Textal: AI-Based Structural Determination for X-ray Protein Crystallography

Tod Romo, Kreshna Gopal, Erik McKee, Lalji Kanbi, Reetal Pai, Jacob Smith, James Sacchettini, and Thomas Ioerger, *Texas A&M University*

A central tenet to biology is that structure implies function. So, knowledge of a protein's structure leads to understanding how it works. You can then use this knowledge to engineer better proteins or solve problems that nature never encountered. Understanding proteins' role in disease can also help us understand the disease itself and design drugs that prevent or cure the disease (see the sidebar "Proteins").

Protein structures are most commonly determined by *x-ray crystallography*, an experimental technique.[1] Crystallography involves many difficult steps, from cloning and crystallization to data collection and phasing, but they can be automated through robotics.[2] The final step is building the protein model to fit the experimental data. This requires an experienced crystallographer and an appreciable time commitment and has been resistant to automation. Building the protein model involves determining the coordinates of the atoms that make up the protein by interpreting an electron density map (derived from x-ray diffraction data by a Fourier transform; see the "X-ray Crystallography" sidebar). In the past, a crystallographer built the model manually by applying biophysical and chemical knowledge to visually interpret the density patterns and by deciding how best to fit the atoms into the density. This traditional approach is both time-consuming and error prone. Time is no longer a luxury crystallographers can afford given the growth of structural genomics initiatives that generate thousands of potential candidates for structural determination.[3] Although several automated methods for model building are available, they're geared toward higher-resolution data.[4–8] They often don't build good models when the x-ray data resolution is around 2.5 angstroms (Å) or lower, as is typical with larger macromolecules, particularly when the data collection occurred in the local laboratory as opposed to a synchrotron.

Textal is a successfully deployed system for automated model-building in protein x-ray crystallography. It represents a novel solution to an important, complex real-world problem using various AI and pattern recognition algorithms. Textal takes a model-building approach based on real-space density pattern recognition, similar to how a human crystallographer would work. This approach potentially tolerates more noise and has been optimized for medium-resolution x-ray data in the 2.4 Å to 3.0 Å range. Textal first tries to predict the coordinates of the alpha-carbon (Cα) atoms in the protein's connected backbone using a neural network. It then analyzes the density patterns around each Cα atom and searches a database of previously solved structures for regions with similar patterns. Textal determines the best match, retrieves the coordinates for that region, and fits them to the unknown density. Textal concatenates these local models into a global model and subjects them to various subsequent refinements to produce a complete protein model automatically. The whole process typically takes between 15 minutes and three hours, depending on the protein's size. Figure 1 shows Textal's results on a test protein, CZRA, which has the Protein Data Bank code 1R1V.

How Textal works

Textal begins with either an electron density map or structure factors (x-ray diffraction data) as input, along with an amino acid sequence and information about the symmetry of the crystal used. It outputs a Protein Data Bank file containing the solution structure. Textal combines AI, pattern matching, and non-AI methods into three principal phases (see figure 2) that mimic how a human crystallographer would solve an unknown structure:

1. Capra, the C-Alpha Pattern Recognition Algorithm, determines where the Cα atoms should be placed via a neural network. It also decides how to connect them to form the protein's core scaffolding by using heuristic search methods and case-based reasoning.[9]
2. Lookup then models the density surrounding each Cα atom determined by Capra. It does this by comparing the unknown density in a sphere about the atom to a known database using pattern matching and nearest-neighbor learning.[9] Lookup concatenates these local models to form the full protein structure. Recently, we incorporated an optimization phase into Lookup based

1541-1672/05/$20.00 © 2005 IEEE
Published by the IEEE Computer Society

Proteins

Proteins are macromolecules consisting principally of amino acids, of which 20 commonly occur in nature.[1,2] A protein's basic organization is a backbone, formed by a regular, repeating series of *peptide bonds* that link the amino acids. Hanging off the backbone are *side chains*, which are different for each amino acid. They connect to the backbone at a specific carbon atom called the *alpha-carbon* (cα). These polypeptides typically fold up onto themselves into a compact 3D structure to become biologically active. A protein's basic topology is often described by the path of its backbone, consisting of helical fragments and parallel strands (this local topology is called the *secondary structure*), all of which pack together into complex shapes. The side chains then fill the space between the backbone elements.

References

1. L. Hunter, *Artificial Intelligence for Molecular Biology*, AAAI Press, 1993.
2. D. Voet and J. Voet, *Biochemistry*, 3rd ed., John Wiley & Sons, 2004.

on the Nelder-Meade simplex. This not only improved the fit of the side chains to their density but also Textal's ability to pick the correct amino acid type.

3. Post-processing encompasses numerous steps that further refine the solution from Lookup by incorporating sequence information (sequence alignment) and further optimize the fit of the modeled side chains with their density (real-space refinement).

AI technology in Textal

AI techniques are well suited to automate the complex decision-making processes involved in electron density map interpretation. Previous AI-related work in this area includes expert systems and molecular scene analysis.[4,6] Textal adopts a divide-and-conquer approach that decomposes the problem into pieces that can be solved independently, some of them by AI methods. The various uses of AI technology in Textal are described next.

Neural network to locate Cα atoms

Textal determines the Cα atoms' locations in the unknown model using a feed-forward neural network, with one layer of 20 hidden units and sigmoid thresholds.[9] Textal feeds 38 numerical features—rotationally invariant descriptions of the local density, such as statistical moments and the moments of inertia—to the network, which outputs the predicted distance between each candidate position and the real Cα. Capra then uses these predictions to pick those most likely closest to the real Cα positions, factoring in constraints such as the expected distances between Cα atoms. The network is trained on a set of candidate points in maps of solved proteins where the distances to the real Cα atoms are known. Textal optimizes the network weights using back-propagation.

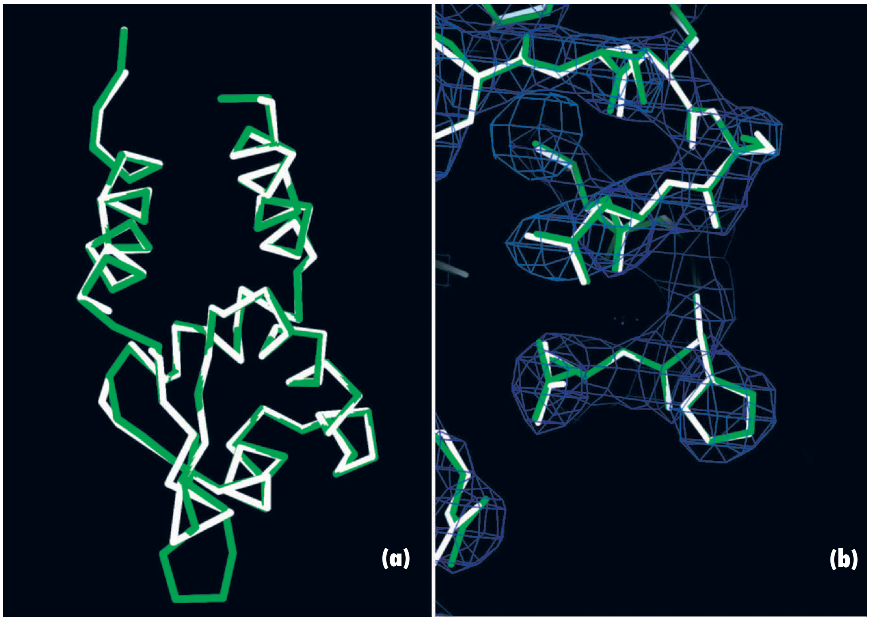

Figure 1. The results of running Textal on a test protein, CZRA, whose true structure was known. The Cα chains Textal found are white, and the true structure is green. (a) There's close agreement except for the loop at the bottom, which is poorly resolved in the electron density map. (b) A close-up of some side chains that Textal found. The agreement is close, not only in side chain type but also in positioning.

Heuristic search to build chains

After Textal determines the Cα positions, it connects them using a heuristic search algorithm that emulates the reasoning experienced crystallographers would use in building their models, such as knowledge of secondary-structural motifs and stereochemical constraints.[9]

Case-based reasoning to connect chains

Noisy data (and maps) make building chains challenging because noise can cause breaks to appear where a chain should be connected or make connections where there should be none. In practice, unconnected chains determined by Capra based on discontinuous density isn't an uncommon problem. It's often a problem near proteins' termini and surface, where the molecule tends to be more mobile and hence poorly resolved. Textal applies case-based reasoning to "stitch" together these split chains using a method similar to that proposed by T. Alwyn Jones and Søren Thirup.[5,9] Textal searches a database of solved density fragments (and their chains) using a window of consecutive Cα atoms by superimposing the known fragments over the potential unknown matches. If a match is sufficiently close and the density sufficiently strong, the missing Cα atoms can be inserted from the database template.

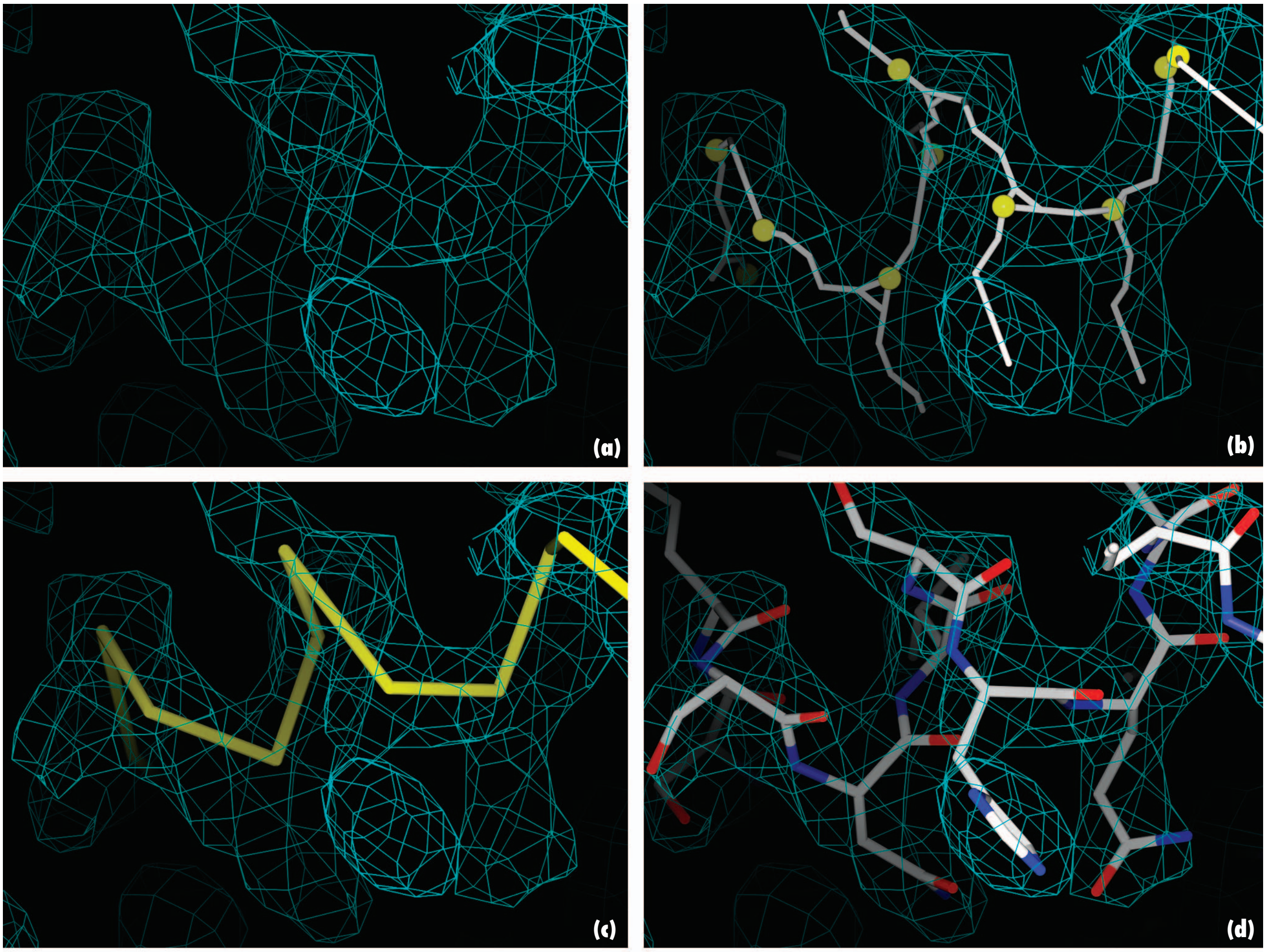

Figure 2. Solving an unknown structure: (a) the raw electron density for a protein without any model built, (b) the medial axis of the density that Capra found (in cyan) along with the predicted Cα atoms (in white), (c) the connected chain (an alpha-helix) that Capra built, and (d) Lookup's results, fitting entire amino acids into the density forming the whole model.

Case-based reasoning to model side chains

Textal models side chains using Lookup, which takes the set of Cα chains (output by Capra) as input and uses case-based reasoning and nearest-neighbor learning to effectively and efficiently retrieve, from a database, spherical regions (of 5 Å radius) that are structurally similar to regions from the unsolved map.[9] Textal compares the 5 Å spherical regions centered on Cα atoms in the backbone model Capra produced to a large database of 50,000 regions from 200 maps of proteins (for which the local structures of the regions are known and cover a wide range of structural motifs in proteins). Textal retrieves and assembles the matching local structures to gradually produce a preliminary model, which we can further refine by postprocessing routines. We find matching regions via a similarity metric that uses 76 numeric features to locally characterize the spherical regions.

Feature weighting for pattern recognition

A key requirement for accurately determining similarity in Lookup is correctly choosing features to be used in the similarity metric. In Textal, experts determined the 76 numeric features on the basis of domain knowledge as well as intuitions on what would be relevant in discriminating electron density patterns. But not all features might be relevant in all situations. Irrelevant features effectively introduce noise in the data and can mislead pattern matching. So, we use Slider, a feature-weighting algorithm that evaluates how well a given feature-based distance measure (that uses the weights) ranks the matches of an instance relative to its mismatches.[10] We do this for a set of instances, and the average ranking of the matches reflects how good the weight vector is.

But the space of continuous weight vectors is exponentially large, and an exhaustive search is clearly intractable. So, we use a heuristic to greedily search only those weights that affect the ranking of matches (if we slide the weight of a feature from 0 to 1, the ranking will change at a weight where an instance is equidistant to a match and a mismatch in Euclidean space). This strategy makes the search efficient and effective.

Linear discriminant analysis to detect disulfide bridges

A disulfide bridge is a covalent bond between the sulfur atoms of two cysteine amino acids from the same or neighboring polypeptide chains. Disulfide bridges occur

X-ray Crystallography

Although a protein's sequence largely determines its 3D structure, accurately predicting the structure from the sequence alone is nontrivial. The Levinthal paradox says that a random-walk folding algorithm would successfully fold a protein only on a cosmological time scale.[1] Considerable progress has been made in ab initio protein structure prediction using methods ranging from molecular dynamics and hidden Markov models to fragment-based approaches. However, none have been able to robustly predict the structure, particularly on proteins with novel sequences and folds.[2–4] Fortunately, we can use experimental methods to determine a protein's structure.

X-ray crystallography is the most widely used technique to accurately determine protein structure.[5] After a protein is purified and crystallized, an x-ray beam is directed through the crystal at various angles. The interaction of the x-ray with the electrons in the protein's atoms, locked in the crystal's rigid lattice, produces a set of diffraction patterns. We can acquire these patterns and process them into a set of amplitudes for Fourier coefficients (known as *structure factors*), representing the Fourier transform of the protein's electron density convolved with the crystal's repeating lattice. By taking the inverse Fourier transform, we can reconstruct the electron density, at least in principle. The diffraction phases are necessary to complete the transform, but these aren't directly available. We can, however, estimate or approximate these using various techniques, such as molecular replacement, multiple isomorphous replacement, multiwavelength anomalous diffraction, and sulfur anomalous diffraction. The data is also collected only to a certain resolution, described in angstroms (Å). The resolution determines how fine a structure can be resolved, owing to experimental constraints.

Once we've generated a map by combining the observed diffraction intensities with the estimated phases, we must build a model for it. Traditionally, this involves a crystallographer sitting for days or weeks at a graphics workstation, visually interpreting the density, placing atoms for the backbone, and fitting side chains into the remaining density. Noise in the experimental data can cause perturbations in the density, such as breaks in the backbone continuity. Additionally, some parts of the protein might be mobile, even in a crystal (such as at the protein's surface), causing the corresponding density to appear diffuse, if not nonexistent. Finally, if the data's resolution is low (that is, greater than 2.5 Å), the density might appear blurred and difficult to resolve, requiring the crystallographer to rely on a great deal of background knowledge about typical protein structures to make decisions on the best way to model the density, such as typical backbone angles, secondary-structure composition, common side-chain configurations, and long-range interresidue contacts.

References

1. R.A. Zwanzig, A. Szabo, and B. Bagchi, "Levinthal's Paradox," *Proc. Nat'l Academy of Sciences of the United States of America*, vol. 89, no. 1, 1992, pp. 20–22.
2. J.R. Bienkowska et al., "Protein Fold Recognition by Total Alignment Probability," *Proteins*, vol. 40, no. 3, 2000, pp. 451–462.
3. M.R. Lee et al., "Molecular Dynamics in the Endgame of Protein Structure Prediction," *J. Molecular Biology*, vol. 313, no. 2, 2001, pp. 417–430.
4. J.A. McCammon and S.C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge Univ. Press, 1987.
5. J. Drenth, *Principles of Protein X-Ray Crystallography*, Springer, 1999.

about once in every four proteins. In Textal, a linear discriminant model detects disulfide bridges, using training that optimizes the model's parameters.[11] Positive and negative training examples of disulfide bridges are represented by the same 76 features used in Lookup. This classification method projects the high-dimensional data onto an optimal line in space along which classification is performed using a single threshold to distinguish between disulfide bridge and nonbridge classes.

Deployment

Initiated in 1998, the Textal project stems from collaboration between researchers from the Departments of Computer Science and Biochemistry and Biophysics at Texas A&M University. Textal now consists of 100,000 lines of C, C++, Perl (Practical Extraction and Report Language), and Python code and runs on various versions of Irix, Linux, Solaris, Mac OS, and Windows. We use the Concurrent Versions Systems (www.nongnu.org/cvs) to coordinate software development and maintenance.

We've deployed Textal in various ways:

- WebTex is a Web-based interface (http://textal.tamu.edu:12321) where registered users can upload their maps. Our servers process them, and the models Textal outputs are automatically sent back in an email. We launched WebTex in June 2002; about 120 users in 70 research institutions from 20 countries currently use it, both from industry and academia.
- Linux and OSX distributions, available since September 2004, are available at http:// textal.tamu.edu:12321.
- Textal is also the structure determination component of the Python-based Hierarchical Environment for Integrated Crystallography (www.phenix-online.org), an integrated crystallographic computing environment. Phenix's other main components are the Computational Crystallography Toolbox, Phaser, Solve, and Resolve. Phenix was released in July 2003 as an alpha test version; Phenix's public release was in April 2005.[8,12,13]

Textal, by necessity, has been an interdisciplinary project requiring immersion in structural biology to understand the scientific problems and to find a common tongue for discussing the issues and their solutions. This has required cross-training students and interaction with the crystallographic community. Participation in crystallographic and structural biology conferences and workshops has been essential to developing intuition about these problems to intelligently design and apply AI techniques. It's also necessary to understand crystallographic data's subtleties, such as data scaling, thermal vibration's effect (B-factors), and other noise sources, which adversely impact pattern recognition. A deeper understanding of the more mundane but practical tasks that crystallographers face, such as handling different data formats and understanding data representation conventions, is also important.

Despite the difficulties in integrating multiple disciplines and expertise, protein crystallography has proved to be rich and rewarding. It's a domain with many problems to which we can apply AI.

Future plans for TEXTAL include expanding the range of resolutions for which it's optimized by constructing different databases and neural networks and selecting the appropriate database depending on the data. We plan to apply the TEXTAL approach to identifying DNA and RNA. We're also working on incorporating automated detection of noncrystallographic symmetry within TEXTAL.

Tod Romo is a research scientist at Texas A&M University's Center for Structural Biology, Institute for Biosciences and Technology. Contact him at tromo@tamu.edu.

Kreshna Gopal is a doctoral student at Texas A&M University's Department of Computer Science. Contact him at kgopal@cs.tamu.edu.

Erik McKee is a doctoral student at Texas A&M University's Department of Computer Science. Contact him at emckee@cs.tamu.edu.

Lalji Kanbi is a postdoctoral researcher at the Department of Biochemistry and Biophysics at Texas A&M University. Contact him at lalji@pompano.tamu.edu.

Reetal Pai is a doctoral student at Texas A&M University's Department of Computer Science. Contact her at reetalp@cs.tamu.edu.

Jacob Smith is a doctoral student at Texas A&M University's Department of Computer Science. Contact him at thechao@neo.tamu.edu.

James Sacchettini is a professor of biochemistry, biophysics, and chemistry and the director of the Center for Structural Biology at Texas A&M University, where he also holds the Wolfe-Welch Chair in Science. Contact him at sacchett@tamu.edu.

Thomas Ioerger is an associate professor at Texas A&M University's Department of Computer Science. Contact him at ioerger@cs. tamu.edu.

References

1. J. Drenth, *Principles of Protein X-Ray Crystallography*, Springer, 1999.
2. D. Hennessy et al., "Statistical Methods for the Objective Design of Screening Procedures for Macromolecular Crystallization," *Acta Crystallographica Section D—Biological Crystallography*, vol. 56, no. 7, 2000, pp. 817–827.
3. S.K. Burley et al., "Structural Genomics: Beyond the Human Genome Project," *Nature Genetics*, vol. 23, no. 2, 1999, pp. 151–157.
4. E.A. Feigenbaum, R.S. Engelmore, and C.K. Johnson, "A Correlation between Crystallographic Computing and Artificial Intelligence Research," *Acta Crystallographica Section A—Foundations of Crystallography*, vol. 33, no. 1, 1977, pp. 13–18.
5. T.A. Jones and S. Thirup, "Using Known Substructures in Protein Model Building and Crystallography," *EMBO (European Molecular Biology Organization) J.*, vol. 5, 1986, pp. 819–822.
6. L. Leherte et al., "Analysis of Three-Dimensional Protein Images," *J. Artificial Intelligence Research*, vol. 7, 1997, pp. 125–159.
7. R.J. Morris, "Statistical Pattern Recognition for Macromolecular Crystallographers," *Acta Crystallographica Section D—Biological Crystallography*, vol. 60, no. 12, 2004, pp. 2133–2143.
8. T.C. Terwilliger, "Maximum-Likelihood Density Modification," *Acta Crystallographica Section D—Biological Crystallography*, vol. 56, no. 8, 2000, pp. 965–972.
9. T.R. Ioerger and J.C. Sacchettini, "The TEXTAL System: Artificial Intelligence Techniques for Automated Protein Model Building," *Methods in Enzymology*, vol. 384, 2003, pp. 244–270.
10. K. Gopal et al., "Determining Relevant Features to Recognize Electron Density Patterns in X-ray Protein Crystallography," *J. Bioinformatics and Computational Biology*, vol. 3, no. 3, 2005, pp. 645–676.
11. T.R. Ioerger, "Automated Detection of Disulfide Bridges in Electron Density Maps Using Linear Discriminant Analysis," *J. Applied Crystallography*, vol. 38, no. 1, 2005, pp. 121–125.
12. P.D. Adams et al., "Recent Developments in PHENIX Software for Automated Crystallographic Structure Determination," *J. Synchrotron Radiation*, vol. 11, no. 1, 2004, pp. 53–55.
13. R. Read, "Pushing the Boundaries of Molecular Replacement with Maximum Likelihood," *Acta Crystallographica Section D—Biological Crystallography*, vol. 57, no. 10, 2001, pp. 1373–1382.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.