MS29-02 R as a tool to check data quality in the context of low resolution crystallography.

Manfred W. Baumstark,[a]
[a]*University Hospital, 79106 Freiburg, Germany*
E-mail: maba@uni-freiburg.de

When working with low resolution data the usual crystallographic software often reaches its limits. Reasons are the low number of reflections, a very high dynamic range of refection intensities, and the violation of assumptions that are generally valid for higher resolution data, i.e. Wilson statistics etc. Therefore one has to be very careful when using standard crystallographic tools and it turned out to be necessary to carefully check the individual steps of data processing. R [1] is a programming language and environment that offers easy to use high level functions for data manipulation, plotting and statistical testing. Plotting possibilities include displaying images and 3D visualisation using openGL. R is fully scriptable and therefore allows a high level of automation. A subset of R functions is supported by a user friendly GUI [2]. Examples where R proved to be useful in our work include the analysis of CCD detector images, procedures to analyse XDS [3] output files (INTEGRATE.HKL, XDS_ASCII.HKL) and the production of various plots to visualize the signals present in the measured data. An automated analysis of XDS output files allowed us to find optimal parameters for integration and scaling of low resolution data sets. Noteworthy, within the R environment non-parametric tests are readily available. Especially (low resolution) intensities are rarely normally distributed and non-parametric tests should be preferred over the parametric equivalent (e.g. Spearman vs. Pearson correlation). An interesting perspective would be to extend the R environment with a package that implements concepts and methods specific to crystallography.

[1] R Development Core Team (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* ISBN 3-900051-07-0, http://www.R-project.org/.

[2] Fox, J. (2005). *Journal of Statistical Software*, **14(9),** 1–42.

[3] Kabsch, W. (2010). *Acta Cryst.* D**66**, 125-132.

Keywords: statistical methods; statistics program R; data quality

MS29-03 Implementation of a B-factor validation protocol for macromolecular structures.

Fabio Dall'Antonia,[a] Jacopo Negroni[a], Garib N. Murshudov[b] & Thomas R. Schneider[a] [a]*European Molecular Biology Laboratory, Hamburg Outstation, Notkestraße 85, 22603 Hamburg, Germany.* [b]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, United Kingdom.*
E-mail: fabio.dallantonia@embl-hamburg.de

In macromolecular crystallography, the diffraction-component precision index (DPI) is calculated from several quantitative accuracy criteria of a structure, such as R-free value and resolution.[1] Atomic coordinate errors on a global scale can be estimated by means of scaling isotropic temperature (B-) factors by the ratio of DPI and average B, thus making structures with different experimental parameters comparable. DPI-based coordinate error estimates are used in the program Escet[2]. However, the B-factors themselves are the result of the refinement procedure and may, in addition to being pure displacement parameters related to thermal motion and conformational flexibility, reflect systematic errors in the model. A validation of observed B-factors against a reference distribution would help to detect refinement problems and support the development of a more general coordinate error estimation method.

As a result of Bayesian inference on data from high-resolution structures, the inverse gamma distribution (IGD) has been proposed as natural distribution for B-factors[3], as it is the conjugate prior of a Gaussian distribution with unknown variance. We have developed a validation protocol for protein and/or nucleic acid structures, assuming a shifted IGD (SIGD). Our procedure consists of a maximum-likelihood estimation of SIGD parameters, embedding the log-likelihood target function into the L-BFGS-B optimization algorithm[4], and a bootstrapped Kolmogorov-Smirnov (KS-) test comparing a large set of variate-sampled model distribution instances with the empirical B-factor data. Numerical as well as graphical validation output is provided. The protocol was first implemented as a script for the statistical R package and was recently ported to C++, using the CCTBX library[5]. Details on the technicalities of the C++ implementation are given as well as concrete validation examples. An attempt to classify cases of macromolecular structure refinement by SIGD parameters and KS-test p-value is presented.

[1] Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583-601.

[2] Schneider, T. R. (2004). *Acta Cryst.* D**60**, 2269-2275.

[3] Dauter, Z., Murshudov, G. N. & Wilson, K. S. (2006). *International Tables for Crystallography*, vol. F, edited by M. G. Rossmann & E. Arnold, pp. 393-402. 1st online edition, Chester, UK: International Union of Crystallography.

[4] Zhu, C., Byrd, R. H., Lu, P. & Nocedal, J. (1997). *ACM Trans. Math. Software* **23**, 550-560.

[5] Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126-136.

Keywords: Temparature factor; refinement; validation