MS 29 P01

Classifying Molecular Geometries: Application of Factor Analysis to Cluster Formation in *d*SNAP.

D. Sneddon, C.J. Gilmore, G Barr, W. Dong, A. Parkin, and C.C. Wilson, *WestChem. Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland.*

Keywords: Cluster analysis, Visualisation techniques, Data mining

Cluster analysis is a well-established tool in statistics, but one that is used surprisingly little in crystallography. We have established its use in analysing the results of database searches using the Cambridge Structural Database (CSD) [1].

CSD searches can produce thousands of 'hits' especially if a simple fragment is used as the search fragment. As a result processing and interpreting the results can be a laborious task and one where mistakes are easy to make. Cluster analysis, involving dendrograms, metric multidimensional scaling and other visualisation tools, can reduce the workload to a few hours in the hand of an experienced user. *d*SNAP [2] is a cluster analysis program that provides these tools.

Factor analysis and other statistical methods are being used, in combination with visualisation techniques, including biplots, to uncover the latent underlying structural properties of the molecular fragments under investigation. This process is being undertaken in concert with investigations into many of the possible descriptions of the geometries of the fragments under study.

Details of this two-pronged approach will be presented, along with its utility as a method to uncover the underlying reason behind the formations of clusters in *d*SNAP. The development of this additional feature within our suite of automated methods for interpreting CSD data should substantially ease the interpretation of the results of cluster analysis.

The application of the factor analysis methods being developed will be discussed, and illustrated with various examples, including rings and hydrocarbon chain fragments.

The *d*SNAP software is available free of charge to all interested researchers, distributed through Bruker-AXS [3].

[1] A, Parkin, G. Barr, W. Dong, C.J. Gilmore and, C.C. Wilson. *CrystEngComm,* 2006, **8,** 257-264

[2] G. Barr, W. Dong, C. J. Gilmore, A. Parkin and C. C. Wilson,. *J. Appl. Cryst*,. 2005, **38**, 833-841

[3] http://www.chem.gla.ac.uk/snap

MS29 P02

Searching the Cambridge Structural Database for the \"best\" representative of each unique polymorph:

Jacco van de Streek, *Institute for Inorganic and Analytical Chemistry, Frankfurt University, Frankfurt am Main, Germany.* E-mail: jacco@chemie.uni-frankfurt.de

Keywords: Cambridge Structural Database, polymorphs

The Cambridge Structural Database (CSD, [1]) is a database containing virtually all organic and metal-organic crystal structures ever published. Because the CSD aims to cover the literature as completely as possible, it cannot be avoided that some crystal structures of questionable quality are incorporated, as are multiple publications of the same crystal structure (redeterminations). The presence of suspicious and duplicate crystal structures could lead to possible outliers when using the CSD for statistical analyses, and a list of unique, basically correct crystal structures would be desirable. When removing duplicate crystal structures, however, care needs to be taken that genuine polymorphs are correctly distinguished from redeterminations and are retained in the list.

This talk is based on a recently published paper[2] describing a computer program that was written to analyse the CSD to produce a list of the best representatives of each unique polymorph. Because of the large number of crystal structures in the CSD—over 400,000—the program was designed to be fully automatic.

The program operates in three stages:

In the first stage, suspicious crystal structures are detected and eliminated based on 14 quality criteria.

In the second stage, the remaining crystal structures are clustered as either polymorphs or redeterminations by calculating the similarities of their simulated powder diffraction patterns using a normalised weighted cross-correlation function[3].

In the third stage, the best representative from every set of redeterminations is selected on the basis of four different sets of criteria.

The results, 243,355 well-determined crystal structures grouped by unique polymorph, are presented, validated and analysed.

[1] Allen, F.H., *Acta Cryst.*, 2002, B58, 380.

[2] Van de Streek, J., *Acta Cryst.*, 2006, B62, 567.

[3] Gelder, R. de, Wehrens, R. & Hageman, J.A., *J. Comput. Chem.*, 2001, 22, 273.