POINTLESS: a program to find symmetry from unmerged intensities

Phil Evans, MRC Laboratory of Molecular Biology, Hills Road, Cambridge
pre@mrc-lmb.cam.ac.uk

POINTLESS reads one or more files containing unmerged intensities, and prepares them for scaling (SCALA) in various ways. It accepts files produced by MOSFLM or COMBAT in MTZ format, and can also read unmerged files from XDS (XDS_ASCII) and unmerged SCALEPACK output.

Its principal functions are:-

- determining the Laue (Patterson) group from the observed symmetry, and making a guess at the space group from systematic absences.
- in cases where there are alternative indexing schemes, making the indexing consistent with a reference dataset.

If multiple files are input, it will force the batch numbers to be unique (a long-standing irritation), and ensure consistent indexing in cases where there are alternatives. Dataset names may be edited. It can thus be used to prepare files for input to Scala, as a simpler alternative to SORTMTZ.

An early version of the program was described in Evans (2006), but more recent versions have additional options and some different scoring schemes. The program uses the clever facilities of Ralf Grosse-Kunstleve's cctbx library (Grosse-Kunstleve *etal.*2002) for its symmetry handling, as well as Clipper (Cowtan 2003) and CCP4libraries.

Determination of Laue group and space group

The true spacegroup cannot be definitely known until the structure has been solved and satisfactorily refined, since it is easy to be misled by pseudo-symmetry or twinning. When the diffraction images are indexed, a lattice & crystal class is chosen based only on the cell dimensions, which is not a reliable guide to the true symmetry (it is perfectly possible for example to have a orthorhombic cell with *a* nearly equal *b*). However, by examining how well intensities which are related by potential symmetry agree with each other, we can make an estimate of the likelihood of possible Laue groups. The Laue group has the symmetry of the diffraction pattern, and is the rotational par tof the space group plus a centre of symmetry (Friedel symmetry) and any lattice centering (this is essentially the same as the Patterson group). For chiral space groups, ie for all macromolecules, there is only one point group corresponding to each Laue group (POINTLESS is general for all space groups, including non-chiral and centrosymmetric ones, but by default will choose only chiral groups.) The hierarchy of symmetry determination is: crystal class(lattice symmetry); Laue group (diffraction symmetry); point group; space group(including translational symmetry such as screw axes).

Normalisation. In order to do any scoring of agreement between intensities, we need to put the reflections at least approximately on the same scale. Proper scaling requires knowledge of the symmetry, which is what we are trying to determine, but we can do a rough job by normalising intensities to E^2, by making $\langle E^2 \rangle = 1$ over all resolution ranges. This done by fitting a B-factor to resolution bin averages, then smoothing the residual deviations from $\langle E^2 \rangle = 1$for each resolution-bin with a spline function. Since radiation damage generally increases the apparent B-factor, a simple-minded correction is made by fitting a B-factor as a linear function of "time" (at present using the batch or image number as a proxy for time): this makes a substantial improvement in the scores in cases with marked radiation damage. The scale& B-factors are determined separately for each "run" (ie sweep of contiguous images.)

Scoring functions. The main scoring function used in matching intensities is the correlation coefficient (CC), since it is relatively insensitive to the unknown scale. Probability estimates are then derived from the correlation coefficients. Multiplicity-weighted R-factors are calculated, but these are sensitive to the unknown scales, so are not used in ranking. Correlation coefficients do assume that the data all arise from the same distribution, which is why the raw intensities need to be normalised to E^2, otherwise a correlation will be observed just from the variation of $\langle I \rangle$ in particular with resolution. This can be seen by noting that the correlation coefficient between eg x_i & y_i is just the least-squares slope of plot o fall the (x_i, y_i) points: since $\langle I \rangle$ is larger at low than at high resolution, if we plot pairs of potential but not symmetry-related intensities, which are necessarily at the same resolution, then we will see an apparent correlation due to eg a strong low resolution intensity matching another intensity which is likely to be strong, and so on.

Symmetry elements. The highest possible lattice symmetry compatible with the cell dimensions, within quite generous limits (by default 2° and the equivalent on lengths), is chosen as the test symmetry, ignoring the symmetry in the input file. Each rotation axis in the lattice group is scored separately using all pairs of observations related by that rotation. A probability of the axis being present is estimated from the correlation coefficient, using an error estimate σ (CC) derived from the distribution of correlation coefficients between unrelated pairs, proportional to 1/v(Npairs): this allows for a larger uncertainty if we only have a small sample of reflections. The distribution is modelled as a Lorentzian function, centred on an expectation or "ideal" value estimated as an average of (i) the CC for the identity or Friedel operator and (ii) an estimate of $\boldsymbol{E}$(CC) allowing for the observed error estimates($\boldsymbol{E}(CC) = Var(E^2)/(Var(E^2) + \langle \sigma^2(E^2) \rangle$)(Read, personal communication). To allow for the possibility of pseudo-symmetry, the expected value of CC if the symmetry is not present is not assumed to be 0, but is modelled as a declining probability from P(CC|no symmetry) = 1 at CC=0 to P(CC) = 0 at CC = $\boldsymbol{E}$ (CC | symmetry present) and integrated out. Normalising this probability, P(symmetry | CC) = P(CC|symmetry)/[P(CC|symmetry)+ P(CC|no symmetry)] gives a reasonably robust scoring of the likelihood of each possible symmetry element, without too much danger of over-confidence from an accidental high score with a very few observations. This means that it is often possible to get a reasonable estimate of the Laue group even from a small wedge of data.

Laue groups. The list of possible Laue groups, sub-groups of the lattice group down to the minimum P-1, can be generated from all possible pairs of symmetry elements, including the identity. An estimate of the likelihood of each group is calculated using the combination of the probabilities of each symmetry element which is either present or absent in each sub-group. In each sub-group, each potential symmetry element *i* is either present e_i = true, or absent e_i = false, and we have a measured CC_i and $P(CC_i \mid e_i)$ for e_i= true or false. Then for each group, $P(CC \mid group) = \Pi_i P(CC_i \mid e_i)$. This probability is used to rank the possible

Laue groups. Various other scores are also listed for each sub-group: correlation coefficient, a "net Z-score" from the CCs, R-factors and a measure of the lattice distortion from the original unit cell (in cases where the test lattice is higher symmetry than the original assignment). If the crystal class is different from that used in the integration, you should reprocess with the correct symmetry, ie with the correct cell constraints.

Systematic absences & space groups. These arise from translational symmetry operators, notably screw axes which lead to absences on axial reflections (in non-chiral crystals, glide planes lead to absences in 2-dimensional zones). They can thus be used to distinguish between different space groups within a chosen pointgroup. However, they are not always a reliable guide to the true space group, because there are relatively few axial reflections, and axes lying close to the spindle rotation axis may be only partly sampled or missing from the dataset altogether, so the information from the absences should be treated with caution. POINTLESS uses a Fourier analysis of I / σ values to estimate the probability of the translational element being present or not. For example, if the chosen point group might have a 2_1 screw dyad along the a axis, this would be indicated by presence of h00 reflections only when h is even(=2n). Then the one-dimensional Fourier transform of I / σ (or I) should peak at 1/2 in Fourier space, and the peak height at 1/2 relative to the origin is a measure of the strength of the screw. A probability of the presence of the screw is then calculated, using an error estimate derived from samples of the same number of observations of non-axial reflections, and again a Lorentzian distribution centred on the ideal value of 1. A similar analysis applies to 3-fold screws, and to glide planes, but 4-fold and 6-fold screws are more complicated, since there are multiple possible Fourier peaks, at 1/4 & 1/2 for a 4-fold, or 1/6, 1/3 & 1/2 for a 6-fold, and these are not independent. These can be treated by using a distribution based on a single deviation from all the 2 (4-fold) or 3 (6-fold) ideal peak values, considering the deviation as a "distance" in 2 or 3 dimensions.

In many cases, combining the probabilities from the rotational symmetry and from the systematic absences gives a unique choice of space group, but often several different space groups may need to be tried in the structure determination. POINTLESS tries to avoid over-confidence in its assignment of likelihood, which works in most cases, but it is occasionally fooled by close pseudo-symmetry.

Alternative indexing schemes

If the Laue symmetry is lower than the lattice symmetry, there are alternative indexing schemes which are different but equally valid, related by the rotational symmetry operators present in the lattice but not in the Laue group. These are the same conditions which allow merohedral twinning. For example, in Laue group P-3 (point group P3) there are four possible indexing schemes: (h,k,l);(-h,-k,l); (k,h,-l); (-k,-h,-l). As well as these exact cases, ambiguities may arise accidentally for special values of cell dimensions: for example, a monoclinic cell with β=90° will appear orthorhombic, leading to an alternative indexing as (-h,-k,l). Less obvious cases can occur with special relationships involving cell diagonals. For some examples, see http://www.ccp4.ac.uk/dist/html/reindexing.html

For the first crystal (or indexing), you are free to choose any of the alternatives, but subsequent indexing must match the original "reference" scheme. POINTLESS can check which scheme matches best in two ways: you can give a reference file HKLREF (which can now be either merged or unmerged), in which case the test data (HKLIN) will be

checked against the reference, and its space group will be assumed to be correct; or if you give multiple test data files (HKLIN) and no HKLREF file is defined, the first one will be treated as a reference for alternative indexing, but the combined data will still be tested for Laue group symmetry.

Examples

Two examples were given in Evans (2006) and they remain valid even though the scoring system has changed. Most crystals give a clear answer: uncertainties generally arise through pseudo-symmetry, including twinning, and the difficult case *(2)* illustrated here is not typical.

(1)Discriminating between orthorhombic groups with systematic absences

This case (Parker, unpublished) gave a clear indication of orthorhombic symmetry, Laue group Pmmm. Fourier analysis of the axial reflections gave a definitive suggestion that the space group was $P22_12_1$ (standard setting $P2_12_12$ with the reindexing operation (k,l,h))(table 1)

Axis	Number	Peak height at 1/2	SD	Probability	Reflection condition
2(1) [a]	39	-0.234	0.242	0.000	h00: h=2n
2(1) [b]	**27**	**0.997**	**0.176**	**0.970**	**0k0: k=2n**
2(1) [c]	**87**	**0.993**	**0.109**	**0.988**	**00l: l=2n**

Table 1 . *Systematic absence analysis*

(2)Pseudo-symmetry from incomplete pseudo-merohedral twinning

The true space group in this case (Sanchez Barrena, unpublished) is $P2_12_12_1$but all three cell lengths are about the same (79.2, 81.3, 81.2 Å) and the crystals have a variable amounts of twinning into the apparent point group 422(twinning operator k,h,-l). Table 2 shows the scores for the possible cubic symmetry operators for two crystals, a native crystal with about 20% twinning(refined twin fraction), and a more highly twinned SeMet crystal. Table 3 shows their scores for the different Laue groups: the native crystal gives the correct Pmmm group, but the program is fooled by twinning in the SeMet data into preferring Laue group P4/mmm.

	Native			SeMet		
Symmetry operator	**Likelihood**	**CC**	**R_{meas}**	**Likelihood**	**CC**	**R_{meas}**
Identity	0.953	0.97	0.066	0.949	0.97	0.076
2-fold (1 0 1)	0.058	0.22	0.466	0.055	0.04	0.605
2-fold (1 0 -1)	0.059	0.23	0.451	0.054	0.14	0.516
2-fold (0 1 -1)	0.063	-0.01	0.667	0.056	0.03	0.653
2-fold (0 1 1)	0.062	-0.01	0.671	0.054	0.05	0.636
2-fold (1 -1 0)	0.052	0.04	0.639	***0.713***	***0.83***	***0.155***
2-fold k (0 1 0)	**0.947**	**0.96**	**0.103**	**0.921**	**0.93**	**0.104**
2-fold (1 1 0)	0.051	0.05	0.631	***0.562***	***0.78***	***0.175***
2-fold h (1 0 0)	**0.944**	**0.95**	**0.138**	**0.916**	**0.92**	**0.111**
2-fold l (0 0 1)	**0.943**	**0.95**	**0.156**	**0.931**	**0.94**	**0.108**
3-fold (1 1 1)	0.059	0.00	0.766	0.058	0.02	0.639
3-fold (1 -1 -1)	0.058	0.01	0.776	0.057	0.03	0.800
3-fold (1 -1 1)	0.058	0.01	0.731	0.053	0.06	0.673
3-fold (1 1 -1)	0.058	0.01	0.798	0.059	0.02	0.838
4-fold h (1 0 0)	0.066	-0.02	0.731	0.064	-0.00	0.686
4-fold k (0 1 0)	0.056	0.21	0.463	0.054	0.05	0.586
4-fold l (0 0 1)	0.052	0.04	0.619	***0.539***	***0.77***	***0.175***

Table2. *Scores for potential symmetry elements for native and SeMet crystals. Both crystals show the dyads for orthorhombic symmetry (***bold***), but the SeMet crystal is more highly twinned (perhaps ~35%) than the native (~20%) also shows the tetragonal operators (****italic bold****)*

			Native					SeMet					
Rank	**Laue group**	**Reindex**	**Prob**	**Z-CC**	**CC+**	**CC-**	**R**	**Prob**	**Z-CC**	**CC+**	**CC-**	**R**	**Rank**
1	Pmmm	[h,k,l]	0.988	9.05	0.96	0.05	0.11	0.209	7.24	0.94	0.22	0.09	2
2	P 1 2/m 1	[k,l,h]	0.004	7.91	0.96	0.17	0.08	0.002	6.57	0.96	0.30	0.08	5
3	P 1 2/m 1	[l,h,k]	0.003	7.96	0.96	0.16	0.10	0.001	6.65	0.95	0.28	0.09	7
4	P 1 2/m 1	[h,k,l]	0.003	7.87	0.96	0.17	0.11	0.001	6.49	0.95	0.31	0.08	6
5	P 4/mmm	[l,k,-h]	0.000	4.58	0.53	0.07	0.32	0.000	2.61	0.54	0.27	0.29	13
6	P 4/m	[l,k,-h]	0.000	4.57	0.63	0.18	0.25	0.000	4.13	0.71	0.30	0.18	11
7	P 4/mmm	[k,h,-l]	0.000	6.53	0.66	0.01	0.24	**0.778**	**8.32**	**0.87**	**0.04**	**0.13**	**1**
8	P 4/m	[k,h,-l]	0.000	5.31	0.70	0.17	0.19	0.002	6.31	0.89	0.26	0.11	4
9	P -1	[h,k,l]	0.000	7.52	0.97	0.22	0.07	0.000	6.38	0.97	0.33	0.08	10
10	P 4/m	[h,l,-k]	0.000	4.35	0.62	0.18	0.25	0.000	3.01	0.63	0.33	0.23	12
11	P 4/mmm	[h,l,-k]	0.000	5.07	0.56	0.05	0.30	0.000	3.43	0.60	0.25	0.26	14

Table3. *Laue group rankings for the same crystals as in table 1. For the native crystal, the top rank solution is the correct Pmmm. For the SeMet crystal, P4/mmm is ranked higher because of the pseudo-symmetry from the twinning. Prob is the likelihood estimate, Z-CC is the net "Z-score" for CC, CC+ is for symmetry operators present in the Laue group, CC- for symmetry operators present in the cubic lattice but not in the Laue group, & R is the R-factor R_{meas}.*

Conclusions

In most cases, POINTLESS will give an unambiguous assignment of the Laue group, and often a good indication of the space group. Nevertheless, the results should always be treated with some caution, because of the possibility of pseudo-symmetry, which is not uncommon. In difficult cases, careful examination of the scores may lead to a decision different to that given by the program.

The options to combine multiple input files (from version 1.2.0) provides a more convenient method than the previous use of SORTMTZ, since it ensures that the files are on the same indexing system, and it adjusts the batch numbers if necessary to ensure that they are unique. POINTLESS is available from the CCP4 pre-release site or by anonymous ftp from ftp.mrc-lmb.cam.ac.uk/pub/pre/, and will be in future releases of CCP4. The program is under active development, the ultimate aim being to replace and extend all the scaling functions of Scala

Acknowledgements

I have been helped in the development of POINTLESS by many useful discussions with many people, including George Sheldrick, Ralf Grosse-Kunstleve, Airlie McCoy, Randy Read, Eleanor Dodson, Kevin Cowtan, Andrew Leslie, and Graeme Winter.

References

The Clipper C++ libraries for X-ray crystallography , Cowtan K. (2003) *IUCrComputing Commission Newsletter,***2**, 4-9

Scaling & assessment of data quality, Evans, P.R., (2006) *Acta Cryst.* D **62**, 82-82

*TheComputational Crystallography Toolbox: crystallographic algorithms in are usable software framework,*Grosse-Kunstleve, R.W., Sauter, N.K., Moriarty, N.W.& Adams, P.D (2002) *J.Appl.Cryst.* **35**, 126-136