Comment on "Data for Refinement and Deposition/Publication"

Mariusz Jaskolski (July 13, 2014)

1. My main motto is that each and every one reflection with meaningfully measured intensity is precious and should be included in the refinement. Therefore, I do not truncate my data only because the high-resolution shell is incomplete. I use all reflections (e.g. from the corners of square detectors) that satisfy other criteria (e.g. $\langle I/\sigma \rangle$ or $CC_{1/2}$) but in Table 1 of my publication I try to estimate the resolution limit realistically.

2. If the high-resolution shell is incomplete, I base my "effective resolution" estimate on the radius of a reciprocal-lattice sphere that would be filled completely with all reflections used in the refinement.

3. My leniency toward completeness at high-resolution does not extend to overall completeness. A data set should be overall complete (>85%) to avoid systematic errors. Poor overall completeness is usually systematic (missing indices) and leads to systematic errors in the model, e.g. ill-determined atomic coordinates in one direction. Poor completeness at low resolution may also hamper structure solution.

4. My main criterion for realistic high-resolution limit is $\langle I/\sigma \rangle=2$, although with proper statistical error model (in maximum-likelihood methods) the threshold is often lowered (even to 1) because inclusion of such reflections (if properly weighted) will cause no harm.

5. "No harm" is probably not a very scientific justification. Instead, we should ask, after Phil Evans, at what resolution inclusion of more reflections still contributes new information.

6. The more liberal approaches tend to use $CC_{1/2}$ correlation instead of the criticized $\langle I/\sigma \rangle$.

7. Concentration on high resolution is justified from the point of view of structure refinement, although even in this case, systematic omission of low-resolution data will introduce systematic model errors (e.g. in solvent model). Omission of low-resolution data can have fatal consequences for structure determination. Therefore, the practice (now disappearing) of low-resolution cut-offs should be discouraged.

8. R_{merge} is not a good criterion. First of all, R_{pim} should be used to eliminate the influence of data redundancy (which in itself is a desired factor). I agree that overall $R_{merge}>0.15$ looks very bad and does not bode well for a stellar structure.

9. I do not pay too much attention to R_{merge} at high resolution, as other criteria are more important. For aesthetical reasons, I do not accept $R_{merge}>0.99$.

10. **And now a general comment about the proposed procedure.** 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.