reflections
  CC>0.08 is significant at p=0.01 for >1000 reflections
- From $CC_{1/2}$, we can analytically estimate **CC of the merged dataset against the true** (usually unmeasurable) **intensities** using

$$CC^* = \sqrt{\frac{2\,CC_{1/2}}{1+CC_{1/2}}}$$

- Under weak assumptions: $CC_{1/2} \sim 1/(1+4/\langle I/\sigma\rangle^2)$

(Karplus and Diederichs (2015) *Current Opinion in Struct.Biol.* **34**, 60-68)

Unmerged *versus* merged

- ***Rule***: “the quality of the data used for refinement can be assessed by $\mathrm{R}_{\mathrm{merge/meas}}$. Data with $\mathrm{R}_{\mathrm{merge/rmeas}}$ > e.g. 60% are useless.”

- Suboptimal result: Wrong indicator - *$R_{merge/rmeas}$ does not predict R_{work}* . Wrong high-resolution cutoff. Wrong data-collection strategy.

Concept: - use an indicator for the precision of the *merged* data if you are interested in the suitability of the data for MR, phasing and refinement.

- Use <I/σ> or <I>/<σ> (but how to calculate σ; and which cutoff??)

- Use $CC^{*}=\sqrt{\frac{2\,CC_{1/2}}{1+CC_{1/2}}}$ if you want to know how high (numerically) $\mathrm{CC}_{\mathrm{work}}$, $\mathrm{CC}_{\mathrm{free}}$ in refinement can become (i.e. how *data quality limits model quality*):
$\mathrm{CC}_{\mathrm{work}}$ larger than CC* implies overfitting, because in that case the model agrees better with the experimental data than the true signal does.

This does not work with R-values because data R-values and model R-values have different definitions!

3rd example: *improper* crystallographic reasoning

apples and oranges

situation: data to 2.0 Å resolution

using all data: $R_{work}=19\%$, $R_{free}=24\%$ (overall)

cut at 2.2 Å resolution: $R_{work}=17\%$, $R_{free}=23\%$

- ***Rule***: “The lower the R-value, the better.”
„cutting at 2.2 Å is better because it gives lower R-values“

- (Potentially) suboptimal result: throwing away data.

- ***Concept***: indicators may only be compared if they refer to the *same* reflections.

Proper crystallographic reasoning

1. A better model has (overall) a lower R_{free}, and a lower R_{free}-R_{work} gap

2. *Comparison* of R_{free} and R_{free}-R_{work} gap-values is only *meaningful* when using the *same* data (sets of reflections)

3. This comparison is done for a model refined with hi-res cutoff, and requires calculation of its R_{free}, R_{work} for the lower-res shells of data.

This is the „*paired refinement technique*“: compare models in terms of their R-values against the *same* data.

P.A. Karplus and K. Diederichs (2012) Linking Crystallographic Data with Model Quality. *Science* **336**, 1030-1033.

Highly controversial?

4th ex.: Resolution of the data

Rules:

1. Worst: cutoff based on $R_{meas}/R_{merge}/R_{sym}$
2. Better: cutoff based on R_{pim} / $<I/\sigma(I)>$ (which value?) merged data
3. Even better: cutoff based on $CC_{1/2}$ (which value?) merged data, no σ

Concepts:

1. “ideally, we would determine the point at which adding the next shell of data is not adding any statistically significant information” (P. Evans)

2. paired refinement method proper comparison
3. only a good model can extract information from weak data external
4. R_{work}/R_{free} of model against *noise* is ~43% (G. Murshudov) validation

Advice: * be generous at the data processing stage
* decide only at the very end of refinement
* deposit the data up to the resolution where $CC_{1/2}$ becomes insignificant!

Highly controversial?

Resolution of the model

Rule:

the resolution of the *model* is the resolution of the data it was refined against

Concepts:

1. the notion “resolution of a model” is misguided – it answers the wrong question!

2. resolution of a *map* is well-defined (Urzhumtsev *et al*): how far are features apart that we can distinguish? depends on Wilson-B

3. better to ask about precision and accuracy of the model
   - precision: reproducibility of coordinates
   - accuracy: which errors are present? much more important!

Summary (of things said until now)

- It is important to understand what the objective of the experiment is. More to the point, what should the “target function” be?
- Accuracy rather than precision
- Merged intensities rather than individual measurements
- More correct model rather than low R values
- Science should be based on logic, not on “what everybody has always been doing”

After all this fundamental (but nevertheless under-appreciated) stuff, here is one practical/specific hint for data processing.

This is relevant not only for XDS but also for DIALS and MOSFLM since Phil Evans recently implemented ISa (which has been in XDS for a long time) in AIMLESS.

AIMLESS's ISa is given at the *beginning* of the logfile.

non-obvious

How do random and systematic *error* depend on the *signal*?

random error obeys *Poisson statistics*
error = square root of signal

Systematic error is *proportional* to signal
error = x * signal   (e.g. x=0.02 ... 0.10)

(which is why James Holton calls it „fractional error“; there are exceptions) non-obvious

Systematic errors (noise)

- beam flicker (instability) in flux or direction
- spindle movement, and/or lack of smooth and accurate rotation
- shutter jitter or lack of synchronization with spinlde
- crystal vibration due to cryo stream
- split reflections, secondary lattice(s), ice
- absorption from crystal and loop
- radiation damage
- detector calibration and inhomogeneity; overload
- shadows on detector
- deadtime in shutterless mode
- imperfect assumptions about the experiment and its geometric parameters in the processing software
- ...

The “error model”

Random error: $\sigma_r(I) \approx \sqrt{I}$

- this is what the integration program calculates

Systematic errors: $\sigma_s(I) \approx I$

- lead to deviations > $\sigma_r(I)$ between sym-related reflections

New $\sigma(I)$ estimate: $\sigma(I) = \sqrt{(a^*(\sigma_r(I)^2+b^*I^2))}$

with constants a,b fitted by scaling program for the dataset

When random error vanishes (“asymptotically”), this results in $I/\sigma(I) = 1/\sqrt{(a^*b)}$

A *proxy* for good data

$(I/sigma)_{asymptotic}$=ISa (reported in CORRECT.LP and AIMLESS logfile) is *a measure of systematic error arising from beamline, crystal, and data processing*.

For a given data set, ISa increases: if the geometric description is improved, and parameters like mosaicity and reflection profiles are correct. In short: when the experimental data are well processed

Maximizing ISa (good values are 30 and higher) *means minimizing systematic errors;*

This usually also optimizes $CC_{1/2}$ *at high resolution*

Thank you for your attention!

References:

Karplus, P.A. and Diederichs, K. (2015) Assessing and maximizing data quality in macromolecular crystallography. *Current Opinion in Struct.Biol.* **34**, 60-68.

Diederichs, K. (2015) Crystallographic data and model quality. in: Nucleic Acids Crystallography (Ed. E. Ennifar), Methods in Molecular Biology **1320**, 147-173.

Karplus, P.A. and K. Diederichs, K. (2012) Linking Crystallographic Data with Model Quality. *Science* **336**, 1030-1033.

(PDFs at https://www.biologie.uni-konstanz.de/diederichs/publications)