Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

FINDMOL: automated identification of macromolecules in electron-density maps

E. W. McKee,[a] L. D. Kanbi,[a] K. L. Childs,[a] R. W. Grosse-Kunstleve,[b] P. D. Adams,[b] J. C. Sacchettini[c] and T. R. Ioerger[a]*

[a]Department of Computer Science, Texas A&M University, 301 H. R. Bright Building, 3112 Texas A&M University, College Station, TX 77843, USA, [b]Lawrence Berkeley National Laboratory, One Cyclotron Road, Building 64R0121, Berkeley, CA 94720, USA, and [c]Department of Biochemistry and Biophysics, Texas A&M University, 103 Biochemistry/Biophysics Building, 2128 TAMU, College Station, TX 77843, USA

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

Received 18 June 2005
Accepted 30 August 2005

Automating the determination of novel macromolecular structures *via* X-ray crystallographic methods involves building a model into an electron-density map. Unfortunately, the conventional crystallographic asymmetric unit volumes are usually not well matched to the biological molecular units. In most cases, the facets of the asymmetric unit cut the molecules into a number of disconnected fragments, rendering interpretation by the crystallographer significantly more difficult. The *FINDMOL* algorithm is designed to quickly parse the arrangement of trace points (pseudo-atoms) derived from a skeletonized electron-density map without requiring higher level prior information such as sequence information or number of molecules in the asymmetric unit. The algorithm was tested with a variety of density-modified maps computed with medium- to low-resolution data. Typically, the resulting volume resembles the biological unit. In the remaining cases the number of disconnected fragments is very small. In all examples, secondary-structural elements such as α-helices or β-sheets are easily identifiable in the defragmented arrangement. *FINDMOL* can greatly assist a crystallographer during manual model building or in cases where automatic model building can only build partial models owing to limitations of the data such as low resolution and/or poor phases.

1. Introduction and background

A key step in automating the determination of macromolecular structures *via* X-ray crystallographic methods involves building a model consisting of atomic coordinates into an electron-density map. Some methods such as *ARP/wARP* (Perrakis *et al.*, 1999) and *RESOLVE* (Terwilliger, 2004) work directly from structure factors, applying symmetry to generate electron density at any arbitrary point in space. Other methods such as *TEXTAL* (Holton *et al.*, 2000), *X-Powerfit* (Oldfield, 2003) and *MAID* (Levitt, 2001) must build within the boundaries of a given map generated for a specific region of space by the user. Although the asymmetric unit (ASU) is commonly used in crystallographic data processing, it is not optimal for model building. The macromolecule can be positioned arbitrarily within the ASU volume and the facets of the conventional ASU volume (Hahn *et al.*, 1984) can cut the molecule into a number of disconnected fragments, with symmetry-related portions appearing on opposite sides of the map. As a result, when building models manually crystallographers must often visually identify a contiguous volume of density which contains the protein macromolecule and define an appropriate mask around a

doi:10.1107/S0907444905027332

symmetrically unique region of space which may be arbitrarily shaped and which may cross one or more ASU boundaries[1].

The ability to automatically identify a region of electron density containing the macromolecule and determining its molecular boundaries without the need for human input/judgement is critical for more fully automating the structure-determination process, especially for high-throughput applications such as *PHENIX* (Adams *et al.*, 2002, 2004). Although methods for protein masking and discrimination from solvent regions (Cowtan & Zhang, 1999; Abrahams & Leslie, 1996; Brünger *et al.*, 1998; Terwilliger, 2004) have been available for some time, it has proven challenging to robustly select the core of a macromolecule and determine the correct packing of domains/subunits.

In this paper, we present a novel algorithm for automated identification of a region of electron density containing the macromolecule. *FINDMOL* starts similarly to other protein-masking methods by creating a solvent-flattened map to enhance density within the protein regions. A skeleton of trace points (Greer, 1974; Ioerger & Sacchettini, 2002) is then created inside the high-density contours, representing occupied regions of space. Cluster analysis is used to divide the trace points into local groups that are either disconnected from each other or have only minimal contact, *e.g.* across non-biological crystal contacts (Carugo & Argos, 1997; Dasgupta *et al.*, 1997; Valdar & Thornton, 2001). Furthermore, crystallographic symmetry operations are used to extend clusters across ASU boundaries, allowing complete macromolecules to be identified by appending adjacent symmetry copies of trace points from other parts of the ASU. Finally, the resultant trace points may be used to generate a mask that can be output for automated map generation and automated model building.

While the routine usually cannot separate subunits packed in biologically relevant complexes, which often have more extensive protein–protein interfaces, tests on a variety of medium-resolution maps (2.0–3.0 Å) show that *FINDMOL* can effectively identify regions of electron density containing a contiguous and unique macromolecule for building. This facilitates automation of structure determination and can also benefit crystallographers as a tool for manual model building. The *FINDMOL* algorithm has been incorporated into *PHENIX*, a comprehensive crystallographic computing platform for structure determination (Adams *et al.*, 2002, 2004).

2. Data preparation

In preparation for running *FINDMOL*, a density-modified electron-density map is produced using a method such as *DM* (Cowtan & Zhang, 1999) and then scaled and skeletonized with *CAPRA* (Ioerger & Sacchettini, 2002). The scaling routine is designed to normalize the magnitudes of the density values. The tracing routine returns a set of trace points (pseudo-atoms), a skeletonized surrogate for the high-density regions of the map occupied by the macromolecule, derived from an interpolated 0.5 Å grid over the region, which approximates the medial axis of a 1σ contour. In order to allow these trace points to span the entire ASU, the electron density is computed over the ASU and a small border. The initial set of trace points is then pruned by elimination of points which lie in the border. The identification of trace points lying entirely in the border is facilitated through use of the `direct_space_asu` class in the *CCTBX* (Grosse-Kunstleve *et al.*, 2002), the open-source crystallographic library component of the *PHENIX* software suite (Adams *et al.*, 2002, 2004). The pruned set of trace points forms the input for the *FINDMOL* algorithm.

3. The *FINDMOL* algorithm

3.1. Definitions

Given a trace point i, another trace point j is said to be a neighbor of i if distance$(i, j) \leq \varepsilon$, where distance is the Euclidean distance and ε is a user-defined parameter. The default value for ε is 6 Å. Adjusting the neighborhood radius in a reasonable range (5–9 Å) may help prevent trace-point arrangements bridging symmetry-related molecules in crystal structures with tightly packed molecules.

The set of all neighbors, the neighborhood, for a point i, $N(i)$, is defined as $N(i) = \{j | j \text{ is a neighbor of } i\}$. The neighborhood density of i, $D(i)$, is defined to be number of neighbors in the set $|N(i)|/(4\pi\varepsilon^3)$.

For the purpose of these definitions, a trace point and all its symmetry-equivalent copies are considered. Essentially, the *FINDMOL* algorithm is a means of selecting one member from each set of symmetry-equivalent trace points so that the selected points are packed together in a contiguous part of space.

3.2. Methods

3.2.1. Rationale and overview. Typically, the neighborhood densities $D(i)$ in the core regions of proteins are higher than those in surface regions. In the solvent regions, trace points are very sparse owing to the prior application of density-modification procedures. A gradual transition from high to low neighborhood density occurs at the protein surface and depends on ε. Furthermore, in most cases crystal contacts involve less surface area than biological contacts (Dasgupta *et al.*, 1997); thus, there will often be a difference in the density of trace points between the two. This will lead to trace points near a crystal contact having a smaller neighborhood density than trace points near a biological contact. Thus, the overall strategy of the *FINDMOL* algorithm is to select trace points in a sequence of decreasing neighborhood density, effectively building up a region starting from the core of the protein and expanding outwards to the surface. In such an ordering, it is expected that biological contacts will be crossed before crystal contacts are considered (Fig. 1).

As input, *FINDMOL* takes a set of structure factors and phases and then reassembles trace points from a skeletonized

[1] A unique and contiguous macromolecule is important for subsequent refinement steps, to avoid steric clashes by symmetry and to allow proper computation of geometric and non-bonded constraints, preventing severe distortions of the macromolecule.

electron-density map generated with data up to a resolution of 2.8 Å. High-resolution data sets may generate maps containing strong side-chain density at crystal contacts, whereas at 2.8 Å resolution side-chain density is typically not so well defined. As a result of limiting the resolution, we obtain fewer trace points near the surface of the protein. This prevents *FINDMOL* from bridging over crystal contacts.

It was also observed that small clusters of trace points can be found on the surface of the molecule, representing surface side chains or ordered solvent. It was found that better results can be obtained if these points are removed prior to the main *FINDMOL* algorithm owing to the reduction in contact between neighboring molecules in the crystal structure. Such points typically manifest themselves as small isolated clusters that are separated from the main cluster of points by a distance greater than one grid unit (0.5 Å) of the electron-density map. Hence, these peripheral points are removed by cluster analysis.

The maximum distance between two adjacent grid points, approximated as a cubic grid, is approximately $(0.5 \times 3)^{1/2}$ Å = 0.877 Å. To locate these points, the `asu_clusters` object of the *CCTBX* is used to tabulate connected components of points within a 1.0 Å radius. This clustering operation can be visualized by letting the trace points be the vertices of a graph. Any two vertices i and j are considered connected if distance$(i, j) < 1.0$. All points which are members of connected components of size < 3 are eliminated. As a consequence, this operation also removes other types of noise which may be found in the map, including isolated regions of density in solvent regions.

3.2.2. Algorithm. The neighborhood calculations are based on the `asu_mappings` and `pair_asu_table` classes provided by the *CCTBX*. These classes were originally developed for the efficient computation of non-bonded interactions in refinement (Grosse-Kunstleve *et al.*, 2004), where similar algorithms are part of the standard repertoire. The `asu_mappings` class uses the space-group symmetry to map the initial set of trace points into the conventional ASU volume plus a border of width ω to account for all possible neigbors of trace points lying near a facet of the ASU volume. The `pair_asu_table` class builds the list of trace-point pairs within radius ε using a highly efficient algorithm that assigns the points stored in the `asu_mappings` object to cubes with

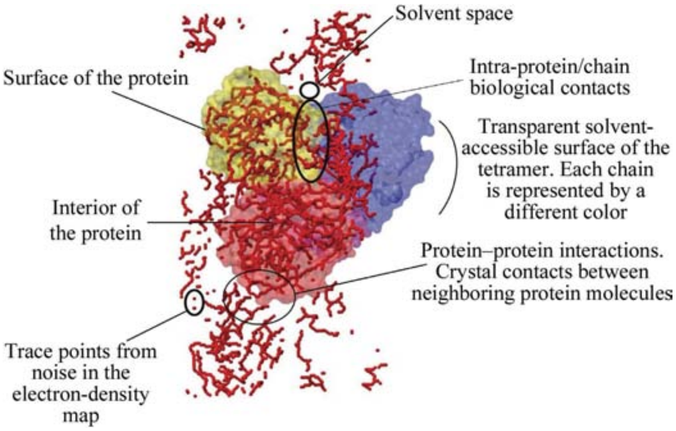
Solvent space
Surface of the protein
Intra-protein/chain biological contacts
Transparent solvent-accessible surface of the tetramer. Each chain is represented by a different color
Interior of the protein
Protein–protein interactions. Crystal contacts between neighboring protein molecules
Trace points from noise in the electron-density map

Figure 1
A comparison of trace=point density in an ASU of α2u globulin. This illustrates four chains with associated 222 symmetry. The transparent solvent = accessible surface of the tetramer is superimposed over the ASU. The trace points are sparser at the crystal contact than at the biological contact, leading to a difference in neighborhood density. Tight extensive interactions exist between each chain of the tetramer. All figures were created with *CHIMERA* (Pettersen *et al.*, 2004).

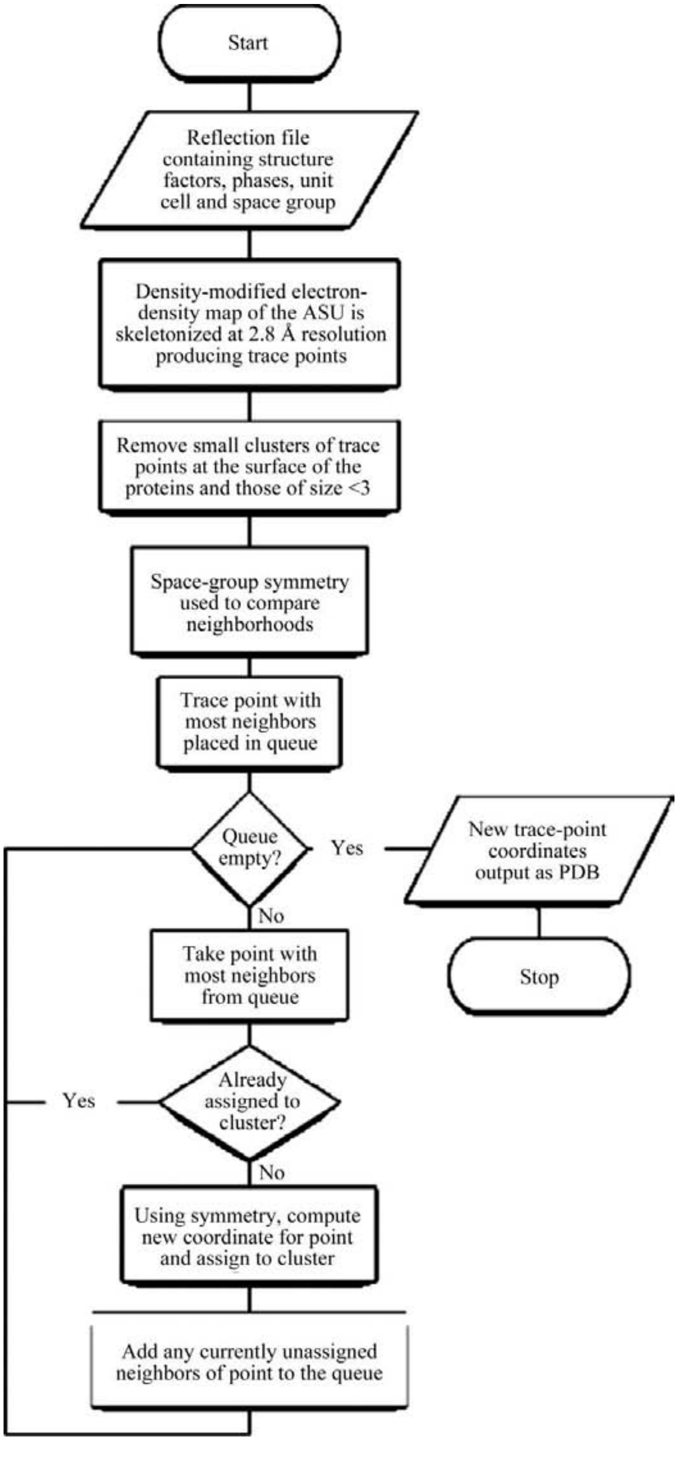
Start
Reflection file containing structure factors, phases, unit cell and space group
Density-modified electron-density map of the ASU is skeletonized at 2.8 Å resolution producing trace points
Remove small clusters of trace points at the surface of the proteins and those of size <3
Space-group symmetry used to compare neighborhoods
Trace point with most neighbors placed in queue
Queue empty?
Yes
New trace-point coordinates output as PDB
Stop
No
Take point with most neighbors from queue
Yes
Already assigned to cluster?
No
Using symmetry, compute new coordinate for point and assign to cluster
Add any currently unassigned neighbors of point to the queue

Figure 2
A schematic diagram illustrating the *FINDMOL* algorithm.

Table 1
Nine experimental electron-density maps were scaled and skeletonized and then defragmented using the *FINDMOL* algorithm.

Seven data sets (i–vii) were fully defragmented with >95% of the trace points repositioned over the protein macromolecule. Isocitrate lyase and calmodulin (viii and ix, respectively) were the only proteins that were only partially defragmented; the trace points over the protein macromolecule were ~73 and 43%, respectively. These proteins are cases that belong to two types of proteins that currently require additional assembly, either manually or later in the model-building process.

Protein (PDB code)	Resolution (Å)	No. of molecules/ No. of chains in the ASU†	Space group	Phasing method	No. of trace points	% of trace points over final protein model or its symmetry-related copy in a neighboring ASU‡			
(i) α2u globulin (2a2u)	2.50	1 (4)	$P2_12_12_1$	MR	4968	99.87	0.10	0.03	—
(ii) NADPH-flavin oxidoreductase (1bkj)	2.50	1 (2)	$P2_1$	SIRAS	3867	99.29	0.52	0.19	—
(iii) Cyanase (1dw9)	2.40	1 (10)	$P1$	MAD	14785	99.76	0.19	0.05	—
(iv) Armadillo-repeat region from β-catenin (3bct)	2.40	1 (1)	$C222_1$	MAD	3698	96.60	2.41	0.94	0.05
(v) Neuronal synaptic fusion complex (1sfc)§	2.40	3 (12)	$I222$	SAD	6077	95.81	3.78	0.18	0.09
(vi) Ornithine aminotransferase (1gbn)	2.30	3 (3)	$P3_221$	MR	10833	95.34	4.58	0.08	—
(vii) HiPIP (1iua)	0.80	1 (1)	$P2_12_12_1$	MR	417	99.51	0.49	—	—
(viii) Isocitrate lyase (1f61)	2.00	1 (2)¶	$P6_522$	MAD	7697	73.25	24.60	2.13	0.02
(ix) Calmodulin (1exr)	1.10	1 (1)	$P1$	SAD	1122	43.60	38.69	15.2	1.01

† Values are listed as functional molecules, with the number of chains in parentheses. ‡ The percentage of trace points over one contiguous protein of the trace points in macromolecule in an ASU. All experimental electron-density maps were generated using a resolution cutoff of 2.8 Å. The columns are sorted by the amount of overlap. § Fusion complex consists of three distinct proteins. ¶ The biological unit spans over two ASUs.

edge length ε. Distance calculations are performed only for points assigned to neighboring cubes, thus reducing the time complexity of the algorithm from $O(N^2)$ to $O(N)$ in the average case.

The next step in the *FINDMOL* algorithm (Fig. 2) is to build clusters by grouping the trace points starting from the conventional ASU volume. The implementation of this step is based on the considerations outlined in the previous section. A queue Q of candidate points is initialized and is empty at the start. All trace points are marked as unassigned. Each point is initially associated with the identity symmetry operation. The trace point with the highest neighborhood density (as provided by the `pair_asu_table` object) is selected as the initial most dense point for the putative core of the molecule and is enqueued in Q along with its current symmetry operator.

Subsequently, neighbors are added in order of decreasing neighborhood density to grow the molecular region by accretion. Each trace point, P_i, in Q is stored as a pair $\{\mathbf{x}_i, \sigma_i\}$, where $\mathbf{x}_i$ is the original coordinate of P_i and σ_i is the symmetry operation required for computation of this symmetric copy of P_i. At each iteration, the algorithm selects the point P_i with the maximum neighborhood density from Q. If this point is already marked as assigned, P_i is discarded and the next iteration is started. Otherwise, P_i is now marked as assigned and the following procedures carried out upon it. Coordinates for P_i are calculated through use of its initial coordinate and its symmetry operation. All neighbors of this point are then examined. Each neighbor P_j which is not yet marked as assigned is enqueued into Q. The required symmetry operation, σ_j is computed by

$$\sigma_j = \sigma_i \sigma_{ia}^{-1} \sigma_{ja}. \qquad (1)$$

In this equation, σ_{ia} and σ_{ja} are the symmetry operations that map P_i and P_j, respectively, from the current location into the ASU volume as tabulated in the `pair_asu_table` and `asu_mappings` objects. Once Q is empty, the algorithm terminates. In some cases of disconnected density the trace points devolve into more than one cluster. In this case, not all points will be marked as assigned after the first pass and the set of unassigned points is used to rerun the algorithm to collect additional clusters. This is repeated until all points are assigned. Each pass generates a cluster rooted in the conventional ASU volume and growing in an arbitrary direction. Thus, the final position of the clusters with respect to one another is arbitrary.

Experiments were also performed to validate that the heuristic of queuing points based on neighborhood density is reasonable. As a comparison, a simple FIFO ('first in first out') queue was substituted for the priority queue. In this way, the algorithm can be envisioned as starting with the point of highest neighborhood density and expanding in concentric shells of trace points until a single symmetry copy of each trace

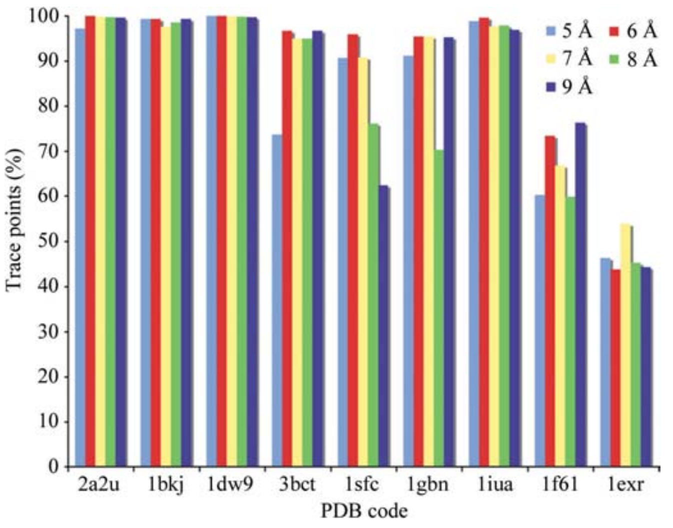

Figure 3
The largest percentage of trace points over the final protein model in an ASU for a given neighborhood radius ε (5, 6, 7, 8 and 9 Å). A key to the PDB codes is given in Table 1.

point is retained. This was demonstrated using our test set of electron-density maps to perform worse than the algorithm employing the priority queue.

4. Results and discussion

Density-modified electron-density maps were computed with a resolution cutoff of 2.8 Å for several experimental data sets for proteins of various shapes, sizes and oligomeric states. These maps were scaled and skeletonized using routines from the *CAPRA* component of *TEXTAL* (Ioerger & Sacchettini, 2002). The resulting sets of trace points were then pruned and passed through the cluster-size filter described previously. A value of $\varepsilon = 6$ Å was used as the default neighborhood radius in the experiments below. The impact of this parameter on the *FINDMOL* output was found to be minor when *FINDMOL* was run on representative maps varying ε from 5 to 9 Å (Fig. 3). Although the amount of a molecule found depends on ε and no ε is consistently best, the variation is usually minor (<10%). The output arrangements were then analyzed by a scoring routine which generates all symmetry copies of the known structure and computes the percentage of output points which corresponded to each of the symmetry copies. This determines the number and size of molecular fragments generated by the *FINDMOL* algorithm. We are primarily interested in recovering a symmetry-unique set of points associated with a single copy of the molecule, regardless of its molecular boundary or position in space. Hence, our metric finds the copy with maximum coverage. Results from a representative sample of runs are summarized in Table 1. These results are taken from a set of *FINDMOL* runs over a collection of 50 phased and density-modified data sets graciously provided by Dr Paul Adams.

The nine experimental data sets that were used to test *FINDMOL* are (i) α2u globulin (PDB code 2a2u; Chaudhuri *et al.*, 1999), (ii) NADPH-flavin oxidoreductase (PDB code 1bkj; Tanner *et al.*, 1996), (iii) cyanase (PDB code 1dw9; Walsh *et al.*, 2000), (iv) armadillo-repeat region from β-catenin (PDB code 3bct; Huber *et al.*, 1997), (v) fusion complex (PDB code 1sfc; Sutton *et al.*, 1998), (vi) human ornithine aminotransferase (PDB code 1gbn; Shah *et al.*, 1997), (vii) HiPIP (PDB code 1iua; Liu *et al.*, 2002), (viii) isocitrate lyase (PDB code 1f61; Sharma *et al.*, 2000) and (ix) calmodulin (PDB code 1exr; Wilson & Brunger, 2000). Over most of the maps (seven out of nine), *FINDMOL* is able to locate large contiguous portions (>95%) of a single copy of a protein/complex. The

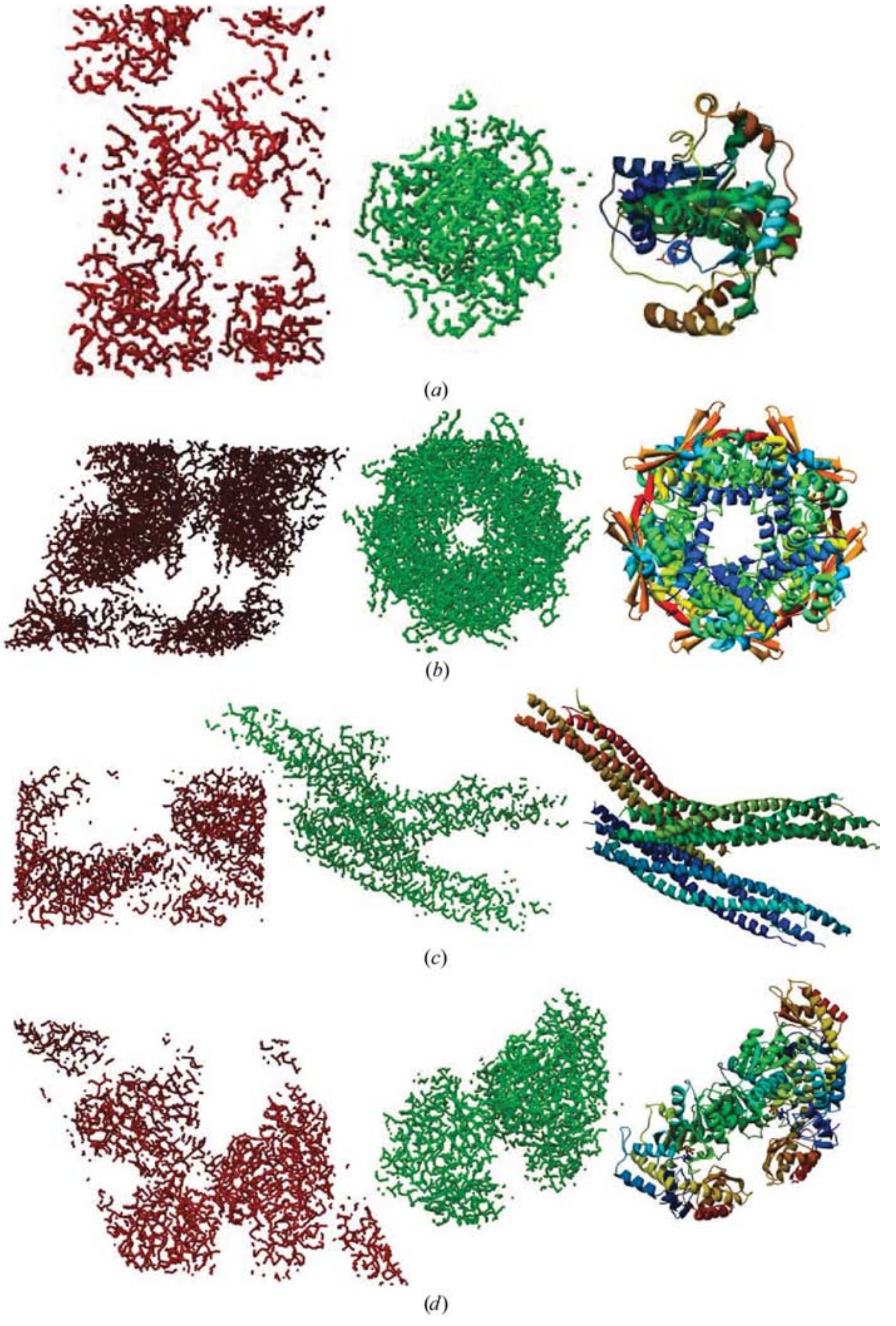

Figure 4
Red trace points, skeletonized electron-density map covering a single ASU (red). Green trace points, skeletonized electron density map defragmented by *FINDMOL*. The rainbow colored ribbons are the published protein model of (*a*) NADPH-flavin oxidoreductase, (*b*) cyanase, (*c*) neuronal synaptic fusion complex and (*d*) human ornithine aminotransferase. *FINDMOL* fully defragments the protein macromolecule (*a*–*d*). All figures were created with *CHIMERA* (Pettersen *et al.*, 2004).

two examples where the defragmentation does not fully succeed are calmodulin and isocitrate lyase (discussed below). However, even in these cases *FINDMOL* outputs just two and three fragments, respectively.

α2u globulin consists of four polypeptide chains of 158 amino acids, with each having a closed eight-stranded β-barrel structure. The four chains form a tetramer (Chaudhuri *et al.*, 1999). Almost all of the trace points (>99%) in the ASU are correctly regrouped over the tetramer by *FINDMOL* (Table 1).

NADPH-flavin oxidoreductase consists of two protein chains of 230 amino acids that form a homodimer in the ASU; upon dimer formation 4834.6 Å^2 of solvent-accessible surface area is lost. NADPH-flavin oxidoreductase reduces FMN to luciferase, which is important for bacterial bioluminescence (Tanner *et al.*, 1996). Almost all of the *FINDMOL* trace points (>99%) are found over the dimer (Table 1). Tight intraprotein packing ensured correct defragmentation by *FINDMOL*.

Cyanase consists of ten polypeptide chains each consisting of 156 amino acids that assemble to form a dimer of pentamers. The result is a decamer that is essential in forming the active site of the enzyme (Walsh *et al.*, 2000). The core of each chain consists of four helices forming a folded leaf arrangement with additional subdomains. Almost 100% of the *FINDMOL* trace points are found over the protein macromolecule owing to the tight protein–protein interactions between neighboring protein chains in the ASU (Fig. 1*a*). Cyanase has an extensive solvent-accessible area of 6345.8 Å^2 that is lost upon decamer assembly in the ASU (Henrick & Thornton, 1998). The oligomer is fractured into several symmetry-related pieces in the conventional ASU. *FINDMOL* fully identifies the protein macromolecule from the electron density extended across several ASUs (Fig. 4*b*).

Armadillo-repeat region from β-catenin is a 42-amino-acid sequence motif from murine β-catenin consisting of one monomer of 457 amino acids. 12 copies of the armadillo repeat are arranged into two curved layers: a right-handed superhelix layer and an α/α layer (Huber *et al.*, 1997). The armadillo repeats have extensive intrachain interactions that enable *FINDMOL* to reorganize ~97% of the trace points over the monomer (Table 1). There is little to no biological interaction between neighboring protein molecules and those that exist can be assigned as crystal contacts, *i.e.* surface contacts which are ≤400Å^2 (Dasgupta *et al.*, 1997).

Neuronal synaptic fusion complex is a three-molecule complex consisting of syntaxin-1A, synaptobrevin-II and SNAP-25A. Each molecule consists of a four-helix bundle (Fig. 4*e*) that forms both a non-globular and unusually twisted oligomer (Sutton *et al.*, 1998). The fusion complex is involved in the fusion of vesicles and membranes and its non-globular nature did not challenge the *FINDMOL* algorithm. It correctly defragments this unusual protein with almost 96% of the trace points being returned over the trimer (Fig. 4*c*).

Human ornithine aminotransferase consists of three polypeptide chains consisting of 402 amino acids and forms a homodimer. The ASU contains three molecules, two of which form a homodimer; the third forms a homodimer with itself across a symmetry axis. The protein has two major domains, one with a three-layer $\alpha/\beta/\alpha$ sandwich, while the second domain is smaller with only a two-layer α/β-sandwich structure (Shah *et al.*, 1997). *FINDMOL* correctly repositions ~95% of the trace points over the two domains (Table 1).

HiPIP is a small iron–sulfur protein (83 amino acids), consisting of a five helices and three β-strands connected mainly by β-turns (Liu *et al.*, 2002). HiPIP has relatively few secondary-structural features and as a globular protein *FINDMOL* manages to reassemble almost all of the trace points correctly over it (Table 1), as expected.

Isocitrate lyase from *Mycobacterium tuberculosis* is a homotetramer in solution. However, isocitrate lyase crystallizes as a dimer of two pairs of entwined monomers in the ASU (Sharma *et al.*, 2000). Each monomer consists of 418 amino acids, consisting of a TIM α/β-barrel fold with additional subdomains that form a domain-swapped dimer. The biological tetramer is formed from two of these dimers bisected by a crystallographic twofold axis. *FINDMOL* only partially identified the protein molecule owing to tight packing between the neighboring symmetrically-related dimers that form the biological tetramer. An extensive solvent-accessible area of 7127.5 Å^2 is lost upon tetramer assembly (Henrick & Thornton, 1998). It is expected that *FINDMOL* would experience difficulties with identifying the biological unit owing to the symmetric redundancy. Isocitrate lyase's 'biological symmetry' coincides with the crystallographic symmetry; hence, its biological unit spans over two ASUs. Only 73.3% of trace points are returned over one dimer and the rest are over neighboring dimers (Table 1). In cases such as this, *FINDMOL* is likely to return protein segments (domains, folds or secondary structure) which can be visually identified and moved to the correct position.

Calmodulin is a case of weak intraprotein interactions; a single disordered helix connects two domains. In this case, interactions with neighboring molecules are more extensive than those found within the protein itself. Calmodulin has a single 146-amino-acid chain consisting of two pairs of EF-hand units, each having two helices connected by a calcium-binding loop (Wilson & Brunger, 2000). The two domains are connected by a long isolated helix, which has relatively weak density [by $D(i)$ calculation], making a 'dumb-bell' shape. *FINDMOL* partially defragments the protein macromolecule, returning approximately 44, 39 and 15% of the trace points over three distinct domains of calmodulin (Table 1) and these can be later reassembled during model building. In this case, *FINDMOL* essentially returns two domains from separate symmetry copies, whose interface is denser than the intermolecular helical linker.

5. Conclusions

In this paper, we present the *FINDMOL* algorithm for the automated identification of macromolecules in medium- to low-resolution density-modified electron-density maps. *FINDMOL* can assist automated model building by defining a contiguous symmetry-unique region of space for model building. The algorithm is based on the cluster analysis of neighborhood densities of trace points obtained *via* skeletonization. *FINDMOL* does not require knowledge of the number of molecules in the ASU and is also independent of sequence information. Our results show that the *FINDMOL* algorithm is able to identify boundaries of complete molecules/complexes in low-resolution electron-density maps, regardless of where they are located with respect to the conventional ASU volume. *FINDMOL* also performs well for unusually shaped protein molecules such as the neuronal synaptic fusion complex and the armadillo-repeat region from β-catenin. This demonstrates that *FINDMOL* is not limited to globular proteins alone. However, the current methodology has two limitations. Firstly, *FINDMOL* cannot easily separate tightly packed protein complexes. Secondly, *FINDMOL* is challenged by macromolecules with biological symmetry that coincides with crystallographic symmetry, *e.g.* a dimer interface across a twofold axis. Proteins with internal regions of weak electron density may also be difficult to assemble. Preliminary experiments indicate that some improvements are possible based on the ideas of ‘dilation and erosion’ established in the field of mathematical morphology (Serra, 1982), but it is clear that tightly packed complexes can be built correctly only if the crystallographic symmetry is reconsidered after advancing the model building to the point of docking the sequence. However, even in these cases *FINDMOL* will be helpful since it is likely to greatly reduce the number of required rearrangement steps.

6. Availability and performance

The *FINDMOL* algorithm described in the paper is implemented as part of the *TEXTAL* suite (Ioerger & Sacchettini, 2003), which is available as a standalone from the corresponding author and also as a component of the *PHENIX* suite (Adams *et al.*, 2002) available from http://phenix-online.org. A map as large as cyanase which contains ten molecules (1560 amino acids) and 14 785 trace points in the *P*1 space group can be scaled, traced and then defragmented by *FINDMOL* to cover the homodecamer within 90 s (AMD Opteron 2.6 GHz).

The authors would like to thank all the members of the *TEXTAL* group, Tod Romo, Reetal Pai, Kreshna Gopal and Jacob Smith, for their valuable criticism and support for this research. This work was supported by NIH grant GM-63210. We especially thank Peter Zwart for his suggestions and feedback in the preparation of this manuscript. Crystal structure data sets were graciously provided to Paul Adams by the Berkeley Structure Genomics Center, A. Brunger, G. Gilliland, E. Gordon, O. Herzberg, J. Sacchettini and J. Smith.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Adams, P. D., Gopal, K., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Pai, R. K., Read, R. J., Romo, T. D., Sacchettini, J. C., Sauter, N. K., Storoni, L. C. & Terwilliger, T. C. (2004). *J. Synchrotron Rad.* **11**, 53–55.

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.

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.

Carugo, O. & Argos, P. (1997). *Proteins*, **28**, 29–40.

Chaudhuri, B. N., Kleywegt, G. J., Bjorkman, J., Lehman-McKeeman, L. D., Oliver, J. D. & Jones, T. A. (1999). *Acta Cryst.* D**55**, 753–762.

Cowtan, K. D. & Zhang, K. Y. (1999). *Prog. Biophys. Mol. Biol.* **72**, 245–270.

Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. & Bell, J. A. (1997). *Proteins*, **28**, 494–514.

Greer, J. (1974). *J. Mol. Biol.* **82**, 279–301.

Grosse-Kunstleve, R. W., Afonine, P. V. & Adams, P. D. (2004). *IUCr Comput. Commun. Newsl.* **3**, 19–36. http://www.iucr.org/iucr-top/comm/ccom/newsletters/2004jan/index.html.

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

Hahn, T., Shmueli, U. & Wilson, A. J. C. (1984). *International Tables for Crystallography*. Dordrecht: Kluwer Academic Publishers Group.

Henrick, K. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 358–361.

Holton, T., Ioerger, T. R., Christopher, J. A. & Sacchettini, J. C. (2000). *Acta Cryst.* D**56**, 722–734.

Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). *Cell*, **90**, 871–882.

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.

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

Liu, L., Nogi, T., Kobayashi, M., Nozawa, T. & Miki, K. (2002). *Acta Cryst.* D**58**, 1085–1091.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.

Oldfield, T. (2003). *Methods Enzymol.* **374**, 271–300.

Serra, J. P. (1982). *Image Analysis and Mathematical Morphology*. New York: Academic Press.

Shah, S. A., Shen, B. W. & Brünger, A. T. (1997). *Structure*, **5**, 1067–1075.

Sharma, V., Sharma, S., Hoener zu Bentrup, K., McKinney, J. D., Russell, D. G., Jacobs, W. R. Jr & Sacchettini, J. C. (2000). *Nature Struct. Biol.* **7**, 663–668.

Sutton, R. B., Fasshauer, D., Jahn, R. & Brünger, A. T. (1998). *Nature (London)*, **395**, 347–353.

Tanner, J. J., Lei, B., Tu, S. C. & Krause, K. L. (1996). *Biochemistry*, **35**, 13531–13539.

Terwilliger, T. (2004). *J. Synchrotron Rad.* **11**, 49–52.

Valdar, W. S. & Thornton, J. M. (2001). *Proteins*, **42**, 108–124.

Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (2000). *Structure Fold. Des.* **8**, 505–514.

Wilson, M. A. & Brunger, A. T. (2000). *J. Mol. Biol.* **301**, 1237–1256.