

## The Use of Protein Homologues in the Rotation Function

BY C. F. AGUILAR, M. P. NEWMAN,\* J. SANZ APARICIO,† J. B. COOPER,  
I. J. TICKLE & T. L. BLUNDELL

Laboratory of Molecular Biology and Imperial Cancer Research Fund Structural Molecular Biology Unit,  
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, England

(Received 20 May 1992; accepted 5 August 1992)

### Abstract

The success of molecular replacement depends, in part, on the degree of similarity of the target and search molecules. We have systematically investigated this effect in cross-rotation functions for members of the aspartic proteinase family of enzymes. The influence of various parameters on peak heights was investigated for six search models using  $|F_{\text{obs}}|$  data for two target enzymes. The beneficial effects of high-resolution data and a large radius of integration are most pronounced when target and search molecules have high-percentage identities. Correction for small differences in domain-domain orientation (typically 4–8°) between search and target structures leads to only a marginal improvement in the rotation-function peak height. There is an almost linear relationship between the structural distance,  $D$ , a parameter used in cluster analysis to define differences between three-dimensional protein structures, and the height of the cross-rotation-function peaks.

### 1. Introduction

The technique of molecular replacement has played a major role in X-ray structural studies of biological macromolecules since its introduction by Rossman & Blow (1962). The potential application of this method in determining the unknown structure of a protein or nucleic acid, by using structure factors calculated from a structurally related molecule, has proved to be a very powerful alternative to the method of isomorphous replacement.

The search model used in molecular replacement will occasionally be derived from the structure of another crystal form of the same enzyme. More usually, it will be a structure of a homologous protein or a model based on one or more homologues. In the latter case, the success of the method will depend critically on the similarity of the search model to the target structure. Chothia & Lesk (1986) have shown that with decreasing sequence similarity there is

an increasing root-mean-square (r.m.s.) deviation between equivalent atoms when the protein structures are optimally superimposed. This divergence in three-dimensional structure implies that homologues become less-satisfactory models for molecular replacement as the percentage sequence identity decreases.

Although homologous proteins have been used widely in molecular replacement, few researchers have reported systematic studies using search models of differing divergence from that of the target structure. Here we report a study in which we have examined the influence of various parameters on the rotation function using the structures of several homologues of differing percentage identities. In particular, we investigate the radius of the sphere of integration in the Patterson synthesis, the effect of the resolution limits of data used, the degree of similarity between the search model and the unknown structure and the effect of rigid-body rotations within the molecule.

For this study, we have used the crystal structures of various aspartic proteinases. Aspartic proteinases comprise a family of proteolytic enzymes found in fungi, higher plants and mammals, which are generally between 320 and 340 residues in length and whose catalytic activity is dependent on two invariant aspartic acid residues. There are ten structures that have been determined by X-ray crystallography. These are porcine pepsin (Sielecki, Fedorov, Boodhoo, Andreeva & James, 1990; Cooper, Khan, Taylor, Tickle & Blundell, 1990; Abad-Zapatero, Rydel & Erickson, 1990), pepsinogen (James & Sielecki, 1986), calf chymosin (Strop *et al.*, 1990; Gilliland, Winborne, Nachman & Wlodawer, 1990; Newman *et al.*, 1991), human renin (Sielecki *et al.*, 1989; Dhanaraj *et al.*, 1992), mouse renin (Dealwis *et al.*, 1993), rhizopuspepsin (Suguna *et al.*, 1987), endothiapepsin (Blundell *et al.*, 1990), penicillopepsin (James & Sielecki, 1983), yeast proteinase A (Aguilar, Badasso, Cooper, Newman & Blundell, 1993) and *Mucor pusillus* pepsin (Newman *et al.*, 1993). The enzymes are composed mainly of  $\beta$ -strands, folded into two lobes that are related by a twofold axis of pseudosymmetry that passes through the middle of the active-site cleft.

\* Present address: Department of Biochemistry and Molecular Biophysics, Columbia University, New York, USA.

† Present address: Instituto de Química Física Rocasolano, Madrid, Spain.

Table 1. *Pairwise sequence identities and pairwise r.m.s. differences of equivalent C<sub>α</sub> positions for optimally superposed proteins for the six aspartic proteinases involved in this study*

	Porcine pepsine	Pepsinogen	Chymosin	Rhizopus	Penicillo	Endothia
Porcine pepsin	-	1.02	0.98	1.41	1.41	1.55
Pepsinogen	98*	-	1.15	1.41	1.55	1.59
Chymosin	59	58	-	1.30	1.45	1.58
Rhizopus	36	35	33	-	1.22	1.27
Penicillo	31	30	28	39	-	1.21
Endothia	26	29	26	38	54	-

\* The propart of pepsinogen was removed in the comparison. The sequence differences between porcine pepsinogen and porcine pepsin arise from two residues that have been misidentified and five residues that were omitted in the pepsinogen three-dimensional structure.

The aspartic proteinases form a diverse family, in terms of both sequence identity and r.m.s. coordinate difference (Table 1). Much of the difference between the molecules is due to variable loops on the surface of the molecule joining  $\beta$ -strands. However, Šali, Veerapandian, Cooper, Foundling, Hoover & Blundell (1989) found that the endothiapepsin molecule exists in two forms that differ in orientation of part of the carboxyl-terminal domain with respect to the rest of the molecule. The same type of structural variability also accounts for some of the differences in the aspartic proteinase family in general (Abad-Zapatero, Rydel & Erickson, 1990; Cooper, Khan, Taylor, Tickle & Blundell, 1990; Sielecki, Fedorov, Boodhoo, Andreeva & James, 1990; Šali, Veerapandian, Cooper, Moss, Hofmann & Blundell, 1992). The first of the two rigid bodies comprises residues 2 to 189 and 303 to 326. The second comprises residues 190 to 302. Clearly, differences in the orientation of domains between the search and target molecules may have significant effects on the rotation and translation function.

A similar analysis has been performed for crystal-line Fab fragments (Cygler & Anderson, 1988) although the Patterson functions were approximated by spherical Bessel functions of maximum order 30 and the number of reflections was limited to 2500. In this paper, we present a systematic study of the effects of radius of integration, resolution, sequence divergence and domain orientation on the rotation function.

We have used a version of the fast rotation function that uses spherical Bessel functions up to order 60 and a maximum of 10 000 reflections. This allows use of a larger radius of integration with high-resolution data.

## 2. Methods

Coordinates of endothiapepsin, penicillopepsin, rhizopuspepsin, hexagonal pepsin, chymosin and pepsinogen (4APE, P2APP, P2APR, P5PEP, P2CMS, P1PSG; available from the Brookhaven Data Bank: Bernstein *et al.*, 1977) were used as search models (Table 2). The search models were placed in a com-

Table 2. *Data-collection statistics for the two aspartic proteinases used as targets in this study*

Structures	R merge (%)	Resolution	% Complete	% > 3 $\sigma$
Chymosin	8.9	2.2	96.6	54.2
Endothia	7.4	2.1	98.8	81.9

mon reference frame by superimposing the coordinates in a pairwise manner on the coordinates for porcine pepsin in an arbitrary orientation. The structurally conserved regions as defined by *MNYFIT* (Sutcliffe, Haneef, Carney & Blundell, 1987) were used in a least-squares program *XS5* (Šali, 1991) to define a rotation and translation matrix that superimposes one molecule on the other.

The structure factors were calculated from these search models using *GENSF*C (in program suite *CCP4*, 1979). The coordinates of the models were placed in a hypothetical orthogonal *P1* lattice with cell edges equal to the dimension of the molecule in each direction plus 20 Å so that most of the cross vectors were larger than the chosen radius of integration (Lifchitz, 1983). The atomic temperature factors of the search models were used in the calculation as these tend to be closely related between structures and their inclusion leads to the generation of more accurate structure factors (Dodson, 1985). It is also generally true that the amino acid residues with highest temperature factors are on the surface and are most often substituted in homologues (for a review see Bajaj & Blundell, 1984). Thus, the effect of the temperature factors is to weigh down the differences between the search model and the target structure.

These calculated structure factors were used to compute cross-rotation functions for the observed structure-factor amplitudes from the crystals of endothiapepsin and chymosin (Table 3). Both the observed and the calculated structure-factor amplitudes were converted into *E* values using *ECALC* (in program suite *CCP4*, 1979), so that the rotation function would not be dominated by a few large intensities (Dodson, 1985; Tickle, 1985).

The cross-rotation searches were performed using the 60-Bessel-function version of *ALMN* (in program suite *CCP4*, 1979), based on Crowther's (1972) fast

Table 3. *Refinement statistics for the six aspartic proteinases involved in this study*

Structures	R factor	Resolution	Coordination error
Porcine pepsin	0.19	2.3	0.11*
Chymosin	0.17	2.2	0.21*
Rhizopus	0.14	1.8	0.16†
Penicillo	0.13	1.8	0.16†
Endothia	0.17	2.1	0.20*†
Pepsinogen	0.17	1.6	-

\* Calculated using the method of Read (1986).

† Calculated using the method of Luzzatti (1952).

rotation function. This program will accept a maximum of 10 000 reflections. The rotation function of endothiapepsin crystals, whose space group is  $P2_1$ , was calculated for a single asymmetric unit of the Eulerian cell defined by the angular ranges  $\alpha = 0-180$ ,  $\beta = 0-180$  and  $\gamma = 0-360^\circ$ . A coarse  $5^\circ$  step size in  $\beta$  was used. For calf chymosin, whose space group is  $I222$ , the search was performed over angular ranges  $\alpha = 0-180$ ,  $\beta = 0-90$  and  $\gamma = 0-360^\circ$  in  $2.5^\circ$  steps. The resolution limits and radii of integration are as specified in each case.

For comparison of the target molecule with the search model, structural similarity was expressed in terms of the structural distance measure  $D$  (Johnson, Sutcliffe & Blundell, 1990; Johnson, Šali & Blundell, 1990). This expression combines both the r.m.s. distance and number of  $\alpha$ -carbon atoms found to be topologically equivalent within a  $3.5 \text{ \AA}$  cutoff distance. It is defined as

$$D = -100 \ln [(1-w)PFTE + wSRMS], \quad (1)$$

where

$$w = (PFTE + SRMS)/2. \quad (2)$$

PFTE, the pairwise fractional topological equivalence, was obtained by dividing the number of equivalences (Johnson, Sutcliffe & Blundell, 1990; Johnson, Šali & Blundell, 1990) by the total number of main-chain  $\alpha$ -carbon positions in the smaller structure. The r.m.s. distance over these equivalences was converted into a similarity measure, SRMS, through the expression  $1 - (\text{RMS}/3.5)$ . The effect of  $w$  is to moderate the influence of PFTE on the function at small distances (close relationships) where the SRMS provides a better representation of the relationship; similarly, at larger distances where PFTE gives a better representation, the purpose of  $w$  is to moderate the contribution of the SRMS to the function. Both the number of topological equivalences and the r.m.s. distance between two structures were obtained from a pair-wise least-squares superposition using *SFIT* (Johnson, Šali & Blundell, 1990) based generally on *MNYFIT* (Sutcliffe, Haneef, Carney & Blundell, 1987).

### 3. Results

#### 3.1. Effect of the radii of integration on the rotation function

The radii of integration used in the rotation function are chosen to include a large proportion of 'self' (intramolecular) vectors but to minimize the number of 'cross' (intermolecular) vectors arising from neighbouring molecules (Blow, 1985). The choice is influenced by solvent content, space group, molecular shape *etc.* Although 'cross' vectors can be avoided in the search-model Patterson by careful choice of the unit-cell size (Lifchitz, 1983), they cannot be completely avoided in the target Patterson function. This is especially true for ellipsoidal molecules such as the aspartic proteinases (major axis 60 and minor axis 30  $\text{\AA}$ ).

To test the influence of the outer radius of integration, rotation functions employing the endothiapepsin and calf chymosin observed structure-factor amplitudes were calculated using the various search models. The value for the outer radius of integration was varied between 15 and 35  $\text{\AA}$ . The inner radius was maintained at 6  $\text{\AA}$ . The resolution range of data was 20 to 3  $\text{\AA}$ . The results are shown in Fig. 1.

The dimensionless ratio of peak height/RMS was used to obtain a scale-invariant measure of the Patterson overlap function. Here RMS is the root weighted mean-square deviation of the rotation-function-map values from their weighted mean, the weight factor  $\sin \beta$  being the metric in Eulerian-angle space (Navaza, 1987).

The position of the highest peak in the rotation function was relatively insensitive to the change in the outer radius of integration. However, the highest values for peak height/RMS were obtained for an outer radius of approximately 30  $\text{\AA}$ .

This is approximately the same size as the smallest dimension of the molecule and about 50% of the greatest dimension. For an approximately spherical molecule, the outer radius of integration should be 75–80% of the diameter of the molecule (Blow, 1985). For an ellipsoidal molecule, it may be preferable to use an outer radius much smaller than the greatest dimension. The ability to integrate over a non-spherical shape function would possibly be advantageous.

Clearly, the expectation that increasing the radius of integration well beyond the separation of the search molecules would rapidly reduce the height/RMS owing to the inclusion of search-model cross vectors does not appear to be borne out by these data.

The highest peak height/RMS was reduced in all cases when the outer radius of integration was lowered to 15  $\text{\AA}$ . The largest variation in height/RMS with resolution occurred for molecules with the greatest percentage sequence identity. When the penicillopepsin search model was used with the

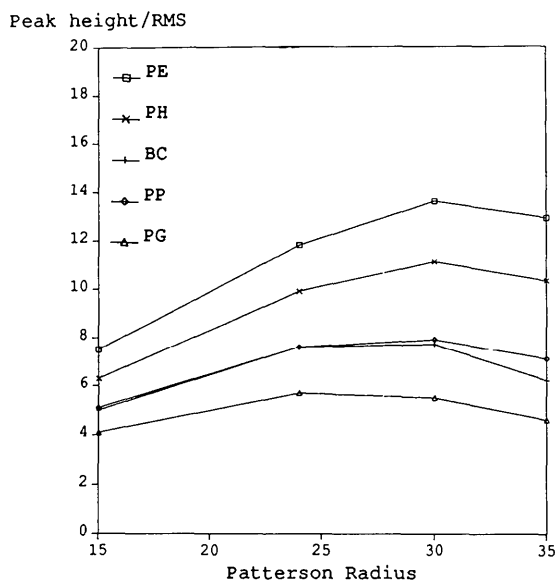
endothiapepsin observed structure-factor amplitudes (sequence identity of 53% between the target molecule and search model; Table 1), the ratio between the peak height at an outer radius of 30 Å and that at 15 Å was 1.8. A similar effect is observed for the

pepsin *versus* chymosin rotation function in which search molecules have 58% sequence identity.

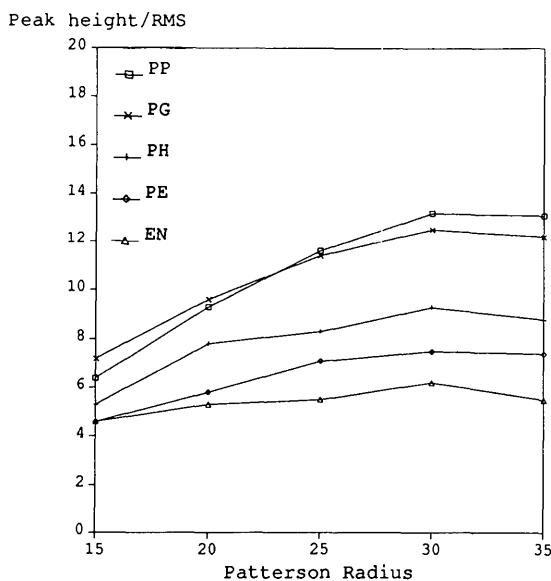
In contrast, for molecules with low percentage identity, varying the outer radius of integration has very little effect on the peak height/RMS. The curve for the endothiapepsin search model against calf chymosin observed X-ray data is almost independent of the outer radius (Fig. 1*b*). In these studies, the inner radius of integration, which removes the origin peak of the Patterson function, was maintained at a constant value.

Although for an average unit-cell shape it is not essential to omit the origin in Patterson space because it serves only to add a constant positive term to the rotation function (Blow, 1985), this may be necessary in the case of very flattened or elongated unit cells. In this case, it can give false overlaps with neighbouring unit cells if the radius of integration is too large.

In the molecular replacement of calf chymosin, an inner radius of 6 Å was found to be optimal (Newman *et al.*, 1991). In contrast, during molecular-replacement studies on the aspartic proteinase A from *Saccharomyces cerevisiae*, two peaks corresponding to two molecules in the asymmetric unit could only be found when no inner cutoff in the Patterson overlap was applied (Aguilar *et al.*, 1993). The search model used for this study was porcine pepsin which has 39% sequence identity with aspartic proteinase A.



(a)



(b)

Fig. 1. Rotation-function peak height/RMS *versus* Patterson function outer radius, calculated using  $E$  values with *ALMN* (Dodson, 1985). The inner radius was 6 Å. The resolution limits of data used were 3 to 20 Å. Rotation functions were calculated using observed structure-factor amplitudes (a) for endothiapepsin (space group  $P2_1$ ) and (b) for calf chymosin (space group  $I222$ ). Search models used are denoted by: PP porcine pepsin; PG porcine pepsinogen; BC calf chymosin; PH rhizopuspepsin; PE penicillopepsin; EN endothiapepsin.

### 3.2. Effect of the resolution limits on the rotation function

The resolution limits control the number of diffraction terms to be included in the Patterson summations. The coordinate difference between the target molecule and the search model necessitates the omission of the high-resolution terms. It has been suggested that the ideal higher-resolution limit is approximately two or three times the expected mean coordinate difference (Blow, 1985).

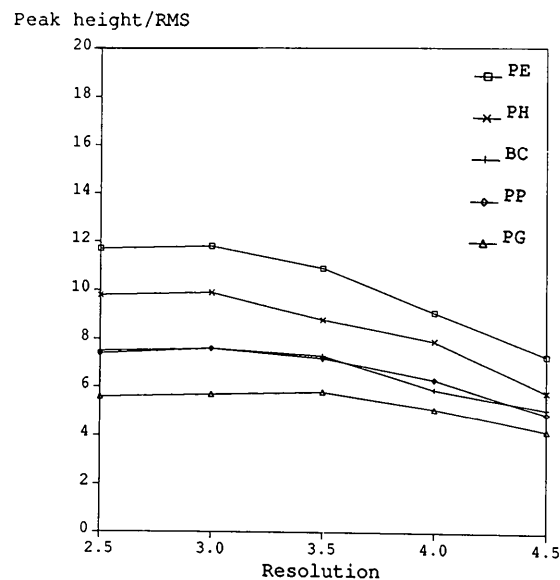
To assess the effect of the resolution limits of the X-ray data on the rotation function, a series of calculations were performed using the various search models against the endothiapepsin and calf chymosin structure-factor amplitudes. The higher-resolution limit was varied between 4.5 and 2.5 Å. The lower-resolution limit was maintained at 20 Å. The inner and outer radii of integration were 6 and 30 Å, respectively, for all calculations.

One clear peak was observed in each of the rotation-function maps. When the higher-resolution limit of the data was varied within the range 4.5–2.5 Å, the position of the peak did not appear to be significantly affected. The highest peak height/RMS was influenced by the higher-resolution limit (Fig. 2). For all rotation functions calculated, better results were obtained by including the higher-resolution data. The peak height/RMS increases by a factor of 1.6 as the

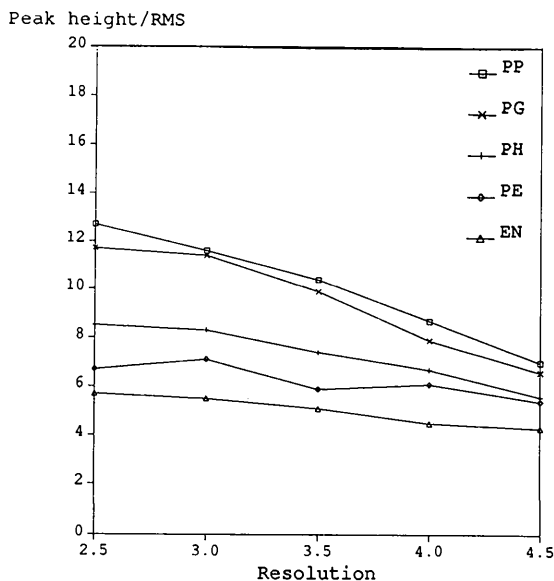
resolution is increased from 4.5 to 3.0 Å for the penicillopepsin search model with the endothiapepsin data.

The influence of the higher-resolution limit appears to be largest when the structural similarity between

the target molecule and the search model is greatest and decreases as the similarity decreases. Variation of the higher-resolution limit has an insignificant effect when endothiapepsin is used as a search model against the calf chymosin target molecule (with only 28% sequence identity; Table 1). This observation is probably due to the fact that differences between homologous structures, such as relative shifts between the secondary-structure elements and variations of side-chain conformation become more apparent at high resolution. Thus, for dissimilar structures, inclusion of high-resolution data serves to highlight these differences and so does not lead to an improvement in Patterson overlap. Fig. 2 indicates no support for the hypothesis that high-resolution terms decrease the peak-to-noise ratio. However, there is no significant increase in height/RMS for any model on increasing the resolution from 3 to 2.5 Å.



(a)



(b)

Fig. 2. Rotation-function peak height/RMS versus higher-resolution limit. Calculated using  $E$  values with *ALMN* (Dodson, 1985). The lower-resolution limit was 20 Å. The radii of integration were 6 to 30 Å. Rotation functions were calculated using observed structure-factor amplitudes (a) for endothiapepsin (space group  $P2_1$ ) and (b) for calf chymosin (space group  $I222$ ). Search-model symbols are as in Fig. 1.

### 3.3. Effect of structural similarity on the rotation function

The success of the rotation function depends heavily on the existence of a closely related search model.

All rotation functions calculated have produced distinct solutions. (Table 4) despite the low percentage of sequence identities ( $\sim 26\%$ ) in some cases. Figs. 3(a) and (b) show plots of highest peak versus  $\beta$ -section for the rotation functions calculated using the endothiapepsin and calf chymosin  $F_{\text{obs}}$  data.

The points belonging to the peak extend over a range of sections. The range was generally less than  $20^\circ$  but extended over  $30^\circ$  for very strong peaks, e.g. chymosin model versus chymosin  $F_{\text{obs}}$  data. Some  $\beta$ -sections containing the highest peaks are shown in Fig. 4.

After the rotation corresponding to the peak maximum was applied, the angular component of the screw transformation needed to superimpose the model on the target coordinates (i.e. the error in the rotation-function solution,  $\Delta$ ) was calculated.

The transformed search models were found to be positioned to within a few degrees of the target coordinates. The search models used with the calf chymosin structure-factor amplitudes gave errors ( $\Delta$ ) that ranged from  $1.2$  to  $2.0^\circ$ . This error is clearly smaller than the  $2.5^\circ$  step size of the Eulerian angles used in the rotation function and should be within the radius of convergence of rigid-body refinement (Sussman, Holbrook, Church & Kim, 1977).

A maximum deviation of  $2.2^\circ$  was observed for the pepsinogen search model in the rotation function calculated for endothiapepsin X-ray data; this may be caused either by the presence of the propeptide segment or by the rigid-body rotation.

The maximum peak height/RMS values obtained in the rotation function increased with increasing

Table 4. *Rotation-function solutions calculated for observed structure-factor amplitudes using various search models*

Radii of integration 6–30 Å. Resolution limits 3–20 Å.  $\alpha$ ,  $\gamma$  are Eulerian angles with a step size of 2.5°.  $\beta$  is an Eulerian angle with step size of 5.0°.  $\Delta$  is the angular component of the screw transformation necessary to superimpose the coordinate of the search model (after having been transformed by the relevant Eulerian angles) to the final refined coordinates of the molecule, *i.e.* the error in the rotation-function solution. Ratio is the ratio of the highest peak to second-highest peak.

## (a) Results for endothiapepsin molecule

Search Model	Highest peak				$\Delta$ (°)	Next-highest peak				Ratio
	$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS		$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS	
Endothiapepsin	55.0	45.0	-65.0	28.1	1.1	97.5	115.0	170.0	4.4	6.3
Penicillopepsin	57.5	45.0	-67.5	15.9	1.4	12.5	155.0	95.0	5.0	3.1
Rhizopuspepsin	57.5	45.0	-67.5	11.8	1.9	152.5	120.0	-157.5	4.2	2.7
Chymosin	57.5	45.0	-67.5	8.1	1.2	77.5	95.0	-130.0	4.8	1.6
Pepsin	50.0	50.0	-67.5	7.9	0.9	17.5	135.0	122.5	4.2	1.8
Pepsinogen	55.0	45.0	-65.0	5.8	2.2	140.0	160.0	145.0	3.8	1.5

## (b) Results for calf chymosin molecule

Search model	Highest peak				$\Delta$ (°)	Next-highest peak				Ratio
	$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS		$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS	
Chymosin	142.5	50.0	-87.5	22.1	1.2	117.5	50.0	-87.5	4.9	4.5
Pepsin	142.5	50.0	-87.5	13.7	1.3	65.0	50.0	-87.5	4.7	2.9
Pepsinogen	142.5	50.0	-87.5	11.6	2.0	65.0	50.0	-85.0	4.3	2.7
Rhizopuspepsin	142.5	50.0	-87.5	9.0	0.8	15.0	17.5	-165.0	4.1	2.2
Penicillopepsin	142.5	52.5	-87.5	7.4	1.7	90.0	25.0	122.5	4.6	1.6
Endothiapepsin	145.0	50.0	-90.0	6.4	1.5	82.5	42.5	-75.0	4.6	1.4

percentage sequence identity between the target and search molecules (Table 4). As expected, the fungal search models gave the best results in the rotation functions using the endothiapepsin observed structure-factor amplitudes (approximate sequence identity 40–50%; Table 1), whilst the mammalian search models were superior for the calf chymosin X-ray data (58% sequence identity between porcine pepsin and calf chymosin; Table 1).

To quantify the similarity between the target molecule and the search model, the structural distance measure  $D$  was used. This had previously been shown to be useful in cluster analysis of proteins on the basis of their three-dimensional structures (Johnson, Sutcliffe & Blundell, 1990).

The structural distance measure  $D$  of each search model with respect to the endothiapepsin and calf chymosin target molecules was calculated and plotted against the maximum peak height/RMS obtained in the corresponding rotation function (Fig. 5). As can be seen, there is a remarkably linear relationship between the rotation-function peak height and the structural distance measure  $D$ . A straight line was fitted to these points by linear regression. The maximum values obtained in the rotation function were those when endothiapepsin or calf chymosin were used as search molecules ( $D = 0$ ).

These results serve to emphasize the importance of structural similarity between the search model and the unknown target molecule in the rotation function. Clearly, use of the parameter  $D$  helps to rationalize rotation-function results obtained with different search models.

3.4. *Effect of rigid-body fitting on the rotation function*

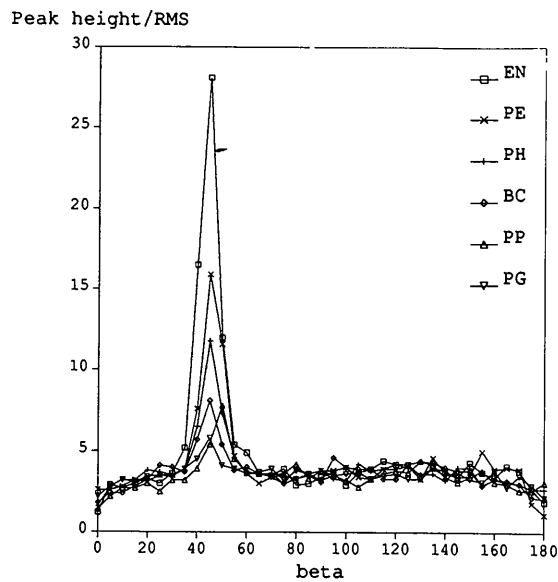
Rigid-body movements of approximately one third of the molecule (residues 190–302) account for some of the structural differences between the aspartic proteinases (Šali *et al.*, 1992). These movements differ in both magnitude and direction (12.7° rigid-body shift between endothiapepsin and hexagonal porcine pepsin but only 3.7° shift between calf chymosin and hexagonal porcine pepsin). To examine the effect of these rigid-body movements on the rotation function, calculations were repeated using search models that had these relative shifts removed. This was achieved by separately superimposing all the search models as two rigid groups on the target endothiapepsin and calf chymosin coordinates, respectively, prior to calculation of the rotation function.

The results are shown in Table 5. Before the rigid-body fitting procedure, the peaks were distributed over a wider range of  $\alpha$ ,  $\beta$  and  $\gamma$  values (See Table 4 for comparison). The effects of rigid-body fitting can also be observed in the general improvement shown by  $\Delta$  values, where  $\Delta$  is the error in the rotation-function solution.

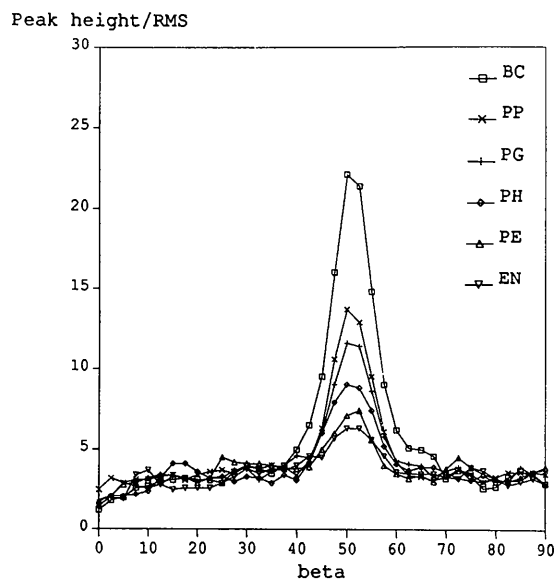
There is only a slight difference in the peak height/RMS when the rigid-body movements are subtracted from the search models, indicating that the Patterson superposition has improved slightly. This may be due to the changes to vectors that occur between rigid bodies 1 and 2.

To assess the effect of having a model that comprises only part of the unknown target molecule, the rotation functions were repeated using either rigid

body 1 or rigid body 2. Both rigid groups had been previously independently superimposed onto either the target endothiapsin or calf chymosin coordinates.

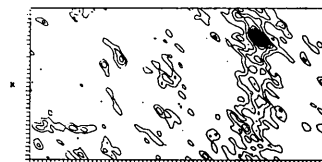


(a)

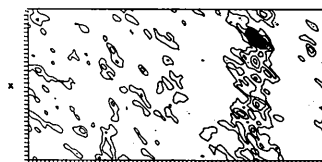


(b)

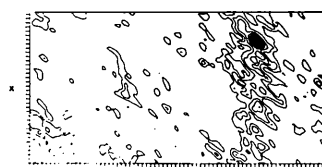
Fig. 3. Plot of the highest rotation function value per  $\beta$ -section. Rotation functions were calculated using  $E$  values with *ALMN* (Dodson, 1985). The radii of integration and resolution limits were 6 to 30 Å and 3 to 20 Å, respectively. These values were found to be optimal for the aspartic proteinases. Rotation functions were calculated using observed structure-factor amplitudes (a) for endothiapsin (space group  $P2_1$ ) and (b) for calf chymosin (space group  $I222$ ). Search-model symbols are as in Fig. 1.



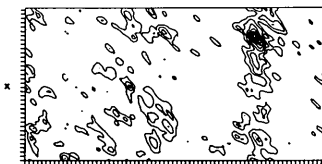
(a)



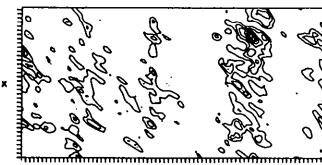
(b)



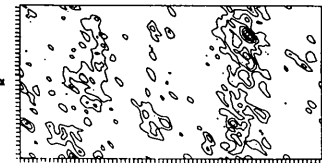
(c)



(d)



(e)



(f)

Fig. 4.  $\beta$ -sections in the rotation-function map containing the highest peak for the calf chymosin target molecule. The search models used were: (a) calf chymosin; (b) porcine pepsin; (c) porcine pepsinogen; (d) rhizopuspepsin; (e) penicillopepsin; (f) endothiapsin. Maps are contoured at the RMS defined using *EULSIG* (I. J. Tickle, unpublished). Conditions are given in Fig. 3 and Table 4.

When rigid body 1 was used, corresponding to approximately two thirds of the molecule, this produced the correct solution for all the rotation-function calculations. All peaks coincided to within a few degrees, but the peak heights/RMS were generally reduced by approximately 20–30% compared with the results obtained with a complete search model. However, rigid body 2 gave only poor results. For example, solutions were obtained when rigid body 2 from hexagonal porcine pepsin and calf chymosin were used as search models against calf chymosin X-ray data. The peak heights/RMS were greatly reduced compared with the entire model. The rotation functions calculated using endothiapepsin X-ray data

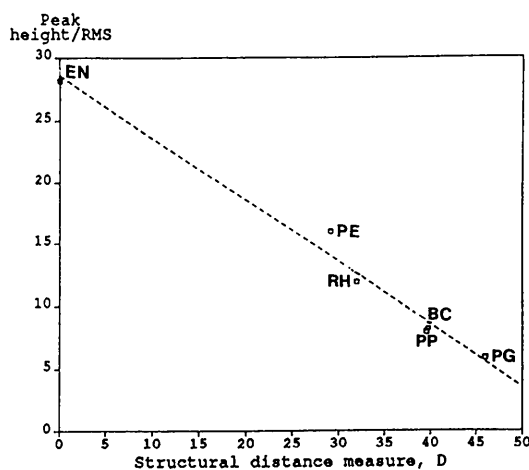
were completely featureless for all rigid body 2 search models. The fact that correct solutions were obtained using rigid body 1 but not rigid body 2 is probably due to the fact that the C-terminal lobe has undergone much greater divergence among the aspartic proteases. In addition, since the rigid group 2 represents only one third of the molecule, the calculated Patterson function will contain approximately only one ninth of the peaks for the full molecule. The rotation-function solution would therefore be expected to have a low signal/noise ratio, as is observed.

#### 4. Concluding remarks

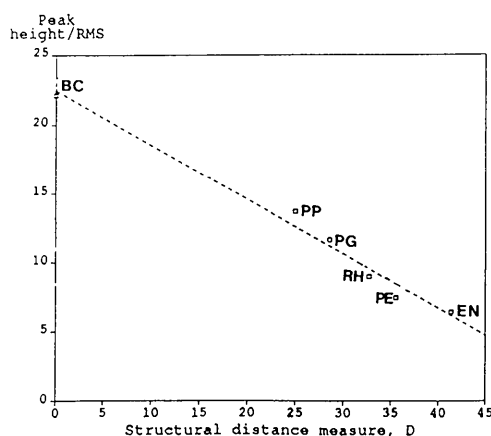
By maintaining the Patterson-function inner radius at 6 Å and varying the outer radius for different search models, we found that the optimum value was around 30 Å. This is approximately equal to half the largest dimension of the molecule. Reflection data in the range 20 to 3.0 Å were used for this analysis. As the radius of integration is increased towards 30 Å, the rotation-function peaks improve more for cases where search and target structures are most similar. The improvement in height/RMS is considerably less notable for enzymes with percentage identities with the target molecule of ~30%. Increasing the radius of integration well beyond the separation of the molecular boundaries in the target lattice, with the inherent danger of incorporating cross vectors in the summation, does not appear to be deleterious, at least in the cases that we have examined. There appears only to be a slight fall off in the height/RMS ratio as the radius of integration increases from 30 to 35 Å.

Maintaining the low-resolution cutoff at 20 Å and varying the outer limit from 4.5 to 2.5 Å gives an optimum value at 3.0 Å in many cases. As the resolution of the data is increased beyond this point, there appears to be no sign of a sharp decrease in height/RMS ratio. In fact, for cases where search and target molecules exhibit a high degree of homology, there are worthwhile improvements in peak height as the outer resolution limit is increased beyond 3.0 Å. This finding lends no support to suggestions that inclusion of high-resolution data may be detrimental to the rotation function (see above) even when search and target structures have low sequence identity. The effects of resolution are likely to be problem-specific since Cygler & Anderson (1988) found that increasing the resolution limit beyond 4.0 Å with Fab fragments leads to no improvement in the peak height. However, limitations in the 30-Bessel-function version of Crowther's program (Crowther, 1972) used in their analysis meant that use of high-resolution data necessitated lowering the outer radius of integration. Hence, their results may not be strictly comparable with ours.

One observation we have made in this and a number of other structural analyses (see, for example, Cooper, Khan, Taylor, Tickle & Blundell, 1990) is that



(a)



(b)

Fig. 5. Peak heights/RMS values obtained in the rotation function are plotted *versus* the corresponding distance measure  $D$ . The plots are (a) for endothiapepsin as the target molecule (correlation coefficient is 99.2%) and (b) for calf chymosin as the target molecule (correlation coefficient is 99.1%). The straight lines were fitted by linear regression. The search-model symbols are as in Fig. 1. Other details and conditions are given in Fig. 3 and Table 4.



Table 5. *Rotation-function solutions calculated for observed structure-factor amplitudes using various search models with the rigid-body movement removed*

Radii of integration 6–30 Å. Resolution limits 3–20 Å.  $\alpha$ ,  $\gamma$  are Eulerian angles with step size of 2.5°.  $\beta$  is an Eulerian angle with step size of 5.0°.  $\Delta$  is the angular component of the screw transformation necessary to superimpose the coordinates of the search model (after having been transformed by the relevant Eulerian angles) to the final refined coordinates of the molecule, *i.e.* the error in the rotation-function solution. Ratio is the ratio of the highest peak to second-highest peak.

## (a) Results for endothiapepsin molecule

Search model	Highest peak				$\Delta$ (°)	Next-highest peak				Ratio
	$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS		$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS	
Endothiapepsin	55.0	45.0	-65.0	28.1	1.1	97.5	115.0	170.0	4.4	6.3
Penicillopepsin	55.0	45.0	-65.0	17.6	1.1	97.5	110.0	170.0	5.0	3.5
Rhizopuspepsin	55.0	45.0	-65.0	11.8	1.1	20.0	150.0	105.0	4.5	2.6
Chymosin	57.5	45.0	-67.5	10.3	1.3	77.5	95.0	-132.5	4.5	2.3
Pepsin	55.0	45.0	-65.0	9.9	1.1	155.0	85.0	-5.0	4.4	2.2
Pepsinogen	57.5	45.0	-67.5	7.1	1.2	117.5	165.0	-60.0	4.6	1.5

## (b) Results for calf chymosin molecule

Search model	Highest peak				$\Delta$ (°)	Next-highest peak				Ratio
	$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS		$\alpha$ (°)	$\beta$ (°)	$\gamma$ (°)	Peak/RMS	
Chymosin	142.5	50.0	-87.5	22.1	1.2	117.5	50.0	-87.5	4.9	4.5
Pepsin	142.5	50.0	-87.5	14.6	1.2	117.5	50.0	-144.0	4.3	3.4
Pepsinogen	142.5	50.0	-87.5	11.5	1.3	155.0	65.0	-100.0	4.2	2.7
Rhizopuspepsin	142.5	50.0	-87.5	10.1	1.2	12.5	17.5	-162.5	4.5	2.3
Penicillopepsin	142.5	52.5	-87.5	8.1	1.5	122.5	40.0	-97.5	4.9	1.6
Endothiapepsin	142.5	50.0	-87.5	7.4	1.2	117.5	47.5	-90.0	5.4	1.4

enzymes that have almost identical percentage sequence identities can give very different results when used as search models in the rotation function. For example, when using the endothiapepsin and penicillopepsin models in the rotation search for chymosin, significantly different peak heights were obtained despite both models having almost equal percentage identity to the target enzyme. Therefore, measures of structural similarity that incorporate information other than the sequence identity are necessary to rationalize the behaviour of some search models in the rotation function. Our finding that the peak heights for different search models are related by a straight line of negative gradient to the structural distance,  $D$ , indicates that this will be an informative parameter in other retrospective analyses.

Correcting for the differences in domain orientation of the target and search molecule was found to cause a slight improvement in the height/RMS ratio. Although domain movements in the aspartic proteinase family are usually quite small (typically 4–8°), the effects would be expected to be large at the periphery of the molecule. The fact that correcting for rigid-body movement does not lead to a major improvement in peak height indicates that the smaller domain (segment 190–302) makes very little contribution to the rotation function. Rigid group 2 alone fails to produce the correct peak when used as the search model. In contrast, rigid group 1 yields essentially the correct solution in every test although the peak height is somewhat lower than with the complete search model. It therefore appears that rigid group 2 makes such a small contribution to the rotation func-

tion that the peak heights obtained with a complete search model are insensitive to the orientation of this domain relative to rigid group 1. This effect may be due not only to the smaller size of this domain but also to the low-percentage sequence identity between the enzymes within rigid group 2.

We are grateful to Professor David Blow, Dr David Moss and Dr Mark Johnson for making helpful comments and suggestions on our manuscript. We thank the UK SERC, AFRC and ICRF for financial support.

## References

- ABAD-ZAPATERO, C., RYDEL, T. J. & ERICKSON, J. (1990). *Proteins*, **8**, 62–81.
- AGUILAR, C. F., BADASSO, M., COOPER, J. B., NEWMAN, M. P. & BLUNDELL, T. L. (1993). In preparation.
- BAJAJ, M. & BLUNDELL, T. L. (1984). *Annu. Rev. Biophys. Bioeng.* **13**, 453–492.
- BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F. JR, BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.
- BLOW, D. M. (1985). In *Molecular Replacement*. Proc. of the Daresbury Study Weekend, edited by P. MACHIN, pp. 2–7. SERC Daresbury Laboratory, Warrington, England.
- BLUNDELL, T. L., JENKINS, J. A., SEWELL, B. T., PEARL, L. H., COOPER, J. B., TICKLE, I. J., VEERAPANDIAN, B. & WOOD, S. P. (1990). *J. Mol. Biol.* **211**, 919–941.
- CCP4 (1979). The SERC Collaborative Computing Project in Crystallography. SERC Daresbury Laboratory, Warrington, England.
- CHOTHIA, C. & LESK, A. M. (1986). *EMBO J.* **5**, 823–826.
- COOPER, J. B., KHAN, G., TAYLOR, G., TICKLE, I. J. & BLUNDELL, T. L. (1990). *J. Mol. Biol.* **214**, 199–222.

- CROWTHER, R. A. (1972). In *The Molecular Replacement Method*, edited by M. G. ROSSMANN, pp. 173–178. New York: Gordon and Breach.
- CYGLER, M. & ANDERSON, W. F. (1988). *Acta Cryst.* **A44**, 38–45.
- DEALWIS, C., FRAZAO, C., BADASSO, M., SIBANDA, B. L., TICKLE, I. J., COOPER, J. B. & BLUNDELL, T. L. (1993). In preparation.
- DHANARAJ, V., DEALWIS, C., FRAZAO, C., BADASSO, M., SIBANDA, B. L., TICKLE, I. J., COOPER, J. B., DRIESSEN, H. P. C., NEWMAN, M. P., AGUILAR, C. F., WOOD, S. P., BLUNDELL, T. L., HOBART, P. M., GEOGHEGAN, K. F., AMMIRATI, M. J., DANLEY, D. E., O'CONNOR, B. A. & HOOVER, D. J. (1992). *Nature (London)*, **357**, 466–472.
- DODSON, E. J. (1985). In *Molecular Replacement*. Proc. of the Daresbury Study Weekend, edited by P. MACHIN, pp. 33–45. SERC Daresbury Laboratory, Warrington, England.
- GILLILAND, G. L., WINBORNE, E. L., NACHMAN, J. & WLODAWER, A. (1990). *Proteins*, **8**, 82–101.
- JAMES, M. N. G. & SIELECKI, A. R. (1983). *J. Mol. Biol.* **163**, 299–361.
- JAMES, M. N. G. & SIELECKI, A. R. (1986). *Nature (London)*, **319**, 33–38.
- JOHNSON, M. S., ŠALI, A. & BLUNDELL, T. L. (1990). *Methods Enzymol.* **183**, 670–690.
- JOHNSON, M. S., SUTCLIFFE, M. J. & BLUNDELL, T. L. (1990). *J. Mol. Evol.* **30**, 43–59.
- LIFCHITZ, A. (1983). *Acta Cryst.* **A39**, 130–139.
- LUZZATTI, V. (1952). *Acta Cryst.* **5**, 802–810.
- NAVAZA, J. (1987). *Acta Cryst.* **A43**, 645–653.
- NEWMAN, M., SAFRO, M., FRAZAO, C., KHAN, G., ZDANOV, A., TICKLE, I. J., BLUNDELL, T. L. & ANDREEVA, N. (1991). *J. Mol. Biol.* **221**, 1295–1309.
- NEWMAN, M., WATSON, F., ROYCHOWDHURY, P., JONES, H., BADASSO, M., CLEASBY, A., WOOD, S. P., TICKLE, I. J. & BLUNDELL, T. L. (1993). In preparation.
- READ, R. J. (1986). *Acta Cryst.* **A42**, 140–149.
- ROSSMAN, M. G. & BLOW, D. M. (1962). *Acta Cryst.* **15**, 24–31.
- ŠALI, A. (1991). PhD thesis, Univ. of London, England.
- ŠALI, A., VEERAPANDIAN, B., COOPER, J. B., FOUNDLING, S. I., HOOVER, D. J. & BLUNDELL, T. L. (1989). *EMBO J.* **8**, 2179–2188.
- ŠALI, A., VEERAPANDIAN, B., COOPER, J. B., MOSS, D. S., HOFMANN, T. & BLUNDELL, T. L. (1992). *Proteins*, **12**, 158–170.
- SIELECKI, A. R., FEDOROV, A. A., BOODHOO, A., ANDREEVA, N. S. & JAMES, M. N. G. (1990). *J. Mol. Biol.* **214**, 143–170.
- SIELECKI, A. R., HAYAKAWA, K., FUJINAGA, M., MURPHY, M. E. P., FRASER, M., MUIR, A. K., CARILLI, C. T., LEWICKI, J. A., BAXTER, J. D. & JAMES, M. N. G. (1989). *Science*, **243**, 1346–1351.
- STROP, P., SEDLACEK, J., STYS, J., KADERABKOVA, Z., BLAHA, I., PAVLICKOVA, L., POHL, J., FABRY, M., KOSTKA, V., NEWMAN, M., FRAZAO, C., SHEARER, A., TICKLE, I. J. & BLUNDELL, T. L. (1990). *Biochemistry*, **29**, 9863–9871.
- SUGUNA, K., BOTT, R. R., PADLAN, E. A., SUBRAMANIAN, E., SHERIFF, S., COHEN, G. H. & DAVIES, D. R. (1987). *J. Mol. Biol.* **196**, 877–900.
- SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S. M. (1977). *Acta Cryst.* **A33**, 800–804.
- SUTCLIFFE, M. J., HANEEF, I., CARNEY, D. & BLUNDELL, T. L. (1987). *Protein Eng.* **1**, 337–384.
- TICKLE, I. J. (1985). In *Molecular Replacement*. Proc. of the Daresbury Study Weekend, edited by P. MACHIN, pp. 22–26. SERC Daresbury Laboratory, Warrington, England.

*Acta Cryst.* (1993). **A49**, 315–324

## Arithmetic Equivalence of Point Groups for Quasiperiodic Structures

BY F. WIJNANDS AND T. JANSSEN

*Institute for Theoretical Physics, University of Nijmegen, 6525 ED Nijmegen, The Netherlands*

(Received 15 May 1992; accepted 17 August 1992)

### Abstract

Necessary and sufficient conditions are formulated for an  $n$ -dimensional arithmetic point group such that it may be the symmetry group of a  $d$ -dimensional quasiperiodic but not periodic, *i.e.* incommensurate, structure with Fourier modulus of rank  $n$ . Only point groups leaving invariant a  $d$ -dimensional subspace (the physical space) are considered. For an arithmetic point group describing an incommensurate structure, all equivalent choices for the internal space are related by the normalizer in  $\text{Gl}(n, \mathbb{Z})$  of the point group. Also, the conditions on arithmetic equivalence of two point groups allowing an incommensurate structure are discussed. These conditions yield a further partition of the arithmetic crystal classes.

### 1. Introduction

A well known problem in crystallography is the determination of nonisomorphic  $n$ -dimensional space groups. According to Ascher & Janner (1965), a space group  $G$  can be interpreted as a group extension of  $\mathbb{Z}^n$  by a finite subgroup  $\Gamma(K) \subset \text{Gl}(n, \mathbb{Z})$ , a faithful representation of a point group  $K \subset O(n)$ . In this formalism, all nonisomorphic extensions of  $\mathbb{Z}^n$  can be obtained by taking one representative  $\Gamma(K)$  of each arithmetic equivalence class, which consists of conjugate subgroups of  $\text{Gl}(n, \mathbb{Z})$ . Note that group extensions for arithmetically nonequivalent point groups are not isomorphic. For each representative arithmetic point group  $\Gamma(K)$ , all nonequivalent extensions can be determined. For this construction,