Patterson positivity combined with statistical matching can estimate unobserved intensities

A. Kadziola

Department of Chemistry, University of Copenhagen, Denmark

kadziola@chem.ku.dk

Many macromolecular data sets suffer from being more or less incomplete mainly as a result of experimental difficulties. Often axes are missing which should be inspected for systematic absences. Even data sets considered complete usually miss very low resolution reflections due to beam stop issues. Low resolution reflections are important as they to a large extent define the protein/solvent boundary. Seriously incomplete data sets can hamper many crystallographic calculations.

A direct space constraint on intensities is the positivity of the Patterson map. All intensities should also make statistical sense and conform to distributions based on observed intensities. Statistics include histograms of normalized and full intensities and scaling as a function of resolution or intensity. Here it is demonstrated how Patterson positivity combined with statistical matching can estimate unobserved intensities. Among applications are space group determination from observation of systematic absences on missing axes and classical rotation functions for molecular replacement. Also, any *ab initio* phasing procedure based on intensities alone is expected to benefit from a complete data set.

The calculations towards a complete data set consist of flipping negative values in the Patterson map followed by histogram match and scaling of the back transformed intensities in order to conform to observed intensities. The generated intensities for observed reflections are, in turn, flipped relative to the true observations while the unobserved reflections are kept as is. The procedure is initiated by fitting the observed intensities as a function of resolution and determine $F(000)$ using knowledge of the solvent content. Initially, fitted intensities are substituted in for unobserved data. The calculations are iterative gradually reducing the flipping factor in direct as well as reciprocal space. For cross validation a free data set with 5 % of the observed intensities are kept aside and treated as unobserved.

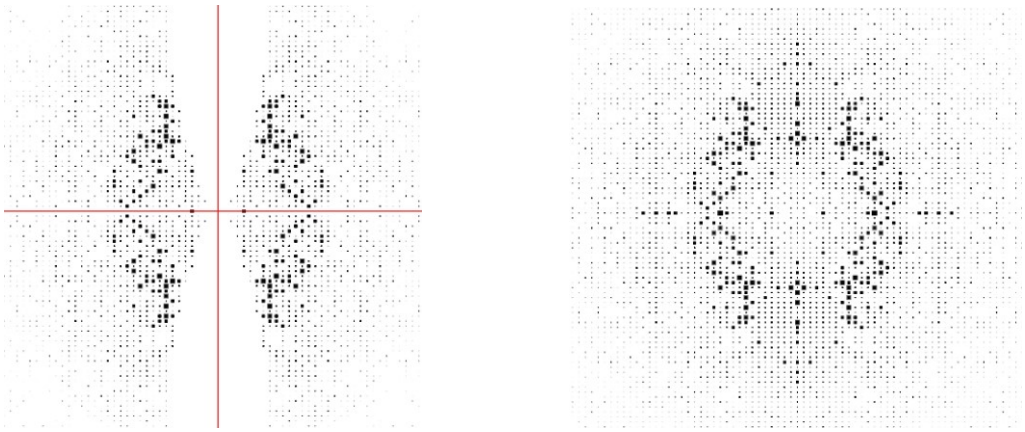

Figure 1. Structure factor amplitudes of *Aspergillus aculeatus* rhamnogalacturonan acetylesterase in space group $P2_12_12_1$

Left: 95.45 % complete data set, *0kl*-section. Right: Data set completed by Patterson positivity and statistical matching. Systematic absenses on missing *l*-axis are clearly visible.

As a further test of the procedure a virtually complete data set is subjected to various omissions. Omissions include: Low resolution cut off (beam stop issues), high resolution cut off (detector misplaced too far), thin shells of resolution (ice rings), the 10 % strongest intensities (overloads) and increasing omissions around axes and planes in reciprocal space.

In the order of minutes a completed data set can be produced with estimates of unobserved intensities along with estimated standard deviations based on how well the free intensities are reproduced by following the ideas of Read (1986) [1].

[1] Read, R. (1986). *Acta Cryst.* A**42**, 140-149

Keywords: incomplete data; Patterson positivity; histogram; scaling

Acknowledgement: The author which to thank the CCP4 community for their open source policy.