Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Determining protein structure from electron-density maps using pattern matching

Thomas Holton,[a] Thomas R. Ioerger,[b] Jon A. Christopher[a] and James C. Sacchettini[a]*

[a]The Center for Structural Biology, Department of Biochemistry and Biophysics, Texas A&M University, Mail Stop 2128, College Station, TX 77843-2128, USA, and [b]Department of Computer Science, Texas A&M University, Mail Stop 3112, College Station, TX 77843-3112, USA

Correspondence e-mail: sacchett@tamu.edu

Received 21 September 1999
Accepted 3 March 2000

TEXTAL is an automated system for building protein structures from electron-density maps. It uses pattern recognition to select regions in a database of previously determined structures that are similar to regions in a map of unknown structure. Rotation-invariant numerical values, called *features*, of the electron density are extracted from spherical regions in an unknown map and compared with features extracted around regions in maps generated from a database of known structures. Those regions in the database that match best provide the local coordinates of atoms and these are accumulated to form a model of the unknown structure. Similarity between the regions in the database and an uninterpreted region is determined firstly by evaluating the numerical difference in feature values and secondly by calculating the electron-density correlation coefficient for those regions with similar feature values. *TEXTAL* has been successful at building protein structures for a wide range of test electron-density maps and can automatically model entire protein structures in a few hours on a workstation. Models built by *TEXTAL* from test electron-density maps of known protein structures were accurate to within 0.6–0.7 Å root-mean-square deviation, assuming prior knowledge of C^{α} positions. The system represents a new approach to protein structure determination and has the potential to greatly reduce the time required to interpret electron-density maps in order to build accurate protein models.

1. Introduction

X-ray crystallography is the most widely used method for determining the atomic structures of proteins and other macromolecules. While there are many steps involved in structure determination, from collecting X-ray diffraction data to calculating phases for electron-density maps, the final process of interpreting an electron-density map is one of the most time-consuming and error-prone tasks. Model building is typically performed by a human crystallographer at a computer graphics terminal, with the help of molecular-visualization software such as *O* (Jones, 1978), and can take weeks to months of the researcher's time examining complex three-dimensional patterns of electron density. This process is complicated by a number of sources of noise that can perturb the density and obscure the underlying structure, such as low-resolution data, errors in phase estimates or inherently disordered regions (Richardson & Richardson, 1985; Brändén & Jones, 1990). As a consequence, model-building is often viewed as a 'labor of love' for the crystallographer and is one of the primary bottlenecks impeding the progress of structural biology. Increased automation in structure determination is critically needed for large-scale structural genomics projects

(Terwilliger & Berendzen, 1999; Bonanno, 1999) which aim to solve a wide range of protein structures in order to increase the understanding of complex biological systems and to explore the space of protein folds. Structure-based drug-design methods would similarly benefit from rapid access to new structures aided by automated methods for protein structure determination. To date, several automated methods have been developed (Jones *et al.*, 1991; Fortier *et al.*, 1997; Kleywegt & Jones, 1997; Leherte *et al.*, 1994; Holm & Sander, 1991; Levitt, 1992). However, these methods are typically limited to high-quality maps, requiring high-resolution data and/or near-perfect phase estimates for suitable success and accuracy in model building.

In this paper, we introduce a new approach to interpreting electron-density maps based on *pattern-recognition* methods, implemented in a program called *TEXTAL*. *TEXTAL* exploits the large number of protein structures previously solved by X-ray crystallography as a source of insight on how to solve new structures. The core principle underlying the pattern-recognition aspect of *TEXTAL* is that regions from two maps with similar patterns of electron density should have similar local molecular structures. We have developed a program which extracts characteristic numerical values that describe the patterns in a local region of an unsolved electron-density map. A related program then efficiently searches for similar patterns in a database of maps of previously solved structures. Regions that have similar density patterns are located in the database and atomic coordinates corresponding to these known regions are retrieved from the database, reoriented and appended to the growing model of the unknown.

The advantage of this pattern-recognition approach is that it can exploit the availability of natural regularities in protein structure (*e.g.* common backbone and side-chain conformations) as they occur in the database. *TEXTAL* also is able to exploit the natural bias in the database towards commonly occurring conformations (and density patterns), which facilitates the interpretation of regions of density in lower resolution maps (∼3 Å) where individual atoms might not be distinguishable, but only the overall shape and orientation of side chains can be seen. Ultimately, *TEXTAL* has the potential to reduce the time required to build complete models of large proteins from weeks to hours and may also enable interpretation of lower quality maps. In the remaining sections, we describe related work in computational crystallography, followed by the methods used by *TEXTAL* in more detail and then the results of several experiments in which *TEXTAL* was used to model proteins from both simulated and real electron-density maps.

2. Related work

Over the past 20 years, numerous computational procedures have been developed to assist the crystallographer throughout the structure-determination process. Methods are available for improving phase estimates, such as *SHARP* for heavy-atom parameter refinement (de la Fortelle *et al.*, 1997; Bricogne, 1997) or *Shake-n-Bake* (Miller *et al.*, 1994), which has been shown to identify accurate phases by direct methods for smaller proteins. Various programs are also available for masking and solvent flattening, Patterson correlation searches *etc*. However, the final process of interpreting a map and building a model for a protein structure remains a significant challenge for automation.

Table 1
Feature types and descriptions.

Feature type	Description	Number
Basic characteristics of spheres of density	Average density, distance from center of sphere to center of mass	2
Moments of inertia	Magnitude of primary, secondary and tertiary moments, ratios among moments	6
Statistical properties of density	Standard deviation, skewness, kurtosis	3
Spokes of density within spheres	Three spoke angles, three radial sums, sum of spoke angles, area of the spoke triangle	8

There are typically two steps to automated model building: skeletonization/main-chain tracing followed by side-chain construction. Skeletonization provides a framework for solving a structure by forming a tentative backbone trace for the initial map. Common skeletonization approaches include Greer's method, which uses a density threshold to define a continuous chain (Greer, 1985), and critical-point analysis (Leherte *et al.*, 1994), which analyzes the gradient in the density to identify likely locations of atoms. Other methods include core-tracing (Swanson, 1994) and *X-AUTOFIT* (Oldfield, 1997). Automated methods for building side chains include template matching of fragments using a database search followed by energy minimization (Jones & Thirup, 1986), an approach recently extended in the *MaxSprout* (Holm & Sander, 1991) and *Segment Match Modeling* (Levitt, 1992) algorithms. A similar approach also uses fuzzy logic to guess the sequence of residues in a region and (where possible) real-space refinement to improve the fit of side chains to density (Oldfield, 1997). These methods have been successful at building protein molecules to a moderate degree of accuracy, especially for high-resolution structures (1.0–1.5 Å r.m.s.d.).

Other approaches assist the crystallographer by positioning entire molecular structures within density. Such methods include template convolution (Kleywegt & Jones, 1997), which searches an unknown map in Fourier space for prototypical α-helices or β-strands, but requires anticipation of the correct prototype fragment and orientation of the structural element in the map, molecular scene analysis (Fortier *et al.*, 1997), which uses computer visual-processing routines to characterize geometric structures within a map, and knowledge-based methods such as *CRYSALIS* (Terry, 1983), which capture heuristics and human expertise about protein structure to interpret electron-density maps. Of particular note is *wARP* (Perrakis *et al.*, 1997), which combines phase improvement with model building by placing pseudo-atoms into the map and adjusting their fit to the density, recalculating

phases, reconstructing the map and iterating. *wARP* has been shown to build accurate models for a wide variety of proteins, although it is limited to interpreting only high-resolution maps ($\leq$2.4 Å resolution). In spite of the progress that has been made in solving structures automatically from high-resolution maps, it is imperative to develop new methods to automate the interpretation of larger proteins in lower quality maps owing to either low resolution and/or imperfect phases.

3. Methods

3.1. Overview

The method *TEXTAL* uses to create a model of a protein from an unsolved density map is briefly described as follows. First, a suitable database containing density maps for a large number ($\sim$200) of previously solved proteins is compiled. Likely positions for C^{α} atoms in the unknown map are identified (or, as in our initial tests, assumed to be known). Numeric values which describe the patterns of density around each α-carbon of the unknown are then calculated. In the pattern-matching field of computer science, the technical term for such numeric representations of patterns is *features* (Fayyad *et al.*, 1996; Asker & Maclin, 1997; Wisniewski & Medin, 1994; Duda & Hart, 1973) and we will use this term for such values throughout the rest of this manuscript. Appropriate features describing a region of electron density are quantities such as the average electron density, the moments of inertia *etc*., and are detailed below. The features of the unknown region are then directly compared with the features for all regions in the database. Regions from the database which have similar features to the unknown potentially have similar structures, so a more detailed comparison is made based on the correlation coefficient of the electron density in the two regions. The region from the database which has the highest correlation to the unknown region is treated as the best-matching region in the database. Atomic coordinates corresponding to the side chain for that best-matching region are then retrieved from the database, suitably oriented and added to the model. The process is then repeated for the next C^{α}, incrementally building a model for the unknown structure.

3.2. Descriptions of the electron-density features

The first step necessary for a pattern-recognition approach such as that employed by *TEXTAL* is the development of a set of features which describe the patterns in the data. Of the many possible features, not all are useful for the task at hand. For example, in a pattern-recognition algorithm designed to determine the make and model of a car from a photograph, the color of the car may not be a useful feature (many models may be painted the same shade of white), while the number of doors, the spacing between the wheels and the shape of the mirrors would be likely to be useful. For interpreting electron-density maps, we first determine a reasonable set of features which are likely to be useful for describing the patterns of density. To do so, an electron-density map is treated as a set of overlapping spheres of density containing information about regions of the protein structure. We have experimented with spheres of density that range from 3 to 6 Å in radius (Ioerger *et al.*, 1999). Using multiple radii permits *TEXTAL* to focus on different aspects of the local structure (*e.g.* side-chain rotamers, common backbone configurations, secondary structural characteristics). Because features are extracted and tabulated only once for regions in each protein in the database and unknown regions are compared to the database initially by feature value alone, *TEXTAL* is able to efficiently search a large database for regions with similar patterns of density. Such regions with similar density patterns are presumed to contain similar local molecular structures, which are then placed into the growing model.

Because protein structural elements can be positioned in any orientation, useful features of the electron density must be *rotation-invariant* (*i.e.* constant even when a pattern is rotated). 15 rotation-invariant numeric features have been developed which characterize patterns in the electron-density maps to be used in recognizing similar regions. There are four major categories of features (Table 1), with several different variants within each category, to give a total of 19 features for each radius. Since each feature is calculated at each of four different radii, there are a total of 76 feature/radius combinations. A description of each of the 19 basic features is given below.

Two features are used to describe the basic characteristics of the spheres. The first is the average density of the region (μ). If two regions of density are similar in structure, the average densities of the regions should also be similar,

$$\mu = (\sum \rho_i)/n,$$

where ρ_i is the density at lattice point i and n is the number of points in the region. Also, while the location of the center of mass of the region is not rotation-invariant, we use the *distance* from the center of mass of a sphere to the center of the sphere, where the center of mass $\langle c_x, c_y, c_z\rangle$ is given by

$$\langle c_x, c_y, c_z\rangle = \langle(\sum \rho_i x_i)/\mu, (\sum \rho_i y_i)/\mu, (\sum \rho_i z_i)/\mu\rangle$$

and

$$d_{\text{center}} = (c_x^2 + c_y^2 + c_z^2)^{1/2},$$

assuming the geometric center of the sphere is translated to $\langle 0, 0, 0\rangle$. These features are independent of orientation and were found to be important in the selection of good matches.

The second category consists of six different features all based on the moments of inertia in a region. The moments of inertia for a given region of density are measurements of the distribution of density in three dimensions. Each pattern of density has exactly one set of moments that describe the distribution of density around its center of mass. The primary moment lies along the path through the sphere around which the density is most widely distributed; the secondary and tertiary moments are orthogonal to each other and the primary moment and describe paths which have progressively narrower density distributions. Since the moments themselves are direction vectors with three components, we take the *magnitude* of the three moments of inertia as separate

features, in sorted order. Moments of inertia are calculated by constructing an inertia matrix,

$$I = \begin{vmatrix} \sum \rho_i y_i^2 + z_i^2 & -\sum \rho_i x_i y_i & -\sum \rho_i x_i z_i \\ -\sum \rho_i x_i y_i & \sum \rho_i x_i^2 + z_i^2 & -\sum \rho_i y_i z_i \\ -\sum \rho_i x_i z_i & -\sum \rho_i y_i z_i & \sum \rho_i x_i^2 + y_i^2 \end{vmatrix},$$

where ρ_i is the density at point i and x_i, y_i and z_i are coordinates of point i relative to the center of mass. The inertia matrix is diagonalized and the diagonal elements (eigenvalues) are sorted by magnitude to obtain the corresponding moments of inertia. The *ratios* of these moments provide additional information about the shape of the density (*e.g.* spherical, ellipsoidal) and are included as three more features.

Statistical properties of the density of lattice points within each sphere form a third category of features. The standard deviation (σ) is a sensitive description that varies widely throughout different distributions of data. Skewness $(1/n)[\sum(\rho_i - \mu)^3/\sigma^3]$ is a measure of the asymmetry in the distribution. Only a perfect Gaussian distribution has a skewness of 0.0; all others are either skewed positively or negatively. The kurtosis $(1/n)[\sum(\rho_i - \mu)^4/\sigma^4]$ describes the peakedness of the statistical distribution. Although difficult to visualize in three dimensions, these features are all rotation-invariant and it is expected that similar regions of density will have similar statistical characteristics.

A fourth category of features which is designed to take advantage of specific knowledge about protein structure describes the geometry of the density within each sphere. Given a sphere of density centered at an α-carbon of a (non-glycine) amino acid, there should be three major 'tubes' of density (like spokes on a wheel) projecting out from this point: one for the side chain and two for either direction of the main chain. The three spokes are defined as the vectors from the center to the surface of the sphere which have the maximum *radial sum*, with the caveat that the spokes be at least 75° apart. The radial sum is calculated as the sum of the densities evaluated at ten evenly spaced points along the length of spoke. Since evaluating all possible spoke directions is impractical, the surface of the spherical region is sampled at 320 evenly spaced regions (which result from successive subdivisions of an icosahedron), so 320 trial spokes are used; the three with the highest radial sum which are also at least 75° apart are defined as the spokes for that region. A higher number of samples on the surface of the sphere was investigated, but did not result in any improvement.

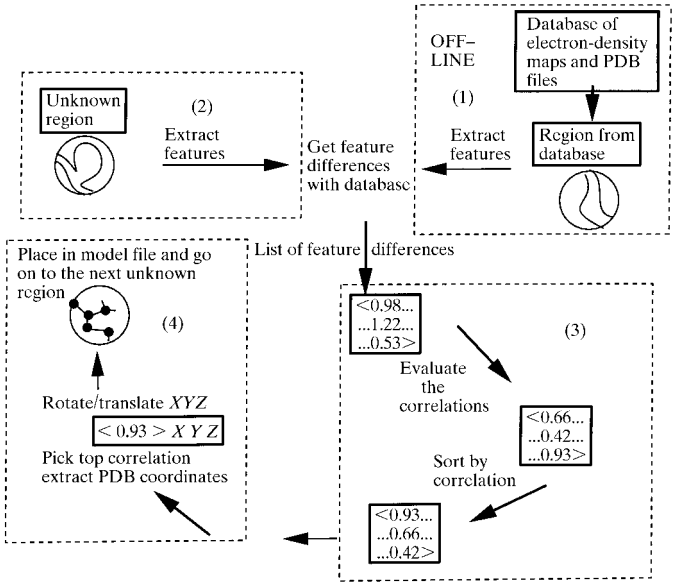

Figure 1
Schematic of the *TEXTAL* process.

To derive rotation-independent information about the arrangement of tubes of density within each sphere, the angles between the spokes are measured and the maximum, median and minimum spoke angles are utilized as features, as there should be similar angles between spokes in similar regions of density. Also, the sum of the angles is an approximate measure of the planarity of the three spokes (since the sum of the angles for co-planar spokes would equal 360°); hence, this allows us to use the sum of the spoke angles as another feature. Three additional spoke features developed for *TEXTAL* are the radial sum of each spoke; the final spoke feature is the area of the triangle with vertices formed by the endpoints of the three spokes.

3.3. Outline of the *TEXTAL* method

The core model-building procedure in *TEXTAL*, illustrated in Fig. 1, involves the following steps: (1) creating a database of known structures and extracting the features from regions in it, (2) for each region in the unknown, searching for regions in the database with matching features to create a list of candidate matches, (3) evaluating the candidate matches by density correlation and (4) assembling the model from the matched regions. The input required for *TEXTAL* is an electron-density map and a database of feature-extracted maps of known structures, which is created off-line from the *TEXTAL* model-building process (step 1 above). For a given test map, *TEXTAL* first extracts the features described above for the region under investigation in the uninterpreted map and compares these features with the pre-tabulated features for each of the regions in the database (step 2, referred to as the *lookup*). In our current implementation, the centers of the regions are selected as the locations of the C^{α} atoms. The similarity between two regions is evaluated by measuring the difference in the feature values for the two regions: ideally, the lower the difference, the more similar the regions. The total feature difference ΔF is defined as the weighted Euclidean distance between the feature values in the two regions,

$$\Delta F(R_i, R_j) = \{\sum w_k[F_k(i) - F_k(j)]^2\}^{1/2}, \qquad (1)$$

where F_k are the individual features and w_k is the weight associated with each feature. The weights are calculated separately and are discussed below.

The program then retains the top N matching regions based on feature evaluation, where N is a user-selectable parameter. These N regions are further analyzed for similarity by calculating the density correlation coefficient (step 3). Since a

density map is a discrete representation of a continuous three-dimensional density function sampled at lattice points i, the correlation coefficient (cc), may be calculated by

$$\mathrm{cc} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{[\sum (x_i - \overline{x})^2]^{1/2}[\sum (y_i - \overline{y})^2]^{1/2}}, \qquad (2)$$

where x_i and y_i are the densities in each region at each lattice point i and $\overline{x}$ and $\overline{y}$ are the average densities in the two regions. Since a region from an unknown map may be rotated relative to the known region, we define the true density correlation as the correlation coefficient obtained when the two regions are optimally aligned. The optimal rotation can be estimated using one of several methods, including a full three-dimensional search, a more directed peak-matching search or by aligning the moments of inertia, each representing a different trade-off of accuracy and efficiency. After investigating these methods, we found the simple approach of aligning the moments of inertia for the two regions to be both fast (up to 70 cc calculations per second) and accurate for regions with similar patterns of density.

This correlation coefficient as a measurement of similarity is appropriate for comparing spherical regions of a general density map. However, for regions centered on C^{α} atoms of amino-acid residues as employed in this study, a large portion of the density in the spherical region is not from the amino acid of interest, but instead from neighboring amino acids and long-range contacts. Therefore, a modification of the correlation-coefficient calculation is introduced which increases the discriminating capability of *TEXTAL* by limiting the density comparison to a cylinder which covers only the side chain and backbone of the current residue. A vector **v** which originates at the α-carbon of each region and points in the direction of the midpoint of the side-chain atoms of the current residue, referred to as the *side-chain axis vector*, was stored as part of the feature database.

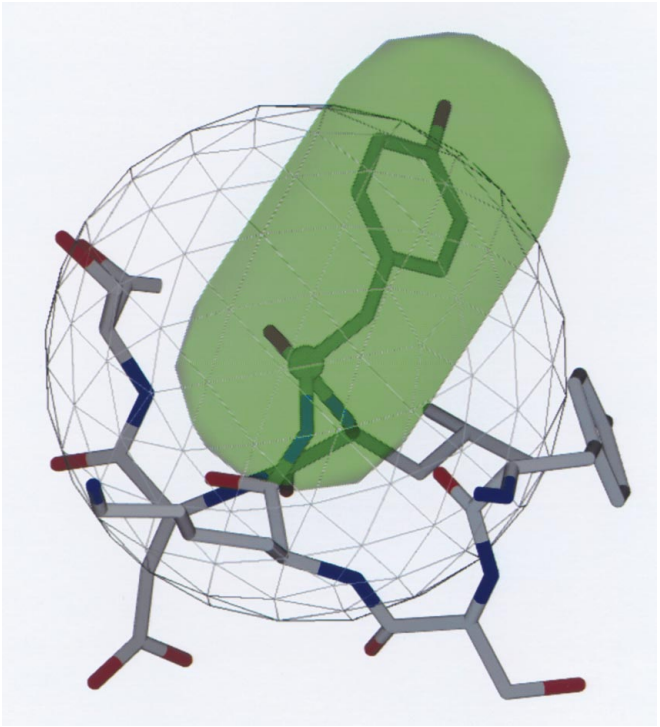

Figure 2
The side-chain axis correlation. The green cylinder is 5.0 Å in length, with spherical 2.5 Å endcaps. The axis lies in the direction from the C^{α} to the center of mass of the side-chain atoms. The wire-frame sphere, also centered at the C^{α}, has a radius of 5.0 Å. Restricting the correlation calculation to this cylinder helps make cc more accurately reflect the similarity between the two regions by focusing on the side-chain and backbone of the residue in question, without being affected by irrelevant differences in the density arising from neighboring residues or discontiguous parts of the structure that also enter the 5 Å sphere.

The *lookup* process involves superimposing regions *via* the moments of inertia. However, the correlation calculation is limited to the density in a round-ended cylinder 2.5 Å in diameter and 5.0 Å in length along the axis defined by this side-chain axis vector **v**, as demonstrated in Fig. 2. Empirical tests indicated that such a region is sufficient to cover the side chain and backbone of a single residue in most cases, which is the true region of interest. Using the side-chain axis correlation calculation significantly improves *TEXTAL*'s selection of candidate regions, presumably because the influence of neighboring disconnected atoms is eliminated. The improvement was found in both higher correlations (an increase of ~0.2) and in the accuracy of the matching regions by sequence identity/similarity (an improvement of between 10 and 20%). All correlation coefficients reported here are side-chain axis correlation coefficients.

The final step (step 4) involves selecting the region from the top N candidate matches which has the highest correlation coefficient. This region is considered to be the most likely region in the database to match the unknown region. The atoms from the amino acid and protein corresponding to this region are extracted from the database and rotated and translated into position using the same transformation matrix which was used to orient the density regions for comparison.

3.4. Database of electron-density maps

TEXTAL requires a database of previously interpreted electron-density maps as a source of example regions that associate density patterns with local molecular structures. Ideally, maps derived from measured structure factors and experimental phases (MAD or MIR) would be used, since they best represent the types of patterns likely to be encountered in solving a new map. These should become widely available in the near future as deposition of structure factors into protein databases such as the PDB continues. In our current experiments, we have used electron-density maps generated from atomic coordinates in PDB files. A set of simulated structure factors was created by Fourier transformation and then back-transformed into a density map using *X-PLOR* (Brünger, 1996).

These maps were produced using reflections from 10 to 2.8 Å resolution to make them representative of patterns in medium/low-resolution maps. Importantly, the transformation process used the temperature factors associated with each atom to produce weakened or diffuse density in less-ordered

parts of the structure such as arginine side chains or glycine main chains, which is often found in real MIR/MAD maps. Each map was created in a $P1$ space group with orthogonal axes, with approximately 1.0 Å grid spacing and a 5.0 Å border around the edges of the protein. Back-transformed maps are currently used to populate the database in *TEXTAL*. For these experiments, our database consisted of the first 200 proteins in the PDB Select, a list of unique well refined structures in the PDB with less than 25% similarity (Holm & Sander, 1993). At each of the C^{α} coordinates for all residues in these 200 structures (54 164 regions in total), features were extracted in spheres of 3, 4, 5 and 6 Å in radius. Below, we report results of using this database to model other back-transformed maps, as well as a real map derived from newly collected X-ray diffraction data.

3.5. Weighting of features

TEXTAL uses a weighted Euclidean feature-difference calculation (1) as an initial measure of similarity. It is important to weight the features in a Euclidean distance metric based on which features are most relevant because irrelevant features can confuse the pattern-matching algorithm (Langley, 1994; Aha, 1998; John *et al.*, 1994). Those features that are better at describing characteristics of density than other features should have a greater weight associated with them. To find appropriate weights that weight the relevant features in *TEXTAL* more heavily, the *Slider* algorithm was developed. We can measure the relevance of a feature by considering how similar it is between known pairs of matching regions, relative to random pairs of mis-matching regions. To quantify this, the *ranking quality* of a feature F is defined as *the average relative rank of matching regions in comparison to non-matching regions*. A set of m pairs of regions $\{(A_i, B_i)\}$ that are known to match, in the sense of having high density correlation (*e.g.* cc > 0.7), is used to estimate this empirically. Then, for each region A_i, a set of n other regions $\{C_{i,j}\}$ that do not match (cc < 0.7) is selected. Given these sets of data, the ΔF score (1) is determined for A_i with B_i and with each of the corresponding mis-matches $C_{i,j}$, then sorted based on ΔF, and the absolute rank r_i of the matching region B_i against all the others is obtained. The relative rank $\hat{r}_i$ is calculated by dividing by the total number of mis-matches, which normalizes it to a range of $[0 \ldots 1]$, and orienting it so that 1 corresponds to ranking the true match at the top (with smallest ΔF): $\hat{r}_i = (n - r_i)/n$. Finally, the ranking quality of the feature is determined by the average relative ranking score over all the matching pairs,

$$\mathrm{RQ}(F) = (1/m) \sum_i \hat{r}_i.$$

The definition of ranking quality extends to arbitrary distance metrics in addition to individual features; hence, we can measure the performance of given set of weights $\mathbf{w}$ by calculating its ranking quality in terms of how the linear combination of feature differences, weighted by $\mathbf{w}$, ranks true matches relative to mis-matches, RQ($\mathbf{w}$).

The approach *Slider* uses to optimize weights is based on a greedy algorithm (Russell & Norvig, 1995) that randomly selects one feature at a time and adjusts its weight against all the others simultaneously, with the aim of increasing ranking quality as much as possible. To begin, we start with a uniform weight vector $\mathbf{w}_0 = \langle \frac{1}{n}, \frac{1}{n} \ldots \rangle$, where n is the number of features. Then, with each iteration i, we randomly select a feature F_j, $1 \leq j \leq n$. Given the current weight vector at that time, $\mathbf{w}_i$, we construct a modified weight vector $\mathbf{w}'_i$ in which the weight for the selected feature F_j is set to 0 and the other weights are increased in proportion to maintain the property of summing to 1,

$$w'_{i,j} = 0 \text{ and } w'_{i,k} = w_{i,k}/(1 - w_{i,j}) \text{ for } k \neq j.$$

Given the selected feature F_j and the modified metric $\mathbf{w}'_i$, *Slider* uses the sets of matches and mis-matches to find the optimal combination of these two metrics as a binary mixture {here, $\Delta F_x(R_a, R_b)$ means $[F_x(R_a) - F_x(R_b)]^2$; we have dropped the square root, but the relative order of the feature differences is not changed},

$$\begin{aligned}\Delta_{\mathrm{mix}}(R_a, R_b) &= u \cdot \Delta F_j(R_a, R_b) + (1 - u) \cdot \Delta_{\mathbf{w}'_i}(R_a, R_b) \\ &= u \cdot [F_j(R_a) - F_j(R_b)]^2 \\ &\quad + (1 - u) \cdot \sum_k w'_{i,k} [F_k(R_a) - F_k(R_b)]^2.\end{aligned}$$

The goal is to find the value for $0 \leq u \leq 1$ that maximizes the ranking quality RQ(mix). Once the optimal value for u is determined, a new and improved weight vector $\mathbf{w}_{i+1}$ can be calculated for the next iteration as follows:

$$w_{i+1,j} = u \text{ and } w_{i+1,k} = w'_{i,k}/(1 + u) \text{ for } k \neq j.$$

The optimal value for u was calculated by solving simple linear equations for comparisons between various matching and non-matching regions. Suppose we have two distance metrics M_1 and M_2, corresponding to F_j and $\mathbf{w}'_i$ above, and we want to find the u that maximizes the ranking quality for the mixture metric $M_{\mathrm{mix}} = u \cdot M_1 + (1 - u) \cdot M_2$. If we consider triplets consisting of each example region A_i, its known matching region B_i and one of the random mis-matching regions $C_{i,j}$, we can easily determine the effect on the ranking of B_i above or below $C_{i,j}$ for all $0 \leq u \leq 1$. Since the mixture is a linear combination, the distances of B_i and $C_{i,j}$ to A_i 'slide' linearly between $\Delta M_1(A_i, B_i)$ and $\Delta M_2(A_i, B_i)$ and between $\Delta M_1(A_i, C_{i,j})$ and $\Delta M_2(A_i, C_{i,j})$, respectively, as u slides between 0 and 1. Hence, there is at most one 'crossover point' at which B_i can switch places with $C_{i,j}$ in the ranking, thus increasing or decreasing the overall ranking quality. If $\Delta M_1(A_i, B_i) > \Delta M_1(A_i, C_{i,j})$ and $\Delta M_2(A_i, B_i) < \Delta M_2(A_i, C_{i,j})$, or $\Delta M_1(A_i, B_i) < \Delta M_1(A_i, C_{i,j})$ and $\Delta M_2(A_i, B_i) > \Delta M_2(A_i, C_{i,j})$, then a switch will take place, in which case the crossover point v can be determined by solving the following linear equation:

$$\begin{aligned}&v \cdot \Delta M_1(A_i, B_i) + (1 - v) \cdot \Delta M_2(A_i, B_i) \\ &\quad = v \cdot \Delta M_1(A_i, C_{i,j}) + (1 - v) \cdot \Delta M_2(A_i, C_{i,j}).\end{aligned}$$

This crossover point is calculated for all triplets of A_i, B_i and $C_{i,j}$ that do cross over and the direction of the switch is recorded as +1 if B_i becomes ranked *more highly* than the mis-

match $C_{i,j}$ as u goes to 1 (thus, increasing the overall ranking quality) or -1 if the ranking of B_i drops below that of $C_{i,j}$.

After all of the possible crossover points are calculated, they are analyzed to determine the single best point v^* which maximizes the number of positive crossovers (with direction +1) while minimizing the number of negative crossovers (with direction -1). This is performed by sorting the crossover points on the values v. An accumulator is then initialized to 0 and swept through the list of crossovers in sorted order, incrementing by 1 for each positive crossover and decrementing by 1 for each negative crossover. The crossover point v^* where the accumulator reaches its maximum value will be exactly the value of u at which the ranking quality of the mixture RQ(mix) [where $M_{\rm mix} = u\cdot M_1 + (1 - u)\cdot M_2$] is most improved.

This core computation only optimizes the weight of one feature at a time against all the others. Hence, it must be repeated to find the best overall combination of weights for the weight vector. We use a greedy search procedure based on a randomized version of hill-climbing (Russell & Norvig, 1995) as described above (*i.e.* select a random feature F_j and recalculate its weight *u etc.*). Features that are relevant (in the context of all the others) will increase in weight when selected and noisier features that tend to interfere with matching will see their weights drift toward 0. This process is iterated until the overall ranking quality of the weight vector stops increasing. It is important to note that as with all greedy search procedures, *Slider* is not guaranteed to find the *globally* optimal weight vector (which is computationally intractable), but only a local optimum. However, by re-running the search multiple times, it can be observed that the resulting ranking qualities are fairly consistent, suggesting convergence. Also, owing to the randomness in the algorithm (*i.e.* the order in which features are selected for re-weighting), the final weight vectors themselves can be different. Hence there is no 'absolute' optimal weight for any individual feature; weights are only meaningful in combinations. For example, if there are two highly correlated features, sometimes one will get a high weight and the other will be near 0, and other times the weights will be reversed.

Like the creation of the feature database, the weighting of the features is performed prior to the *TEXTAL* model-building process. A given set of weights is specific for a particular database, so that a new database consisting of different proteins would require re-evaluation of the feature weights.

3.6. Evaluation of the accuracy in the *TEXTAL* models

To evaluate the accuracy of the *TEXTAL* method, electron-density maps of three different proteins of known structures were modeled. Because *TEXTAL* does not require or utilize any amino-acid sequence information when choosing the best match for a region, the amino-acid types in the generated model may differ from those in the actual protein. The accuracy of the *TEXTAL* model is first evaluated by an amino-acid sequence-identity comparison. The similarity of the side-chain structures, evaluated using a 'similarity matrix', was also used as a measure of the models' accuracy (see Table 3). The similarity matrix treats residues as similar if they are from the same category of amino acids, where the categories are defined as listed in Table 5. The categories are formed from residues that are (i) identical, (ii) isosteric (*e.g.* threonine and valine) or (iii) structurally similar up to the 6.0 Å cutoff (*e.g.* the aromatics). This similarity matrix was chosen since the features are extracted over a maximum radius of 6.0 Å and the difference between a phenylalanine and a tyrosine, for example, may be apparent only beyond 6.0 Å, thus the features will be unable to distinguish them. All aromatic residues are considered structurally similar in this matrix and histidine is also included with the aromatics, so that a His match for a Phe is considered similar. It is important to note that the *shapes* of the density in 6.0 Å radius spheres may be very similar for amino acids that have little or no *chemical* or *physical* similarities (such as Leu and Asp or Ser and Cys). The percentage similarity based on this matrix is the total number of similar residues divided by the number of positions. The similarity matrix reveals structural accuracy not reflected by the sequence-identity measurement.

The difference between the model and the real structure is further evaluated by measuring the r.m.s.d. between the atomic coordinates of the model built and the true coordinates. The r.m.s.d. is the average of all of the differences between the atoms in one region with the nearest corresponding atoms in the other region. If the number of atoms differ between two regions, which can easily happen when the model and the correct structure are of different residue types, some atoms in one region will not have counterparts in the other region. The r.m.s.d. calculation takes into account only those atoms that have reasonable counterparts, *i.e.* within 3 Å.

Table 2
Feature weights used in the *TEXTAL* experiments, listed in order of relevance.

Both the weights and the optimal radii were calculated using the *Slider* algorithm.

Feature	Weight	Radius (Å)
Distance to center of mass	0.183103	5.0
Ratio of moments 1 and 3	0.153815	4.0
Ratio of moments 1 and 3	0.136521	5.0
Skewness	0.080056	3.0
Skewness	0.055487	6.0
Ratio of moments 1 and 2	0.055124	4.0
Median spoke angle	0.052865	6.0
Minimum spoke angle	0.051494	4.0
Skewness	0.049710	5.0
Ratio of moments 1 and 2	0.038135	5.0
Maximum spoke angle	0.037110	4.0
Ratio of moments 1 and 3	0.031616	3.0
First moment of inertia	0.022025	6.0
Median spoke angle	0.019343	4.0
Minimum spoke angle	0.015231	6.0
Distance to center of mass	0.008193	3.0
Spoke triangle area	0.007749	3.0
Maximum spoke angle	0.001490	3.0
Median spoke angle	0.000618	3.0
Kurtosis	0.000196	6.0

Table 3
Correlation coefficients of the models built by *TEXTAL* and the original structures.

PDB ID	Number of residues	Average correlation	Maximum correlation	Minimum correlation	Sequence identity (%)	Structural similarity (%)
1gcn	29	0.88	0.95	0.73	41.4	77.6
1tup	196	0.87	0.98	0.64	40.0	63.3
1fnb	296	0.89	0.98	0.67	50.0	69.4

Table 4
R.m.s.d. of corresponding atoms.

Overall r.m.s.d. calculation when side chains were included was performed by determining the closest neighbor in the known structure for each candidate atom and measuring the distance between all of these pairs. No information regarding the identity of the atoms was involved and if a pair of atoms was separated by more than 3 Å they were not included in the calculation. Flipped residues were excluded in the r.m.s.d. calculations.

	R.m.s.d. (Å)					
Protein	Overall (incl. side chains)	C—C	O—O	N—N	No. flipped	% flipped
1gcn	0.70	0.80	0.92	0.37	5	17
1tup	0.69	0.45	0.76	0.52	18	9
1fnb	0.62	0.33	0.72	0.39	15	5

4. Results

To evaluate the utility of pattern matching for interpreting electron-density maps, we have taken the preliminary step of modeling regions in 'unknown' maps with the assumption that the α-carbon positions are known *a priori*. This limitation will eventually be removed, but it allows us presently to evaluate the pattern-matching capabilities of the *TEXTAL* method separate from the issue of locating C^{α} atoms. We also report an analysis of the sensitivity to errors in C^{α} coordinate estimates.

4.1. Feature weights

The algorithm *Slider* was used to determine appropriate weights for the 76 features (19 unique features for four different radii) for matching regions in *TEXTAL*. To determine the optimal weights for our features, it is first necessary to obtain a representative set of regions containing highly similar density and other regions that do not match. We selected a randomly chosen subset of 500 pairs of regions from our database that had high density correlations and 500 other random regions for each that acted as non-matches. Based on visual inspection of the results, we observed that a density correlation of 0.7 or greater typically indicates a sufficiently similar pattern of density; hence, we use cc $\geq$ 0.7 as the definition of a good match.

Using the *Slider* algorithm, the mixture of weights for the feature set was optimized to discriminate between matching (cc $\geq$ 0.7) and non-matching (cc $\leq$ 0.7) regions. The features that contributed to improving the overall ranking quality of the matches were returned with the best weights associated with them and these features are shown in Table 2. Several features, such as the ratio of the first and third moment of inertia, were found to be relevant at different radii (*e.g.* 3, 4 and 5 Å), each contributing unique information. This mixture of weights gives a ranking quality of 0.865, meaning that true matches were ranked by feature differences among the top 13.5% of all candidates on average. We note that owing to randomization in the *Slider* algorithm, it is possible that there are other mixtures of features that give equally high performance.

4.2. Modeling unknowns

Three different proteins from the Protein Data Bank were used as 'unknowns' for our experiments, representing varying levels of complexity. Glucagon, 1gcn, contained the fewest residues (29) and possesses the simplest secondary structure: a single α-helix. The other two proteins, ferredoxin reductase (1fnb; 296 residues) and p53 tumor repressor (1tup; 196 residues), are considerably more complex structures, having both α and β structure as well as stretches of random coil.

The features of each 'unknown' map were extracted at each C^{α} position of the unknown structure in spheres which ranged from 3 to 6 Å in radius. The features of each region in the unknown were then compared with all of the 54 164 regions in the database and a feature difference (ΔF) score was calculated. For each of the top $N = 2000$ most highly ranked regions, the known electron density of the region from the database was rotated and superimposed onto the unknown region using the moment-of-inertia alignment described above and the correlation coefficient cc was calculated within the side-chain axis cylinder. The region with the highest density correlation was selected as the best match and the corresponding rotation and translation was applied to atoms retrieved from the matched region in the known structure to create *TEXTAL*'s estimation of the local structure in the unknown region.

As shown in Table 3, *TEXTAL* was able to identify regions with a high overall correlation to the electron density around the correct structures for the unknowns. The modeling was performed entirely automatically by *TEXTAL* (*i.e.* there was no manual intervention or post-processing) and took around 30 s per residue (with N set to 2000) on an SGI Origin 2000. Although there is a fairly wide range of correlations for the three proteins, the average in all cases was near 0.9, well above our observations of a minimum cutoff for a reasonable match. These results indicate that our matching process is able to identify similar regions of electron density.

The accuracy of the *TEXTAL* models was measured in terms of the amino-acid identity and structural similarity (*via* the similarity matrix defined above). Sequence identity was 41.4% for 1gcn, 40.0% for 1tup and 50% for 1fnb. The similarity matrix evaluation showed that our model of 1gcn was 77.6% similar to the actual structure, while the 1tup and 1fnb models were 63.3 and 69.4% similar, respectively, to the correct structures.

In addition to frequently recognizing side-chain types, the generated structures also had conformations similar to the correct ones. Shown in Table 4 is the r.m.s.d. between corre-

sponding coordinates in the actual 'unknown' structure and the model built by *TEXTAL*. Note that during the model-building process, the matching region found in the database for a given unknown region could be associated with the region in the opposite backbone configuration where the carbonyl C atom is mapped onto the backbone N atom and *vice versa* or where the side-chain atoms are mapped into backbone positions. We refer to these residues as 'flipped' positions; the number of such flipped residues is also reported in Table 4. Although no information was provided regarding the orientation of the polypeptide chain during pattern matching, the occurrence of flipped positions was relatively low (5–17%). Flipped residues dramatically misrepresent the accuracy of the model when included in the r.m.s.d. calculation, but they can easily be repaired in the model by post-processing (*i.e.* by enforcing a consistent directionality to the peptide chain). Also shown in Table 4 are the r.m.s.d.s of the main-chain atoms (besides C^{α}); the flipped positions were not included in this measurement.

4.3. Examples of regions modeled in the test maps

Glucagon contains 29 residues in a single α-helix. When *TEXTAL* modeled 1gcn, it was able to match structurally similar side chains 77.6% of the time and only five of the 29 positions had flipped backbones. Because this is a different type of similarity matrix than is customarily used in biochemistry (a *structural* similarity as opposed to *chemical* similarity), it is instructive to consider what a similarity score of this magnitude indicates. For comparison, the average structural similarity as measured by this matrix is only 15% for four randomly picked proteins of an equal number of amino acids. Therefore, 77.6% similarity is well above random chance. A superposition of 1gcn with the *TEXTAL* model is shown in Fig. 3. Examples of similar amino-acid matches include placing a His for Phe6, a Leu for Asp21 and a Glu for Gln24. Some of the other matches that were returned by *TEXTAL* contain one more or one fewer C atoms, for example, Gln for Asn28.

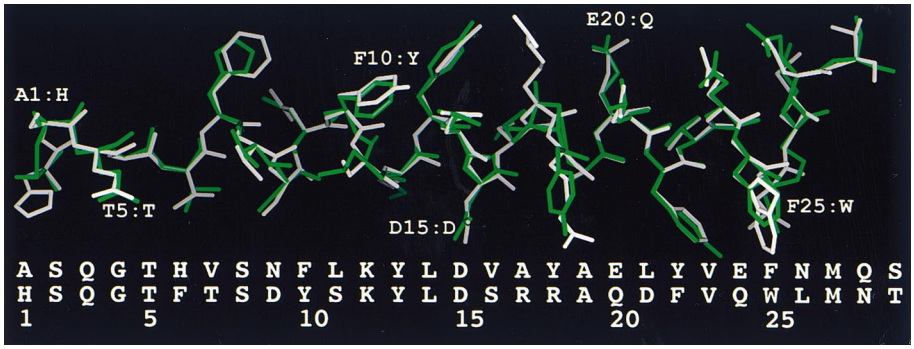

Figure 3
Superposition of 1gcn and model built by *TEXTAL*. This figure and all color figures were produced using the program *SPOCK* (Christopher, 1998). The sequence on top is for the model built by *TEXTAL*, as is the structure in green.

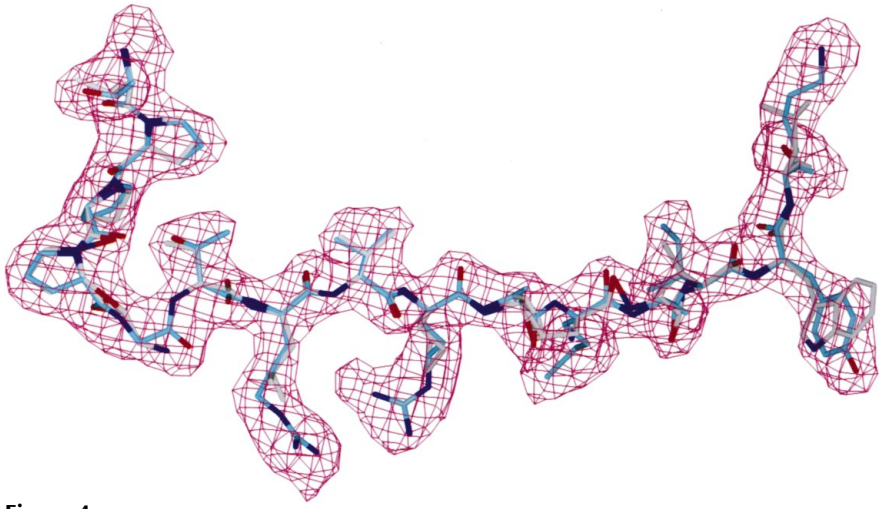

Figure 4
Region in 1tup compared with the model built by *TEXTAL*. Shown are residues 147–155. The figure includes a 1σ contour of the electron-density map of 1tup. In this and in Fig. 5, the *TEXTAL* model C atoms are in white and the C atoms of the original PDB structure (here, 1tup.pdb) are in light blue; all N atoms for both are dark blue, all O atoms are red and all S atoms are yellow.

The modeling results for the p53 tumor repressor (1tup) were equally encouraging, especially given the relative complexity of the secondary structure. 1tup contains three α-helices and a total of 11 β-strands. Out of the 196 residues, *TEXTAL* identified 40% of the amino acids exactly and 63.3% were structurally similar. The model also contains several fairly long contiguous segments where *TEXTAL* performed quite well, producing many structurally similar and identical matches. The longest stretch of good matches is between residues 147 and 155 (except one position containing a match differing by one carbon). This region, which also contains three consecutive prolines successfully matched by *TEXTAL*, is shown in Fig. 4. There are several such highly accurate regions in the 1tup model; however, these regions were interspersed with regions where *TEXTAL* was not able to produce such good matches. These long matching regions were neither consistently found in the interior or the exterior of the protein, nor were they consistently in the same secondary structure (*i.e.* all α-helical or β-strand).

The ferredoxin reductase structure contains 296 amino-acid residues. The average cc for all positions in the *TEXTAL* model is slightly higher (0.89)

Table 5
Confusion matrix of similarities in 1fnb.

The leftmost column shows the amino-acid groups in the actual structure and the number of occurrences of this group in the structure. (One-letter amino-acid names are used.)

TEXTAL model	FYWH:28	QE:29	DNL:54	SC:26	ITV:59	RKM:34	G:20	A:30	P:14
PDB:No.									
FYWH:36	25	2	3	1		3		1	1
QE:31	2	16	1			9		3	1
DNL:52		1	40	4	5	2			
SC:20			2	12	3	2		1	
ITV:47		1		1	45				
RKM:54	1	9	8	7	6	18		5	
G:26							20	6	
A:17				1				16	
P:13									13

than for 1tup (0.87) and the percentage of flipped residues is lower (5% for 1fnb and 9% for 1tup). 50% of the 296 residues were matched with identical amino-acid matches and 69.4% of the matches were similar by the matrix score. An example region of well matched positions in the *TEXTAL* model of 1fnb is shown in Fig. 5. The atomic r.m.s.d. for this region was only 0.27 Å.

Shown in Table 5 is a more detailed description of *TEXTAL*'s ability to match similar regions. The table is a confusion matrix which shows the actual amino-acid classes in the rows, while the columns are the amino-acid classes found by *TEXTAL*. For example, there are 47 residues of the short-branched type (I, T or V). Of the 47 residues, 45 were matched with this type of residue, one with either S or C and one with Q or E. These are the same categories of matches used in the similarity matrix score. Qualitatively, of the nine categories, *TEXTAL* appears to have the most difficulty modeling the amino acids with long flexible side chains (RKM). This could arise from a combination of their reaching beyond the radii of feature calculations and/or increased degrees of freedom.

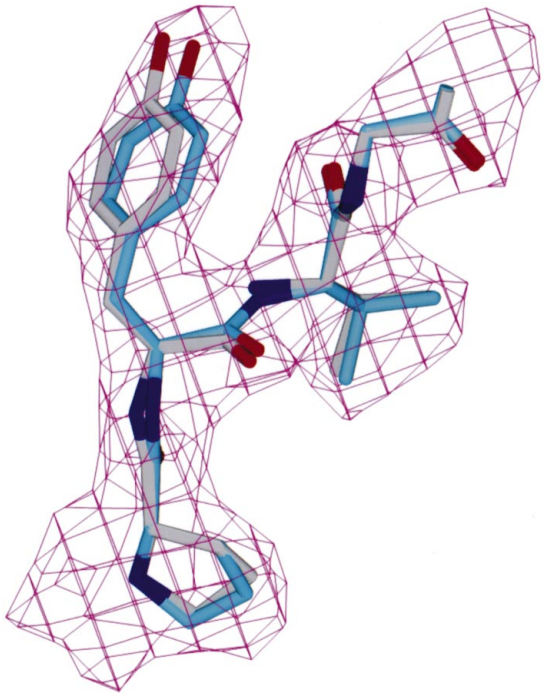

Figure 5
Superpositions of 1fnb with the *TEXTAL* model. A (1σ) contour of the electron-density map of 1fnb.pdb is included. The region shown contains residues 37–40, where all positions were matched correctly in amino-acid type: Pro, Tyr, Val, Gly.

4.4. *TEXTAL* modeling of rat intestinal fatty acid binding protein from newly collected X-ray data

X-ray diffraction patterns were collected from crystals of recombinant rat intestinal fatty acid binding protein (iFABP). The data were collected under cryogenic conditions using MacScience dual image-plate system on a Rigaku RU-200 generator. Indexing and scaling of the data was performed using the software packages *SCALEPACK* and *DENZO* (Otwinowski, 1993) and the map was calculated and refined using *CNS* (Brunger *et al.*, 1998). The data were 91% complete to 1.62 Å; however, the electron-density map used on *TEXTAL* was calculated to the medium resolution of 2.8 Å in order to match the database. The crystal structure for rat iFABP was previously published (Scapin *et al.*, 1992; PDB code 1ifc) with an R factor of 16.9% from diffraction data to 1.19 Å resolution. The new cryogenic data were used to obtain a map quickly (from X-ray diffraction data but not optimally refined) for testing the accuracy of *TEXTAL*. Based on these new data, the model was refined to an R factor of 21.1% ($R_{\rm free}$ 24.9%) (including 121 new cryo-data water molecules).

The automated model construction for iFABP using *TEXTAL* followed the same method used for modeling the test proteins. Features were extracted in spheres centered at

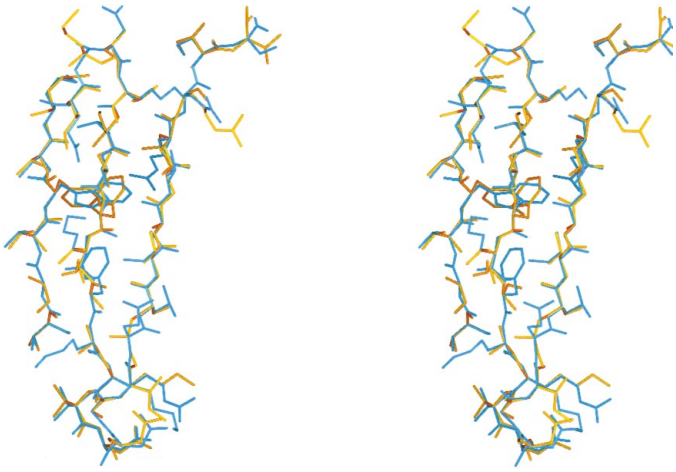

Figure 6
Superpositions of *TEXTAL* model and the known structure of 1ifc.pdb (intestinal fatty acid binding protein). The structure in green is the model built by *TEXTAL*.

Table 6
TEXTAL modeling of iFABP.

Number of residues	Average correlation	R.m.s.d. (Å)	Percent flipped	Sequence identity (%)	Structural similarity (%)
130	0.834	0.74	11	35.4	54.6

130 C^{α} atoms obtained from the refined structure and the density correlations were calculated. For each match, 2000 density regions were retrieved by feature difference and the one region that matched with highest correlation was returned and incorporated into the model. *TEXTAL* matched 35.4% of the regions with the correct amino acid and 54.6% of the selections were of a similar type (Table 6). The model matched the original refined structure to 0.74 Å r.m.s.d. Only 14 of the 130 positions (10.7%) were returned in a flipped configuration. Shown in Fig. 6 is a region of the model built by *TEXTAL* superimposed with the original refined structure. This test shows that *TEXTAL* can produce accurate models with a map calculated from X-ray diffraction data and does not require artificially generated maps to be successful.

4.5. Relationship between electron-density patterns and molecular structure

Since *TEXTAL* uses the density correlation coefficient as its final measure of similarity, it is important to establish the relationship between cc and atomic r.m.s.d. Fig. 7 plots density cc *versus* atomic r.m.s.d. between corresponding regions in the original structures for all three test proteins and the models built by *TEXTAL*. The relationship between high cc and low r.m.s.d. suggests that looking for regions with similar patterns of density is a reasonable strategy for finding good matches in terms of atomic structure. However, because some high-cc regions also show relatively high r.m.s.d., not all regions in the test maps contain well matching regions in the database. Still, over 75% of the regions had matching regions with density correlations greater than 0.81 and matched with an r.m.s.d. of less than 0.9 Å.

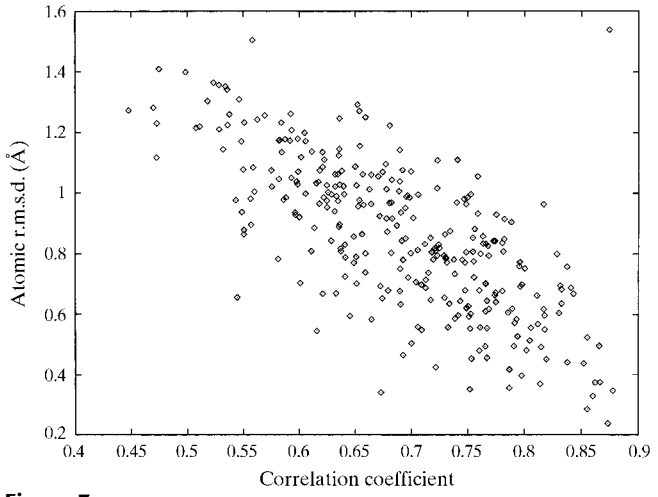

Figure 7
Plot of atomic r.m.s.d. as a function of density correlation coefficient between regions. The error bars show standard error.

4.6. Sensitivity of models to errors in C^{α} coordinates

One of the limitations of the experiments described so far is that the locations of C^{α} atoms were presumed to be known *a priori*; *i.e. TEXTAL* was used to model regions centered on C^{α} coordinates derived from a PDB file. This information will clearly not be available in real uninterpreted maps. However, one of several methods could be used to estimate the locations of C^{α} atoms in a new map. For example, a skeletonization algorithm such as *BONES* could be used to pick coordinates along the main chain that are likely candidates for C^{α} positions (such as at branch points). Alternatively, we are developing a pattern-recognition routine to accurately identify C^{α} atoms in a map by training a neural network to use features of the local pattern of electron density to predict how far away a given lattice point is from a true C^{α} atom.

Regardless of the approach to predicting C^{α} positions in a map, there will almost certainly be some error in the estimates of the coordinates. Such errors could potentially cause problems for *TEXTAL*, since it will be attempting to model regions whose centers are offset from a true C^{α} atom with a database of regions precisely centered on C^{α} atoms. Therefore, we tested *TEXTAL*'s sensitivity to errors in the C^{α} coordinate estimates. In this experiment, we selected 5000 regions randomly from our database of 200 back-transformed maps. Each of these was centered on a C^{α} atom. We ran *TEXTAL* on these regions to determine the highest density correlation that could be achieved by any other match in the database. A random vector was then added to offset the center of the region from the C^{α} (uniform sampling of $-1.5 \ldots +1.5$ for *X*, *Y* and *Z)*, producing errors of 0–1.9 Å in arbitrary directions. Finally, *TEXTAL* was run a second time on each of these regions to find the match with the highest density correlation, given the random shift.

Fig. 8 shows the ratio of the maximum density correlation of each region with the offset to that for the unshifted region as a function of the magnitude of the offset vector. While there is a great deal of variation in how much the offset decreases the quality of matches, it can be observed that the general trend is that regions with a high density correlation ($\geq$90%) are retrieved for regions offset by up to around 0.8 Å. Beyond this point, regions not centered on C^{α} atoms increasingly fail to have high-quality matches in the database. This quantifies *TEXTAL*'s tolerance for errors in the initial estimates of C^{α} locations.

5. Conclusions

In the work reported here, we have described a new approach for interpreting electron-density maps, implemented in a system called *TEXTAL*. *TEXTAL* was used to model three test proteins from their electron-density maps and the coordinates of their C^{α} atoms. The potential of the method is demonstrated by the high structural similarity of the *TEXTAL*

models to the original protein structures. These preliminary tests show that *TEXTAL* is able to build models that are structurally similar to the original proteins regardless of whether the amino-acid identity is correct. The r.m.s.d. scores are low, in the range 0.6–0.7 Å. The unknowns used in our tests contained a wide range of secondary structures and *TEXTAL* was able to model these differing structures. Except for the modeling of iFABP, the electron-density maps used in testing and contained in the database are calculated by Fourier transformation from known structures at medium resolution (∼2.8 Å). A real MIR map for iFABP at a medium resolution of 2.8 Å was also successfully modeled by *TEXTAL*.

These results validate the usefulness of pattern recognition as a technique for electron-density map interpretation and suggest that *TEXTAL* could be an important tool for building protein structures. *TEXTAL* is able to build fairly accurate models of proteins from medium-resolution density maps (∼2.8 Å) in a few hours using only the initial estimates of the α-carbon positions; the requirement for knowing the α-carbon coordinates will be removed in a future version. This automated approach represents an important advance in X-ray crystallography. Other computational methods currently available for the interpretation of electron-density maps include graphical or mathematical density-analysis programs (Jones *et al.*, 1991; Fortier *et al.*, 1997; Kleywegt & Jones, 1997; Leherte *et al.*, 1994), along with fragment-fitting approaches (Holm & Sander, 1991; Levitt, 1992). *TEXTAL* is distinct from these previous methods because it uses pattern matching of the electron density to recognize and model regions in an unknown map. It assigns atomic coordinates to the unknown map based on similar regions in a database of previously determined structures. This approach is a form of instance-based or nearest-neighbor learning (Aha *et al.*, 1991), where the regions in the database provide example cases from which to model unknowns. This form of pattern recognition has not been applied to X-ray crystallography; previously the potential for building accurate protein structures is supported by our preliminary results.

There are several ways in which *TEXTAL* could be improved, such as by developing new features to improve the fidelity of the matching or by specializing the database for specific secondary or side-chain types (*e.g.* rotamer classes). Another significant means for improving the overall accuracy is to add post-processing procedures to refine the models. For example, the method by which the best match is chosen from a list of candidates is a prime target for improving *TEXTAL*. In these preliminary experiments, the choice for the best match in the database was based solely on the region which gave the highest density correlation. Many of the other matches for the same region have similar cc's, but correspond to different amino-acid types. Post-processing steps that consider several of the top matches could improve model accuracy, such as by using consensus among the top matches to choose the best residue type. The predicted structures for some neighboring regions may also aid in choosing the best match, for example, by rejecting flipped candidate residues based on the orientation of the neighboring residues' backbone atoms. Also, energy minimization (*e.g.* real-space refinement) could be applied to the final models to improve the r.m.s.d. by regularizing the structures and adjusting the fit to the density. Finally, our current experiments make no use of the amino-acid sequence, although it is typically known. Future experiments could exploit knowledge of the amino-acid sequence in order to determine which of the best-correlated regions to choose by enforcing consistency with the identities of neighbors, for example, by using a dynamic programming algorithm (Baxter *et al.*, 1996). However, even in its current implementation, the *TEXTAL* approach is able quickly to build fairly accurate models of entire proteins from medium- to low-resolution data in only a few hours.

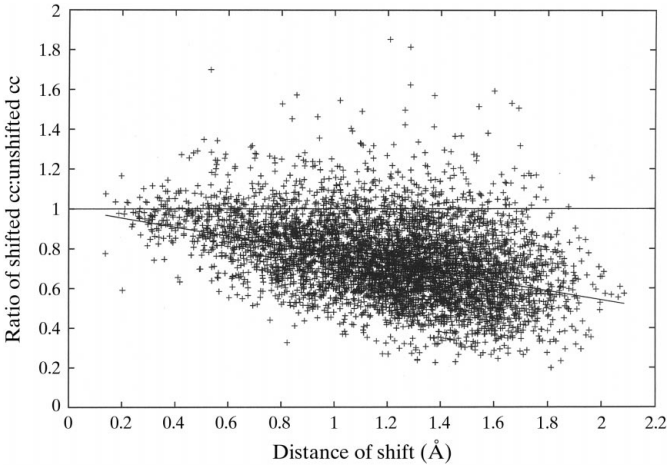

Figure 8
Quantitative analysis of the ability of *TEXTAL* to tolerate errors in C^{α} estimates. The line at $y = 1.0$ is included to aid in interpretation with the line representing the exponential fit to the data.

This work was supported by the the Robert A. Welch Foundation and the National Institutes of Health Grants GM 45859 and GM 59398.

References

Aha, D. W. (1998). *Feature Extraction, Construction and Selection: A Data-Mining Perspective*, edited by H. Liu & H. Motoda, pp. 1–20. Norwell, MA: Kluwer.
Aha, D. W., Kibler, D. & Albert, M. K. (1991). *Mach. Learn.* **6**, 37–66.
Asker, L. & Maclin, R. (1997). *Proceedings of the 14th International Conference on Machine Learning*, edited by D. H. Fisher Jr, pp. 3–11. San Mateo, CA: Morgan Kaufmann.
Baxter, K., Steeg, E., Lathrop, R., Glasgow, J. & Fortier, S. (1996). In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, edited by D. J. States, P. Agarwal, T. Gaasterland, L. Hunter & R. F. Smith. Menlo Park, CA: AAAI Press.
Bonanno, J. (1999). *Curr. Biol.* **9**(23), R871–872.
Brändén, C. & Jones, T. (1990). *Nature (London)*, **343**, 687–689.
Bricogne, G. (1997). *Methods. Enzymol.* **276**, 362–423.
Brünger, A. T. (1996). *X-PLOR: A System for X-ray Crystallography and NMR, Version* 3.851. Yale University, New Haven, CT, USA.

Brunger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Christopher, J. A. (1998). *SPOCK: The Structural Properties Observation and Calculation Kit.* Center for Macromolecular Design, Texas A&M University, College Station, TX, USA.
Fortelle, E. de la, Irwin, J. J. & Bricogne, G. (1997). In *Crystallographic Computing 7: Proceedings of the Macromolecular Computing School*, edited by P. E. Bourne & K. Watenpaugh. Oxford University Press.
Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* New York: John Wiley & Sons.
Fayyad, U. M., Djorgovski, S. G. & Weir, N. (1996). *AI Mag.* **17**(2), 51–66.
Fortier, S., Chiverton, A., Glasgow, J. & Leherte, L. (1997). *Methods Enzymol.* **277**, 1–141.
Greer, J. (1985). *Methods Enzymol.* **115**, 206–224.
Holm, L. & Sander, C. (1991). *J. Mol. Biol.* **218**, 183–194.
Holm, L. & Sander, C. (1993). *J. Mol. Biol.* **233**, 123–138.
Ioerger, T. R., Holton, T. R., Christopher, J. A. & Sacchettini, J. C. (1999). *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, edited by T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes & R. Zimmer, pp. 130–137. Menlo Park, CA: AAAI Press.
John, G. H., Kohavi, R. & Pfleger, K. (1994). *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 121–129. San Mateo, CA: Morgan Kaufmann.
Jones, T. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.
Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268.
Jones, T. A., Zou, J., Cowan, S. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.
Langley, P. (1994). In *Proceedings of the AAAI Symposium on Relevance*, edited by R. Greiner. New Orleans, LA: AAAI Press.
Leherte, L., Fortier, S., Glasgow, J. & Allen, F. (1994). *Acta Cryst.* D**50**, 155–166.
Levitt, M. (1992). *J. Mol. Biol.* **226**, 507–533.
Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.
Oldfield, T. J. (1997). *Crystallographic Computing 7: Proceedings from the Macromolecular Computing School*, edited by P. E. Bourne & K. Watenpaugh. Oxford University Press.
Otwinowski, Z. (1993). *Proceedings of the CCP4 Study Weekend: Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.
Perrakis, A., Titia, K. S., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.
Richardson, J. & Richardson, D. (1985). *Methods Enzymol.* **115**, 189–206.
Russell, S. & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach.* New Jersey: Prentice Hall.
Scapin, G., Gordon, J. I. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 4253–4269.
Swanson, S. M. (1994). *Acta Cryst.* D**50**, 695–708.
Terry, A. (1983). *The CRYSALIS Project: Hierarchical Control of Production Systems.* Technical Report HPP-83–19, Department of Computer Science, Stanford University, USA.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Wisniewski, E. J. & Medin, D. L. (1994). In *Machine Learning: A Multistrategy Approach*, Vol. IV, edited by R. Michalski & G. Tecuci. San Francisco: Morgan Kaufmann.