TEXTAL™: Artificial Intelligence Techniques for Automated Protein Structure Determination

Kreshna Gopal, Reetal Pai, Thomas R. Ioerger
Department of Computer Science
Texas A&M University
College Station TX 77843-311
{kgopal, reetalp, ioerger}@cs.tamu.edu

Tod D. Romo, James C. Sacchettini
Department of Biochemistry and Biophysics
Texas A&M University
College Station TX 77843-311
{tromo, sacchett}@tamu.edu

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 3D image of the electron density cloud surrounding the protein. This is often done manually by crystallographers and is very time-consuming and error-prone. The difficulties stem from the fact that the domain knowledge required for interpreting electron density data is uncertain. Thus crystallographers often have to resort to intuitions and heuristics for decision-making. The problem is compounded by the fact that in most cases, data available is noisy and blurred. TEXTAL™ is a system designed to automate this challenging process of inferring the atomic structure of proteins from electron density data. It uses a variety of AI and pattern recognition techniques to try to capture and mimic the intuitive decision-making processes of experts in solving protein structures. The system has been quite successful in determining various protein structures, even with average quality data. The initial structure built by TEXTAL™ can be used for subsequent manual refinement by a crystallographer, and combined with post-processing routines to generate a more complete model.

X-ray Protein Crystallography

Proteins are very important macromolecules that perform a wide variety of biological functions. Knowledge of their structures is crucial in elucidating their mechanisms, understanding diseases, designing drugs, etc. One of the most widely used methods for determining protein structures is X-ray crystallography, which involves many steps. First, the protein has to be purified and then grown into a crystal. The protein crystals are usually small, fragile and may contain as much solvent as protein. The crystal is exposed to beams of X-rays, and the position and intensity of diffracted waves are detected to make up a diffraction pattern sampled at points on a 3D lattice. But the diffraction pattern contains information only about the intensity of the diffracted wave; the phase information cannot be experimentally determined, and must be estimated by other means. This is referred to as the classic *phase problem*. Furthermore, the sample of points on which intensities are collected is limited, which effectively limits the resolution i.e. the degree to which atoms can be distinguished. Resolution is usually measured in Angstrom (Å), where 1Å = 10^{-10}m.

A Fourier transform of the observed intensities and approximated phases creates an *electron density map*, which is an image of a unit cell of the protein structure in the form of density of the electron cloud around the protein at various (x, y, z) coordinates. An initial approximate structure can be determined from the typically noisy and blurred electron density map, with the help of 3D visualization and model-building computer graphic systems such as O (Jones, Zou, and Cowtan 1991). The resulting structure can be used to improve the phases, and create a better map, which can be re-interpreted. The whole process can go through many cycles, and the complete interpretation may take days to weeks. Solving an electron density map can be laborious and slow, depending on the size of the structure, the complexity of the crystal packing, and the quality and resolution of the map.

Automated Map Interpretation: Issues and Related Work

Significant advances have been made toward improving many of the steps of crystallography, including crystallization, phase calculation, etc. (Hendrickson and Ogata 1997; Brünger et al. 1998). However, the step of interpreting the electron density map and building an accurate model of a protein remains one of the most difficult to improve.

This manual process is both time-consuming and error-prone. Even with an electron density map of high quality, model-building is a long and tedious process. There are many sources of noise and errors that can perturb the appearance of the map (Richardson and Richardson 1985). Moreover, the knowledge required to interpret maps (such as typical bond lengths and angles, secondary structure preferences, solvent-exposure propensity, etc) is uncertain, and is usually more confirmatory than predictive. Thus, decisions of domain experts are often made on the basis of what is most reasonable in specific situations, and generalizations are hard to formulate. This is quite inevitable since visible traits in a map are highly dependent on its quality and resolution.

AI-based approaches are well suited to mimic the decisions and choices made by crystallographers in map interpretation. Seminal work by (Feigenbaum, Engelmore, and Johnson 1977) led to the CRYSALIS project (Terry 1983), which was an expert system designed to build a model given an electron density map and the amino acid sequence. It is based on the *blackboard model*, where independent experts interact and communicate by means of global data structure, which contains an incrementally built solution. CRYSALIS maintains domain-specific knowledge through a hierarchy of production rules.

Molecular scene-analysis (Glasgow, Fortier, and Allen 1993) was proposed based on using computational imagery to represent and reason about the structure of a crystal. Spatial and visual analysis tools that capture the processes in mental imagery were used to try mimicking visualization techniques of crystallographers. This approach is based on geometrical analysis of *critical points* in electron density maps (Fortier et al. 1997). Several other groups including (Jones, Zou, and Cowtan 1991; Holm and Sander, 1991) have developed algorithms for automated interpretation of medium to high-resolution maps using templates from the PDB (Protein Data Bank).

Other methods and systems related to automated model-building include template convolution and other FFT-based approaches (Kleywegt and Jones 1997; Cowtan 1998), combining model building with phase refinement (Terwilliger 2000; Perrakis et al. 1999; Oldfield 1997) and database search (Diller et al. 1999), MAID (Levitt 2001), and MAIN (Turk 2001).

However, most of these methods have limitations, like requiring user-intervention or working on maps of high quality only (i.e. with resolution of around 2.3Å or better).

Overview of the TEXTAL™ system

TEXTAL™ is designed to build protein structures automatically from electron density maps, particularly those in the medium to poor quality range. The structure determination process is iterative and involves frequent backtracking on prior decisions. The salient feature of TEXTAL™ is the variety of AI and pattern recognition techniques used to address the specificities of the different stages and facets of the problem. It also attempts to capture the flexibility in decision-making that a crystallographer needs.

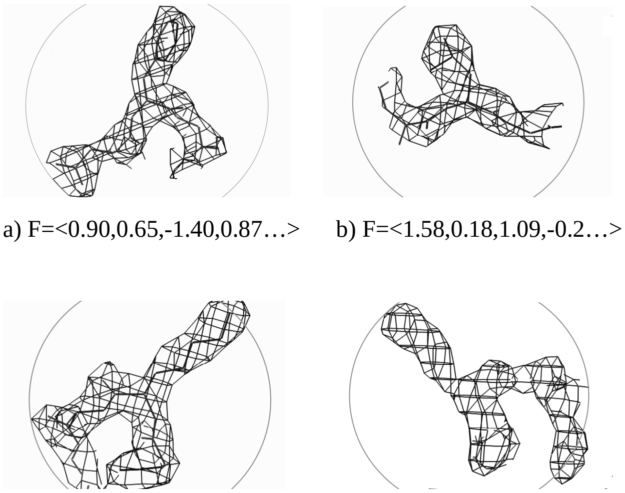
TEXTAL™ is primarily a *case-based reasoning* program. Previously solved structures are stored in a database and exploited to interpret new ones, by finding suitable matches for all regions in the unknown structure. The best match can be found by computing the *density correlation* of the unknown region with known ones. But this "ideal" metric is computationally very expensive, since it involves searching for the optimal rotation between two regions [see (Holton et al. 2000) for details], and has to be computed for many candidate matches in the database. This is a very common problem in case-based reasoning systems. In TEXTAL™, we use a filtering method where we devise an inexpensive way of finding a set of potential matches based on feature extraction, and use the more expensive density correlation calculation to make the final selection. The method to filter candidate matches is based on the Nearest Neighbor algorithm, which uses a weighted Euclidean distance between features as a metric to learn and predict similarity. There are two central issues related to this method: (1) identification of relevant features, and (2) weighting of features to reflect their contributions in describing characteristics of density.

As in many pattern-recognition applications, identifying features is a challenging problem. In some cases, features may be noisy or irrelevant. The most difficult issue of all is interaction among features, where features are present that contain information, but their relevance to the target class on an individual basis is very weak, and their relationship to the pattern is recognizable only when they are looked at in combination (Ioerger 1999). While some methods exist for extracting features automatically (Liu and Motoda 1998), currently consultation with domain experts is almost always needed to determine how to process raw data into meaningful, high-level features that are likely to have some form of correlation with the target classes, often making manual decisions about how to normalize, smooth, transform or otherwise manipulate the input variables (i.e. "feature engineering").

Applying these concepts of pattern recognition to TEXTAL™, we want to be able to recognize when two regions of electron density are similar. Imagine we have a spherical region of around 5Å in diameter (which is large enough to cover one side-chain), centered on a Cα atom and we have a database of previously solved maps. We would like to extract features from the unmodeled region that could be used to search the whole database efficiently for regions with similar feature values, with the hope of finding matching regions with truly similar patterns of density. This idea of feature-based retrieval is illustrated in Figure 1.

However, there is one important issue: candidates in the database with matching patterns for a region we are trying to interpret might occur in any orientation in 3D (rotational) space, relative to the search pattern. In principle, methods like 3D Fourier transforms could have been used to extract features, but they would have been sensitive to the orientation, which would have required the database to be much larger, to contain examples of every pattern in every orientation. Therefore, one of the initial challenges in TEXTAL™ was to develop a set of numeric features that are *rotation-invariant*. Statistical properties like average density and standard deviation are good examples of rotation-invariant features.

Once the features were identified, it was important to weight the features according to relevance in describing patterns of electron density; irrelevant features can confuse the pattern-matching algorithm (Langley and Sage 1994; Aha 1998). The SLIDER algorithm (Holton et al. 2000) was developed to weight features according to relevance by considering how similar features are for pairs of matching regions relative to pairs of mismatching regions. A more detailed discussion on SLIDER is provided later.



a) F=<0.90,0.65,-1.40,0.87…> b) F=<1.58,0.18,1.09,-0.2…>

c) F=<1.72,-0.39,1.04,1.55…> d) F=<1.79,0.43,0.88,1.52…>

Fig. 1. Illustration of feature-based retrieval. In the four panels above are shown examples of regions of density centered on Cα atoms. In panels (a), (b), and (c) are shown representative density for amino acids Phenylalanine, Leucine, and Lysine respectively (the circle indicates 5Å-radius sphere). In panel (d) is a region of unknown identity and coordinates [actually a Lysine, but oriented differently from (c)]. Feature values like average density, standard deviation of density, distance to center of mass in region, moments of inertia, etc. can be used to match it to the most similar region in the database.

Rotation-invariant features were extracted for a large database of regions within a set of electron density maps for proteins whose structures are already known. Hence, the atomic coordinates for a region in a new (unsolved) map could be estimated by analyzing the features in that region, scanning through the database to find the closest matching region from a known structure (a procedure referred to as LOOKUP), and then predicting atoms to be in analogous locations. To keep the LOOKUP process manageable, the regions in the TEXTAL™ database are restricted to those regions of an electron density map that are centered on known Cα coordinates.

Clearly, the effectiveness of this approach hinges on the ability to identify candidate Cα positions accurately in the unknown map, which become the centers of the probe regions for LOOKUP. In TEXTAL™, this is done by the CAPRA (C- Alpha Pattern Recognition Algorithm) sub-system. CAPRA uses the same rotation-invariant features as LOOKUP, though it employs a neural network to predict which positions along a backbone trace are most likely to be closest to a true Cα in the structure. CAPRA uses a heuristic search technique to link these putative Cα atoms together into linear chains; the remaining backbone and side-chains atoms are filled in by performing LOOKUP on each Cα-centered region along the predicted chains.

Thus, TEXTAL™ is intended to simulate the kind of intelligent decision-making that crystallographers use to interpret electron density maps, following the basic two-step approach of main-chain tracing followed by side-chain modeling. The AI and pattern-recognition techniques employed approximate many of the constraints, criteria, and recognition processes that humans intuitively use to make sense out of large, complex, 3D datasets and produce coherent models that are consistent with what is known about protein structure in general. The model obtained from TEXTAL™ can be edited by a human crystallographer or used to generate higher quality maps through techniques like *reciprocal space refinement* or *density modification*.

In the next three sections, we provide a more detailed description of the main stages of TEXTAL™: CAPRA, LOOKUP and post-processing routines (Figure 2).

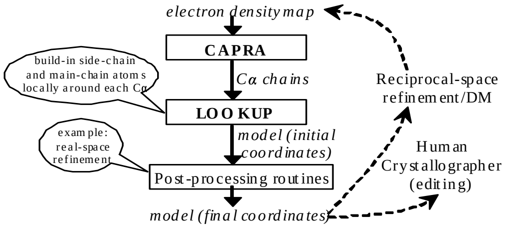

electron densitymap
CAPRA
build-in side-chain and main-chain atoms locally around each Cα
Cα chains
LOOKUP
example: real-space refinement
model (initial coordinates)
Post-processing routines
model (final coordinates)
Reciprocal-space refinement/DM
Human Crystallographer (editing)

Fig. 2. Main stages of TEXTAL™.

CAPRA: C-Alpha Pattern Recognition Algorithm

CAPRA (Ioerger and Sacchettini 2002) is the component that constructs Cα backbone chains; it operates in essentially four main steps (see Figure 3): first, the map is scaled to enable comparison of patterns between different maps. Then, a "trace" of the map is made. Tracing is done in a way similar to many other *skeletonization* algorithms (Greer 1985; Swanson 1994). The trace gives a connected skeleton of "pseudo-atoms" that generally goes through the medial axis of the contours of the density pattern. Note that the trace goes not only along the backbone, but also branches out into side-chains.

CAPRA picks a subset of the pseudo-atoms in the trace (which we refer to as "way-points") that appear to represent Cα atoms. To determine which of these pseudo-atoms are likely to be near true Cα's, CAPRA uses a standard feed-forward network. The goal is to learn how to associate certain characteristics in the local density pattern with an estimate of the proximity to the closest Cα. The network consists of one input layer of 38 feature values (19 features for 2 different radii), one layer of hidden units with sigmoid thresholds, and one output node: the predicted distance (unthresholded). The hidden layer has 20 nodes and the network is fully interconnected between layers.

The neural network was trained by giving it examples of feature vectors for high-density lattice points in maps of sample proteins at varying distances from known Cα atoms, ranging from 0 to around 6Å. The weights in the network are optimized on this dataset using back-propagation.

Given these distance predictions, the set of candidate Cα's (i.e. way-points) is selected from all the pseudo-atoms in the trace. Preference is given to those that are deemed to be closest to Cα's (by the neural network). The selection of way-points is also based on domain knowledge about constraints on distance between Cα's.

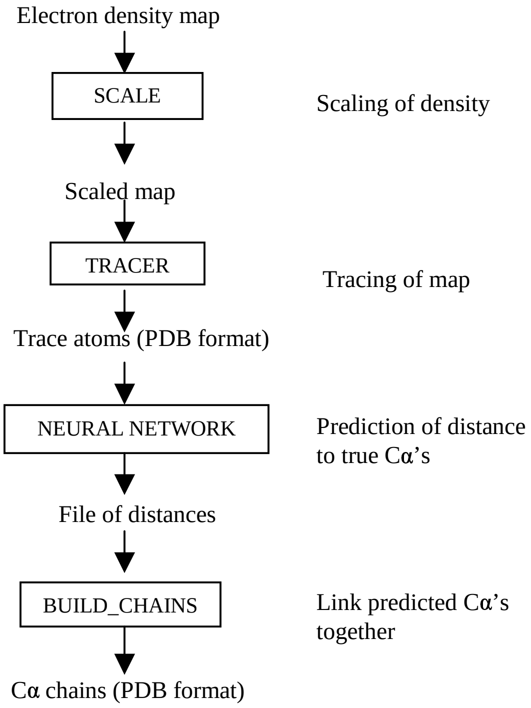


Fig. 3. Steps within CAPRA.

The final step is to link Cα's together into linear chains. This is accomplished by BUILD_CHAINS. Finding correct assignment of Cα atoms into chains is difficult because there are often many false connections in the density. Note that the trace is not generally linear, but is a graph with branches, and often contains many cycles. BUILD_CHAINS integrates a variety of intuitive criteria to try to make intelligent decisions about how to identify the most reasonably linearized sub-structure of the graph, including chain length, quality of predictions by neural network, and geometry (attempting to follow chains with common secondary structure characteristics). Wherever possible, BUILD_CHAINS does an extensive search of all possible ways of building up chains and chooses the best one according to a scoring function. In situations where an exhaustive search will be inefficient, BUILD_CHAINS uses heuristics to guide the search.

LOOKUP: The Core Pattern-Matching Routine

LOOKUP predicts the coordinates of local side-chain and backbone atoms of an amino acid, given an estimate of its Cα location (output from CAPRA). A pattern-matching approach is used to retrieve atoms from regions with similar patterns of density from a database using rotation-invariant features. Before describing the details of this database-search approach, we need to describe the features extracted to represent density patterns.

Feature Extraction

TEXTAL™ relies heavily on the extraction of numerical features to help determine which regions have similar patterns of density. As we described above, it is important to have features that are rotation-invariant i.e. they have a constant value for the same region rotated into any orientation in 3D space.

In previous work, we have identified four classes of features, each with several variations (Holton et al. 2000). For example, statistical features, such as the mean, standard deviation, skewness and kurtosis of the density distribution, can be calculated from the density values at grid points that fall within a spherical region. Another class of features is based on moments of inertia. We calculate the inertia matrix and recover the eigenvalues for the three mutually perpendicular moments of inertia. The eigenvalues themselves can be used as features, but we have also found that it is especially useful to look at ratios of the eigenvalues, which give a sense of the way that the density is distributed in the region. Another useful feature, in a class by itself, is the distance to the center of mass, which measures how balanced the region is. Finally, there is a class of features based on the geometry of the density.

While many other features are possible, we have found these to be sufficient. In addition, each feature can be calculated over different radii, so they are parameterized; currently, we use 3Å, 4Å, 5Å and 6Å, so every individual feature has four distinct versions capturing slightly different information.

Searching the Region Database Using Feature Matching

Given an estimate of the coordinates of a Cα atom from CAPRA, LOOKUP predicts the coordinates of the other (backbone and side-chain) atoms in the vicinity, using a database-lookup approach. The database consists of feature vectors extracted from regions within previously solved maps. The features for the new region to be modeled are calculated and used to identify the region in the database with the most similar pattern of density; since coordinates of atoms are known for these regions, they can be translated and rotated into position in the new region as a prediction (model) of local structure (the overall process is illustrated in Figure 4).

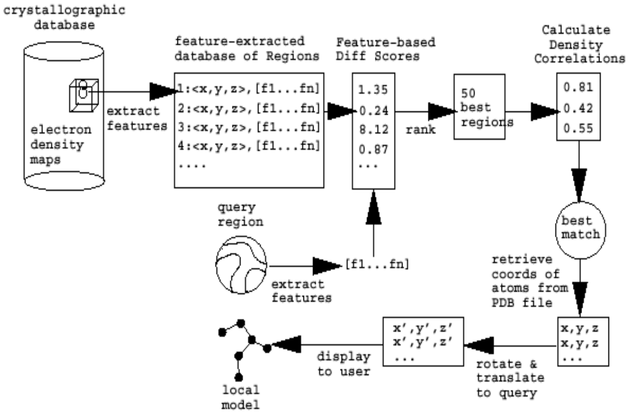


Fig. 4. The LOOKUP process.

The database of feature-extracted regions that TEXTAL™ uses is derived from maps of 200 proteins from PDBSelect (Hobohm et al. 1992). The maps are re-computed at 2.8Å resolution (to simulate medium resolution maps). Features are calculated for a 5Å spherical region around each Cα atom in each protein structure for which we generated a map, producing a database with ~50,000 regions.

To find the most similar region in the database for a given region in a new map, we use a three-step process. First, the features for a new region are calculated. Then they are compared to the feature vectors for each of the regions in the database. The comparison we use is a weighted Euclidean distance, given by the following: $\Delta F(R_1,R_2) = \{\sum(w_i[F_i(R_1) - F_i(R_2)]^2\}^{1/2}$, where i ranges over the features, and R_1 and R_2 are the two regions being compared. The weights w_i are intended to reflect the relevance or utility of the features, and are determined by a specialized feature-weighting algorithm called SLIDER (which is described in the next section). This distance measure is calculated from the probe region to all the regions in the database, and the top K=400 (with smallest distance values) are selected as candidates.

The calculation of feature-based distance, however, is not always sufficient; there could be some spurious matches to regions that are not truly similar. Hence we use this selection initially as a filter, to catch at least some similar regions. Then we must follow this up with the more computationally expensive step of further evaluating the top K candidate matches by density correlation, and choosing the best one.

The final step is to retrieve the coordinates of atoms from the known structure for the map from which the matching region was derived, specifically, the local side-chain and backbone atoms of the residue whose Cα is at the center of the region, and apply the appropriate transformations to place them into position in the new map. The resulting side-chain and backbone atoms are written out in the form of a new PDB file, which is the initial, unrefined model generated by TEXTAL™ for the map.

Weighting of Features: The SLIDER Algorithm

TEXTAL™ uses a weighted Euclidean feature-difference calculation, $\Delta F(R_1,R_2)$ (defined earlier), as an initial measure of similarity between regions R_1 and R_2. It is important to weight the features according to relevance in describing patterns of electron density. The SLIDER (Holton et al. 2000) algorithm was developed to weight features by considering how similar features are for pairs of matching regions relative to pairs of mismatching regions.

While there are a variety of methods that have been proposed in the pattern-recognition literature for optimizing feature weights for classification problems (Aha 1998), our goal is slightly different: to optimize feature-based retrieval of similar matches from a database, where true similarity is defined by an objective distance metric (density correlation). SLIDER works by incrementally adjusting feature weights to make matches (similar regions) have a closer apparent distance than mis-matches. As data for this empirical method, a set of regions is chosen at random. For each region, a match (with high density correlation) and a mis-match are found, forming 3-tuples of regions. These 3-tuples are used to guide the tuning of the weights. Suppose the set F of all features is divided into two subsets, A and B. Each subset can be used to compute distances between examples. Good subsets of features are those that rank the match for a region higher (with lower distance) than the mis-match, on average over the 3-tuples. Furthermore, the subsets of features can be mixed together by linear combination to form a composite distance metric, $\Delta_{A+B}(R_1,R_2) = \lambda\Delta_A(R_1,R_2) + (1-\lambda)\Delta_B(R_1,R_2)$, with the parameter λ. As λ changes from 0 to 1, it may cause the match for a region to become closer or farther relative to the mis-match for each 3-tuple. The point at which the distance from a region to its match becomes equal to the distance to the mis-match is called a 'cross-over'. The weights can be optimized by finding the value of $(0 \leq \lambda \leq 1)$ that produces the most positive crossovers among the set of 3-tuples. Then the process can be repeated with different (random) divisions of the overall set of features until it converges (the number of positive cross-overs reaches a plateau). This approach bears some resemblance to wrapper-based methods (Kohavi, Langley, and Yun 1997), but replaces a grid-search through the space of weight vectors with a more efficient calculation of optimal cross-over points.

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.

Post-Processing Routines

There are a number of ways in which the initial protein model output by LOOKUP might be imperfect. For example, because residues are essentially modeled independently (based on regions most likely coming from entirely different molecules in the region database), the backbone connections do not necessarily satisfy optimal bond distance and angle constraints. Often, TEXTAL™ identifies the structure of the side-chains correctly, but makes errors in the stereochemistry. There are a number of possible refinements that can be applied to improve the model by fixing obvious mistakes during post-processing. Currently, there are three important post-processing steps we use.

The first post-processing step is a simple routine to fix residues whose backbone atoms are going in the wrong direction with respect to their neighbors. It determines chain directionality by a voting procedure, and then re-invokes LOOKUP to correct the residues whose side-chain or backbone atoms are pointed in a direction inconsistent with the rest of the chain.

The second post-processing step is real-space refinement (Diamond 1971), which tends to move atoms slightly to optimize their fit to the density, while preserving geometric constraints like typical bond distances and angles.

Another post-processing step involves correcting the identities of mis-labeled amino acids. Recall that, since TEXTAL™ models side-chains based only on local patterns in the electron density, it cannot always determine the exact identity of the amino acid, and occasionally even predicts slightly smaller or larger residues due to noise perturbing the local density pattern. However, up to two-thirds of the time in real maps used, TEXTAL™ calls a residue that is at least structurally similar to the correct residue. We could correct mistakes about residue identities using knowledge of the true amino acid sequence of the protein, if we knew how the predicted fragment mapped into this sequence. The idea is to use *sequence alignment* techniques (Smith and Waterman 1981) to determine where each fragment maps into the true sequence; then the correct identities of each amino acid could be determined, and another scan through the list of candidates returned by LOOKUP could be used to replace the side chains with an amino acid of the correct type at each position.

Results

TEXTAL™ was run on a variety of real electron density maps, which cover a range of medium resolutions (2-3Å), and include a variety of α-helices and β-sheet structures. All these maps have been obtained through a variety of data collection methods, and have had some sort of density modification applied (using CNS). Table 1 summarizes the details of these 12 test cases. The maps were re-computed at 2.8Å, the resolution at which TEXTAL™ has been optimized for. The results of CAPRA and LOOKUP are presented in Table 2. Figure 5 shows the Cα chains obtained for MVK, and Figure 6 shows a fragment of CzrA to illustrate the result of LOOKUP.

The r.m.s error in the Cα coordinates predicted is typically less than 1Å, compared to manually-built and refined models. CAPRA usually builds 80-95% of the backbone, creating several long Cα chains with a few breaks. The paths and connectivity that CAPRA chooses are often visually consistent with the underlying structure, only occasionally traversing false connections through side-chain contacts. It tends to produce Cα atoms correctly spaced at about 3.8Å apart, and corresponding nearly one-to-one with true Cα's, leaving a few skips and spurious insertions.

Given the Cα chains from CAPRA as input, the side-chain coordinates predicted by LOOKUP matched the local density patterns very well, and the additional (non-Cα) atoms in the backbone were also properly fit. The all-atom r.m.s. error of TEXTAL™ models compared to manually-built and refined ones is close to 1Å, and the mean density correlation of residues is close to 0.8. These suggest a rather good superposition of the predicted Cα's as well as side-chain atoms. Although LOOKUP does not always predict the correct identity of the residue in each position, it can find *structurally similar* residues with reasonable accuracy (typically 30-50%). It should be noted that low similarity score is often related to diffused density of residues at the surface. Furthermore, these results were obtained without sequence alignment, which is still being tested. Please refer to (Ioerger and Sacchettini 2003) for an in-depth discussion of the performance.

Discussion

TEXTAL™ has the potential to reduce one of the last major bottlenecks standing in the way of high-throughput Structural Genomics (Burley et al. 1999). By automating the final step of model building (for noisy, medium-to-low resolution maps), less effort and attention will be required of human crystallographers. The protein structures constructed by TEXTAL™ from electron density maps are fairly accurate. The neural network approach to recognize Cα's, coupled with heuristics for linking them together, can accurately model the backbone. The feature-based method enables efficient filtering of good matches from the database. The case-based reasoning strategy exploits solved structures and enables fairly accurate modeling of side chains. For poor quality maps, the relationship between density and structure is weak, and modeling necessitates a knowledge-based approach. In TEXTAL™ this knowledge is encoded in the database of solved structures.

There are many additional ideas that can or are being tested to improve TEXTAL™'s accuracy, such as adding new features, clustering the database or integrating model building with other computational methods, such as *phase*

Table 1. Proteins used in this study. All the maps were rebuilt at 2.8 Å using CNS. The proteins studied cover a range of sizes (as seen from the number of residues in each protein) and were obtained from different map generation routines. The proteins also cover the major secondary structure classes.

Name of protein	Abbreviation	Method of map generation	Original map resolution (Å)	Secondary structure	No. of residues
α2u-globulin	A2u-globulin	MR+NCS	2.50	β	158
β-Catenin	Armadillo	MAD	2.40	α	469
Cyanase	Cyanase	MAD	2.40	α+β	156
Sporulation Regulatory Protein	Gere	MAD	2.70	α	66
Granulocyte Macrophage Colony-Stimulating Factor (GM-CSF)	GM-CSF	MIRAS+NCS	2.35	α	118
N-Ethylmaleimide Sensitive Factor	Nsf-d2	MAD	2.40	α/β	370
Penicillopepsin	Penicillopepsin	MIR	2.80	β	323
Postsynaptic Density Protein	Psd-95	MAD	2.50	α/β	294
G-Protein Rab3a	Rab3a	MAD	2.60	α/β	176
Haloalkane Dehalogenase	Rh-dehalogenase	MIRAS	2.45	α/β	290
Chromosome-determined Zinc-responsible operon A	CzrA	MAD/MR	2.30	α	94
Mevalonate kinase	MVK	MAD	2.40	α/β	317

* The data for the first ten proteins were collected from various researchers and processed by Dr. Paul Adams (Lawrence Berkeley National Lab). The original references for each structure are available upon request (Email: ioerger@cs.tamu.edu).

Table 2. Results of CAPRA & LOOKUP. § The ratio of the structure built compared to the manually built and refined model. ¶ The r.m.s error of the Cα predictions relative to the refined model. ¥ The mean density correlation between the regions of the protein and their corresponding matches retrieved from the database. Ŧ The r.m.s. error of all the atoms relative to the manually built and refined model. Ψ The structural similarity between the residue selected for a region and the actual residue.

Protein	No. of chains output	Length of longest chain	Mean length of output chains	% of structure built §	Cα rms error (Å) ¶	Mean residue density corr. ¥	All-atom rms error (Å) Ŧ	% Side chain structural similarity Ψ
A2u-globulin	2	88	68.5	85	0.851	0.84	0.99	48.9
Armadillo	9	217	46.7	89	0.979	0.82	N.A.	43.7
Cyanase	6	94	32.0	94	1.099	0.79	1.03	42.7
Gere	2	44	30.5	90	0.854	0.83	1.00	30.0
GM-CSF	4	46	25.0	82	0.911	0.84	0.94	28.9
Nsf-d2	6	79	39.5	92	0.963	0.83	1.13	33.5
Penicillopepsin	13	58	25.0	91	1.136	0.78	1.09	41.9
Psd-95	8	58	31.8	94	1.000	0.82	1.04	34.7
Rab3a	8	30	20.5	90	0.905	0.82	1.06	30.5
Rh-dehalogenase	8	66	36.5	97	0.924	0.83	0.99	54.6
CzrA	3	57	33.3	94	1.054	0.82	1.15	39.1
MVK	10	58	28.7	88	0.833	0.82	1.00	44.5

refinement (Murshudov, Vagin, and Dodson 1997; Brünger et al. 1998). But even in its current state, TEXTAL™ is of great benefit to crystallographers.

Although the output model may still need to be edited and refined (especially in places where the density itself is poor), generating an initial model that is approximately correct saves an enormous amount of crystallographers' time.

Currently, access to TEXTAL™ is being provided through a website (http://textal.tamu.edu:12321), where maps can be uploaded and processed on our server. Since its release in June 2002, an average of 2 maps have been regularly submitted to the TEXTAL™ website every week. The development of TEXTAL™ started in 1998, and currently the system consists of ~72,000 lines of C/C++ code, with a few programs in Fortran, Perl and Python. The system is currently being incorporated as the automated structure determination component in the PHENIX crystallographic computing environment currently under development at the Lawrence Berkeley National Lab (Adams et al. 2002). The alpha release of PHENIX is planned for March 2003.

This work is supported by grant P01-GM63210 from the National Institutes of Health.

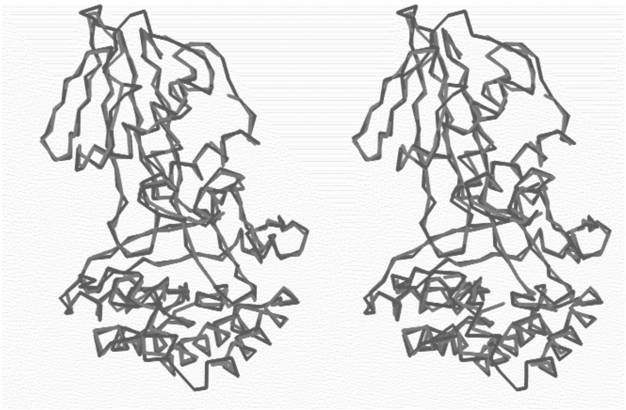


Fig. 5. CAPRA chains for MVK (in green or light grey), with Cα trace of manually built model superimposed (in purple or dark grey).

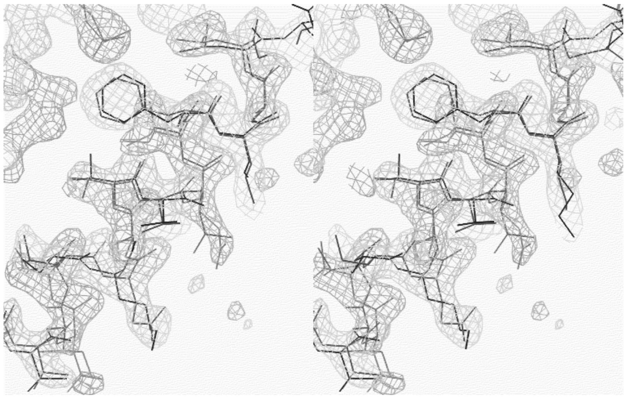


Fig. 6. A fragment of an α-helix in CzrA is shown where LOOKUP guessed the identities of four consecutive residues correctly, and put the atomic coordinates in extremely good superposition of the model built by hand.

References

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., and Terwilliger, T.C. 2002. PHENIX: Building new software for automated crystallographic structure determination. *Acta Cryst*. D58:1948-1954.

Aha, D.W. 1998. Feature Weighting for Lazy Learning Algorithms. In Liu H., and Motoda, H. eds. *Feature Extraction, Construction and Selection: A Data Mining Perspective.* Boston, MA: Kluwer.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nigles, M., Pannu, N.S., Read, R.J., Rice, L.M., Simmonson, T., and Warren, G.L. 1998. Crystallography & NMR System. *Acta Cryst*. D54:905-921.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, W., and Swaminathian, S. 1999. Structural genomics: Beyond the Human Genome Project. *Nature Genetics* 232:151-157.

Cowtan, K. 1998. Modified phased translation functions and their application to molecular fragment location. *Acta Cryst*. D54:750-756.

Diamond, R. 1971. A real-space refinement procedure for proteins. *Acta Cryst*. A27:436-452.

Diller, D.J., Redinbo, M.R., Pohl, E., and Hol, W.G.J. 1999. A database method for automated map interpretation in protein crystallography. *PROTEINS: Structure, Function, and Genetics* 36:526-541.

Feigenbaum, E.A., Engelmore, R.S., and Johnson, C.K. 1997. A correlation Between Crystallographic Computing and Artificial Intelligence Research. *Acta Cryst*. A33:13-18.

Fortier, S., Chiverton, A., Glasgow, J., and Leherte, L. 1997. Critical-point analysis in protein electron density map interpretation. *Methods in Enzymology* 277:131-157.

Glasgow, J., Fortier, S., and Allen, F. 1993. Molecular scene analysis: Crystal structure determination through imagery. In Hunter, L., ed. *Artificial Intelligence and Molecular Biology*. Cambridge, MA: MIT Press.

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

Hendrickson, W.A. and Ogata, C.M. 1997. Phase determination from multiwavelength anomalous diffraction measurements. *Methods in Enzymology* 276:494-523.

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. 1992. Selection of a represetative set of structures from the Brookhaven Protein Data Bank. *Protein Science* 1: 409-417.

Holm, L. and Sander, C. 1991. Database algorithm for generating protein backbone and side-chain coordinates from a Cα trace. *J. Mol. Biol*. 218:183-194.

Holton, T.R., Christopher, J.A., Ioerger, T.R., and Sacchettini, J.C. 2000. Determining protein structure from electron density maps using pattern matching. *Acta Cryst*. D46:722-734.

Ioerger, T.R. 1999. Detecting feature interactions from accuracies of random feature subsets. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 49-54. Menlo Park, CA: AAAI Press.

Ioerger, T.R. and Sacchettini, J.C. 2002. Automatic modeling of protein backbones in electron-density maps via prediction of C-alpha coordinates. *Acta Cryst.* D5:2043-2054.

Ioerger, T.R and Sacchettini, J.C. 2003. The TEXTAL system: Artificial Intelligence Techniques for Automated Protein Model-Building, In Sweet, R.M. and Carter, C.W. eds. *Methods in Enzymology*. Forthcoming.

Jones, T.A., Zou, J.Y., and Cowtan, S.W. 1991. Improved methods for building models in electron density maps and the location of errors in these models. *Acta Cryst*. A47:110-119.

Kleywegt, G.J. and Jones, T.A. 1997. Template convolution to enhance or detect structural features in macromolecular electron density maps. *Acta Cryst*. D53:179-185.

Kohavi, R., Langley, P., and Yun, Y. 1997. The utility of feature weighting in nearest-neighbor algorithms. In *Proceedings of the European Conference on Machine Learning*, Prague, Czech Republic, poster.

Langley, P. and Sage, S. 1994. Pruning irrelevant features from oblivious decision trees. In *Proceedings of the AAAI Fall Symposium on Relevance*, 145-148. New Orleans, LA: AAAI Press.

Levitt, D.G. 2001. A new software routine that automates the fitting of protein X-ray crystallographic electron density maps. *Acta Cryst*. D57:1013-1019.

Liu, H. and Motoda, H. eds. 1998. *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Boston, MA: Kluwer.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. 1997. Refinement of macromolecular structures by the Maximum-Likelihood method. *Acta Cryst.* D53:240-255.

Oldfield, T.J. 1996. A semi-automated map fitting procedure. In Bourne, P.E. and Watenpaugh, K. eds. *Crystallographic Computing 7, Proceedings from the Macromolecular Crystallography Computing School*. Corby, UK: Oxford University Press.

Perrakis, A., Morris, R., and Lamzin, V. 1999. Automated protein model-building combined with iterative structure refinement. *Nature Structural Biology* 6:458-463.

Richardson, J.S. and Richardson, D.C. 1985. Interpretation of electron density maps. *Methods in Enzymology* 115:189-206.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol*. 147:195-197.

Swanson, S.M. 1994. Core tracing: Depicting connections between features in electron density. *Acta Cryst.* D50:695-708.

Terry, A. 1983. The CRYSALIS Project: Hierarchical Control of Production Systems. Technical Report HPP-83-19, Stanford University, Palo Alto, CA.

Terwilliger, T.C. 2000. Maximum-likelihood density modification. *Acta Cryst.* D56:965-972.

Turk, D. 2001. Towards automatic macromolecular crystal structure determination. In Turk, D. and Johnson, L. eds. *Methods in Macromolecular Crystallography*. NATO Science Series I, vol. 325, 148-155.