Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps

Johan Hattne and Victor S. Lamzin*

European Molecular Biology Laboratory, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany

Correspondence e-mail: victor@embl-hamburg.de

Received 10 April 2008
Accepted 13 May 2008

A pattern-recognition-based method for the detection of planar objects in protein or DNA/RNA crystal structure determination is described. The procedure derives a set of rotation-invariant numeric features from local regions in the asymmetric unit of a crystallographic electron-density map. These features, primarily moments of various orders, capture different aspects of the local shape of objects in the electron density. Feature classification is achieved using a linear discriminant that is trained to optimize the contrast between planar and nonplanar objects. In five selected test cases with X-ray data spanning 2.0–3.0 Å resolution, the procedure identified the correct location and orientation for almost all of the double-ring and a majority of the single-ring planar groups. The accuracy of the location of the plane centres is of the order of 0.5 Å, even in moderately noisy density maps.

1. Introduction

In macromolecular X-ray crystallography, electron-density map interpretation is a key step towards determination of the final structure. The procedure is generally time-consuming and subjective (Kleywegt & Jones, 1996; Mowbray *et al.*, 1999), since the crystallographer has to apply biophysical and chemical knowledge to visually interpret three-dimensional patterns of electron density in terms of an atomic model. However, given the growth of structural genomics initiatives (Burley *et al.*, 1999) and the desire to solve structures of increasing complexity (Brown & Ramaswamy, 2007), time and error tolerance is limited. Automated model-building procedures can speed up the structure-determination process considerably and also have the potential to minimize the amount of error in the modelling.

Prior stereochemical knowledge is invaluable for crystallographic model building, in particular with low-resolution data. The presence of planar groups of atoms is one important stereochemical characteristic of a macromolecular structure which may supply additional restraints during the refinement (see, for example, Dodson *et al.*, 1976). Apart from the planarity of the main-chain peptide units, many amino acids contain a planar atomic arrangement within their side chains. Four of these, histidine, phenylalanine, tryptophan and tyrosine, have pronounced planar aromatic rings and account for over 10% of all amino acids in proteins in the Swiss-Prot database (Boeckmann *et al.*, 2003). In addition, every nucleotide in DNA and RNA contains a planar nitrogenous base. During the process of model building, the electron densities of planar aromatic side chains, *e.g.* that of tryptophan, are often easily identifiable owing to their size and

pronounced shape, even if the individual atoms are not resolved. This allows the crystallographer to anchor the sequence once the C^{α} backbone has been traced (see, for example, Pavelcik, 2004). In contrast to the protein peptide unit, which is misshapen owing to the steric influence of adjacent side chains (MacArthur & Thornton, 1996), the delocalization of the π electrons in aromatic ring systems permits only miniscule distortions from planarity. Studies of refined structures show that the r.m.s.d. from planarity is about 0.04 Å in proteins (Hooft *et al.*, 1996) and less than 0.001 Å in accurately determined high-resolution small-molecule structures containing nitrogenous bases (Clowney *et al.*, 1996).

Automated means of plane detection have already been researched outside the field of structural biology. Sarti & Tubaro (2002) used the Hough transform (Illingworth & Kittler, 1988) to find planar fractures in rocks. However, the Hough transform is ill-suited to finding aromatic rings in macromolecular electron density because the size of the rings is small compared with the size of the whole molecule. A common alternative is template matching, in which the local electron density is compared against a library of known structural fragments. *ARP*/*wARP* (Lamzin & Wilson, 1997), for example, uses density templates for the identification of planar peptide units in protein chain tracing.

In this paper, we present a pattern-recognition-based method for the direct detection of planar fragments during the interpretation of electron density. The method uses the relationship between the pattern of atomic arrangement to be recognized and the shape of the electron density surrounding it. We extract a set of numerical values, called features, from spherical regions in a density map. As the number of features is much smaller than the number of density points from which they are derived, a feature vector serves as a compact representation of the local electron-density shape. Matching a feature vector from a density region to a corresponding feature vector derived from a training set provides the interpretation of the planar density shape. This can in turn be exploited for subsequent model building, either manual or automated.

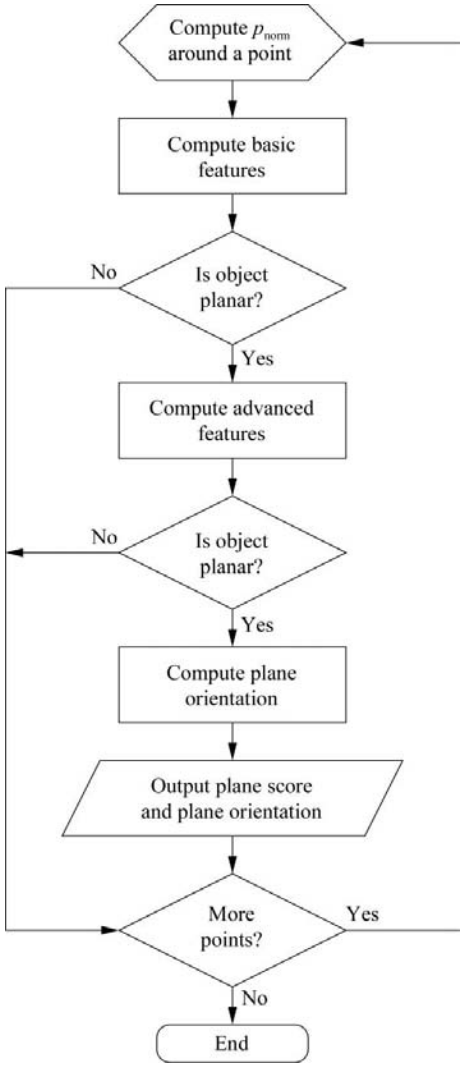

Figure 1
Flowchart of the plane-detection procedure. For each grid point in a density map, a score corresponding to the centre of a planar object is computed.

2. Methods

2.1. Search volume

The method described here finds the location and the orientation of planar aromatic ring structures in three-dimensional space and is outlined in the flowchart in Fig. 1. The required input is an electron-density map covering the crystallographic asymmetric unit. The electron density is interpolated on a cubic grid with a default spacing of 0.6 Å. The grid spacing reflects a trade-off between desired accuracy and computational cost and can be reduced for maps computed at a resolution higher than about 1.5 Å. To relieve the pattern-recognition algorithms from dealing with crystallographic symmetry, the map is extended around the asymmetric unit by 3.0 Å, the radius of the spherical region from which the features are extracted.

2.2. Normalization of the raw density values

The electron-density map is usually computed on an arbitrary scale with its mean value set to zero. The density value at any particular point only approximates the actual number of electrons per unit volume. The extreme values, both positive and negative, are not bounded and are often erroneous owing to computational peculiarities. Thus, these raw density values are not directly suitable for pattern-recognition methods.

Given the nature of macromolecules and the way that X-rays interact with the atoms in the unit cell, it is reasonable to assume that the higher the value of the electron density at a certain point, the higher the likelihood that the point belongs to the region containing ordered atoms. The reverse assumption that lower electron density signifies lower likelihood does not necessarily hold. Low density values can be found both in between ordered atoms and in the disordered solvent region. Hence, the likelihood that low electron-density values belong

to the ordered region can also be high (Urzhumtsev *et al.*, 1989). Abrahams & Leslie (1996) have also utilized this fact, although in a different formulation, for calculating a mask that encompasses the molecule. The calculation of the mask is based on the variation of the density in the ordered region, which differs from the more uniform distribution of the disordered solvent (see example in Fig. 2*a*).

We make use of the above for the normalization of the electron density $\rho(\mathbf{x})$ as follows. Firstly, we create the density histogram, shown in Fig. 2(*b*), which gives the frequency $p(\rho)$ at which a certain density value ρ occurs. The same information can also be provided by a cumulative density distribution or the integral probability, $P(\rho)$, which gives the frequency of observing raw density values less than ρ (Fig. 2*c*). We now define the normalized density value $p_{\rm norm}$ as the fraction of grid points with a density value ρ and an integral probability $P(\rho)$ that lie in the ordered region (Fig. 3). These dependencies were obtained from about 10 000 solved structures with different resolutions and solvent contents and allowed estimation of the values of $p_{\rm norm}$ with an average accuracy of 5%.

The quantity $p_{\rm norm}$ has the properties of a suitable weight for pattern-recognition techniques: it has bounds, it is non-negative and it reflects the significance of a certain map point for building an atomic model. We also note the relatively high values of $p_{\rm norm}$ at low values of the integral probability and that the strength of this effect depends on the resolution of the data.

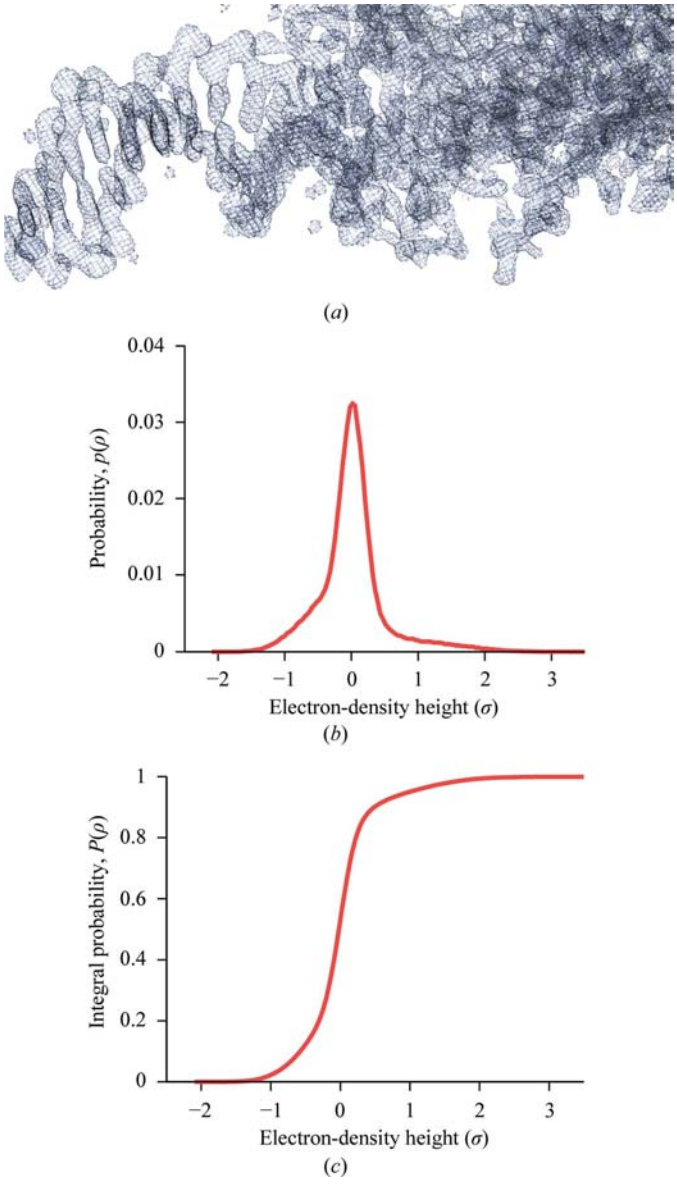

Figure 2
Electron density and its distributions for the training case (PDB code 1mur). (*a*) Part of the electron density contoured at 1.5σ above the mean; (*b*) the histogram, a probability function, of the electron-density height, $p(\rho)$; (*c*) its integral probability function, or cumulant, $P(\rho)$.

2.3. Feature selection

Within a certain radius around each point inside the asymmetric unit, rotation-invariant scalar features are computed from the normalized density values and stored as a feature vector. Features are trained on the local statistical and geometric properties expected for density patterns containing aromatic rings. The features are divided into two categories, basic and advanced, where the advanced features are calculated only if the basic features are inconclusive (see Table 1 and below).

2.3.1. Moments and moment invariants of the density distribution. The variations in the electron density can be exploited by the use of moments. It is known that a well behaved probability density function can be uniquely described by exactly one infinite set of spatial moments (Hu, 1962). This forms the basis of the pattern-recognition technique employed here: if an unknown region of electron density yields moments similar to those computed from a known pattern, the atomic structure underlying that region

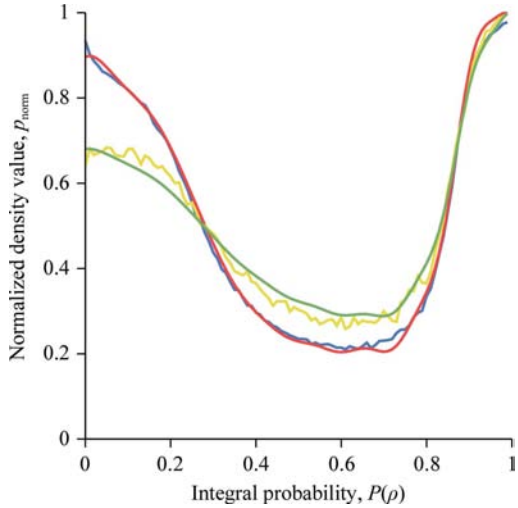

Figure 3
Likelihood of a point belonging to the ordered region as a function of its integral probability. Experimental and modelled curves for a protein at 1.9 Å resolution with solvent content 59% (PDB code 1j4a) are shown in blue and red, respectively. The corresponding curves for a protein at 2.5 Å resolution with solvent content 40% (PDB code 1onl) are shown in yellow and green.

Table 1
The features used for plane detection.

The discriminative power for the advanced features is computed from data first filtered using the basic features as described in the text. Owing to the correlation between the features, the relative discrimination values do not sum to 100%.

Feature type	Description	Discriminative power (7)	Relative discrimination (%)
Basic features	First and second order central moments of the normalized density (1) and (2)	0.01	4
	Radial moments of order $1 \leq q \leq 4$ (3)	0.03	13
Advanced features	Squared distance from the centre of mass to the centre of the search volume (Holton *et al.*, 2000)	0.03	11
	Isotropically weighted 12 moment invariants (4) (Lo & Don, 1989)	0.12	44
	Eigenvalues λ_1, λ_2 and λ_3 and their ratios λ_3/λ_1, λ_3/λ_2 and λ_2/λ_1	0.08	30
	Anisotropically weighted 12 moment invariants (4) (Lo & Don, 1989)	0.15	56

Table 2
Classification of the planar objects.

Class	Members	Description
C_1	His	Small single-ring structures
C_2	C, T, U, Phe, Tyr	Large single-ring structures
C_3	A, G, Trp	Double-ring structures
C_0	Anything else	Noise

can be regarded as similar to that of the known pattern. A general expression for the qth-order moment of $p_{\rm norm}$ over the search volume V is

$$m_q = \int_V p^q_{\rm norm}(\mathbf{x}) g(\mathbf{x})\, {\rm d}\mathbf{x}, \quad (1)$$

where $g(\mathbf{x})$ is a weighting factor. The local variance, which serves as a powerful numeric feature to separate zones of protein and solvent (Abrahams & Leslie, 1996; Terwilliger & Berendzen, 1999), is a special case of (1) with $g(\mathbf{x}) = 1$,

$$m_2 = \sigma^2 = \int_V [p_{\rm norm}(\mathbf{x}) - \overline{p}_V]^2\, {\rm d}\mathbf{x}, \quad (2)$$

where $\overline{p}_V$ denotes the mean normalized density in V. The local variance does not contain information about the spatial arrangement of the electron density in a volume V. One simple way to introduce shape-dependence is to weight the density with the radial distance by setting $g(\mathbf{x}) = ||\mathbf{x} - \mathbf{x}_c||^q$, where $\mathbf{x}_c$ is the centre of the search volume,

$$m_q = \int_V p^q_{\rm norm}(\mathbf{x}) ||\mathbf{x} - \mathbf{x}_c||^q\, {\rm d}\mathbf{x}. \quad (3)$$

Now let us consider the case of $q = 1$, $g(\mathbf{x}) = x^l y^m z^n$ in (1). For any non-negative integers l, m and n, the equation

$$m_{lmn} = \int_V p_{\rm norm}(\mathbf{x}) x^l y^m z^n\, {\rm d}\mathbf{x} \quad (4)$$

defines the raw spatial moments of order $l + m + n$. While these moments capture the spatial characteristics of the electron density inside the search volume, they are not invariant under rotation. Sadjadi & Hall (1980) developed techniques to transform second-order three-dimensional central moments to moment invariants. A set of 12 moment invariants up to order three can be derived using group-theoretic techniques. As the derivations are too unwieldy to be reproduced here, the reader is referred to Lo & Don (1989) for further details. These latter moment invariants are used as features for recognition of planar aromatic rings (Table 1).

2.4. Classification

Once regions of density are represented as numeric feature vectors, we use a linear discriminant analysis (Fisher, 1936) to assign them to an appropriate class. A linear discriminant is a vector $\mathbf{w} = [w_1\ w_2\ \ldots\ w_n]^T$ such that, when applied to any normally distributed feature vector $\mathbf{f}$, the scalar product

$$g(\mathbf{f}) = \mathbf{w}^T\mathbf{f} + w_0 \quad (5)$$

predicts its class [see, for example, Bishop (1995) and Morris (2004) for details]. In essence, Fisher's discriminant maximizes the ratio of the square of the distance between the class means to the within-class variances along the direction $\mathbf{w}$. For convenience, an additive bias parameter w_0 is introduced into (5) to set the mean value of $g(\mathbf{f})$ to zero.

We designed three discriminant vectors $\mathbf{w}$ for each type of planar object, *i.e.* planes with one small ring, one large ring and two rings (Table 2). The discriminants were trained on data from a single protein–DNA complex as discussed in §2.6.

2.5. Finding the plane orientation

To find the plane orientation, we use the eigen decomposition of the estimated covariance matrix,

$$\mathbf{COV} = \begin{pmatrix} m_{200} & m_{110} & m_{101} \\ m_{110} & m_{020} & m_{011} \\ m_{101} & m_{011} & m_{002} \end{pmatrix}, \quad (6)$$

computed around the centre of mass of the volume V. The three eigenvalues, $0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3$, measure the variance of the normalized density along the corresponding orthogonal eigenvectors. As planar objects mainly vary in two orthogonal directions, one would expect $\lambda_1 << \lambda_2 \leq \lambda_3$, which means that the ratios of the eigenvalues can also be used as features (Table 1). Furthermore, the eigenvector $\mathbf{v}_1$ corresponding to the smallest eigenvalue λ_1 defines the normal of the best (in a least-squares sense) fitting plane through the density.

Some planar objects in the structure may be located close to each other. For example, the electronic arrangement of the bases in nucleic acids favours stacking of the nucleotides along a strand (Fig. 4). Stacked bases are arranged parallel to each other along the direction of the strand, with a base-to-base distance of around 3.5 Å. It becomes clear that a sphere large enough to encompass the density of a whole purine base will also include the density from one of the neighbouring stacked

bases. This in turn may confuse the solutions of the eigen decomposition and the direction of minimum variance, $\mathbf{v}_1$, will not necessarily be along the plane normal. Therefore, we first compute a plane estimate by exponential downweighting, $\exp(-\|\mathbf{x}_i - \mathbf{x}_c\|^2)$, of the electron density at each point $\mathbf{x}_i$, where $\mathbf{x}_c$ is the centre of the sphere. The covariance matrix (6) is then recomputed, with the downweighting applied anisotropically, $\exp[-\|\hat{\mathbf{v}}_1\cdot(\mathbf{x}_i - \mathbf{x}_c)\|^2]$, along the unit normal of the plane estimate, $\hat{\mathbf{v}}_1 = \mathbf{v}_1/\|\mathbf{v}_1\|$ (Fig. 4).

2.6. Training

During training of the discriminant, spherical volumes centred within 1 Å of the true plane centre are classified as signal. The weights for Fisher's projection (5) for all three classes (Table 2) are derived from an (F_{obs}, φ_{calc}) map of the previously solved structure of Tn*5* transposase bound to an outside-end DNA duplex (PDB code 1mur; Lovell *et al.*, 2002). This is a protein–DNA complex refined to 2.5 Å resolution, with 455 amino acids including ten histidines, ten phenylalanines, 13 tryptophans and ten tyrosines. The 40 nucleotides of double-stranded DNA are distributed as eight GC pairs and 12 AT pairs. Prior to training, the structure was subject to ten cycles of stereochemically restrained refinement with *REFMAC* (Murshudov *et al.*, 1997) using default refinement parameters and no cutoff on the X-ray data.

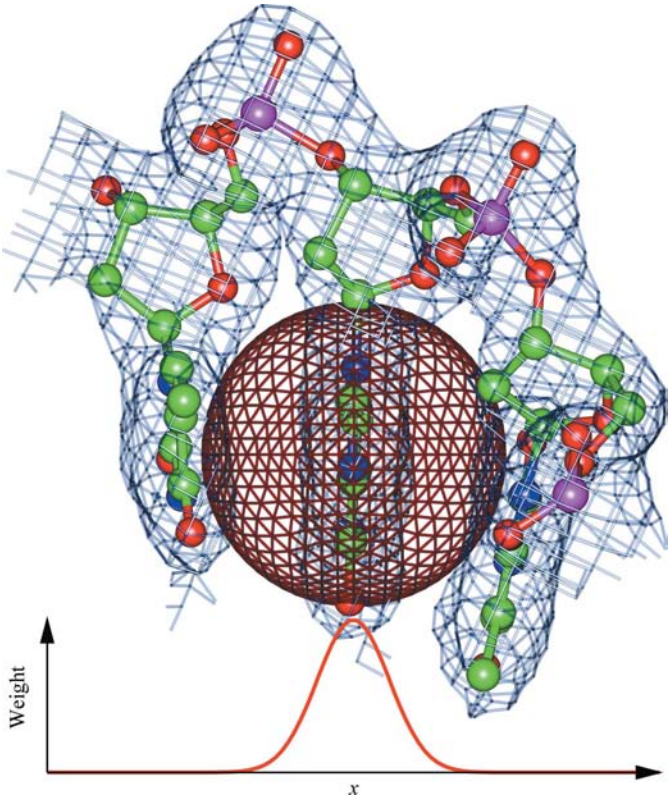

Figure 4
A sphere of 3 Å radius, shown as a red wire frame, centred on the guanine of a d(TGT) triplet. The electron density, shown in blue, is contoured at 1.7σ above the mean. The sphere intersects the density from the stacked bases. The red curve shows the anisotropic weight applied to the points inside the sphere in the direction along the normal of the base.

3. Results and discussion

3.1. The discriminant and the optimum search radius

Spherical volumes of different radii capture different information about the underlying pattern. Intuitively, the optimum diameter of the search volume should be equal to the size of the largest search pattern, *i.e.* 6.6 Å in the case of guanine. However, owing to the specifics of a macromolecular structure, *e.g.* the base stacking shown in Fig. 4, large search volumes may result in poorer discrimination of planar objects. Therefore, we determined the optimum radius empirically by evaluating its discriminative power,

$$D = \int_{-\infty}^{\infty} h(g)\frac{N\cdot h(g)}{N\cdot h(g) + N_0\cdot h_0(g)}\,\mathrm{d}g, \qquad (7)$$

where $h(g)$ and $h_0(g)$ are the probability density distributions for the projection value (5) for all signal classes taken together and for the noise class, respectively. N and N_0 are the sizes of the signal classes and the noise class.

Fig. 5 presents the results of the search for the optimum radius for all three signal classes. The discriminative power increases until the radius reaches the value of 3.0 Å. For larger radii the discriminant decreases sharply. We think that small volumes are inconclusive for plane detection, while volumes with radius higher than 3.0 Å start including surrounding electron density, which confuses the pattern recognition.

3.2. Filtering out the signal

3.2.1. Interpretation of the discriminant analysis. For a protein with a molecular weight of 100 kDa, the feature vectors are computed at over one million grid points of the density map, but only about 0.1% of these points lie in the vicinity of a planar fragment. Thus, the task of identifying a planar object is to find one correct solution among about 1000

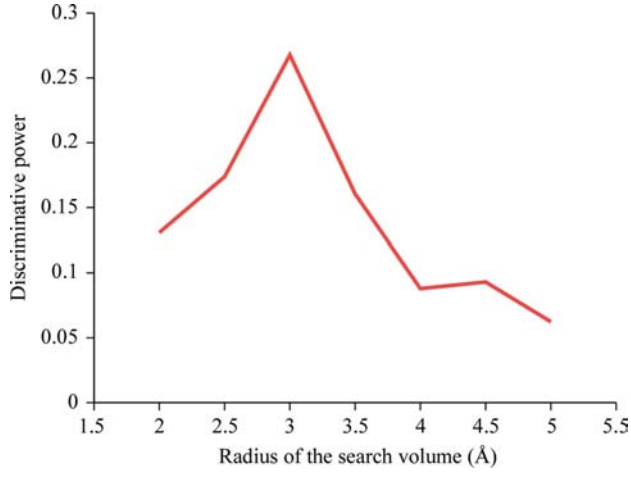

Figure 5
Discriminative power (7) for the advanced features as a function of the radius of the search volume.

candidates. A random classifier would provide a discrimination of no better than $N/(N + N_0) = s/(s + 1)$, where s is the signal-to-noise ratio. For convenience, hereafter we use $N_0/(N + N_0)$, the noise level, which ranges from zero to one, rather than the signal-to-noise ratio, which can vary from zero to infinity.

The Fisher discriminant described in §2.4 uses the features to filter the signal from the noise. Usually, over 80% of the mostly noise points are discarded after inspection of the basic features (Table 1). This already yields a fivefold increase in the relative amount of signal, bringing the noise level to 0.995. Subsequent use of the advanced features provides a further increase.

The final result is a trade-off between the completeness of the signal (the fraction of the signal retained in the output) and the amount of accepted noise. If, for example, we would like to have all signal points in the output, we have to accept a high amount of noise, also called false positives. Indeed, if we choose all solutions with a Fisher projection above −0.05, then all true class C_3 planes (green curve in Fig. 6) will be selected and thus the completeness of the solution will be 1.0. However, the amount of accepted noise (red curve) will be 150 times the amount of signal, corresponding to a noise level of 0.993, which is not much better than the 0.995 obtained from the use of the basic features alone. If instead we were to choose solutions with the value of the Fisher projection above 0.25 (Fig. 6) this would entirely eliminate the noise. However, the completeness of the signal would only be about 0.006. Clearly, none of these ways of filtering out the signal are satisfactory.

We therefore use a compromise approach and choose a threshold on the Fisher projection that corresponds to some small value of the activation function, shown in blue in Fig. 6, which provides a solution with reasonably high completeness and low noise level.

3.2.2. Kernel smoothing and peak picking. The result is a set of points in three-dimensional space which generally cluster around the centres of the true planar objects. To reduce these clusters to single solutions, each point is spatially smoothed using a three-dimensional Gaussian kernel with a total variance of 1.5 Å^2. A peak search is then performed in the smoothed 'map' with the constraint that no two peaks are allowed to be closer than 2.5 Å to each other. Each solution is given a score, which is the ratio of its height to the height of the top solution.

The completeness and noise level as a function of the solution score are shown in Fig. 7. The two pairs of curves corresponding to the two different cutoffs on the activation function (black vertical lines in Fig. 6) are essentially identical. This shows that the method is almost insensitive to the precise value of the activation-function cutoff. As can be seen from Fig. 7, there is a sharp drop in the completeness for scores above 0.6. Hence, 0.6 is used as a second cutoff which determines the final set of solutions.

3.3. Test cases

Several test cases were used to assess the performance of the method under various conditions. All models, except for case 5, were refined for ten cycles with *REFMAC* using default parameters and no resolution cutoff before density-map computation and the plane search were attempted. In order to evaluate the accuracy of the method, the detected planar objects were compared with those in the deposited models. The results are summarized in Table 3.

3.3.1. Case 1, a protein–RNA complex. The first test case is a complex of a sarcin homologue bound to an analogue that

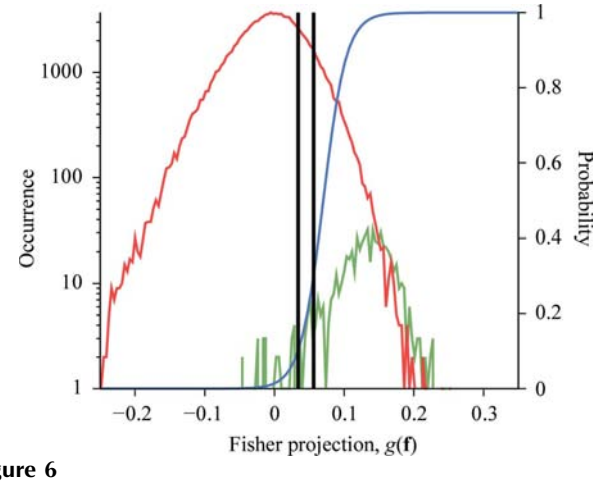

Figure 6
Threshold selection for discrimination of double-ring planar objects, class C_3, in the training case (PDB code 1mur). The green and red curves show the distribution of the signal and the noise, respectively. The total fraction of the signal is only 0.005. The blue curve shows the activation function, which reflects the probability of a point belonging to the signal class as a function of $g(\mathbf{f})$, under the assumption that the signal and noise classes are of equal size. The solid vertical lines, corresponding to values of the activation function of 0.1 and 0.3, indicate suitable thresholds for further signal filtering.

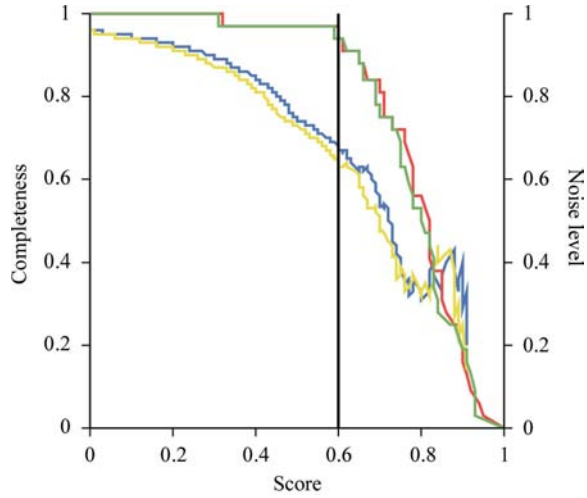

Figure 7
Completeness and noise level as a function of the score for the C_3 class in the training case (PDB code 1mur). The red and green curves show the completeness for the thresholds of the activation function at 0.1 and 0.3, respectively. The blue and yellow curves show the corresponding noise level. The solid vertical line indicates a suitable compromise between completeness and noise level for selecting the final solutions.

Table 3
Performance of the plane-detection method.

For each class, C_1, C_2 and C_3, the number of detected planar fragments and the total number in the final model is quoted. The accuracy of the detected planar objects for the protein and the DNA/RNA regions is also given.

										R.m.s.d. from the centres of the located planes to those in the final model (Å)		Angular difference between the located planes and those in the final model (°)	
Test case	Resolution (Å)	Wilson B (Å^2)	Solvent content	C_1	C_2	C_3	False positives	Completeness	Noise level	DNA/RNA	Protein	DNA/RNA	Protein
Case 1, 1jbs	2.0	29	0.47	11/16	36/52	38/38	28	0.80	0.25	0.5	0.3	7	8
Case 2, 1p7h	2.6	59	0.59	34/48	91/98	37/38	356	0.88	0.69	0.5	0.4	5	10
Case 3, 1p7h (no DNA)	2.6	59	0.59	20/48	70/98	16/38	355	0.58	0.77	0.6	0.5	8	15
Case 4, 1xok	3.0	62	0.57		10/20	16/18	34	0.68	0.57	0.6	0.3	7	21
Case 5, 1xbr	2.5	45	0.56	7/14	46/52	28/32	86	0.83	0.51	0.6	0.5	8	13

mimics the target loop of rat 28S RNA (PDB code 1jbs; Yang *et al.*, 2001). The structure was determined at 2.0 Å resolution with two NCS-related copies in the asymmetric unit. Each of the two RNA 29-mers contains seven adenine, nine cytosine, nine guanine and four uracil nucleotides. The protein part has 2×149 residues with 58 planar side chains.

The procedure correctly identified all double-ring planes (class C_3). The completeness of the identification of large planes was lower at 0.69. Overall, 80% of the planes were identified and the noise level was very small (Table 3).

3.3.2. Case 2, a protein–DNA complex. This test case is a complex of the NFAT1 dimer bound around 15 base pairs of double-stranded DNA at 2.6 Å resolution (PDB code 1p7h; Giffin *et al.*, 2003). There are two copies of the 286-residue protein part in the asymmetric unit containing 124 planar side chains. The DNA part contains 16 AT pairs and 14 GC pairs.

162 planes of 184 in the structure were identified, yielding an overall completeness of 0.88. The histidines (class C_1) were the poorest identified, with a 71% success rate. The recognition of large planes was noticeably better at 93%. Essentially all of the double-ring objects were found. The overall noise level was high at 0.69. Most of the planes which are not found were the same in NCS-related parts of the model. This indicates that the method is dependent on the local quality of the density map.

3.3.3. Case 3, a partial model of a protein–DNA complex. Here the model from case 2 was used but the DNA part and the solvent were removed. 312 solvent sites were then iteratively added using *ARP*/*wARP* (Lamzin & Wilson, 1997). The purpose of this experiment is to assess the performance of plane detection in the DNA region before the nucleic acids are included in the model. Thus, case 3 mimics a real-life structure determination from molecular replacement.

Fewer large-ring planes (71%) were built since the density corresponding to the DNA region was unbiased and less clear. An even smaller amount of the small and double-ring planes were detected. The overall completeness was reduced to 0.58 and the noise level was higher at 0.77.

3.3.4. Case 4, a protein–RNA complex. Here, we use the structure of an RNA–peptide complex of the alfalfa mosaic virus (PDB code 1xok; Guogas *et al.*, 2004) determined at 3.0 Å resolution. The RNA part consists of 36 modelled nucleotides, which are split into two strands. There are 33 amino acids in two protein chains, where the only residues with planar side chains are two tyrosines.

Half of the large-ring planes and almost all double-ring planes were identified. In spite of the 3.0 Å resolution of the data, the overall performance of the plane detection was almost as good as in case 1, which is also a protein–RNA complex but at 2.0 Å resolution. This can probably be attributed to the fact that there are almost no planes in the protein part, which are the most difficult ones to determine using the presented method.

3.3.5. Case 5, initial map for a protein–DNA complex. This is a structure of the T domain from *Xenopus laevis* complexed with a 24-nucleotide palindromic DNA duplex (PDB code 1xbr; Müller & Herrmann, 1997). The 367-residue protein part contains 50 planar side chains. The structure was solved at 2.5 Å resolution by multiple isomorphous replacement using three iodinated DNA derivatives and one selenomethionine derivative. After twofold NCS averaging, solvent flattening and histogram matching, the map correlation coefficient was 79%. This map was input to the plane-detection procedure before any model building was attempted and thus this case represents a real-life example of structure determination.

The method correctly identified 83% of the planes, with almost all double-ring objects being found. This is similar to cases 1 and 2. The histidines (class C_1) have a lower success rate; their completeness is only 0.50. The overall noise level is about 0.50.

3.4. Performance of the method

The procedure uses a weighted combination of six different types of pattern-recognition features, where each feature provides its own contribution (Table 1). Of the advanced features, the most powerful are the rotation-invariant moments and, to a lesser extent, the eigenvalues and their ratios. The squared distance from the centre of mass is less informative. Taken together, the advanced features are responsible for most of the discriminative power of the procedure.

The completeness of the method, *i.e.* the fraction of the signal retained in the output, decreases slightly as the reso-

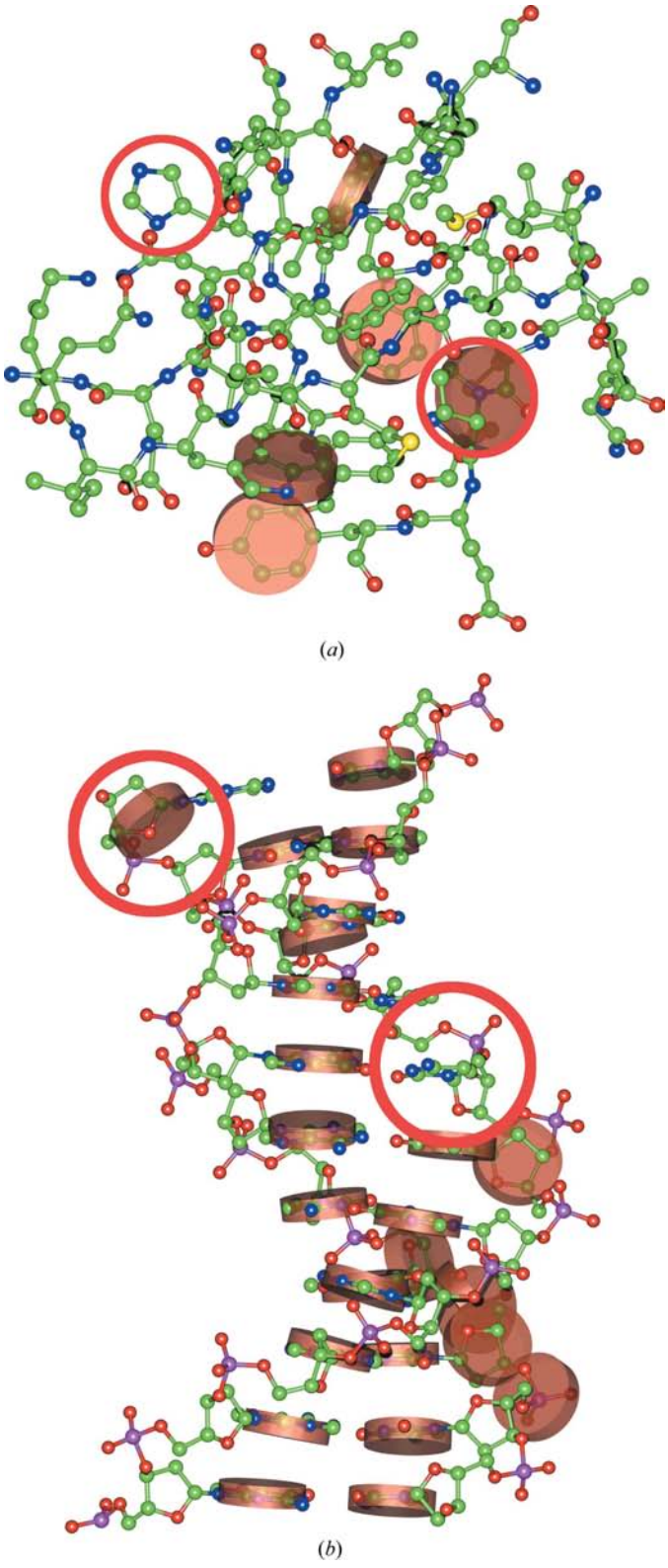

Figure 8
Plane detection for case 2 (PDB code 1p7h). The locations and orientations of the detected planes are indicated by the brown discs. (*a*) Part of the protein region. Circles in red point to an undetected histidine (top left) and an erroneously detected proline ring (middle right). (*b*) A region around the base pairs of double-stranded DNA. Red circles show an undetected cytosine (middle right) and a detected ribose which is not a part of any of the signal classes (top left).

Table 4
Distribution of false positives for the signal classes.

The average from the five test cases in Table 3 is shown.

Class	Planar protein side chains: Arg, Asn, Asp, Gln, Glu (%)	Prolines (%)	Peptides (%)	Riboses (%)	Remaining noise (%)	Total (%)
C_1	38	10	6	6	40	100
C_2	0	7	5	45	43	100
C_3	5	5	10	40	40	100

lution of the data is lowered (Table 3). The noise level increases correspondingly, as can be seen from the number of false positives. However, these results will need to be revisited when larger training sets, ideally corresponding to different resolutions and Wilson *B* factors of the data, are employed. This is discussed in more detail in §4.

As can also be seen from Table 3, the accuracy in the detected plane centre is relatively insensitive to resolution. We attribute this to the fact that the chosen pattern-recognition features capture general aspects of local electron density such as symmetry, planarity and volume. The planes in the protein region are located with a lower r.m.s.d. (∼0.4 Å) compared with those in nucleic acids (∼0.6 Å). At the same time, their orientation is less accurate (about 10–20°), while the accuracy of the orientation for the planes in DNA/RNA is less than 10°. This is presumably related to the smaller size of the planar objects in proteins, which contain more single-ring and fewer double-ring planes. Indeed, as planes increase in size it becomes more difficult to define their centre, but once this is achieved their orientation is more precisely determined.

The relatively high noise level contained in the final solutions, rising from 0.25 in test case 1 to 0.77 in case 3, is not as bad as it may look. We thoroughly examined all these false positives and found that many of them do correspond to planar or near-planar objects, which in this work were not explicitly defined as signal classes (Table 4). The classification for C_1 (histidine residues) shows a tendency to pick up other small planar side chains and puckered proline rings, which together account for about 50% of the classified noise. The classes C_2 and C_3 (large and double rings) often identify the ribose in nucleic acids. Examples of false positives as well as false negatives are shown in Fig. 8.

4. Outlook

The chosen features provide a flexible means of pattern recognition and are able to detect planes of three different types. Some inflexibility comes from the fixed radius of the search volume, which is optimized for the overall discrimination of all three signal classes simultaneously (Fig. 5). Tuning the radius of the sphere to each signal class could perhaps provide higher performance, but the associated increased CPU expenses will have to be assessed.

In many density-modification approaches, the natural difference in local variation of electron density is exploited in order to improve the phase quality. For example, the solvent

region tends to be flat, while the region of ordered matter exhibits high deviation from the local mean. However, noise is also characterized by high local variance. While the density normalization employed in this method captures some of the information contained in the variance, other design choices were made to explicitly introduce noise tolerance. For example, moment invariants of order higher than three are currently avoided, although their potential to capture high-frequency variations needs thorough investigation.

In the current implementation, a poor initial orientation of the plane will not improve by iterating the eigen decomposition. Therefore, a localized implementation of the Hough transform (Illingworth & Kittler, 1988) could be revisited since the number of planar objects in a search volume suitable for detection of planes is small enough for the Hough transform to produce a ranked list of candidate orientations.

As could have been expected from a stereochemical point of view, no difference in the performance is observed for plane detection in DNA and RNA structures. Nevertheless, the noise level is dependent on the data quality and resolution; thus, the differences in the diffraction properties of DNA and RNA may affect the method's real-life performance.

The current serial implementation, where the plane score at each grid point is computed independently from that at any other grid point, delivers plane orientations and locations within minutes on a modern workstation. A future parallel implementation as well as further factorization could speed up the execution of the search considerably.

Advanced construction of the training set is another implementational topic of the scientific problem described in this manuscript. Nevertheless, we should comment that our training set was based on one structure only. Therefore, it cannot be truly applicable to structures with a large diversity in data resolution, solvent content or phase quality. In addition, the training set is somewhat biased towards nucleic acids, which explains the noticeably lower success rate for histidines (class C_1). However, even for the class of double-ring planes the problem is underdetermined, as the Fisher weights for all features (37 parameters) are derived from only 33 class C_3 contributors in the training set. The method trained on a larger set will certainly perform better. Furthermore, if we were to extend the signal classes to carboxylates, peptides, prolines and riboses, the noise level in the final solutions would be considerably reduced. The noise level could also be reduced if the outcome of this method is used for subsequent model building of nucleotides, proteins or ligands, for example, where additional stereochemical considerations may come into play. However, thorough discussion on planarity-based model building is beyond the scope of this paper.

We thank Dr Christoph Müller for the provision of the isomorphous replacement data for 1xbr.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003). *Nucleic Acids Res.* **31**, 365–370.

Brown, E. N. & Ramaswamy, S. (2007). *Acta Cryst.* D**63**, 941–950.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Šali, A., Studier, F. W. & Swaminathan, S. (1999). *Nature Genet.* **23**, 151–157.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 509–518.

Dodson, E. J., Isaacs, N. W. & Rollett, J. S. (1976). *Acta Cryst.* A**32**, 311–315.

Fisher, R. A. (1936). *Ann. Eugen.* **7**, 466–475.

Giffin, M. J., Stroud, J. C., Bates, D. L., von Koenig, K. D., Hardin, J. & Chen, L. (2003). *Nature Struct. Biol.* **10**, 800–806.

Guogas, L. M., Filman, D. J., Hogle, J. M. & Gehrke, L. (2004). *Science*, **306**, 2108–2111.

Holton, T., Ioerger, T. R., Christopher, J. A. & Sacchettini, J. C. (2000). *Acta Cryst.* D**56**, 722–734.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *J. Appl. Cryst.* **29**, 714–716.

Hu, M.-K. (1962). *IEEE Trans. Inf. Theory*, **8**, 179–187.

Illingworth, J. & Kittler, J. (1988). *Comput. Vis. Graph. Image Process.* **44**, 87–116.

Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 829–832.

Lamzin, V. S. & Wilson, K. S. (1997). *Methods Enzymol.* **277**, 269–305.

Lo, C.-H. & Don, H.-S. (1989). *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 1053–1064.

Lovell, S., Goryshin, I. Y., Reznikoff, W. R. & Rayment, I. (2002). *Nature Struct. Biol.* **9**, 278–281.

MacArthur, M. W. & Thornton, J. M. (1996). *J. Mol. Biol.* **264**, 1180–1195.

Morris, R. J. (2004). *Acta Cryst.* D**60**, 2133–2143.

Mowbray, S. L., Helgstrand, C., Sigrell, J. A., Cameron, A. D. & Jones, T. A. (1999). *Acta Cryst.* D**55**, 1309–1319.

Müller, C. W. & Herrmann, B. G. (1997). *Nature (London)*, **389**, 884–888.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Pavelcik, F. (2004). *Acta Cryst.* D**60**, 1535–1544.

Sadjadi, F. A. & Hall, E. L. (1980). *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 127–136.

Sarti, A. & Tubaro, S. (2002). *Signal Processing*, **82**, 1269–1282.

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 501–505.

Urzhumtsev, A. G., Lunin, V. Yu. & Luzyanina, T. B. (1989). *Acta Cryst.* A**45**, 34–39.

Yang, X., Gérczei, T., Glover, L. & Correll, C. C. (2001). *Nature Struct. Biol.* **11**, 968–973.