

T5. Two examples of chemometrics application in protein crystallography

Andrey Bogomolov¹, Gleb Bourenkov², and Alexander Popov¹

¹ *European Molecular Biology Laboratory (EMBL), Hamburg, Germany*

² *Max-Planck-Arbeitsgruppen für Strukturelle Molekularbiologie, Hamburg, Germany*

X-ray crystallography is the most powerful experimental technique for the determination of three-dimensional structures of biological macromolecules. The way from an expressed protein to the solved 3D structure is a complex multi-stage process, including math-intensive data processing and modeling. Multivariate methods and approaches can be successfully applied to solving various tasks related to protein crystallization, crystallographic data acquisition, and subsequent structure solution.

Two application examples of Partial Least Squares (PLS) regression are presented. In the first example, a quantitative relationship has been established between the curve of spherically averaged diffraction data (Wilson plot) and the protein secondary structure. The resulting PLS model built on a representative data set of over 500 protein structures taken from the Protein Data Bank (PDB) can be used for the prediction of the fraction of two main secondary structure elements, α -helix and β -sheet, from the x-ray data of an unknown protein.

In another study, Principal Component Analysis (PCA) and PLS regression were applied to reveal the fundamental principles of the phenomenon known as crystal radiation damage and find a quantitative model for its characterization.