Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automated crystallographic ligand building using the medial axis transform of an electron-density isosurface

Jun Aishima,[a,b] Daniel S. Russel,[c] Leonidas J. Guibas,[c] Paul D. Adams[d] and Axel T. Brunger[a,b]*

[a]Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences and Stanford Synchrotron Radiation Laboratory, USA, [b]Howard Hughes Medical Institute, USA, [c]Stanford University Computer Science Department, USA, and [d]Lawrence Berkeley National Laboratory, USA

Correspondence e-mail: brunger@stanford.edu

Received 27 May 2005
Accepted 20 July 2005

Automatic fitting methods that build molecules into electron-density maps usually fail below 3.5 Å resolution. As a first step towards addressing this problem, an algorithm has been developed using an approximation of the medial axis to simplify an electron-density isosurface. This approximation captures the central axis of the isosurface with a graph which is then matched against a graph of the molecular model. One of the first applications of the medial axis to X-ray crystallography is presented here. When applied to ligand fitting, the method performs at least as well as methods based on selecting peaks in electron-density maps. Generalization of the method to recognition of common features across multiple contour levels could lead to powerful automatic fitting methods that perform well even at low resolution.

1. Introduction

During manual crystallographic model-building procedures, models are placed within an electron-density map by examining equal-density contours (isosurfaces) and placing the model in the middle of the contour (Swanson, 1994). Alternatively, automated model-building programs use either peaks in electron-density maps [*ARP/wARP* (Zwart *et al.*, 2004), *X-LIGAND* (Oldfield, 2001) and *MAID* (Levitt, 2001)] or the entire three-dimensional maps themselves [*RESOLVE* (Terwilliger, 2002), *ESSENS* (Kleywegt & Jones, 1997) and *FFFEAR* (Cowtan, 1998)] to help position models within the maps. At modest resolutions (worse than 2 Å resolution) and in lower quality electron-density maps, however, the points of maximum electron density may not necessarily correspond to the center of the isosurface, preventing accurate placement of the model. In order to overcome this problem, we have used a representation of the central axis of the isosurface, specifically an approximation of the medial axis. In our system, the approximate medial axis is computed for isosurface segments calculated from contoured electron-density maps. The medial axis is then thinned and atoms of the ligand model are matched to medial axis points. As a first application, we have used the medial axis transform method to address the problem of ligand fitting.

Structure determination of ligand–protein complexes is commonplace in the pharmaceutical drug-development process (Card *et al.*, 2005). Typically, the structures of many different ligand-bound complexes have to be solved in order to determine the details of the interaction between a compound and its target. Once the structure of the holoenzyme has been solved (through the use of molecular replacement, automated model-building software or manual methods), ligand fitting is often performed manually, requiring significant labor when multiple crystal structures need to be

doi:10.1107/S0907444905023152

solved. Only recently has automated ligand-building software become available with the release of *X-LIGAND* (Oldfield, 2001) and *ARP/wARP* (Zwart *et al.*, 2004).

Both *X-LIGAND* and *ARP/wARP* use electron-density peak heights to compare the ligand model and the experimental electron-density map. *X-LIGAND* generates trial conformations of the ligand to compare against a candidate region of the electron-density map by using moment-of-inertia calculations (Oldfield, 2001). *ARP/wARP* places trial atoms within the electron-density map at positions based on regions of high electron density, then attempts to interpret the trial atoms as ligand atoms in order to build a ligand structure (Zwart *et al.*, 2004). Other methods for model building, such as Greer's method (Greer, 1985), molecular-scene analysis (Leherte *et al.*, 1997) and critical-point analysis (Fortier *et al.*, 1997), also use peaks, ridges, saddle points and other regions of high values within the electron-density map to locate regions to place the model. In contrast, *ESSENS* (Kleywegt & Jones, 1997), *FFFEAR* (Cowtan, 1998) and *RESOLVE* (Terwilliger, 2002) examine a region of the map in real or reciprocal space and attempt to find places where a template matches the three-dimensional map.

The medial axis transform (Blum, 1967) has been applied to a wide variety of geometry-processing and object-recognition problems. The medial axis of a closed surface is the set of points inside the shape that have more than one closest neighbor on the surface of the shape. Equivalently, these points are the centers of maximal balls contained in the surface. A ball is considered maximal if it is not completely contained in any other ball interior to the surface. Recent applications include recognition of helical features in electron-density maps (Amenta, Choi, Jump *et al.*, 2002) as well use to plan flythroughs for virtual endoscopies (Paik *et al.*, 1998). In both applications, the key feature is that the medial axis provides a thin set of points which follows the center of the shape. The algorithms we use to compute the medial axis take a set of points sampled from a surface (the isosurface in our case) and approximate the medial axis of the sampled surface using a subset of the Voronoi vertices (points equidistant from four surface points) (Amenta, Choi, Dey *et al.*, 2002; Amenta *et al.*, 2001). These algorithms come with guarantees that the computed medial axis converges to the true medial axis as the density of points sampled from the surface increases and, if the density is high enough, the computed medial axis is guaranteed to have the same topology as the true medial axis.

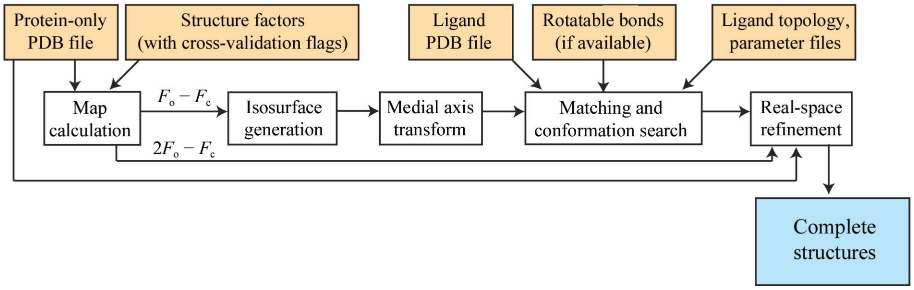

Figure 1
Outline of the medial axis procedure.

Here, we describe the application of the medial axis transform to automated ligand building. Graph matching is performed between the thinned medial axis and a graph generated from the ligand molecular model. Conformational searches and real-space refinement are then used to generate a ranked series of models that optimally fit the electron-density maps as assessed by the local real-space map correlation coefficient. We find that the medial axis method performs at least as well as methods based on electron-density peaks.

2. Methods

The application of the medial axis method to ligand fitting is described in Fig. 1. Electron-density maps are calculated and isosurfaces are extracted from an $F_o - F_c$ difference map. The program *Power Crust* (Amenta *et al.*, 2001) is used to compute a triangulated surface approximating the medial axis. This triangulated surface is then thinned to one-dimensional graphs with vertices spaced at approximately interatomic distances. Segments of similar size to the ligand are matched to the graph of the ligand molecular model with a branch-and-bound matching algorithm. The best matching ligands are then improved with torsion-angle conformational searches and real-space refinement. These steps are described in more detail below.

2.1. The medial axis transform applied to an isosurface of the electron-density map

We first compute an approximation of the medial axis of a set of points on a particular isosurface of the $F_o - F_c$ difference map. The isosurface is calculated from the electron-density map grid with the marching-cubes algorithm as implemented in *PyMOL* (DeLano, 2002). After the point set has been divided into segments separated by at least $3^{1/2}$ times the map grid spacing, the medial axis of each isosurface segment is calculated using *Power Crust* (Amenta *et al.*, 2001). The medial axes of the segments are then passed, separately, to the later stages of the computational protocol (Fig. 1).

2.2. Topology-preserving thinning of the medial axis

The medial axis produced by *Power Crust* has too many points and edges to be efficiently matched to a molecule. We simplify the medial axis using the *simplify* program which is provided with *Power Crust*. The program works by removing medial axis points whose maximal spheres only touch points on the surface which are closer than some threshold as well as medial axis points that are too close to another medial axis point. In *simplify*, we set the noise parameter to 0.5 and the redundancy parameter to 1.2 in order to obtain medial axis poles that are approximately 1.2 Å apart. This spacing is similar to the distance between non-H atoms in macromolecules, with the goal of obtaining a graph similar in size to the ligand-bond graph.

Further simplification of the medial axis is performed by decomposing the triangular faces that comprise medial axis segments into a set of vertices connected by edges. Because the simplified medial axis segments can still contain vertices that are too close to one another, segments are further thinned by iteratively combining the vertices of the shortest edges until there are no edges less than 1.0 Å in length.

2.3. Selecting candidate ligand sites

The asymmetric unit of a crystal generally contains many features that are not accounted for when modelled by macromolecules alone, including ligands, ions, ordered water molecules and unresolved regions of the macromolecules. These features are typically seen as strong regions in difference maps. To prevent unnecessary computation and incorrect fitting of ligands, all subsequent steps are performed on thinned medial axis segments whose number of vertices is between 0.5 to 2 times the number of ligand atoms. An absolute minimum size for the medial axis segments (four vertices) prevents very small segments from undergoing matching.

2.4. Overview of the graph-matching algorithm

Matching of the ligand to the medial axis is performed using two search stages in a manner that is reminiscent of the *ConfMatch* procedure (Wang, 2000). The first stage finds a correspondence between the ligand atom and the medial axis. This matching is constrained by the local connectivities of the medial axis and ligand atoms, as well as bond angles. The second search tries to find a conformation of the ligand with low r.m.s.d. (root-mean-square difference) to a given correspondence between the ligand atoms and the medial axis.

Both search stages employ the 'branch-and-bound' method, a technique for speeding up the search for the optimal solution in a discrete optimization problem without sacrificing optimality. In the following, we illustrate the method for vertex matching between two graphs, A and B, in order to minimize an error function. The error function is the sum of positive error terms for each corresponding pair of vertices. If we order the vertices in A, then a matching between A and B is an ordered list of nodes from B. Each such matching can be viewed as a path through a tree, where the leaves are complete matchings of A to B and shared prefixes in matches result in shared paths from the root of the tree. As the optimization algorithm searches this tree using a depth-first search, at each node it visits (with an associated partial matching M), it must explore all the matchings beginning with M plus the current vertex of A matched with the first available vertex in B, then all matchings beginning with M plus the current vertex of A matched with the second available vertex of B and so on. This iterative testing of all possible additions to the current matching is the 'branch' phase of branch and bound. If, however, the error associated with M is higher than that of some known solution, then none of the leaves of the subtree of the current node can be optimal (since the error on M is a lower bound on the error of any complete matching containing M) and the entire subtree can be skipped. This skipping is the 'bound' part of branch and bound.

2.4.1. Local geometry search. In the first stage we use a branch-and-bound search to find a good matching between points of the medial axis and ligand atoms (see Fig. 2 for an overview of the procedure). The set of candidate vertices for matching to each ligand atom is constrained by requiring that bonded atoms in the ligand be matched to vertices sharing an edge in the medial axis. Since not all ligand atoms necessarily correspond to vertices of the medial axis, we also allow ligand atoms to be matched against nothing; this is called a null match.

The error function comprises the following factors in order to obtain a vertex-to-vertex correspondence that maximizes the similarity between the ligand atoms and medial axis vertices.

(i) Null matches: extra vertices in either graph can be compensated with null matches. This factor is weighted strongly to maximize the number of vertices matched between the medial axis and ligand graphs.

(ii) Difference of vertex degree: compares the number of covalent bonds of the ligand atom to the number of edges of the medial axis vertex.

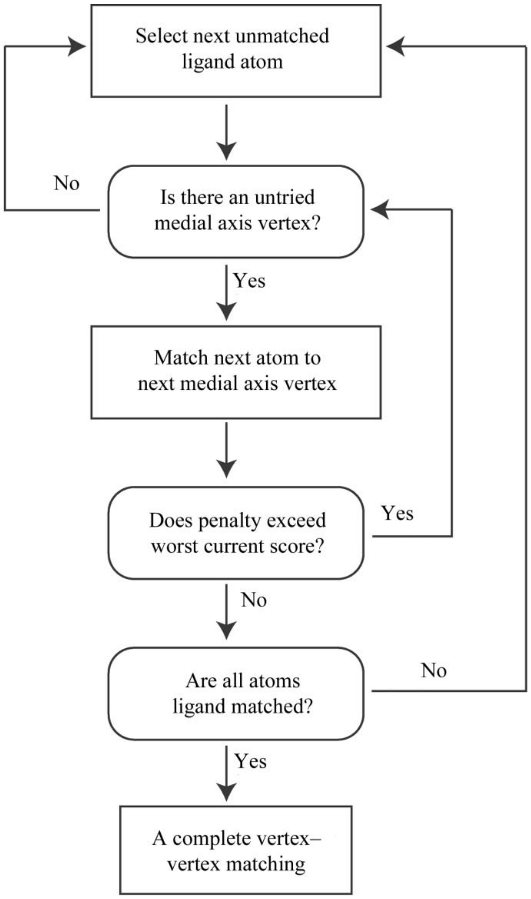
Select next unmatched ligand atom
No
Is there an untried medial axis vertex?
Yes
Match next atom to next medial axis vertex
Does penalty exceed worst current score?
Yes
No
Are all atoms ligand matched?
No
Yes
A complete vertex–vertex matching

Figure 2
Outline of the local geometry search.

(iii) Difference in bond angle between two adjacent edges: angles that are very different between a set of atoms and a corresponding region of the medial axis may indicate that the ligand atoms and medial axis are incorrectly matched.

The error score is comprised of a weighted sum of terms,

$$g = w_{\rm null}N + w_{\rm degree}\sum_{\rm all\ vertices}\Delta_{\rm degree} + w_{\rm angle}\sum_{\rm all\ angles}(\Delta_{\rm angle})^2, \qquad (1)$$

where N is the number of null vertices in the correspondence between ligand atoms and medial axis vertices, $\Delta_{\rm degree}$ is the absolute difference between the number of covalent bonds to the atom and number of vertices connected to the corresponding vertex and $\Delta_{\rm angle}$ is the bond-angle difference (in degrees) calculated for all bond angles defined by three connected vertices/atoms. We chose the following empirical weights: $w_{\rm null}$ = 29 000, $w_{\rm degree}$ = 2000 and $w_{\rm angle}$ = 10.

2.4.2. Conformation search. The matching described in the previous section does not directly give a physically valid conformation of the ligand, so one must search for a nearby valid conformation. This search is performed over the freely rotatable torsion angles in the ligand in order to optimize the fit to the atom-to-vertex correspondences obtained in the previous step.

The conformation-search procedure is outlined in Fig. 3. The ligand is first divided into rigid groups, which consist of the atoms between freely rotatable bonds. These rigid groups are added one at a time. Each added group is then rigidly aligned to the corresponding medial axis vertices by a torsion-angle search with a sample spacing of 10°. A composite error score is calculated and used to rank the conformation of the ligands,

$$c = gd, \qquad (2)$$

where g is the error score from the matching step (1) and d is the overall r.m.s.d. of the conformation. The r.m.s.d. is calculated with the Kearsley method (Kearsley, 1989) as implemented in the Bioinformatics Template Library (Pitt *et al.*, 2001). The branch-and-bound method is used to limit the number of conformations searched. After a ligand is completely built in a particular conformation, its conformation is ranked by the composite error score (2).

2.5. Real-space refinement

Refinement is used to optimize the fit of the best ligand conformations to the electron-density map. Using a $2F_{\rm o} - F_{\rm c}$ electron-density map, conjugate-gradient real-space refinement with *RSRef*2000 (Korostelev *et al.*, 2002) is performed for each ligand conformation. Only positional refinement of the ligand is performed by *RSRef*2000; neither *B*-factor refinement nor occupancy refinement are performed and the protein is held fixed. Because the refinement is performed against a small region of the electron-density map, real-space refinement is generally faster than reciprocal-space refinement using all structure factors. After real-space refinement, duplicate ligands that have identical positions are consolidated to yield a set of unique solutions.

2.6. Evaluating built ligands

To evaluate the fit between ligands and the electron-density map, we compute the local map correlation coefficient,

$$\mathrm{CC}_{\rm ligand} = \frac{\overline{\sum_{\rm ligand}(\rho_{\rm map1} - \overline{\rho_{\rm map1}})(\rho_{\rm map2} - \overline{\rho_{\rm map2}})}}{\left[\sum_{\rm ligand}(\rho_{{\rm map}_1} - \overline{\rho_{\rm map1}})^2\sum_{\rm ligand}(\rho_{\rm map2} - \overline{\rho_{\rm map2}})^2\right]^{1/2}},$$

where $\rho_{\rm map1}$ is the $2F_{\rm o} - F_{\rm c}$ electron-density map and $\rho_{\rm map2}$ is the calculated electron-density map of the ligand–protein complex. The correlation coefficient is calculated for the region that is defined by the union of all van der Waals spheres of the ligand's atoms. Because a correctly positioned ligand in the right conformation should correlate better to the electron density than a ligand either positioned incorrectly or in the wrong conformation, we use the local map correlation coefficient to rank the resulting solutions.

2.7. Automated procedure for medial axis ligand building

A Python script was written to carry out the entire ligand-building procedure automatically. *CNS* v.1.1 (Brünger *et al.*, 1998), *Power Crust* v.1.2 (Amenta *et al.*, 2001), *PyMOL* (DeLano, 2002) and *RSRef*2000 (Korostelev *et al.*, 2002) are executed by the script. *RSRef*2000 input files were modified

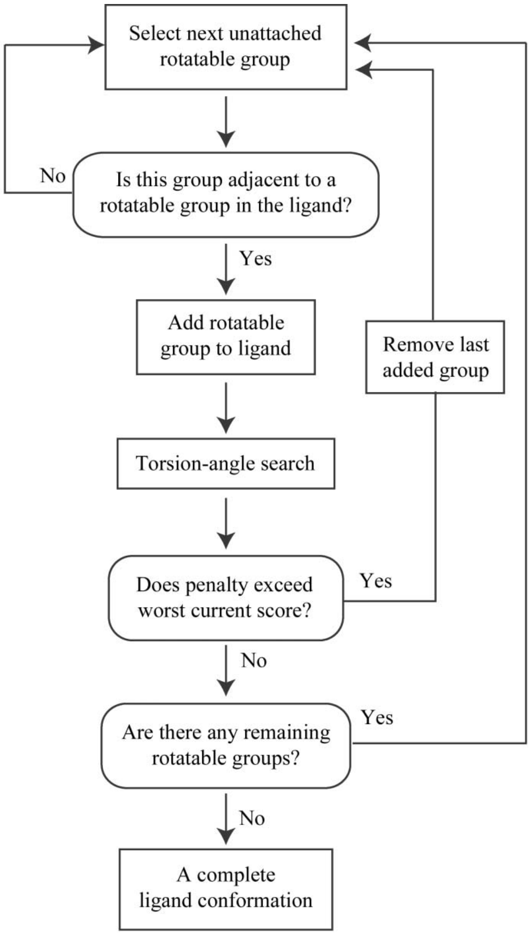
Select next unattached rotatable group
No
Is this group adjacent to a rotatable group in the ligand?
Yes
Add rotatable group to ligand
Remove last added group
Torsion-angle search
Does penalty exceed worst current score?
Yes
No
Are there any remaining rotatable groups?
Yes
No
A complete ligand conformation

Figure 3
Outline of the conformation search.

Table 1
Ligands used in test cases.

Ligand names: SIL, sildenafil; CLR, cholesterol; AMP, adenosine monophosphate; RRP, rolipram; MTX, methotrexate; TRP, tryptophan; RCL, ricinoleic acid; GMP, guanosine monophosphate; HDS, hexadecanesulfonate; IBP, ibuprofen; ATP, adenosine triphosphate; CHD, cholic acid; U49, inhibitor 49; STR, progesterone; REA, retinoic acid; 13P, dihydroxyacetone phosphate; TNT, propamidine. Resolution is the reported resolution. $N_{\rm torsion}$ is the number of rotatable torsion bonds and $N_{\rm atom}$ the number of non-H atoms.

PDB code	Ligands	Resolution (Å)	$N_{\rm torsion}$	$N_{\rm atom}$	Reference
1tbf	SIL	1.3	7	33	Zhang *et al.* (2004)
1lri	CLR	1.5	4	28	Lascombe *et al.* (2002)
1tb7	AMP	1.6	3	23	Zhang *et al.* (2004)
1tbb	RRP	1.6	4	20	Zhang *et al.* (2004)
1ra3	MTX	1.8	10	33	Sawaya & Kraut (1997)
1i1q	TRP	1.9	3	14	Morollo & Eck (2001)
1fk7	RCL	1.9	16	21	Han *et al.* (2001)
1t9s	GMP	2.0	4	24	Zhang *et al.* (2004)
1lpp	HDS	2.1	15	19	Grochulski *et al.* (1994)
1tb5	AMP	2.2	3	23	Zhang *et al.* (2004)
1eqg	IBP	2.6	4	15	Selinsky *et al.* (2001)
1obd	ATP, AMP	1.4	8	31	Antonyuk *et al.* (2001)
1ee2	CHD	1.5	4	29	Adolph *et al.* (2000)
1ld8	U49	1.8	0	33	Bell *et al.* (2002)
1a28	STR	1.8	1	23	Williams & Sigler (1998)
1cbs	REA	1.8	9	22	Kleywegt *et al.* (1994)
1ok4	13P	2.1	4	9	Lorentzen *et al.* (2003)
102d	TNT	2.2	8	23	Nunn & Neidle (1995)

for each ligand to include the corresponding parameter and topology files, resolution limits and ligand-selection commands. Other libraries used include *CCTBX* (Grosse-Kunstleve *et al.*, 2002), *Boost* v.1.29 (Siek *et al.*, 2002) and an alpha test version of *PHENIX* (Adams *et al.*, 2002).

2.8. Coordinate-file preparations

In order to reduce bias, the ligands and protein PDB files shown in Table 1 were prepared for ligand-building tests in the following manner. Firstly, all water molecules, ions and other ligands were removed from the PDB file, resulting in a ligand-removed PDB file. The torsion angles of the ligand were randomly modified such that they no longer resembled the conformation in the deposited structure. Out of 93 rotatable ligand torsion bonds, 63 torsion angles were rotated by an average of 66°. Topology and parameter files for the ligands were obtained from the HicUp database (Kleywegt & Jones, 1998) v.8.2, or provided by the authors (Zhang *et al.*, 2004).

2.9. Ligand building with ARP/wARP

The *ARP/wARP* program was run in its default mode with *CCP4i* (Potterton *et al.*, 2003). *CCP4* v.5.0.2 (Collaborative Computational Project, Number 4, 1994) and *ARP/wARP*

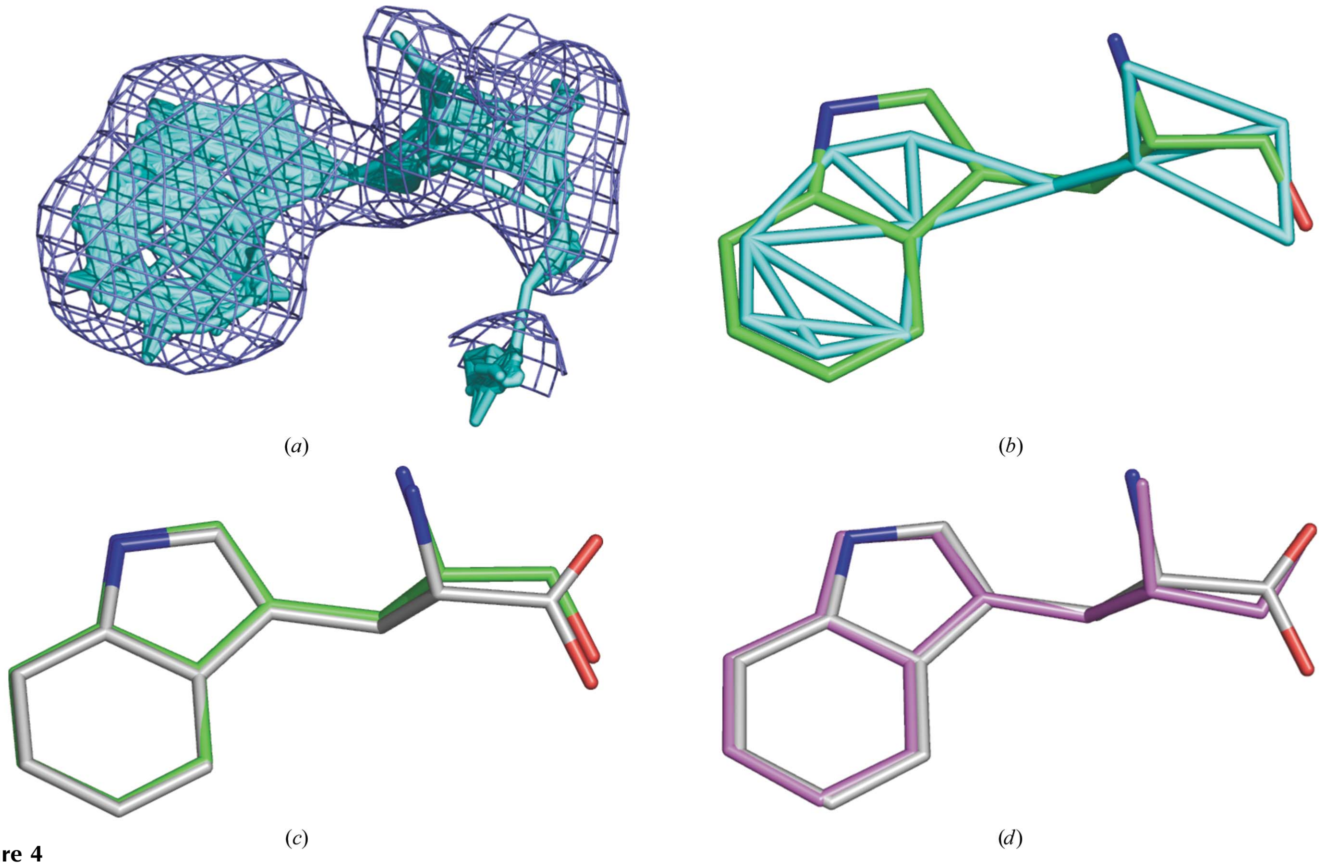

Figure 4
The $F_{\rm o} - F_{\rm c}$ electron-density map for a tryptophan ligand bound to anthranilate synthase (PDB code 1i1q). (*a*) Prior to thinning, the full medial axis (cyan) is a wide flat shape inside the isosurface (blue mesh, contoured at the 3σ level). Both the full medial axis and individual medial axis segments are too complex for direct graph matching against ligand-coordinate graphs. After thinning the medial axis by using *simplify* from the *Power Crust* software package, the medial axis is converted to edges and vertices and thinned to obtain vertices (cyan) separated by approximately atom-to-atom distances. (*b*) The thinned medial axis (cyan) has many vertices near the atom positions from the best model (green). (*c*) The best model (green) aligns well with the deposited structure (gray). (*d*) The *ARP/wARP* model (magenta) also aligns well with the deposited structure (gray).

v.6.1.1 (Lamzin *et al.*, 2001) were used. When multiple *ARP*/*wARP* runs were required to obtain a ligand in the correct region of the electron-density map, the best-fitted ligand from all previous iterations were combined with the original protein-only file and used as the starting file for the next *ARP*/*wARP* cycle.

2.10. Evaluating performance

The solutions found by the two methods were evaluated by comparing the r.m.s.d. between corresponding atoms of the deposited ligand and the built ligand. To better account for artificial misalignments such as reversed building of symmetric molecules, the r.m.s. deviation between each atom of the built ligand to the nearest atom of the deposited ligand is calculated regardless of chemical atom type (referred to as r.m.s.d.$_{\rm any}$).

3. Results

3.1. Testing the method

To test the medial axis method with realistic electron-density maps, PDB entries of ligand–protein cocrystal structures containing deposited structure factors were used as test cases. A total of 18 test structures containing 27 ligands were used (Table 1). The ligands in the test structures contain 9–44 atoms and 0–16 freely rotatable bonds and the structures were solved to between 1.3 and 2.6 Å resolution. It should be noted that these tests represent a proof of concept of the medial axis method, so relatively high-resolution structures were used.

The desired ligands were extracted from the coordinate files and the *B* factors for all atoms were set to 1.5 times the average *B* factor of the refined protein. To reduce the bias from having the search ligand in the deposited conformations, the rotatable ligand torsion angles were randomly modified. Identical starting coordinates were used in the tests of the medial axis method and *ARP*/*wARP* for both the ligands and the proteins.

$F_{\rm o} - F_{\rm c}$ maps were used for calculating isosurfaces. The map isosurface level was set to 3.0σ except for three cases (PDB codes 1fk7, 1tb5 and 102d), where the level was lowered until a thinned medial axis of the appropriate size for graph matching was obtained. Real-space refinement was performed against $2F_{\rm o} - F_{\rm c}$ electron-density maps. All *Power Crust* parameters and thinning programs were used with their default values except for the starting multiplier (Table 2). *ARP*/*wARP* was also run with default parameters.

3.2. The thinned medial axis transform produces molecular shapes

We illustrate the various stages of the method by building a tryptophan molecule bound to anthranilate synthase (1.9 Å

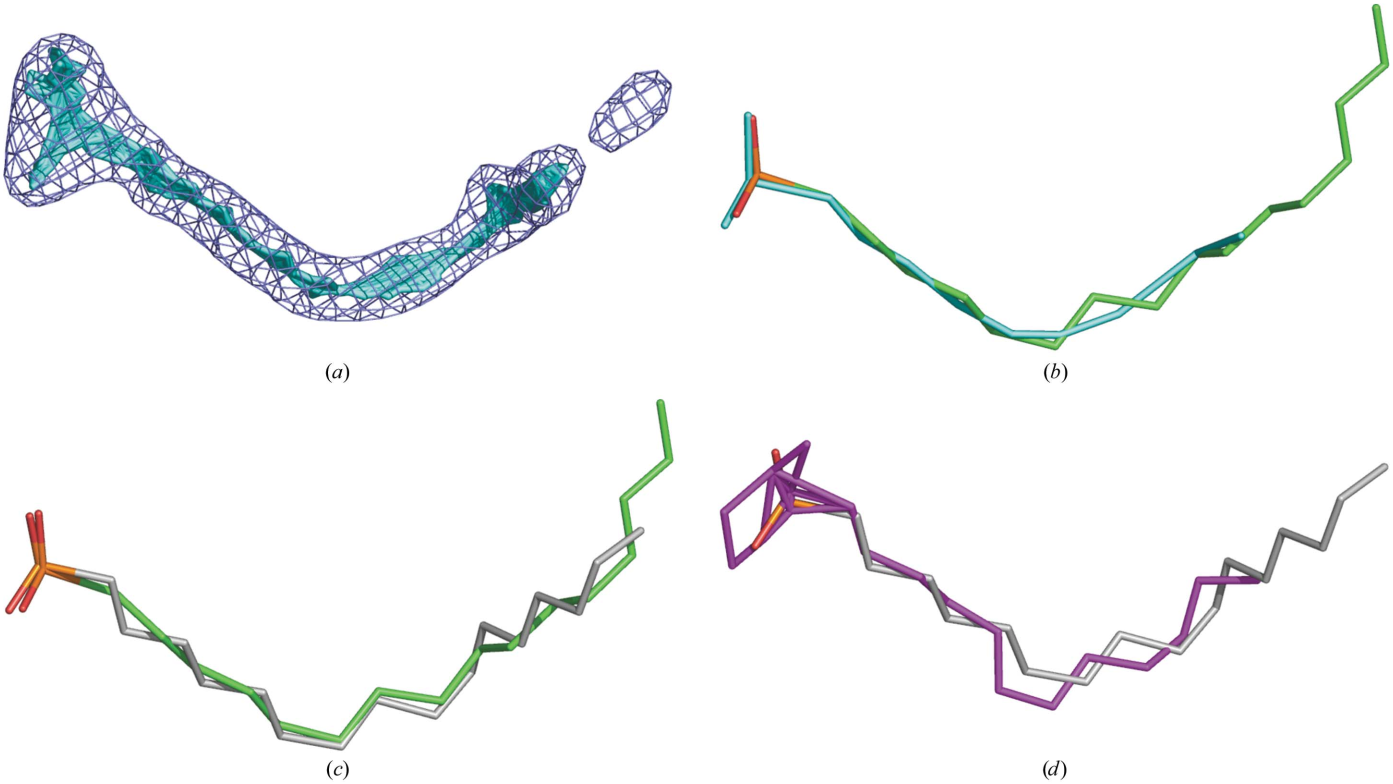

Figure 5
The $F_{\rm o} - F_{\rm c}$ electron-density map showing one of two HDS ligands bound to lyase (PDB code 1lpp). (*a*) The medial axis (cyan) is clearly visible through the head group and much of the aliphatic region at the 3σ level. (*b*) The thinned medial axis (cyan) has a branched region for the head group and a linear region for the hydrocarbon chain; the best model (green) closely follows the medial axis. (*c*) The best structure from the medial axis method (green) matches the deposited structure (gray) throughout the region where the medial axis was visible. (*d*) The *ARP*/*wARP* model (magenta) is built correctly over only part of the ligand (deposited structure, gray).

Table 2
Parameters for the electron-density map calculation, *Power Crust* program and the matching/conformational search programs.

Description	Value
Electron-density map grid spacing	0.25 × map resolution
Initial map isosurface level	3.0σ
Initial *Power Crust* multiplier	100000
Simplify (*Power Crust*) redundancy threshold	1.2
Simplify (*Power Crust*) noise threshold	0.5
Thin: distances below which vertices are combined	1.0
Geometry-matching stage: No. of solutions to keep	200
Conformation-search stage: No. of solutions to keep	50

Table 3
Performance of *ARP*/*wARP*.

The number of iterations required for a ligand to be built within 7 Å r.m.s.d.$_{\rm any}$ are shown in the 'Iterations' column. Two numbers are shown where two ligands were present in the deposited structure. An asterisk indicates ligands that could not be found.

PDB code	Ligand	Iterations
1tbf	SIL	1
1lri	CLR	1
1tb7	AMP	1, 2
1tbb	RRP	7, *
1ra3	MTX	3
1i1q	TRP	1
1fk7	RCL	1
1t9s	GMP	1, 2
1lpp	HDS	3, 8
1tb5	AMP	1, 2
1eqg	IBP	*, *
1obd	ATP, AMP	2, *
1ee2	CHD	4, 5
1ld8	U49	1
1a28	STR	1, 2
1cbs	REA	*
1ok4	13P	1
102d	TNT	1

resolution; PDB code 1i1q). From the calculated electron-density maps, the isosurface clearly resembles a wide flat shape expected of a tryptophan residue in a high-quality electron-density map (Fig. 4*a*). After thinning, the complex full medial axis has been reduced to a much simpler graph that resembles an atomic model (Fig. 4*b*). Another example of the medial axis transform and the thinning procedure is shown for the ligand structure of HDS bound to *Candida rugosa* lipase (2.1 Å resolution; PDB code 1lpp) in Figs. 5(*a*) and 5(*b*).

The locations of the vertices of the simplified medial axis are close to the atoms of the deposited ligand molecular structure. For the tryptophan case, the best model after graph matching and conformation search has an r.m.s.d. of all atoms of 0.18 Å to the deposited structure (Figs. 4*c* and 6) and the r.m.s.d.$_{\rm any}$ of the best model from building HDS is 1.02 Å (Figs. 5*c* and 7), clearly demonstrating that the medial axis method builds ligands in the correct location and position.

3.3. Error-tolerance of the graph-matching algorithm

Fig. 5 highlights one of the strengths of the medial axis and graph-matching methods. While the isosurface (Fig. 5*a*) indicates that the electron-density map is well ordered through C13 of the hexadecanesulfonate ligand, the electron-density map is broken for the rest of the ligand. The full medial axis (Fig. 5*a*) shows a wide region for the head group and a long path for the hydrophobic chain. Thinning the medial axis results in a branched structure with a long tail, as expected for a long linear molecule (Fig. 5*b*). The best built ligand, despite being slightly misaligned at the head group, clearly follows the path of the thinned medial axis throughout its length and then continues on where the electron-density isosurface has a break, similar to the way that the ligand was built in the deposited structure (Fig. 5*c*). The r.m.s.d.$_{\rm any}$ for this ligand is 1.31 Å. In contrast, *ARP*/*wARP* was unable to build the entire ligand so that it fit into the ordered region of the electron-density map (Fig. 5*d*). This case illustrates that our graph-matching algorithm is tolerant of errors (in this case by allowing null matches), making robust building of ligands possible even in cases when the electron-density maps have significant errors, such as disconnected electron density.

3.4. Comparing models built by the medial axis method and *ARP*/*wARP*

The correctness of structures may be classified by their location (whether the ligand is built into the region of the electron-density map where the ligand is located) and the match of the ligand to the deposited structure (whether the ligand has been built correctly). Both methods were able to build most ligands into the correct region. As a result of the refinement and geometry-optimization procedures (real-space refinement by the medial axis method, least-squares minimization by *ARP*/*wARP*; Agarwal, 1978; Murshudov *et al.*, 1997), many structures were built with positions close to that of the deposited structures (with r.m.s.d.$_{\rm any}$ of less than 0.3 Å). The full comparison of all ligands tested are shown in Figs. 6 and 7, and Tables 3, 4 and 5.

3.4.1. Locating the correct region for the ligand. The medial axis method was able to locate the correct region of the electron-density map in 15 out of the 18 test cases (Table 4). No ligands were successfully built into 1obd, where the electron-density maps show weak density for the AMP and ATP ligands. The cases of 1fk7 and 1lpp presented a different problem. In the building of both RCL and HDS, while the second highest correlation coefficients corresponded to the locations of properly built ligands, the ligands with the highest correlation coefficient were built into a region of weak electron density. These are the only two cases where the best ligand obtained by the medial axis method was in an incorrect region. The improper placement of these ligands was found to occur during real-space refinement, where ligands that were close to the thinned medial axis before refinement were found to move into low nearby electron density during real-space refinement.

ARP/*wARP* builds the ligand into the largest block of unaccounted for electron density. As a result, the program must sometimes be run multiple times before a ligand is built into the correct region of the map. In 11 of the 18 structures

Table 4
Performance of the medial axis method.

The numbers indicate correct solutions ranked by the local map correlation coefficient. Two numbers are shown where two ligands were present in the deposited structure. An asterisk indicates ligands that could not be found.

PDB code	Ligand(s)	Correct solution
1tbf	SIL	1
1lri	CLR	1
1tb7	AMP	1, *
1tbb	RRP	1, 2
1ra3	MTX	1
1i1q	TRP	1
1fk7	RCL	2
1t9s	GMP	1, 2
1lpp	HDS	2, *
1tb5	AMP	1, *
1eqg	IBP	1, 2
1obd	ATP, AMP	*, *
1ee2	CHD	1, 2
1ld8	U49	1
1a28	STR	1, 2
1cbs	REA	1
1ok4	13P	1
102d	TNT	1

examined, *ARP*/*wARP* required only the minimum number of runs to identify the correct ligand location(s) in maps (see Tables 3 and 5). Repeated runs of the program eventually lead to all ligand location(s) being found in four additional cases and one of two ligands in two of the seven remaining test cases. In two cases (REA and IBP), none of the correct ligand locations were found (Figs. 6 and 7, respectively). In each of the cases where no ligand was found with *ARP*/*wARP*, the structures contained additional bound ligands, such as heme, β-octylglucoside, *N*-acetylglucosamine and NAD. Poor results were also obtained in building the second HDS from PDB code 1lpp and AMP from PDB code 1tb5, where the r.m.s.d.$_{\rm any}$ values are worse than 5 Å.

3.5. Conformation of the ligands

Both methods cover a range of accuracies of the ligand conformations, with an r.m.s.d.$_{\rm any}$ between 0.1 and 7 Å (Figs. 6 and 7). With the medial axis method, once the correct region of the electron-density map has been located, a solution similar to the deposited structure was generally found within the best ten conformations as ranked by the local map correlation coefficient. The best solutions for the 27 ligands tested were all within r.m.s.d.$_{\rm any}$ of 2 Å to the deposited coordinates (Figs. 6 and 7). Meanwhile, *ARP*/*wARP* also typically built a ligand model that matched the correct solution once it was able to locate the correct electron-density region. *ARP*/*wARP* found 22 ligands that were located within r.m.s.d.$_{\rm any}$ of 7 Å of their deposited positions. Of these 22 ligands, 17 were built with an r.m.s.d.$_{\rm any}$ of less than 2 Å.

3.6. Problem cases

Both methods encountered problems building nucleotide phosphates (such as ATP). The test cases included one molecule each of ATP and AMP in PDB code 1obd, two GMP in PDB code 1t9s and two AMP molecules each in PDB codes 1tb5 and 1tb7. The medial axis method built four of eight of this class of molecules with an r.m.s.d.$_{\rm any}$ of less than 2 Å (one of two AMP from 1tb7, both GMP from 1t9s, one of two AMP from 1tb5 and neither ATP nor AMP from 1obd). The remaining molecules could not be located. *ARP*/*wARP* had similar problems with these test cases. *ARP*/*wARP* was able to successfully build four of eight molecules with an r.m.s.d.$_{\rm any}$ of less than 2 Å (both AMP from 1tb7, both GMP from 1t9s, none of two from 1tb5 and neither ATP nor AMP from 1obd), but was similarly unsuccessful in building the remaining molecules with r.m.s.d.$_{\rm any}$ worse than 2 Å.

Table 5
Number of ligands built correctly.

The number of ligands built with r.m.s.d.$_{\rm any}$ to the deposited structures of less than 7 and 2 Å is also shown for each method.

	No. of ligands	No. of ligands built within 7 Å r.m.s.d.$_{\rm any}$	No. of ligands built within 2 Å r.m.s.d.$_{\rm any}$
Medial axis method	27	22	22
ARP/*wARP*	27	22	17

The electron-density maps corresponding to the nucleotide phosphates contained features that made them difficult to interpret correctly. In one case, two O atoms on the phosphate groups in a GMP molecule are bridged by water molecules and ions. In addition, disconnected electron density is often observed for ATP and AMP molecules, especially between the ribose and phosphate groups.

4. Discussion and conclusions

Fitting flexible nonrepetitive ligands is fundamentally limited by the quality of the electron-density maps. Existing methods for fitting ligands perform best with density that is well connected throughout the ligand, has recognizable chemical features and shows no erroneous connectivity to other parts of the structure. These conditions are met when the ligand is well ordered, fixed and has high occupancy in the crystal. Both the medial axis method as implemented here and *ARP*/*wARP* perform best when electron-density maps meet these criteria. As automatic model building becomes more difficult with poor electron-density maps, manual model building also becomes more difficult.

The medial axis transform and accompanying thinning routines have been effective at producing graphs that can be matched against ligand-coordinate graphs. While calculating the medial axis transform requires more computation than the skeletonization of highest electron-density points, the use of centers of isosurfaces may position vertices better than peak-searching techniques, especially at lower resolutions where atomic positions do not necessarily correspond to peaks in the electron density. For example, in the electron-density map of the tryptophan ligand (PDB code 1i1q) calculated using diffraction data artificially truncated to 3.5 Å resolution, the Bones skeleton (Kleywegt, 1996) appears as a linear piece in the center of the electron density with no molecular detail

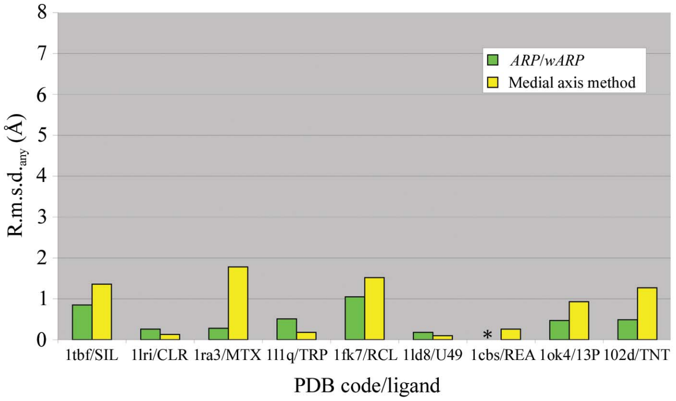

Figure 6
Comparison of built models for structures containing one ligand. R.m.s.d.$_{any}$ is used to evaluate the *ARP/wARP* model closest to the deposited structure and the best model obtained by the medial axis method. An asterisk indicates ligands that were not built to within 7 Å r.m.s.d.$_{any}$.

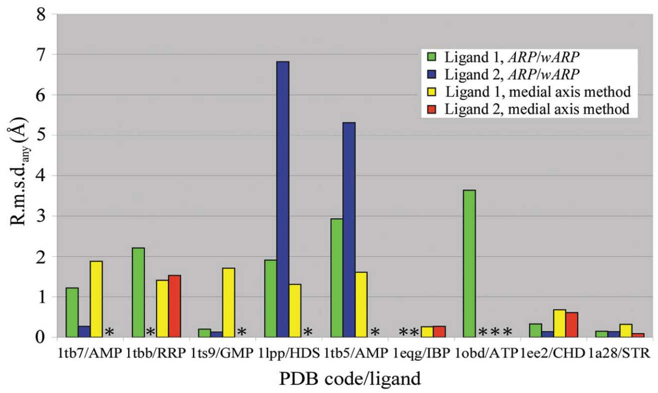

Figure 7
Comparison of built models for structures containing two ligands. R.m.s.d.$_{any}$ is used to evaluate the *ARP/wARP* model closest to the deposited structure and the best model obtained by the medial axis method. An asterisk indicates ligands that were not built to within 7 Å r.m.s.d.$_{any}$.

(Fig. 8*a*), while the thinned medial axis has a shape that better resembles the tryptophan ligand (Fig. 8*b*).

The thinning procedure of the medial axis that we used in this work preserves the topology of the full medial axis and derives a single medial axis graph with individual vertices weighted equally. Improved methods could be devised for thinning the full medial axis to extract more information from the associated electron-density map. For example, medial axes calculated from multiple isosurface contour levels may be used to detect more stable regions of a ligand, suggesting parts of a medial axis that could be weighted more strongly during the matching and conformation-search steps.

Instead of time-consuming exhaustive conformational search algorithms, we used an efficient two-step branch-and-bound matching and search algorithm. In our tests, a solution close to the deposited structure is usually found within the 20 top solutions from the matching stage and after the torsion-angle conformation search the correct conformation is usually among the top ten. Use of additional criteria could further constrain the graph-matching processes, permitting more rapid searches, and ultimately improve the accuracy of the searches. Improvements in the graph-matching process could allow the use of medial axis transform in other pattern-matching applications, such as moderate-resolution model building, structure comparisons and structure classifications.

A major advantage of the medial axis method is that filtering candidate medial axis segments by graph size is effective in reducing the number of segments to undergo computationally intensive comparisons by eliminating regions of the electron-density map that are smaller or larger than the target molecule or ligand. The positions of medial axis vertices are generally close to the atomic positions of the deposited ligand structures and are used to generate multiple solutions that are ranked by the local map correlation coefficient. Thus, the medial axis method is well suited to cases where ligands are bound in multiple conformations to the protein or when the region corresponding to the ligand is smaller than other electron-density map features.

Improved algorithms will have to be developed for handling the more difficult cases where electron-density maps are broken or are not well connected, such as those observed

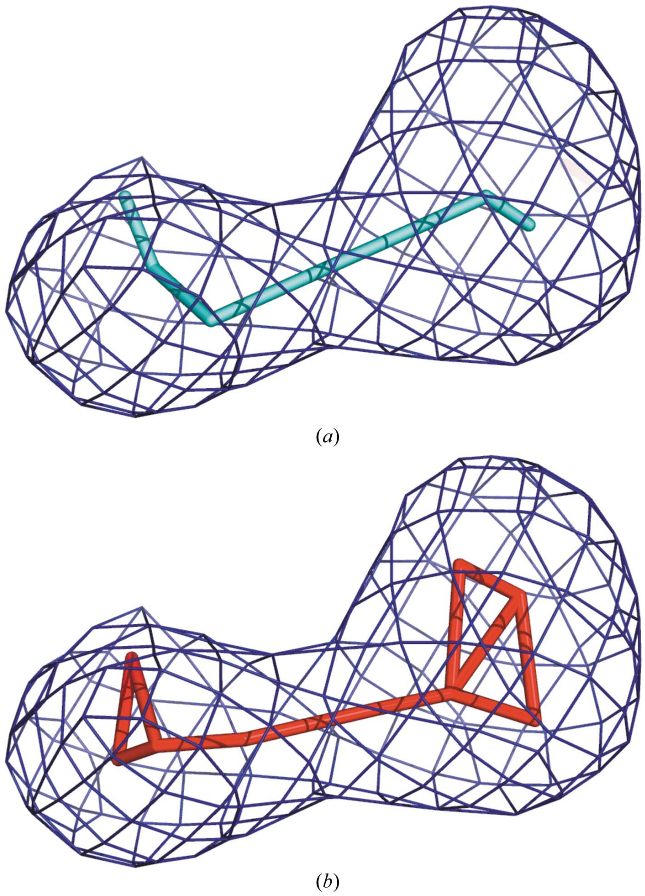

Figure 8
(*a*) Bones skeletonization (Kleywegt, 1996) (cyan) shown within a contour from the $F_o - F_c$ difference map artificially truncated to 3.5 Å resolution calculated from PDB code 1i1q (blue mesh). (*b*) Simplified medial axis (red) shown within the contoured difference map (blue mesh). In both cases, the map was calculated at 2.7σ.

when fitting nucleotide phosphates. For example, analysis of the persistence of features at different isosurface contour levels of the electron-density map may reduce the requirement for strong electron density throughout a molecule. Another challenging problem is building ligands whose identities are unknown.

We have presented here the first successful application of the medial axis algorithm to electron-density map fitting of ligands. The method produces similar or slightly better results than methods based on selecting highest electron-density peaks. Thus, we have shown that the information contained in a single isosurface of the electron density is sufficient for automated model building. The method could be extended to any molecular-fitting problem. Generalization of the method to recognition of common features across multiple contour levels could lead to powerful automatic fitting methods that perform well even at low resolution.

We thank R. Grosse-Kunstleve, N. Moriarty, N. Sauter, P. Kasson and T. Fenn for discussions and the Bio-X Interdisciplinary Initiatives Program at Stanford University and Stanford Synchrotron Radiation Laboratory for generous funding to ATB. Leonidas Guibas and Daniel Russel wish to thank the National Science Foundation for support on this effort through grants ITR-0086013 and CARGO-0310661.

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., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Adolph, H. W., Zwart, P., Meijers, R., Hubatsch, I., Kiefer, M., Lamzin, V. & Cedergren-Zeppezauer, E. (2000). *Biochemistry*, **39**, 12885–12897.

Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.

Amenta, N., Choi, S., Dey, T. K. & Leekha, N. (2002). *Int. J. Comput. Geom. Appl.* **12**, 125–141.

Amenta, N., Choi, S., Jump, M. E., Kolluri, R. & Wahl, T. (2002). Report TR-02–27. University of Texas at Austin, USA.

Amenta, N., Choi, S. & Kolluri, R. (2001). *Sixth ACM Symposium on Solid Modeling and Applications, June 4–8, 2001*, pp. 249–260. Ann Arbor, MI, USA: ACM Press.

Antonyuk, S. V., Grebenko, A. I., Levdikov, V. M., Urusova, D. V., Melik-Adamyan, W. R., Lamzin, V. S. & Wilson, K. S. (2001). *Kristallografiya*, **46**, 687–691.

Bell, I. M. *et al.* (2002). *J. Med. Chem.* **45**, 2388–2409.

Blum, H. (1967). *Models for the Perception of Speech and Visual Form*, edited by W. Whaten-Dunn, pp. 362–380. Cambridge, MA, USA: MIT Press.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Card, G. L., Blasdel, L., England, B. P., Zhang, C., Suzuki, Y., Gillette, S., Fong, D., Ibrahim, P. N., Artis, D. R., Bollag, G., Milburn, M. V., Kim, S. H., Schlessinger, J. & Zhang, K. Y. (2005). *Nature Biotechnol.* **23**, 201–207.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. (1998). *Acta Cryst.* D**54**, 750–756.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* http://www.pymol.org.

Fortier, S., Chiverton, A., Glasgow, J. & Leherte, L. (1997). *Methods Enzymol.* **277**, 131–157.

Greer, J. (1985). *Methods Enzymol.* **115**, 206–224.

Grochulski, P., Bouthillier, F., Kazlauskas, R. J., Serreqi, A. N., Schrag, J. D., Ziomek, E. & Cygler, M. (1994). *Biochemistry*, **33**, 3494–3500.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.

Han, G. W., Lee, J. Y., Song, H. K., Chang, C., Min, K., Moon, J., Shin, D. H., Kopka, M. L., Sawaya, M. R., Yuan, H. S., Kim, T. D., Choe, J., Lim, D., Moon, H. J. & Suh, S. W. (2001). *J. Mol. Biol.* **308**, 263–278.

Kearsley, S. K. (1989). *Acta Cryst.* A**45**, 208–210.

Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 826–828.

Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994). *Structure*, **2**, 1241–1258.

Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.

Kleywegt, G. J. & Jones, T. A. (1998). *Acta Cryst.* D**54**, 1119–1131.

Korostelev, A., Bertram, R. & Chapman, M. S. (2002). *Acta Cryst.* D**58**, 761–767.

Lamzin, V. S., Perrakis, A. & Wilson, K. S. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 720–722. Dordrecht: Kluwer Academic Publishers.

Lascombe, M. B., Ponchet, M., Venard, P., Milat, M. L., Blein, J. P. & Prangé, T. (2002). *Acta Cryst.* D**58**, 1442–1447.

Leherte, L., Glasgow, J., Baxter, K., Steeg, E. & Fortier, S. (1997). *J. Artif. Intell. Res.* **7**, 125–159.

Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.

Lorentzen, E., Pohl, E., Zwart, P., Stark, A., Russell, R. B., Knura, T., Hensel, R. & Siebers, B. (2003). *J. Biol. Chem.* **278**, 47253–47260.

Morollo, A. A. & Eck, M. J. (2001). *Nature Struct. Biol.* **8**, 243–247.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Nunn, C. M. & Neidle, S. (1995). *J. Med. Chem.* **38**, 2317–2325.

Oldfield, T. J. (2001). *Acta Cryst.* D**57**, 696–705.

Paik, D. S., Beaulieu, C. F., Jeffrey, R. B., Rubin, G. D. & Napel, S. (1998). *Med. Phys.* **25**, 629–637.

Pitt, W. R., Williams, M. A., Steven, M., Sweeney, B., Bleasby, A. J. & Moss, D. S. (2001). *Bioinformatics*, **17**, 729–737.

Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *Acta Cryst.* D**59**, 1131–1137.

Sawaya, M. R. & Kraut, J. (1997). *Biochemistry*, **36**, 586–603.

Selinsky, B. S., Gupta, K., Sharkey, C. T. & Loll, P. J. (2001). *Biochemistry*, **40**, 5172–5180.

Siek, J. G., Lee, L.-Q. & Lumsdaine, A. (2002). *The Boost Graph Library: User Guide and Reference Manual.* New York: Addison–Wesley.

Swanson, S. M. (1994). *Acta Cryst.* D**50**, 695–708.

Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1937–1940.

Wang, C. E. (2000). *Acta Cryst.* D**56**, 1591–1611.

Williams, S. P. & Sigler, P. B. (1998). *Nature (London)*, **393**, 392–396.

Zhang, K. Y., Card, G. L., Suzuki, Y., Artis, D. R., Fong, D., Gillette, S., Hsieh, D., Neiman, J., West, B. L., Zhang, C., Milburn, M. V., Kim, S.-H., Schlessinger, J. & Bollag, G. (2004). *Mol. Cell*, **15**, 279–286.

Zwart, P. H., Langer, G. G. & Lamzin, V. S. (2004). *Acta Cryst.* D**60**, 2230–2239.