Error Estimation and Pattern Recognition Techniques in Protein Crystallography
Zwart, P.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Appendix III

III-1 Sphere and hypersphere point picking

Generating point uniformly on a sphere with a unit radius can be carried out by a method developed by Marsaglia (1972) that consists of sampling 2 random numbers, \( A \) and \( B \), distributed independently and uniformly on \((-1, 1)\). Pairs of \((A, B)\) for which \(A^2 + B^2 < 1\), can be used to construct a vector \((x, y, z)\) that is distributed uniformly on a sphere:

\[
\begin{align*}
x &= 2A \sqrt{1 - A^2 - B^2} \\
y &= 2B \sqrt{1 - A^2 - B^2} \\
z &= 1 - 2(A^2 + B^2)
\end{align*}
\]  

(III-5) \hspace{1cm} (III-6) \hspace{1cm} (III-7) \hspace{1cm} (III-8)

Sampling points \((a_0, a_1, a_2, a_3)\) on a 4 dimensional sphere with radius 1, can be carried out in a similar way. Four random numbers \((A, B, C, D)\) are drawn independently from a uniform distribution on \((-1, 1)\). Random numbers for which the pairs \((A, B)\) and \((C, D)\) satisfy \(A^2 + B^2 < 1\) and \(C^2 + D^2 < 1\), are used in the following transformation:

\[
\begin{align*}
a_0 &= A \\
a_1 &= B \\
a_2 &= C \sqrt{\frac{1 - A^2 - B^2}{C^2 + D^2}} \\
a_3 &= D \sqrt{\frac{1 - A^2 - B^2}{C^2 + D^2}}
\end{align*}
\]  

(III-9) \hspace{1cm} (III-10) \hspace{1cm} (III-11) \hspace{1cm} (III-12) \hspace{1cm} (III-13)

The vector \((a_0, a_1, a_2, a_3)\) is then uniformly distributed on a 4 dimensional sphere with radius 1. This vector can be considered as a quaternion

\[ q = a_0 + a_1 i + a_2 j + a_3 k \]  

(III-14)

and can be used to reorient a molecular fragment (Weisstein, 1999). More efficient sampling methods based on the correspondence of the rotation group \(SO(3)\) and a 4-dimensional sphere are described by Bricogne (1997c).
Recommendations for further research

The non-central Maxwell distribution and/or the nearest neighbour distance distribution presented in the second chapter of this thesis can be used as a versatile tool in modelling the distribution of distances between atoms. Possible future applications lie in the use of detecting similarities between subsets of atoms, such as initial searches for non- or cross-crystallographic symmetry or in substructure comparisons. Currently available methods seem to lack a reasonable way of incorporating positional error estimates. A probabilistic treatment might be fruitful.

The practical use of $2\omega$ as a classifier for the behaviour of X-ray data sets needs to be thoroughly tested; initial implementations are being carried out. Its use in constructing a more elaborate scoring function for the contribution of electron density during the proposed ligand building procedure is being considered and preliminary results show that histograms of difference density can be modelled on the basis of a function of $2\omega$. Its ease of computation and reasonably objective measure of the properties of the electron density of a protein with given characteristics of the X-ray data set seems obvious, but awaits the response of the crystallographic community to 'yet another resolution measure'.

The use of the Wilson plot to estimate parameters of the error model might be enhanced by a maximum likelihood target function rather than a least squares function. Taking into account an error model of the atomic displacement parameters could enhance the overall performance, especially when linked to existing $\sigma_A$ estimation methods. Taking into account the effects of restraints in the analysis would result in a more generally applicable of the proposed estimation procedure.

The model building method presented in the fourth chapter are applicable to any structure but is currently limited by the speed of the search algorithm. The time gained by an improved search algorithm can be used to search for larger fragments which could be easier detectable. Efficient ways of handling partial disorder and specific geometric properties of the search fragment in the search strategy is challenging and likely to enlarge the applicability of the method.