TEXTAL: A Pattern Recognition System for Interpreting Electron Density Maps

Thomas R. Ioerger[1], Thomas Holton[2], Jon A. Christopher[2], and James C. Sacchettini[2]
[1]Department of Computer Science
[2]Center for Structural Biology
Texas A&M University

Abstract

X-ray crystallography is the most widely used method for determining the three-dimensional structures of proteins and other macromolecules. One of the most difficult steps in crystallography is interpreting the electron density map to build the final model. This is often done manually by crystallographers and is very time-consuming and error-prone. In this paper, we introduce a new automated system called TEXTAL for interpreting electron density maps using pattern recognition. Given a map to be modeled, TEXTAL divides the map into small regions and then finds regions with a similar pattern of density in a database of maps for proteins whose structures have already been solved. When a match is found, the coordinates of atoms in the region are inferred by analogy. The key to making the database lookup efficient is to extract numeric features that represent the patterns in each region and to compare feature values using a weighted Euclidean distance metric. It is crucial that the features be rotation-invariant, since regions with similar patterns of density can be oriented in any arbitrary way. This pattern-recognition approach can take advantage of data accumulated in large crystallographic databases to effectively learn the association between electron density and molecular structure by example.

Introduction

Interpreting electron density maps is one of the most challenging and time-consuming aspects of X-ray crystallography. There are several steps in solving the structure of a protein or other macromolecule by crystallography (Stout & Jensen 1989). First, the molecule must be purified and crystallized. Then X-rays are passed through the crystal, and diffraction patterns are collected at various angles. The diffraction patterns represent the Fourier transform of the electron density in the unit cell, so, in principle, the inverse Fourier transform of the diffraction pattern could be used to reconstruct the electron density pattern. However, this is complicated by two facts. First, the quality of the data (number of reflections, amount of noise) can limit the resolution of the reconstructed density. Second, the diffraction patterns themselves only provide amplitudes; phases are also needed for the inverse Fourier transform, but cannot be directly measured. Hence a number of experimental techniques such as MIR, MAD, and molecular replacement, along with recent advances in computationally-intensive direct methods (Chang *et al.* 1997), have been devised to make initial inferences of approximate phases. Phases are usually also iteratively refined to improve the fit of a model to the data (Brünger, Kuriyan, & Karplus 1987).

Once an electron density map is generated, it must be interpreted to produce a molecular model with atomic coordinates. This procedure is often done manually with 3D visualization software on a graphics workstation (Jones *et al.* 1991), and relies heavily on the knowledge and expertise of the crystallographer. It can take up to several months and may not even be feasible for very large structures. While in general the electron density should fall along the structure of a molecule, there are a number of sources of disturbances that can make the density appear less representative. Errors in phases can cause density to appear or disappear in random places; highly mobile sidechains or backbones can result in weak density; and the density in low resolution maps is naturally more dispersed. As a result of these effects, crystallographers often make mistakes in structure determination (Jones & Kjelgaard 1997).

Because interpreting electron density maps is such a time-consuming and error-prone process, there is a great need to automate the process. Several approaches have been proposed for automating the interpretation of electron density maps. One class of approaches focuses on trying to identify likely positions of atoms within a map. One of the earliest examples is Greer's (1985) skeletonization algorithm. A more recent method, called critical-point analysis (Fortier *et al.* 1997), examines the gradient of the density to estimate the locations of atoms. Another class of approaches generally uses skeleton atoms (or even manually-picked C_α atoms) as a starting point

and tries to build the rest of the structure from them. Fragment-fitting (Jones & Thirup 1986) uses a sequence of several consecutive C_α's to look up candidate structures, including side-chain atoms, in a database to add to the model. Holm and Sander (1991) and Levitt (1992) each extend this idea with more sophisticated search and conflict-resolution strategies. Glasgow, Fortier, and co-workers have proposed an approach called Molecular Scene Analysis, in which computational imagery routines would be used to match geometric patterns of density to a database of prototypes (Fortier *et al.* 1993). Finally, CRYSALIS (1983) is an expert system that takes into account a variety of knowledge sources and constraints, such as the amino acid sequence (if known), the preference for hydrophobic residues in the core of the protein, etc., in order to construct a plausible model from an initial set of pseudo-atoms.

Many of these approaches have been demonstrated to work well on small proteins with high resolution maps. However, new methods are needed for generating more accurate models for larger proteins with medium- to low-resolution maps. Speeding up the interpretation of electron density maps with automated systems will be especially important to various large-scale Structural Genomics efforts that have recently been discussed (Gaasterland 1998), which aim to solve a wide range of protein structures to quickly increase our knowledge of fold-space, essentially by brute force.

In this paper, we describe a new system, called TEXTAL, for automating the interpretation of electron density maps. TEXTAL is based on *pattern recognition*. Isolated regions (e.g. spheres of 5Å radius) in a map are matched against a database of regions in other maps whose structures have already been solved. When a match is found, the local structure is inferred by translation and rotation of atomic coordinates from the matched region of the known protein. Hence, the non-trivial relationship between electron density and molecular structure can be learned from examples. What is unique about our approach is using the electron density itself as a basis for the matching. To accomplish this, we extract rotation-invariant features of the density in a region and use these features to look for candidate regions with similar patterns of density. In the remainder of this paper, we give an overview of the TEXTAL system, and we describe some results of using it to construct models for both artificial and real electron density maps to demonstrate the effectiveness of this pattern-matching approach.

Methods

Outline of the TEXTAL Program

In the TEXTAL program, an electron density map is treated as a series of overlapping spheres of density containing information about regions of the protein structure. The size of the spheres is chosen to be 5Å to exploit the significant amount of repetition in protein structures at this scale. For example, a 5Å sphere can usually cover about one side-chain and some adjacent backbone. While there is great diversity among protein structures, individual side-chains often adopt one of a few canonical conformations (rotamers), and backbone angles are often restricted to a small set of predictable combinations, depending on the local secondary structure.

In order to efficiently search a large database of spherical regions for similar patterns of density, we extract characteristic features and use them for pattern matching. Because matched regions in other proteins can be positioned in any arbitrary orientation, useful features of the electron density must be rotation-invariant (i.e. constant, even if the density in the region is rotated around the center) to detect similarities. We developed fifteen rotation-invariant numerical features that characterize aspects of the patterns in electron density. The features were capable of associating similar regions of electron density among different maps.

The overall method of TEXTAL involves the following steps for modeling a region in an unknown map: 1) feature matching between unknown regions and regions in the database, 2) evaluating the candidate matches by calculating density correlation, and 3) building the model (see Figure 1). The input information required for TEXTAL is an "unknown" electron density map and a database of feature-extracted maps. TEXTAL first extracts the features of the region under investigation in the unknown map, and this region is compared with all of the regions in the database in terms of the feature values alone. We evaluate how similar two regions are in terms of features by measuring the difference in the feature values for the two regions using a weighted Euclidean distance formula,

$$d(R_1, R_2) = \sqrt{\sum\ w_i\ (F_i(R_1) - F_i(R_2))^2} \qquad (1)$$

where $F_i(R_1)$ and $F_i(R_2)$ are the features values for the unknown region and a region in the database, respectively. This step is implemented as a simple look-up procedure through the database. Similar regions should have low differences between the feature values.

The program then retains the top K matching regions by feature comparison, where K is a user-selectable parameter. These K regions are further analyzed for similarity by calculating a density correlation coefficient. The correlation coefficient measures how similar two regions are in terms of their patterns of electron density. Since electron density maps are a discrete representation of a continuous 3D function, sampled at a finite number of evenly-spaced lattice points, the density correlation can be calculated by:

$$cc = \frac{\sum\ (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\ (x_i - \overline{x})^2}\sqrt{\sum\ (y_i - \overline{y})^2}} \qquad (2)$$

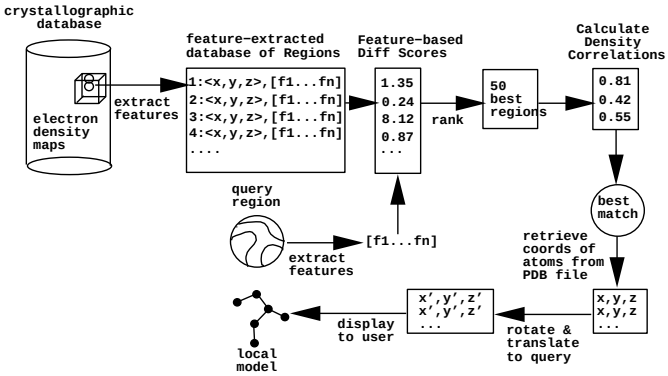

Figure 1: Architecture of TEXTAL.

where x is the density value in one region that is compared with the density value in another region y over all lattice points i, and $\overline{x}$ and $\overline{y}$ are the average densities in each region. However, since the similarity between two regions is required to be independent of orientation, the true measure is the *maximum* density correlation over all possible superpositions between two regions. Because an exhaustive search could require evaluating thousands of candidate rotations, we use a heuristic procedure called *peak-matching* to find the approximately optimal rotation. This method picks the top n peaks or lattice points with the highest densities that are at least 1Å apart in each sphere and superimposes them in all reasonable combinations. Nine peaks ($n = 9$) were found to be sufficient for our experiment, allowing the optimal rotation to be identified to within about 6°. Typically, evaluation of only a few hundred candidate rotations (based on peak combinations) were required to find the maximum density correlation. Other, more efficient, methods for finding the optimal rotation and calculating the density correlation are being developed. We observed that regions with a density correlation of 0.7 or higher often appeared to have highly similar patterns, and often had similar local structures as well. Hence we used a cutoff of $cc > 0.7$ as the definition of a match between regions, though a more quantitative analysis is on-going.

Database of Electron Density Maps

To accumulate a sufficient number of maps for the initial development and evaluation of TEXTAL, we used a database of artificial maps generated from 38 proteins in the PDB, spanning a wide range of α and β fold classes. These maps were created by placing a Gaussian distribution of density around each atom (using *Spock*, http://quorum.tamu.edu/spock), scaled such that the density was 1.0 at the van der Waals radius. The resulting maps have a uniform scaling, with an average density of around 2.0 and a standard deviation of around 0.9. All of our maps were created in the P_1 space group (with orthogonal axes).

Table 1: Feature types and descriptions and number.

Feature Type	Description	N
Basic characteristics of spheres of density	1. average density 2. distance from center of sphere to center of mass	2
Moments of inertia	magnitude of primary, sec,, and tertiary moments; ratios among moments	6
Statistical properties of density	standard deviation skewness, kurtosis	3
Spokes of density within spheres	min, mid, max, and sum of angles between 3 spokes	4

These maps are ideal in that they contain the most accurate representation of density for each amino acid, without any noise. Still, the maps in our database represent a wide range of structures within real proteins, and thus contain a great diversity of density patterns with which to match unknown regions. The unknown maps we modeled were prepared in the same way, which is important to ensure that the database is representative and contains relevant matches.

Feature Extraction

All of the maps in the database are interpolated onto an orthogonal 1Å-spaced grid to facilitate the feature calculations. The maps for all of the proteins in the database were feature extracted by calculating the values of various features over a sample of spherical regions. Currently there are four major types of features (see Table 1) and these are further differentiated to make a total of fifteen. All of these features are rotation-invariant. A brief description of the feature types follows; a more detailed description of the calculation of the features can be found elsewhere (Ioerger, Holton, Christopher, and Sacchettini; manuscript in review).

We use two features to express basic characteristics of the patterns of density in each extracted region. The first is the average density of the region, and the second is the distance from the center of the sphere to the center of mass. If two regions of density are similar in overall pattern, their average densities should be similar. Also, the center of mass for a region of density should be at a similar distance to the center of the sphere for a region with a similar pattern of density, and this does not depend on orientation.

The second set of features is based on the moments of inertia in a region. The moments of inertia for a region characterize the distribution of density in three dimensions. Each pattern of density has unique moments that describe its symmetry around its center of mass. Moments of inertia are calculated by constructing the inertia matrix (various density sums, weighted by lattice-point coordinates). The eigenvectors of the matrix define the inertial axes. So the matrix is diagonalized to obtain the eigenvalues, which are the corre-

sponding moments of inertia. The ratios of these moments provide additional information about the shapes of the density (spherical, ellipsoidal, etc.) and are included as three more features.

Statistical properties of the distribution of density within each sphere are used as features as well. The standard deviation describes the variation in the values of densities for each candidate sphere. The third and fourth moments of the distribution of data (skewness and kurtosis) are also features. Skewness 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 describes the "peakedness." The distributions that are sharply peaked will have less representation at the limits of the distribution while broad peaks may be over-represented. If two regions have a similar overall pattern of density, the statistical measurements of the distribution of density values should also be similar, regardless of orientation.

The last category of features attempts to describe the geometry of the density within each sphere. Given a sphere of density centered at an alpha carbon, for example, we expect that 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 spokes are representations of these tubes of density and are calculated by computing sums over the density, weighted by proximity to directional vectors originating at the center of the sphere, and taking the three strongest vectors at least 75° apart. By measuring the angles between these spokes, we are able to extract orientation-independent information about the arrangement of tubes of density within each sphere. There are expected to be similar angles between the spokes in similar regions of density. Also, the sum of the angles is an approximate indicator of the planarity of the three spokes and should be similar to other regions that have similar patterns of density.

The extraction of features from the database of maps is done separately (offline) from the model-building process. Since our experiments focused on building models for regions centered on C_α atoms, we extracted features for all regions around all C_α's in our database of 38 proteins. Each of the features was calculated for each region over four different radii: 3, 4, 5, and 6Å. This expanded the feature set to 60, allowing for the possibility that different features might be more effective when calculated over different radii (for example, to increase stability by covering a larger area, or to reduce sensitivity to noise by covering a smaller area). We leave the choice of radius for each feature up to the feature-weighting algorithm, described below.

During model-building, the features are extracted from a new region in an unknown map, and the feature comparison is performed to each region in the database using the pre-calculated feature values. Given the simplistic nature of the current features, feature values may occasionally be spuriously similar between an unknown region and a database region, even when the actual density patterns are dissimilar. However, the true measure of similarity is the density correlation. Therefore, we use regions ranked highly by small feature differences to determine a list of candidate regions for the computationally more expensive but more accurate density correlation.

Feature Weighting

Because some features may be more useful than others for distinguishing patterns of electron density, a weight was applied to each feature. The determination of weights was made using the *Slider* algorithm, which is described in detail elsewhere (Ioerger; submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*). The core algorithm involves finding the optimal mixture of two features at a time that maximizes relative rankings of a set of matches against a set of non-matches in a database of examples. Estimations of changes in rankings are made by solving simple linear equations. The method is extended to larger combinations of features by optimizing each against the rest in random order. We used pairwise matches (with $cc > 0.7$) among 263 C_α-regions in an electron density map for 1eny (enoyl acyl carrier protein reductase, InhA), to optimize the weights (and radii) of the features with *Slider*. For example, the four most highly weighted features were: ratio of first to third moments of inertia over 4Å (w=0.158), distance to center of mass over 4Å (w=0.148), average density over 5Å (w=0.145), and minimum spoke angle over 4Å (w=0.118). These weights only need to be computed once for a given set of rotation-invariant features, and were then incorporated as the w_i's in Eq. 1 to calculate distance scores for looking up matches in the database for each new region.

Modeling Experiments

To evaulate the potential of pattern matching for interpreting electron density maps, we used TEXTAL to build a model for an "unknown" protein, 1udi (uracil-DNA glycosylase). 1udi is a medium-sized protein with 244 residues (only 227 with coordinates defined), containing both α-helices and β-sheets. An artificial electron density map was generated for 1udi using the Gaussian-density procedure described above. The goal of this experiment was to model the local structure (determine the atoms and their coordinates) around each of the C_α atoms in 1udi using TEXTAL, and then compare these predictions to the known structure.

The database for this experiment consisted of the 60 feature values extracted for 8,055 regions centered on all C_α atoms in the 38 proteins in our database (computed offline). Then, for each region in the unknown (1udi), the following steps were taken:

1. Extract its 60 feature values.
2. Calculate the feature-based distance to each of the regions in the database, using Eq. 1 with the feature weights determined by Slider.
3. Rank the regions and keep the top $K = 50$ candidates.
4. Compute the density correlation to each of the candidate regions, using the peak-matching routine to determine the optimal rotation.
5. Identify the best match, with maximum density correlation.
6. Retrieve the coordinates for the backbone and side-chain atoms from the original protein in the database for the best match.
7. Translate the atoms to the origin and rotate them by multiplying their coordinates by the optimal rotation matrix found in calculating the density correlation (Step 4).
8. Translate the atoms into the new model, superimposing the C_α atom on the center of the region in the unknown.

This procedure was carried out on all 227 C_α positions in the map for 1udi. The resulting set of atomic coordinates were concatenated to construct the model. Both the identities of the matched residues and the similarities in atomic coordinates were compared to the original structure for 1udi.

Results

TEXTAL was able to identify high correlation matches for almost all of the regions in 1udi. The average density correlation for the top match to each region was 0.701, which is consistent with our informal observations of regions with visually similar patterns of density. Thus: 1) the database in this simplified context is large enough to contain adequate examples for finding matches for unknown regions, and 2) the feature-based lookup process is effective in filtering those matches to the top so they can be identified more quickly by evaluating density correlations. Not all of the residues had matches of equally high quality. Table 2 shows a breakdown of the average density correlations by residue type. Residues such as alanine, proline, and isoleucine had the highest correlation matches ($\overline{cc} > 0.73$), while cysteines had the lowest quality matches on average ($\overline{cc} = 0.59$). These trends are probably a combination of the effects of: 1) relative frequency of the individual residues in the overall database (e.g. cysteines are more rare), and 2) flexibility of side-chains in a protein context (e.g. alanine and valine are small and hydrophobic, with few degrees of freedom, increasing the structural similarity among instances, and hence providing a higher frequency of common density patterns).

Table 2: Average correlation coefficient of the best matches for residues of each type in 1udi.

Ala	Ser	Gly	Leu	Lys	Val	Thr
0.730	0.692	0.670	0.704	0.726	0.716	0.698
Pro	Glu	Asp	Asn	Ile	Gln	Arg
0.732	0.723	0.722	0.702	0.746	0.682	0.701
Phe	Tyr	Cys	His	Met	Trp	
0.655	0.676	0.587	0.670	0.707	0.641	

```
DWTTFRRVFLIDDAWRPLMEPELANPLTAHLLAEYNRRCQTEEVLPPRED
    ::: ::|: || |:: |:||:|| | |||||         | |||:
TELAYMMMYNINKAWHPNKMPQLADPLNAKLLAEYFDLVRSLRKLNPREL

VFSWTRYCTPDEVRVVIIGQDPYHHPGQAHGLAFSVRANVPPPPSLRNVL
 | :  |:||:  |:||:| :|  :||   || : |  : |||  | :::
RFQYKLYSTPNRNRIVITGRNPECYPGRPTGLRYLVYSDNPPPVRLEDIN

AAVKNCYPEARMSGHGCLEKWARDGVLLLNTTLTVKRGAAASHSRIGWDR
| | : :| |  || | | |:|: |:|   :|||  :|  |||  :||:
ACVDLQHPHATQSGQGLLDKHAMIGTLFVHVTLTYEMGSEASHANTGWTE

FVGGVIRRLAARRPGLVFMLWGTHAQNAIRPDPRVHCVLKFSHPSPLSKV
||||:   |||: |  :| | |||| :  |  |  | ||  ||| | | :
FVGGILQELAAKDPCEIFLLCGTHAMDKNRKEPVMHTVLTCSHPAPESYT

PFGTCQHFLVANRYLETRSISPIDWSV
|:||  :: |||||| |  : |:|:||
PYGTAKWHHVANRYLDTLEVTPVDHSV
```

Figure 2: Alignment of the sequence for the model built by TEXTAL (bottom) in comparison to the original sequence for 1udi (top). Exact matches are indicated with vertical bars, and *structurally* similar matches are indicated with a colon. Structural similarity was based on the following partition of the amino acids: A, G, CS, P, TVI, LDN, EQ, KRM, FWYH.

Even though matches with high density correlation were found for most of the regions in 1udi, the matching region was not constrained to have the same residue type. Nonetheless, identical residues were retrieved almost 42% of the time. Figure 2 shows an alignment of the sequence of the model built by TEXTAL to the original sequence for 1udi. There are many places where TEXTAL was able to recognize the exact amino acid based only on similarity in local electron density patterns. When a different amino acid was retrieved by TEXTAL, it often had at least a similar structure (61% of the time). This is reasonable, since residues such as glutamate and glutamine or valine and threonine are essentially indistinguishable based on density patterns, and residues such as valine and isoleucine or phenylalanine and histidine look so much alike that occasional mismatches are inevitable. In fact, the density patterns of all of the aromatic residues - His, Phe, Tyr, and Trp - generally differ only beyond the boundaries of the 5Å surface. Table 3 shows the number of

Table 3: Matches in 1udi by residue type. The first row gives the total number of occurrences of each residue in 1udi, the second row give the number of those that were matched by an identical residue in the model, the third row gives the number that were matched by a residue with a similar structure, and the last row gives the percentage of structurally-similar matches.

	Ala	Cys	Asp	Glu	Phe	Gly	His
# in 1udi	19	6	10	10	9	13	10
# ident.	12	0	1	2	3	12	4
# similar	12	1	5	3	8	12	6
% similar	63	17	50	30	89	92	60
	Ile	Lys	Leu	Met	Asn	Pro	Gln
# in 1udi	8	5	22	3	8	20	5
# ident.	2	1	13	0	1	17	0
# similar	6	1	17	1	6	17	0
% similar	75	20	77	33	75	85	0
	Arg	Ser	Thr	Val	Trp	Tyr	
# in 1udi	23	12	12	20	7	5	
# ident.	4	5	6	7	2	3	
# similar	9	5	7	13	5	4	
% similar	39	42	58	65	71	80	

residues of each type in 1udi that were exactly identified by TEXTAL, and the number that were matched to a residue with a similar structure. We note that a post-processing procedure that exploits knowledge of the sequence of the unknown protein might be able to resolve ambiguities of amino acid identity, and in almost all cases, a residue of the right type was found within the top 50 matches with only slightly lower density correlation.

In addition to correctly identifying many amino acids, TEXTAL often retrieved matches that had a similar molecular configuration. For example, residues in the model often had similar chi-angles in the side-chains and phi/psi-angles in the backbone. Figure 3 shows a striking example of the ability of TEXTAL to create a reasonable model for a segment of six residues in 1udi. Notice how Pro-43 is matched by another Pro, and how Glu-39 is matched by an almost identical rotamer of Gln. Generally speaking, TEXTAL correctly re-oriented the matched regions so that the atoms in the side-chain mapped onto the position of the side-chain in the unknown, and the backbone mapped onto the backbone. The orientation of the backbone was occasionally reversed ("flipped"). However, this occurred in only 11 out of 227 residues, and could also potentially be addressed using a post-processing routine.

To quantitatively assess the ability of TEXTAL to predict atomic coordinates, we measured the root-mean-square (RMS) distance between various subsets of atoms in the model and the original structure. The RMS for backbone atoms between 1udi and the model built by TEXTAL (excluding C_α atoms, which are guaranteed to have 0.0 RMS by the procedure) was 0.42Å (this calculation did not include the 11 residues

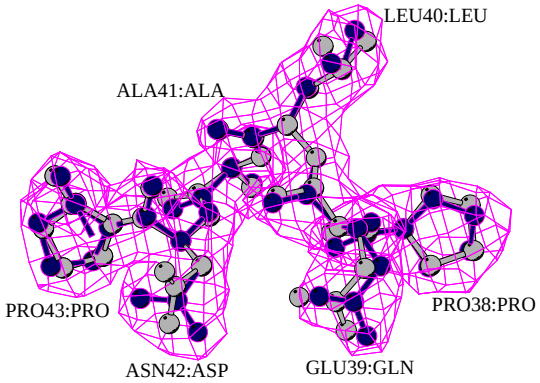

Figure 3: Comparison of a fragment in 1udi with the model built by TEXTAL. Amino acid positions 38 - 43 are shown. 1udi is shown in the darker shade, and the model is shown in the lighter shade. The density surface is a 1σ contour of the Gaussian map. (Image created using *MolScript*, P.J. Kraulis, 1991.)

with reversed backbones). Much of the variance in backbone coordinates was due to the carbonyl oxygens (RMS=0.74), which are notoriously difficult to place in density. The RMS for backbone nitrogens was 0.30Å. Comparing atoms in side-chains was more difficult because the matched residues did not always have the same structure. However, by heuristically pairing-up nearest neighbors (atoms within at least 3Å of an atom in the other region), the RMS among side-chain atoms was found to be 0.64Å, with over 89% of the atoms assigned to a partner.

Preliminary Results on Real Maps

To evaluate the potential of the approach described in this paper for solving real structures in a laboratory, we used TEXTAL to build models from two experimentally-derived electron density maps. These maps were constructed from the original structure factors for these proteins deposited in the PDB by using X-PLOR to take the inverse Fourier transform of $|F_{obs}|$, with phases calculated from the model, ϕ_{calc}. Thus, these maps have all the noise associated with poorly defined atoms in real proteins, though minimal phase error.

In order to construct reasonable models for these maps, we had to make several modifications to the TEXTAL program. First, the initial database contained maps constructed from Gaussian densities, which are not necessarily representative of the patterns of electron density found in real maps. Therefore, we constructed a new database of maps by generating structure factors from the atomic coordinates of the original protein structures directly (i.e. $|F_{calc}|$), and then back-transformed those for 50 proteins in the PDB. We also used several additional features based on simple geometric concepts, and we increased K to

Table 4: Results on using TEXTAL to solve real maps, and comparison to other methods. SMM = Segment Match Modeling (Levitt, 1992). MaxSprout (Holm and Sander, 1991). "all-a"=all-atom; "ma-chn"=main-chain; "si-chn"=side-chain.

method	protein	all-a RMS	ma-chn RMS	si-chn RMS
TEXTAL	crambin	2.21	1.88	2.80
	flavodoxin	2.5	1.9	3.2
SMM	crambin	1.51	0.64	2.17
	flavodoxin	1.71	0.54	2.39
MaxSprout	crambin			2.12
	flavodoxin	1.57	0.48	2.19

400 to improve the quality of the matching.

Maps for crambin (1ab1, 46 residues) and flavodoxin (1ag9, 198 residues) were constructed at a (fairly low) resolution of 2.8Å, with a 1Å grid-spacing, as were the maps in the database. Table 4 shows the results of modeling these two proteins. Between the models built by TEXTAL and the original structures, the RMS scores for all atoms are 2.2Å for crambin and 2.5Å for flavodoxin. By comparison, Levitt (1992) got around 1.5Å for crambin and 1.7Å for flavodoxin using Segment Match Modeling, and Holm and Sander (1991) got 1.6Å for flavodoxin using MaxSprout (all-atom RMS for crambin not reported).[1] Both of these other methods assume prior knowledge of the C_α positions, as we do, but they also assume the identity of each residue is known before modeling. Therefore, to make a fair comparison, we filtered the regions retrieved from the database by TEXTAL down to only those identical to the residue being modeled (which was not done in the experiments described in previous sections). Since our models are guaranteed to have the same sequence as the original structures, we can calculate the RMS in the standard way (without having to match-up neighboring atoms), and these are the scores reported in the table.

The overall RMS scores produced by TEXTAL are not quite as good as for other methods. However, the feature types and weights have not been optimized for pattern-matching in this database yet. When these RMS scores are broken down into their components, it can be observed that the RMS is slightly better for main-chain than for side-chain, as is commonly observed in other methods. The difference between main-chain and side-chain RMS is slightly lower than

[1]It should be noted that comparable results are not available from the Fortier and Glasgow group, since their electron-density map interpretation approach based on critical point analysis is aimed at main-chain tracing. Their results are hard to compare to ours since: a) we assume the C_α positions are known a priori, and b) they do not report RMS scores for side-chains, which is one of the priorities for modeling in TEXTAL.

Table 5: Results for real maps, without filtering out non-identical matches. The RMS calculation pairs up neighboring atoms and takes the average distance between such pairs.

protein	all-atom RMS(Å)	main-chain RMS(Å)
crambin	0.92	0.95
flavodoxin	0.95	0.92

reported in (Levitt 1992) and (Holm & Sander 1991). Our unusually high main-chain RMS could be due to the occurrence of flipped residues in the model, which we did not repair in this experiment (but could be removed via a post-processing routine). We have not yet added this or any other post-processing routines, such as energy minimization, which could improve the accuracy of our models in the future.

This experiment was conducted to compare models built for real maps by TEXTAL to results reported in the literature for other methods. However, it does not adequately reflect the accuracy of TEXTAL as it was intended to be used. In particular, we feel that requiring matches selected from the database to have the same identity as the residue being modeled biases our results negatively. For example, there might be a good match of a Gln to an Asp with high density correlation that would be rejected in favor of the best Asp in the database, which might have a much lower correlation. Therefore, in Table 5 we show results for constructing models for the real maps of crambin and flavodoxin as we intended for TEXTAL. The models constructed now do not have the same amino acid sequences as the original structures. Hence, the RMS scores we report are for the method of matching-up nearest-neighbor atoms and calculating the average distance between them, regardless of atom type. This is much less sensitive to main-chain flips, and also gives good scores for different side-chains superimposed in the same conformation in space. The all-atom RMS's using this methods are 0.92Å for crambin and 0.95Å for flavodoxin - much more representative of the quality of models that can be constructed by TEXTAL. Interestingly, the main-chain RMS scores are almost identical to the all-atom scores, and hence the side-chain RMS's too. This reflects the fact that the pattern recognition process in TEXTAL treats all local atoms around a C_α equally, and does not give special attention to main-chain atoms, as other methods do.

Discussion

Pattern recognition has proven quite effective in our initial experiments, and has the potential to make the automatic interpretation of electron density maps both fast and accurate. The modeling experiment we ran only took a few minutes per region being modeled (on an SGI O2 computer workstation, without any effort

to optimize), and most of the time was spent in computing the density correlations for the top 50 matches. This illustrates the importance of extracting features and using them to rank the large database of candidates.

Before extending the method to more realistic settings, one of the main limitations that we must address is how to eliminate the need to know C_α positions *a priori*, which are usually not available in a true unknown map. We could use a skeletonization algorithm as a pre-processing step to identify likely locations of C_α atoms. Alternatively, we could use pattern recognition itself to locate these positions, perhaps by training a neural network to learn to discriminate between C_α and non-C_α positions based on the rotation-invariant features. An important question would be how sensitive the pattern matching is to having the spheres centered exactly on a C_α.

There are several ways in which we could potentially improve TEXTAL. First, the performance of matching in TEXTAL could be enhanced by introducing new rotation-invariant features. There are many possible sources for new features, for example based on geometric shape analysis, arrangement of pseudo atoms in regions, contour surface areas, spherical harmonics, etc. Another possible improvement is to try to make more intelligent decisions based on the observation that some residues are easier to match than others (e.g. due to high prevalence or structural rigidity). If the best match for an unknown region is to a residue type that matches frequently, we might be more conservative and explore other candidates, perhaps by taking a vote among the top 50 matches.

We could also improve TEXTAL significantly by adding various post-processing routines to integrate the local models into a consistent global model. The prediction of residues with reversed main-chains could be addressed by rejecting such matches when the neighboring residues disagree. A similar constraint-processing procedure could be used to help disambiguate the identity of the amino acid being matched, given knowledge of the protein's sequence. Finally, after the global model is built, energy minimization could be applied to regularize the structure and hopefully reduce the atomic RMS.

Acknowledgements

This work was supported in part by a grant from the Interdisciplinary Research Initiatives Program, adminstered by the Office of the Vice President for Research and Associate Provost for Graduate Studies at Texas A&M University.

References

Brünger, A.; Kuriyan, J.; and Karplus, M. 1987. Crystallographic R-factor refinement by molecular dynamics. *Science* 235:458–460.

Chang, C.; Weeks, C.; Miller, R.; and Hauptman, H. 1997. Incorporating tangent refinement in the Shake-and-Bake formalism. *Acta Crystallographica* A53:436–444.

Fortier, S.; Castleden, I.; Glasgow, J.; Conklin, D.; Walmsley, C.; Leherte, L.; and Allen, F. 1993. Molecular scene analysis: The integration of direct methods and artificial intelligence for solving protein crystal structures. *Acta Crystallographica* D49:168–178.

Fortier, S.; Chiverton, A.; Glasgow, J.; and Leherte, L. 1997. Critical point analysis in protein electron density map interpretation. *Methods in Enzymology* 277:1–141.

Gaasterland, T. 1998. Structural genomics: Bioinformatics in the driver's seat. *Nature Biotechnology* 16:625–627.

Greer, J. 1985. Computer skeletonization and automatic electron density map analysis. *Methods in Enzymology* 115:206–224.

Holm, L., and Sander, C. 1991. Database algorithm for generating protein backbone and side-chain coordinates from a C-alpha trace. *Journal of Molecular Biology* 218:183–194.

Jones, T., and Kjelgaard, M. 1997. Electron-density map interpretation. *Methods in Enzymology* 277:173–208.

Jones, T., and Thirup, S. 1986. Using known substructures in protein model building and crystallography. *EMBO Journal* 5:819–822.

Jones, T.; Zou, J.; Cowan, S.; and Kjeldgaard, M. 1991. Improved methods for building protein models in electron density maps and the locations of errors in these models. *Acta Crystallographica* A47:110–119.

Levitt, M. 1992. Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology* 226:507–533.

Stout, G., and Jensen, L. 1989. *X-ray Structure Determination: A Practical Guide.* John Wiley and Sons.

Terry, A. 1983. The CRYSALIS project: Hierarchical control of production systems. Technical Report HPP-83-19, Stanford University, Department of Computer Science.