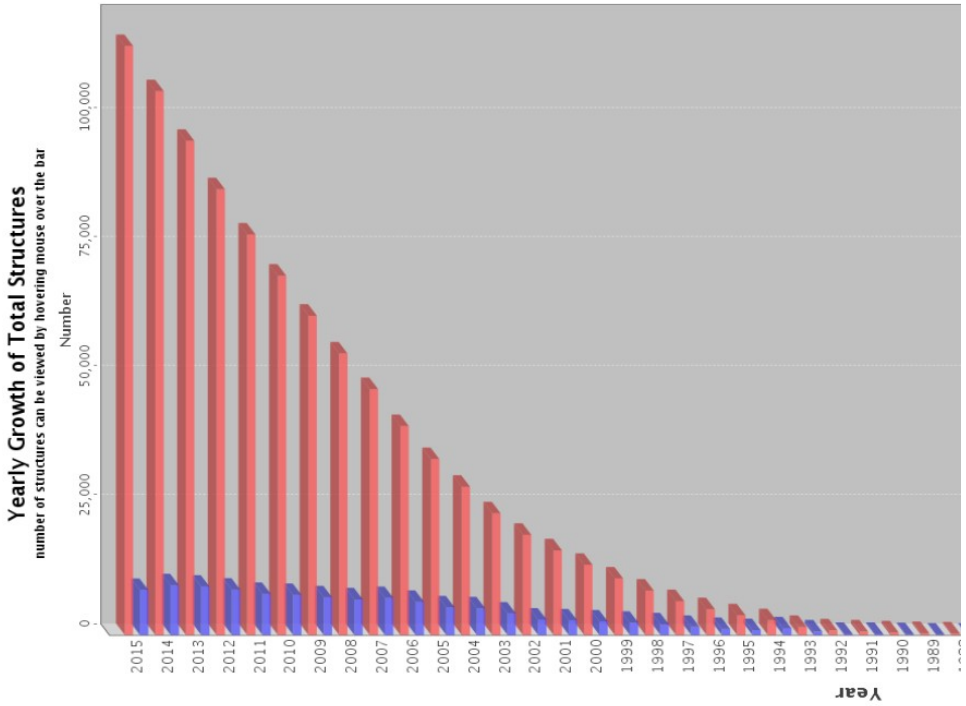
Data quality statistics - and their pitfalls

Kay Diederichs

Protein Crystallography /
Molecular Bioinformatics
University of Konstanz, Germany

Crystallography has been extremely successful

Protein Data Bank on 2016-10-21 : 123.456 entries



Could it be any better?

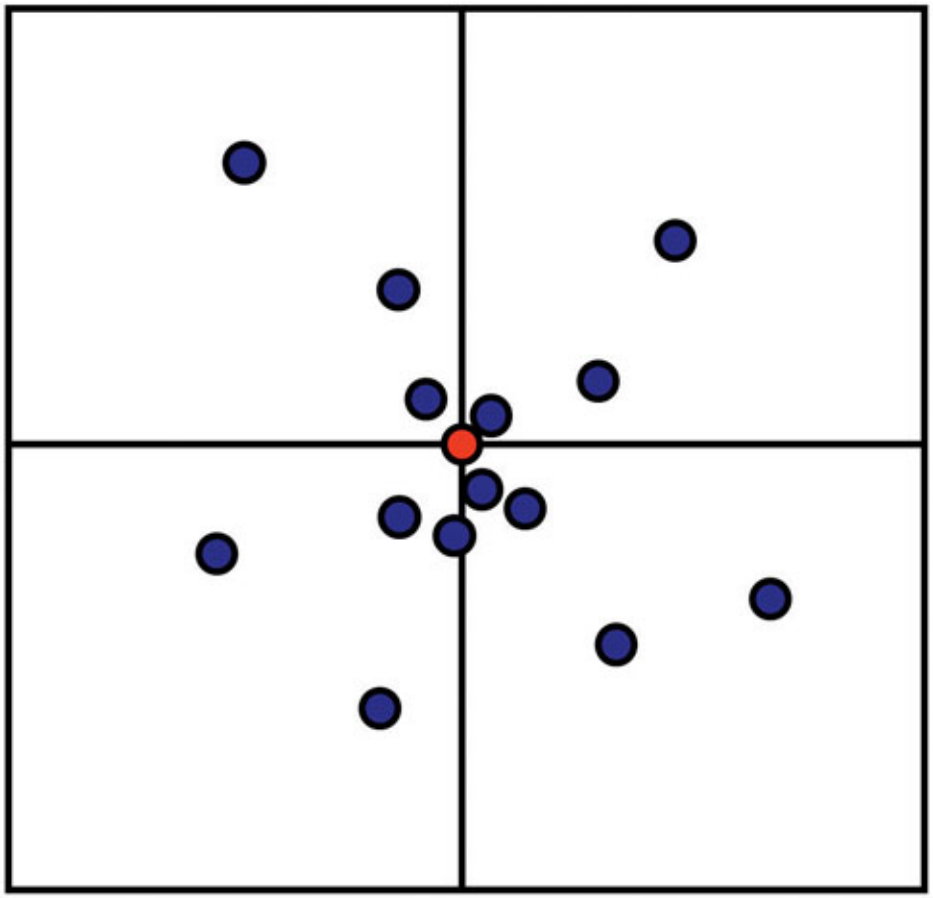
Four examples for

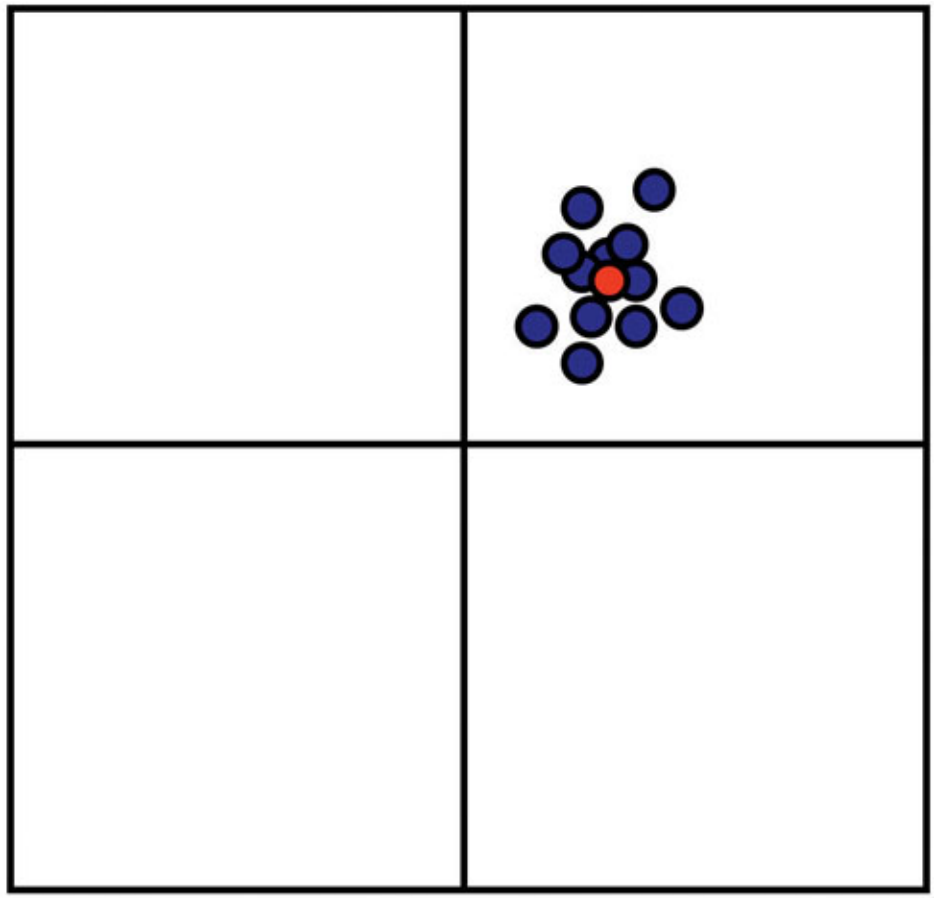
- *Rules* that may have been useful in the past under different circumstances, but are still very commonly used today and which result in wrong decisions
- *Concepts* resulting from first principles that would, if applied, deliver the information to allow the correct decision

1) Accuracy *versus* precision
2) Unmerged *versus* merged indicators
3) Apples *versus* oranges – how to compare
4) Resolution

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

"Quality"





© Garland Science 2010

B. Rupp, Biomolecular Crystallography

Precision – how different are *measurements*?

Accuracy – how different from the *true value*?

Numerical example

Repeatedly determine π=3.14159... as 3.1, 3.2, 3.0 :
observations have low precision, low accuracy

Precision= normalized absolute deviation from average value=
(0.04159+0+0.05841+0.14159)/(3.1+3.2+3.0) = 2.6%

Accuracy= normalized absolute deviation from true value: (3.14159 - 3.1)/3.14159 = 1.3%

R_{merge} formula!

Repeatedly determine π=3.14159... as 2.718, 2.716, 2.720 :
observations have high precision, low accuracy.

Precision= normalized absolute deviation from average value=
(0.002+0+0.002)/(2.718+2.716+2.720) = 0.049%

Accuracy= normalized absolute deviation from true value=
(3.14159-2.718) / 3.14159 = 13.5%

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

What is the “true value“?

- if only **random error** exists, accuracy = precision (on average)
- if unknown **systematic error** exists, true value cannot be found from the data themselves
- consequence: precision can easily be calculated, but not accuracy
- accuracy and precision differ by the unknown systematic error

All data quality indicators estimate *precision* (only), but YOU (should) want to know *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
  - data massaging: rejecting many "outliers", throwing away negative or weak data

➔ ***Concepts***:

- Data processing logfiles report the *precision* of the data, *not* their accuracy.
- averaging increases accuracy unless the data repeat systematic errors
- outliers may be correctly or incorrectly identified. Rejecting too many may *increase* the precision, but *decreases* accuracy!
- accuracy shows by good agreement with intensities from a different origin (model)!

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_{meas}

R_{pim}

R_{cum}

Calculating the precision of unmerged (individual) observations

$\langle I_i/\sigma_i \rangle$ (σ_i from error propagation, i=individual)

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

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

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

Calculating the precision of merged data

using the √n law of error propagation (Wikipedia "weighted arithmetic mean"):

<I/σ(I)>

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

$R_{pim} \sim 0.8 / \langle I/\sigma \rangle$

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

H,K,L	I_i in order of measurement	Assignment to half-dataset	Average I of 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 R-factor (D&K1997) or correlation coefficient $CC_{1/2}$ (K&D 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 n>100; CC>0.08 is significant at p=0.01 for n>1000)
- Apply this idea to crystallographic intensity data: use "random half-datasets" → $CC_{1/2}$ (called CC_Imean by SCALA/aimless, now $CC_{1/2}$)
- From $CC_{1/2}$, we can analytically estimate **CC of the merged dataset against the true** (usually unmeasurable) **intensities** using

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

- (Karplus and Diederichs (2012) *Science* **336**, 1030)

- ***Rule***: “the quality of the data that I use for refinement can be assessed by R_{merge}/R_{meas} . Data with R_{merge}/R_{meas} > e.g. 60% are useless.”

- Suboptimal result: Wrong indicator. 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 an indicator for the precision of *unmerged* data for special purposes like spacegroup determination, and a radiation damage estimate.

- Use $CC^* = \sqrt{\frac{2\,CC_{1/2}}{1+CC_{1/2}}}$ if you want to know how high (numerically) CC_{work} , CC_{free} in refinement can become (i.e. how *data quality limits model quality*).

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

3rd example: *improper* crystallographic reasoning

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

…. requires three concepts:

1. Better data allow to obtain a better model
2. A better model has a lower R_{free}, and a lower R_{free}-R_{work} gap
3. *Comparison* of model R-values is only *meaningful* when using the *same* data

Taking these together, this leads us to 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_{sym} (which value?)
2. Better: cutoff based on $\langle I/\sigma(I) \rangle$ (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 ~42% (Evans&Murshudov) validation

Advice: be generous at the data processing stage, and
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* (Urzhumtsev *et al*) is well-defined: 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

- Crystallographic decisions are often based on *rules* of (if anything) only historical interest. These rules frequently lead to *improper shortcuts* being taken
- “make everything as simple as possible, but not simpler” (attributed to A. Einstein)
- Rules may be needed in expert systems; however, humans should rather learn, apply and further develop the underlying *concepts*
- Change the way we think (and teach)
- Crystallography is a Science, not just “applied technology”

Thank you for your attention!

Three recent references:

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

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

Assmann, G., Brehm, W. and Diederichs, K. (2016) Identification of rogue datasets in serial crystallography (2016) *J. Appl. Cryst.* **49**, 1021-1028.

(PDFs at http://cms.uni-konstanz.de/strucbio/diederichs-group/publications)