electronic reprint

Acta Crystallographica Section D

Biological Crystallography

ISSN 0907-4449

Editors: **E. N. Baker and Z. Dauter**

Improving amino-acid identification, fit and C^{α} prediction using the Simplex method in automated model building

Tod D. Romo, James C. Sacchettini and Thomas R. Ioerger

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Improving amino-acid identification, fit and C^{α} prediction using the Simplex method in automated model building

Tod D. Romo,[a] James C. Sacchettini[a,b] and Thomas R. Ioerger[c]*

[a]Texas A&M Center for Structural Biology, Institute for Biosciences and Technology, Houston, TX 77030, USA, [b]Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA, and [c]Department of Computer Science, Texas A&M University, College Station, TX 77843, USA

Correspondence e-mail: ioerger@cs.tamu.edu

Received 15 June 2006
Accepted 24 August 2006

Automated methods for protein model building in X-ray crystallography typically use a two-phased approach that involves first modeling the protein backbone followed by building in the side chains. The latter phase requires the identification of the amino-acid side-chain type as well as fitting of the side-chain model into the observed electron density. While mistakes in identification of individual side chains are common for a number of reasons, sequence alignment can sometimes be used to correct errors by mapping fragments into the true (expected) amino-acid sequence and exploiting contiguity constraints among neighbors. However, side chains cannot always be confidently aligned; this depends on having sufficient accuracy in the initial calls. The recognition of amino-acid side-chains based on the surrounding pattern of electron density, whether by features, density correlation or free atoms, can be sensitive to inaccuracies in the coordinates of the predicted backbone C^{α} atoms to which they are anchored. By incorporating a Nelder–Mead Simplex search into the side-chain identification and model-building routines of *TEXTAL*, it is demonstrated that this form of residue-by-residue rigid-body real-space refinement (in which the C^{α} itself is allowed to shift) can improve the initial accuracy of side-chain selection by over 25% on average (from 25% average identity to 32% on a test set of five representative proteins, without corrections by sequence alignment). This improvement in amino-acid selection accuracy in *TEXTAL* is often sufficient to bring the pairwise amino-acid identity of chains in the model out of the so-called 'twilight zone' for sequence-alignment methods. When coupled with sequence alignment, use of the Simplex search yielded improvements in side-chain accuracy on average by over 13 percentage points (from 64 to 77%) and up to 38 percentage points (from 40 to 78%) in one case compared with using sequence alignment alone.

1. Introduction

One of the significant challenges in automated construction of protein models from electron-density maps is accurate identification of amino-acid side chains. There are a number of reasons why individual amino-acid side chains may be difficult to recognize in electron-density maps, ranging from noise caused by phase error to diffusiveness arising from high *B* factors to structural similarities among the amino acids that cause ambiguity. In some cases, sequence alignment to the true (or expected) amino-acid sequence can be used to determine the identity of a given fragment (or chain) and thus correct the mistakes among its residues (Terwilliger, 2003; Cohen *et al.*, 2004). However, this cannot always be performed reliably and

is limited by the raw accuracy of the initial side-chain calls. Typically, at least 25–30% of amino acids in a chain need to be correct in order to accurately determine its location in the true sequence. The potential for recognition errors is exacerbated by inaccuracies in the estimated coordinates of putative C^{α} atoms. In this paper, we discuss an application of the Simplex search algorithm to enhance the accuracy of amino-acid side-chain identification (prior to sequence alignment) by a local rigid-body real-space refinement of candidate side chains (including translation of the C^{α}) in the process of selecting the best match and this can ultimately improve the amino-acid identity of models built using sequence alignment.

Most automated model-building methods are comprised of two principal stages. The first stage predicts C^{α} coordinates and constructs a preliminary backbone, typically using a skeletonization or tracing algorithm (Jones *et al.*, 1991; Ioerger & Sacchettini, 2002; Oldfield, 2003). The second stage determines the amino-acid type for each predicted C^{α} and builds it into the nearby density. For example, *MAID* uses a template-matching approach, picking the best rotamer from a library (Levitt, 2001). Once the best rotamer is identified, it is then optimized by torsion-angle Powell minimization where the main chain and hence the C^{α} coordinates are fixed. *ARP*/*wARP* 'docks' an amino-acid sequence onto its initial backbone by examining the connectivity vectors of free atoms in the vicinity of the estimated C^{α} atom (Cohen *et al.*, 2004). After the amino-acid type assignments are made, the side chains are modeled using a rotamer library followed by torsion-angle real-space refinement using the Simplex algorithm (although the only backbone parameter manipulated is φ). *RESOLVE* takes a different approach involving convolution of average side-chain densities for the 20 amino-acid types and uses a Bayesian approach to dock the amino-acid sequence onto the pre-built backbone (Terwilliger, 2003). Once the amino-acid types have been identified, the best rotamer is built into the model based on the previously built and fixed C^{α} coordinates. Finally, *TEXTAL* uses a library of solved prototypic density regions extracted from the PDB (Ioerger & Sacchettini, 2003). This method, which is the middle stage of *TEXTAL*, is referred to as 'LOOKUP'. Firstly, the library is filtered based on rotation-invariant 'features' calculated from the density in the neighborhood of the predicted C^{α} atoms. Each of the remaining high-probability matches is then examined in more detail by superimposing the density region from the library onto the observed density around the predicted C^{α} and evaluating the local density correlation. The side-chain model for the best fitting region is then extracted from the library and used to build the final model.

Each of these methods for identifying and modeling side chains is approximate and depends to a varying extent upon the initial determination of the backbone C^{α} positions. If the predicted C^{α} is offset sufficiently from its true position, then the surrounding density could look significantly different, affecting the identification of its associated residue. For example, if the predicted C^{α} is shifted into the side-chain density, then it may appear to be a shorter side chain than it actually is. If the C^{α} is shifted laterally along the backbone (away from a branch point), then no side chain might fit the density well. This sensitivity of amino-acid type identification to the initially constructed backbone can be addressed by using the Nelder–Mead Simplex algorithm to perform a local optimization of the library region to the density as part of the side-chain identification process.

The Nelder–Mead Simplex algorithm is a classic search algorithm for the optimization of multidimensional functions (Nelder & Mead, 1964). Features of the Simplex algorithm include a large radius of convergence, an ability to adapt to and avoid local maxima/minima and the lack of a need to compute derivatives. Applications of Simplex search include determining optimal weighting of energy terms in sequence threading (Russell & Torda, 2002; Torda *et al.*, 2004), the fitting of simple approximations to complex potential energy surfaces (Marun *et al.*, 2004), the superposition of small ligands for 3D-QSAR studies (Melani *et al.*, 2003), as well as in docking studies (Exner *et al.*, 2002; Hu & Shelver, 2003) and in semi-interactive rigid-body refinement in *Coot* (Emsley & Cowtan, 2004). It has also been used previously in automated model building for optimizing the fit of modeled side chains to their corresponding density, such as in the torsion-angle real-space side-chain refinement in *ARP*/*wARP* (Cohen *et al.*, 2004). However, the use of the Simplex method (or any other form of real-space refinement applied to individual residues) typically occurs after the amino-acid type of the side chain has been inferred.

Here, we introduce a novel application of the Simplex algorithm to improve the correct identification of side-chain types by using it to optimize the fit in matching electron-density patterns with the density to be modeled. By allowing the candidate density regions to rotate and translate, we are able to find better matches from a library of solved regions that are less dependent upon the accuracy of their initial superposition, which is determined by the initial backbone construction. In the *TEXTAL* automated model-building system, this optimization not only improves the raw accuracy of amino-acid identification, but also leads to increased amino-acid identity of resultant models when sequence alignment is applied.

2. Methods

Amino-acid identification and modeling in *TEXTAL* begins with retrieving the spherical region (5 Å radius) of electron density from the *TEXTAL* database (∼50 000 regions extracted from maps of previously solved models) that best matches the density surrounding each predicted C^{α} in a given map. The initial retrieval of matching regions is based on comparison of rotation-invariant features that characterize local electron-density patterns (Ioerger & Sacchettini, 2003). A limited number of candidate matches (K = 400 is typically used as a filter) are selected and re-ranked based on local density correlation in order to identify the best match. The density correlation is calculated over a cylindrical region 5 Å long by 2.5 Å in radius that covers the known side chain (from

the database). The two regions are aligned by rotating them so that 'spokes' of density emanating from the centers of each region (representing the direction of the side chain and the C- and N-terminal directions of the backbone) are optimally superposed (Ioerger & Sacchettini, 2003). The local coordinates for the side-chain atoms from the best-matching region in the *TEXTAL* database are retrieved and rotated into position using the same rotation matrix as the superposition for the correlation calculation.

2.1. Simplex optimization in *TEXTAL*

To enhance the selection of this initial local model, we incorporated a real-space optimization strategy that adjusts the superposition (including both rotation and translation) between two regions to improve the determination of the quality of fit (correlation). We chose to use the Nelder–Mead Simplex algorithm (Nelder & Mead, 1964) as the optimizer, which does not require Fourier synthesis (*i.e.* calculation of Fourier coefficients, as is typically performed in reciprocal-space refinement procedures; Murshudov *et al.*, 1997).

The Simplex optimizer was incorporated into *TEXTAL*'s LOOKUP routine as shown in Fig. 1. The target function is the *TEXTAL* density-correlation function and the dependent variables are the six degrees of freedom of the known region:

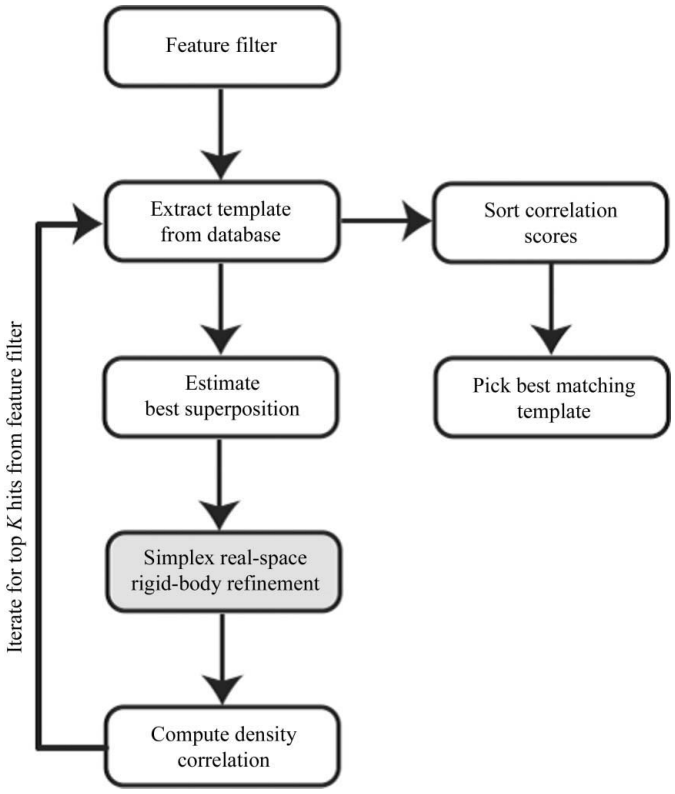

Figure 1
The workflow for LOOKUP is shown, along with where the Simplex real-space refinement stage was introduced. The feature filter first selects K regions from the *TEXTAL* database, which are then extracted, superimposed and the density correlation computed. Finally, the top-scoring regions are considered and the best match to the unknown density region is picked.

three Euler angles and three Cartesian coordinates for the translation. The initial simplex was constructed as described in Mistree & Shoup (1987), with 7 (*i.e.* $N + 1$) vertices and characteristic lengths of 10° for the Euler angles and 0.5 Å for the translation. The Simplex algorithm cannot itself pick a side chain; it merely optimizes the fit between the probe region (and the side-chain contained therein) against the library region that LOOKUP was already considering. This is then used to score the database regions using density correlation and make better selections than when only a coarse orientation optimization is used. It is important to note that the optimization of each unknown region is completely independent of the optimization of every other unknown region, since there are neither stereochemical nor through-space restraints placed on the minimizer.

If the Simplex method is unable to locate a region with better density correlation than that found by the non-Simplex method, then the Simplex result is rejected in favor of the original LOOKUP result. Since the location of the simplex set is unconstrained in the six-dimensional parameter space, it is possible that the probe region could drift sufficiently far from the original anchor point that it is actually fitting another residue nearby. It is also possible for the Simplex routine to push the region far enough away that the best fit is now to the main-chain density and not the unknown side chain. To prevent these errors from occurring, the Simplex result is also rejected if the C^{α} shift is greater than 2.0 Å.

It is important to correctly determine when an optimum has been obtained with the Simplex algorithm or whether no optimum can be found, which taken together constitute the stopping criterion. The former is addressed by checking for convergence of the algorithm, *i.e.* when improvements in the density correlation become smaller than a pre-determined tolerance. A default tolerance of 0.001 was selected for our experiments. To decide when no optimum can be found, a cutoff of 5000 density-correlation evaluations was chosen. Such a large number will permit outlier cases to still be optimized while not unduly extending the execution time of the program. In practice, the average number of evaluations observed during a typical LOOKUP run is around 200 per C^{α}.

2.2. Evaluating model-building performance

Two primary metrics were used to gauge *TEXTAL*'s performance with Simplex optimization: *CAPRA*'s performance was evaluated by the r.m.s.d. (root-mean-square deviation) of C^{α} placement compared with the refined structure, while LOOKUP was evaluated by the percentage of correct amino acids assigned based on both identity and structural similarity. In both cases, the *TEXTAL* model was compared against a model solved and refined by a crystallographer, *i.e.* a 'hand-crafted' model. A structural alignment of the two models was made by finding the closest C^{α} in the *TEXTAL* model to each true C^{α}. This alignment was used to calculate the C^{α} r.m.s.d. as well as to determine what amino-acid assignment should have been made.

Table 1
Data sets used in this study.

The resolution indicated is for the experimental structure factors used, which may differ from the final resolution reported for the refined structures. Phase error (at native resolution) in comparison to phases calculated from the final refined structure is reported. Map correlation is between experimental electron-density maps and σ_A-weighted $2F_o - F_c$ maps calculated at 2.8 Å.

Protein	Resolution (Å)	Phase error (°)	Map correlation
CzrA	2.3	18.1	0.95
If5a	2.1	36.8	0.91
MVK	2.4	42.8	0.84
ICL	3.0	44.1	0.81
PcaA	2.8	54.2	0.73

Since some residues simply look too similar to be distinguished from each other based on their electron density (*e.g.* Val and Thr), a looser notion of identity was also used based on the structural similarity of side chains. Each amino acid is assigned a 'group code' based on ten sets of residues with similar shapes (four groups are unique and only contain one member): {A, G}, {D, L, N}, {Q, E, M}, {F, Y, H}, {K, R}, {S, T, V}, {C}, {I}, {P}, {W}. These are used to compute a 'structural similarity' score in the tables, which is sometimes more reflective of the side-chain modeling accuracy than strict amino-acid identity.

3. Results and discussion

A test suite of experimental electron-density maps for five different proteins in the Protein Data Bank was used to compare the performance of *TEXTAL* with and without Simplex. The test proteins, each originally solved by multi-wavelength anomalous diffraction (MAD), were CzrA (zinc response protein; Eicken *et al.*, 2003), ICL (isocitrate lyase; Sharma *et al.*, 2000), MVK (mevalonate kinase; Yang *et al.*, 2002), If5a (translation initiation factor 5a; Peat *et al.*, 1998) and PcaA (mycolic acid cyclopropane synthase; Huang *et al.*, 2002). The five electron-density maps, generated from MAD phases after solvent flattening but prior to any model-based refinement, spanned a range of resolutions (2.1–3.0 Å; see Table 1) and quality, from relatively high quality (clear backbone and side-chain density) to low quality (*i.e.* high phase error). In each case, the experimental map was re-calculated at 2.8 Å prior to submission to *TEXTAL*, since *TEXTAL* was trained for pattern recognition in 2.8 Å maps. The five proteins and their *TEXTAL* model-building results are summarized in Table 1.

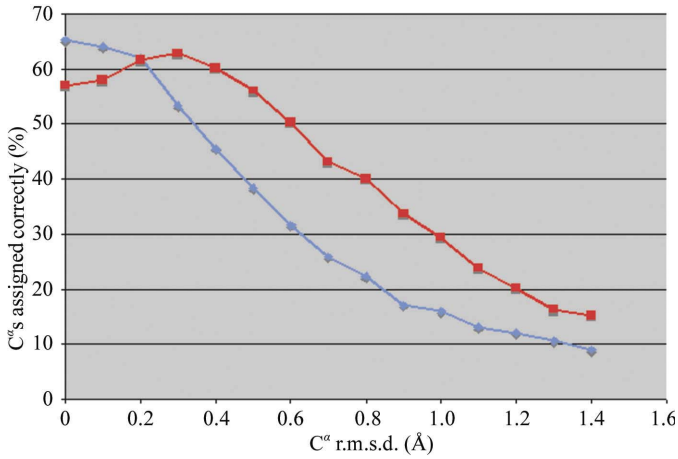

Figure 2
The effect of perturbations in the C^α prediction on the LOOKUP results for CzrA is shown here by displacing the ideal C^α coordinates from the hand-refined structure by vectors of increasing magnitude and random directions. The results from LOOKUP where only the spoke correlation method was used is shown in blue. The Simplex LOOKUP result is shown in red.

Table 2
Comparison of *TEXTAL*'s model-building accuracy over the test suite of five proteins.

The feature-filter cutoff was $K = 400$. Identity is the strict amino-acid identity both before and after sequence-alignment correction. The 'similarity' of side chains is based on the following structural equivalences: {A, G}, {D, L, N}, {Q, E, M}, {F, Y, H}, {K, R}, {S, T, V}, {C}, {I}, {P}, {W}.

	Average	CzrA	ICL	If5a	MVK	PcaA
K = 400, non-Simplex						
Mean residue CC	0.785	0.820	0.771	0.816	0.777	0.740
C^α r.m.s.d.	0.876	0.753	1.030	0.802	0.877	0.916
Identity (without alignment)	25.5	40.0	23.5	30.2	18.1	15.6
Similarity (without alignment)	45.4	65.6	39.9	51.2	39.4	31.1
Identity (with alignment)	64.1	94.4	55.3	92.2	40.1	38.7
Similarity (with alignment)	69.0	96.6	59.0	93.0	49.5	46.7
Run time (s per C^α)	0.70	0.77	0.68	0.71	0.71	0.72
Perfect C^αs, non-Simplex						
Mean residue CC	0.887	0.936	0.851	0.919	0.897	0.833
Identity (without alignment)	41.7	65.3	28.5	51.5	38.2	25.2
Similarity (without alignment)	60.4	81.1	48.3	73.5	53.9	45.0
K = 400, Simplex						
Mean residue CC	0.918	0.937	0.896	0.941	0.924	0.892
C^α r.m.s.d.	0.791	0.577	0.972	0.601	0.741	1.063
Identity (without alignment)	32.5	47.8	26.0	38.8	30.8	19.3
Similarity (without alignment)	52.8	73.3	43.4	63.6	49.6	34.0
Identity (with alignment)	77.5	93.3	76.4	93.0	77.6	47.4
Similarity (with alignment)	80.5	94.4	79.9	95.3	80.9	52.1
Run-time (s per C^α)	1.90	2.00	1.86	1.97	1.90	1.94

3.1. Effects of C^α accuracy on amino-acid identification

Before investigating the improvements that can be achieved with the Simplex algorithm, we start by characterizing the baseline accuracy of *TEXTAL* using the default density-correlation method in LOOKUP to recognize and identify side chains. While the C^α backbones built by *CAPRA* are very good, with C^α placements often within 1 Å r.m.s.d. of correct positions, the initial amino-acid identity from LOOKUP is often low, in the neighborhood of 30% for high-quality maps (*e.g.* with low phase error), or lower for worse ones. This is not entirely unexpected since some residues are similar structurally, such as glutamate and glutamine, and are impossible to distinguish at these resolutions. In addition to structural degeneracy, side-chain recognition can be complicated by noise arising from phase error, high *B* factors, improper masks for density modification *etc.*, which can further decrease the ability to discriminate side-chain identities.

Through a sequence-alignment routine in *TEXTAL*, it is often possible to correct the amino-acid prediction from LOOKUP and bring the overall amino-acid identity of the *TEXTAL* model to above 80%. Sequence alignment in *TEXTAL* is implemented using a traditional dynamic programming alignment algorithm (Needleman & Wunsch, 1970; Gotoh, 1982) for global gapped alignments (without end-gap penalties) to dock fragments from the map into the true sequence, even in the presence of substitutions (mis-identities) and insertion/deletions (small gaps arising from extraneous or missing C^{α} atoms in the predicted backbone chains); LOOKUP can then be re-invoked to replace incorrect side chains with corrected identities (Ioerger & Sacchettini, 2003).

We hypothesize that imprecision in the initial C^{α} placement might be contributing to the inaccuracy in LOOKUP's amino-acid assignments. This is in fact easy to demonstrate by using the C^{α} coordinates from the final refined structures as the predicted C^{α} atoms for LOOKUP. These coordinates can then be randomly perturbed and the effect on LOOKUP results examined, as shown in Fig. 2 for CzrA. Using the 'perfect' hand-crafted C^{α} coordinates, LOOKUP is able to correctly assign approximately 65% of the residues in CzrA without the help of sequence alignment. As the artificially introduced error in C^{α} placement increases, the accuracy of LOOKUP drops rapidly. In fact, this curve is a good predictor of expected LOOKUP performance for other maps. For example, Table 2 shows MVK has a C^{α} r.m.s.d. of 0.877 Å, for which Fig. 1 predicts an average expected identity of around 18–20%. The raw identity (without Simplex or sequence alignment) turns out to be 18.1% (see Table 2). This shows that the accuracy of predicted C^{α} coordinates has a direct influence on the accuracy of side-chain identification.

If the bulk of the error in LOOKUP can be attributed to imprecision in the C^{α} placement, then how well can LOOKUP perform on average if this is factored out? Table 2 also shows the results of running LOOKUP on the test suite where the *CAPRA* C^{α}s have been replaced with those from the true structure, *i.e.* the 'perfect C^{α}s'. On average, the initial identities LOOKUP assigned were 41.7% correct, compared with 25.5% when using the C^{α} predicted by *CAPRA*, and the mean side-chain correlation coefficient improved by 0.10 (from 0.79 to 0.89). These results can be taken as representing an upper bound on the best performance that can be expected from LOOKUP (without sequence alignment).

3.2. LOOKUP with Simplex

The addition of the Simplex optimizer to LOOKUP gives it a fully fledged search capability when comparing regions. LOOKUP is now free to rotate and translate the probe regions to find the best correlation, thereby mitigating the effect of poor initial C^{α} placement on LOOKUP's side-chain calls. This is demonstrated quite clearly in Fig. 2, which shows the effect Simplex search has on LOOKUP's dependence on C^{α} placement accuracy. The Simplex-augmented version of LOOKUP is 15–20% more accurate over most of the range (of artificially introduced C^{α} errors) for CzrA. It is not surprising that the Simplex curve dips slightly from 0–0.3 Å r.m.s.d. (see Fig. 2) since the initial Simplex used to seed the search is constructed from small perturbations to the orientation and position of the probe region.

The performance of LOOKUP using the Simplex search is summarized in the bottom half of Table 2. This use of the Simplex search increases the initial accuracy of amino-acid identities seven percentage points, from 25.5 to 32.5% (without sequence alignment). More importantly, it boosts the average accuracy with sequence alignment to 77.5%, which is 13% better than can be achieved with sequence alignment alone (64.1%). In individual cases, the improvement can be even more dramatic. For example, use of Simplex increases the percentage identity for mevalonate kinase (MVK) almost 38%, from 40.1 to 77.6% (both estimates with sequence alignment applied). This can be interpreted intuitively as Simplex pulling the raw sequence identity out of the 'twilight zone', at which point sequence alignment becomes more

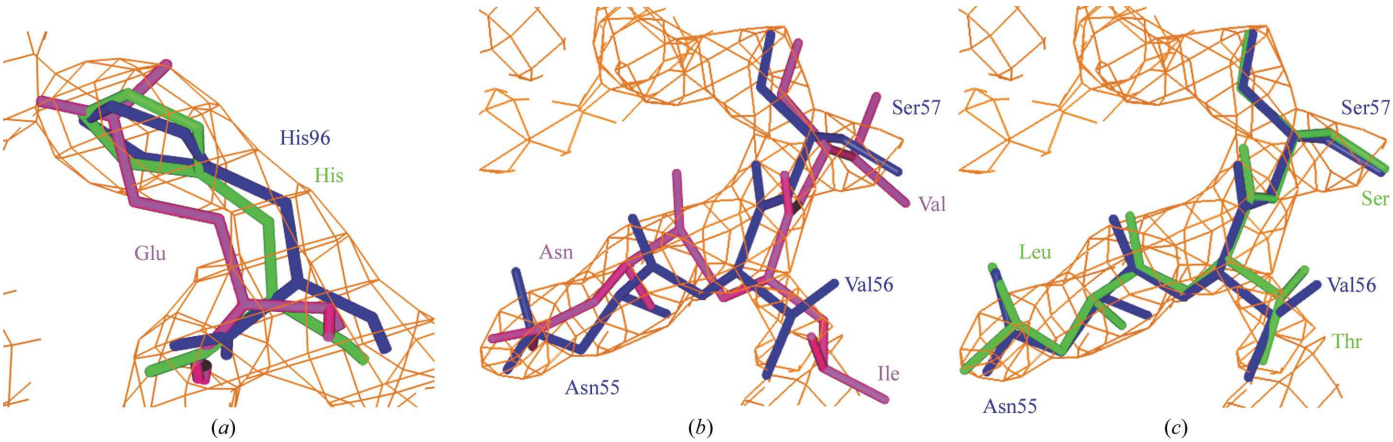

Figure 3
This figure shows representative regions comparing the original *TEXTAL* model (in magenta) against the Simplexed LOOKUP version (in green) and the 'true' or hand-crafted model (in blue). (*a*) shows His96 in CzrA, which was originally modeled incorrectly as a Glu (without Simplex), but was correctly recognized as a His when Simplex shifted the C^{α} closer to its true location. (*b*) and (*c*) show a small fragment built by *TEXTAL* (residues 55–57) without Simplex (*b*, magenta) and with Simplex (*c*, green) compared with the refined coordinates (blue).

effective. Although the Simplex optimization during LOOKUP moved the C^{α} atoms about 0.65 Å on average from where *CAPRA* originally predicted them, this made only a minor improvement in the C^{α} r.m.s.d. compared with the C^{α} atoms of the refined structure (by approximately 0.1 Å, from 0.876 Å in the non-Simplex case to 0.791 Å with Simplex). Note that the improvements made by the Simplex search do not come at a very dramatic computational cost; the new method takes ∼2.0 s on average to model each residue, compared with around 0.7 s per residue for the non-Simplex mode.

The improved C^{α} placement, as well as fit to density, can be observed graphically by examining the output of LOOKUP using the Simplex algorithm (but without identity corrections arising from sequence alignment). Fig. 3 shows two representative regions from CzrA, comparing the output of Simplex LOOKUP, shown in green, with the original *TEXTAL* model (without Simplex), shown in magenta, and the true (hand-crafted) model, shown in blue. Fig. 3(*a*) shows a close-up view illustrating the effect of improved C^{α} placement on His96. The initial C^{α} coordinate predicted by *CAPRA* was offset by 0.99 Å and this caused the residue to be incorrectly recognized by LOOKUP as a Glu (magenta). When Simplex optimization was turned on, the C^{α} shifted 0.5 Å closer to its true location (0.53 Å error) and this permitted the side-chain density to be correctly recognized and modeled as a His (green). Figs. 3(*b*) and 3(*c*) show a comparison of residues 55–57 with and without Simplex optimization. Without Simplex (Fig. 3*b*), there are significant errors in the C^{α} coordinates (0.60–1.12 Å), causing several side chains to be modeled incorrectly. For example, Val56 is modeled as a larger residue, Ile. With Simplex optimization (Fig. 3*c*), the C^{α} coordinates have become considerably more accurate (0.05–0.40 Å), along with the identities and fit of the side chains. On average, Simplex optimization improved the accuracy of the C^{α} coordinates from 0.80 to 0.27 Å over these three residues. The resulting residues are either identical to (Ser57) or isosteric (Leu for Asn55; Thr for Val56) with those in the true structure.

4. Conclusion

In this paper, we have described a novel application of the Nelder–Mead Simplex algorithm to improving the identification and real-space fitting of amino acids based on local patterns in electron density. In contrast to traditional rigid-body real-space refinement, which is typically applied to enhance the fit of side chains to density after their identity has been inferred, our approach uses Simplex optimization during the identification process itself to enhance the evaluation of quality of fit and hence the selection of the best local match. The primary advantage arises from allowing translation as well as rotation during the matching of regions (superposition to maximize local density correlation), which can compensate for the effect of errors in predicted C^{α} coordinates on the recognition of side-chain identities. This was shown to improve both the accuracy of side-chain recognition, as well as estimates of predicted C^{α} coordinates, in the *TEXTAL* automated protein model-building system. The increase in raw amino-acid identity was enough to boost the final sequence identity (after sequence alignment post-processing) from an average of 64% to nearly 78% across the test suite of experimental maps. Yet the Simplex search is not prohibitively inefficient, increasing the run-time by only a factor of approximately threefold over the original LOOKUP implementation.

This work was supported in part by grant P01-63210 from the National Institutes of Health, along with the Welch Foundation (JCS).

References

Cohen, S. X., Morris, R. J., Fernandez, F. J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V. S., Kleywegt, G. J. & Perrakis, A. (2004). *Acta Cryst.* D**60**, 2222–2229.
Eicken, C., Pennella, M. A., Chen, X., Koshlap, K. M., VanZile, M. L., Sacchettini, J. C. & Giedroc, D. P. (2003). *J. Mol. Biol.* **333**, 683–695.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Exner, T. E., Keil, M. & Brickmann, J. (2002). *J. Comput. Chem.* **23**, 1176–1187.
Gotoh, O. (1982). *J. Mol. Biol.* **162**, 705–708.
Hu, X. & Shelver, W. H. (2003). *J. Mol. Graph. Model.* **22**, 115–126.
Huang, C. C., Smith, C. V., Glickman, M. S., Jacobs, S. & Sacchettini, J. C. (2002). *J. Biol. Chem.* **277**, 11559–11569.
Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.
Ioerger, T. R. & Sacchettini, J. C. (2003). *Methods Enzymol.* **374**, 244–270.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.
Marun, R. A. B., Coronado, E. A. & Ferrero, J. C. (2004). *J. Comput. Chem.* **26**, 523–531.
Melani, F., Gratteri, P., Adamo, M. & Bonaccini, C. (2003). *J. Med. Chem.* **46**, 1359–1371.
Mistree, F. & Shoup, T. (1987). *Optimization Methods with Applications for Personal Computers.* Englewood Cliffs, NJ, USA: Prentice–Hall,.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.
Nelder, J. A. & Mead, R. (1964). *Comput. J.* **7**, 308–313.
Oldfield, T. (2003). *Methods Enzymol.* **374**, 271–300.
Peat, T. S., Newman, J., Waldo, G. S., Berendzen, J. & Terwilliger, T. C. (1998). *Structure*, **6**, 1207–1214.
Russell, A. J. & Torda, A. E. (2002). *Proteins*, **47**, 496–505.
Sharma, V., Sharma, S., Hoener zu Bentrup, K., McKinney, J. D., Russell, D. G., Jacobs, S. & Sacchettini, J. (2000). *Nature Struct. Biol.* **7**, 663–668.
Terwilliger, T. C. (2003). *Acta Cryst.* D**59**, 45–49.
Torda, A. E., Procter, J. B. & Huber, T. (2004). *Nucleic Acids Res.* **32**, W532–W535.
Yang, D., Shipman, L. W., Roessner, C. A., Scott, A. I. & Sacchettini, J. C. (2002). *J. Biol. Chem.* **277**, 9462–9467.