new algorithms workshop

Interpretation of very low resolution X-ray electron-density maps using core objects

Philipp Heuser,* Gerrit G. Langer and Victor S. Lamzin

Hamburg Unit, European Molecular Biology Laboratory, c/o DESY, Notkestrasse 85, Hamburg 22603, Germany

Correspondence e-mail: philipp.heuser@embl-hamburg.de

A novel approach to obtaining structural information from macromolecular X-ray data extending to resolutions as low as 20 Å is presented. Following a simple map-segmentation procedure, the approximate shapes of the domains forming the structure are identified. A pattern-recognition comparative analysis of these shapes and those derived from the structures of domains from the PDB results in candidate structural models that can be used for a fit into the density map. It is shown that the placed candidate models can be employed for subsequent phase extension to higher resolution.

1. Introduction

Macromolecular crystallography (MX) has been the main source of three-dimensional structural information at an atomic level. MX has provided over 85% of all entries in the Protein Data Bank (PDB; Berman *et al.*, 2000) and over 95% of those for proteins or complexes larger than 80 amino acids. Although many crystallographic three-dimensional structures of biological macromolecules can be determined, as exemplified by the continuous exponential growth in the number of PDB entries, many challenging projects cannot be pursued further. One reason is that crystals of large proteins and/or complexes diffract to low resolution. Indeed, even after semi-high-throughput sample screening, the crystals of currently studied projects diffract on average to about 4 Å resolution on modern synchrotron beamlines (Holton, 2005). Since the development of methods for solving macromolecular structures has largely been focused on high-resolution data, only a small fraction of currently measured X-ray data result in a structure that is deposited in the PDB.

Perhaps the most advanced algorithms for the interpretation of low-resolution maps have been developed for cryo-electron microscopy (EM) data. These include approaches for the automatic segmentation of EM images of macromolecules using proximity and grey-level similarity between pixels, in conjunction with an eigen decomposition (Frangakis & Hegerl, 2002), fuzzy logic principles (Garduno *et al.*, 2008) or the multiseeded fast marching method (Baker *et al.*, 2006). However, the extraction of accurate boundaries of structural constituents still remains a major challenge. For example, Baker and coworkers reported successfully segmented EM maps at 6.5 and 11.5 Å resolution, while for maps at 20 Å resolution or lower their method was able to identify the oligomeric subcomplexes but not individual protein subunits. Other methods require extensive user interaction (Garduno *et al.*, 2008) or the presence of internal symmetry (Yu & Bajaj,

2005) to be used as a constraint. Making structural models based on EM maps relies either on the identification of secondary-structure elements (Jiang *et al.*, 2001; Kong & Ma, 2003; Dror *et al.*, 2007; Baker *et al.*, 2007) for which a resolution higher than 15 Å is generally required or, if the subunits and their atomic structures are known, on sophisticated fitting methods that use, for example, local or cross-correlation, normal-mode analysis or molecular-dynamics assisted flexible fitting (Chiu *et al.*, 2005; Fabiola & Chapman, 2005; Chacon & Wriggers, 2002; Jolley *et al.*, 2007; Rossmann, 2000; Tama *et al.*, 2004; Velazquez-Muriel *et al.*, 2006; Velazquez-Muriel & Carazo, 2007; Wu *et al.*, 2003; Roseman, 2000; Trabuco *et al.*, 2008).

In cases where it is not possible to obtain an accurate structural model from a 20 Å map, a model constructed of fragments determined at high resolution can help to answer many biological questions. The intermolecular and intramolecular interactions can be proposed based upon such a model. Indeed, successful X-ray structure determination of large molecular machines (*e.g.* ribosomes and fatty acids) has been based on the use of these fragment-based models derived from low-resolution maps. Low-resolution maps can also be used successfully for molecular replacement (Navaza, 2008; Xiong, 2008).

Here, we introduce a novel approach to obtaining structural information from macromolecular X-ray data extending to a resolution of 20 Å. In essence, we address the problem of interpreting low-resolution data *via* the segmentation of a density map into a predefined number of core objects, so that each structural motif (domain) contained in the structure is represented by one such core. No detailed knowledge about the composition of the low-resolution complex is required. Segmentation is followed by a pattern-recognition-based identification of the structure of each core segment, in which it is slid through a database of shapes derived from the PDB and potential matches are identified. The best matched shapes and their PDB structures are superimposed on the corresponding map segments. The method is able to a certain extent to retrieve the boundaries of the structural motifs (*i.e.* the domains) and the matched structures provide a sufficient number of large building blocks to reconstruct a putative model of the whole low-resolution structure. The domains that are placed into the map can then be used for further phase extension.

2. Methods

2.1. Data

As a test example, we chose the structure of a bacterial genotoxin (PDB code 1sr4; Nesic *et al.*, 2004). This toxin causes cell-cycle arrest and subsequent cellular distension in epithelial cells and a rapid death by apoptosis in many
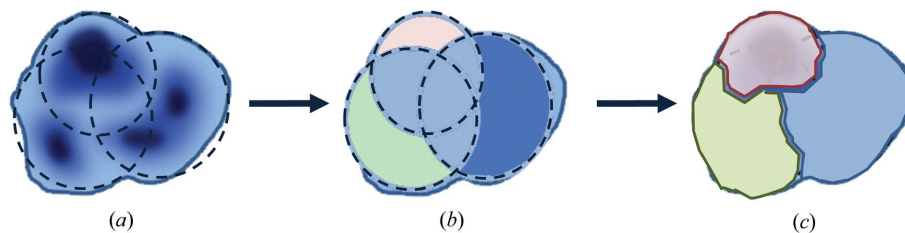
Figure 1
Schematic representation of the map-segmentation process. (*a*) Spheres are placed in the map (coloured blue; darker blue indicates higher density); (*b*) the areas belonging to only one sphere (light brown, green and bright blue) are used as initial building blocks for the segments; (*c*) all segments are generated.

lymphocytes. The molecule is arranged as a heterotrimer composed of three chains, *A*, *B* and *C*, of molecular masses 23, 29 and 18 kDa, respectively. The actual toxic part of the trimer is chain *B*, while chains *A* and *C* are required to deliver the *B* subunit into cells. This bacterial genotoxin structure has properties that make it particularly suitable as a test case: the structure is a heterotrimer of medium size and the constituent subunits are more-or-less globular single domains.

The bacterial genotoxin structure was placed into an artificial crystallographic cell with *P*1 space-group symmetry. The unit-cell parameters were set to yield a cell with an axis length equal to the longest side of the minimal rectangular bounding box around the molecule. A 50% margin was added in order to avoid interference with neighbouring molecules. The density map was computed on a cubic grid (1 Å spacing) from the atomic coordinates using a five-Gaussian approximation (*International Tables for X-ray Crystallography*, 1974). The structure-factor amplitudes and their associated phases were computed from this map using the *CCP*4 program *SFALL* (Collaborative Computational Project, Number 4, 1994). The resulting complete data were truncated to a maximum resolution of 20 Å. No *B*-factor correction was applied, since at 20 Å its effect on the overall amplitude falloff of the data is negligible. These data were used to compute electron-density maps on an orthogonal grid (2.5 Å spacing) using both the error-free phases and phases with a modest uniformly distributed error of $20°$ on average.

2.2. Map segmentation and core objects

We define the protein region as the map area with density values of 1σ or higher above the mean. All grid points that have a density of 2σ or higher above the mean are used as seed points for the map segmentation, in which small identical spheres are placed at each seed point. The radius of each sphere is then increased stepwise by one grid unit until the sphere contains a maximum of 1% of the points outside the protein region. Subsequently, the number of spheres is then reduced based on pairwise comparison of neighbouring spheres in which a distance between their centres is calculated; if this distance is smaller than a quarter of the sum of their radii then the smaller sphere is eliminated. When both spheres have exactly the same radius, the one with the lower average

density is removed. Finally, the N largest spheres are selected, where N is the expected number of domains (see Fig. 1a).

The areas of the protein region that belong to only one of the N spheres are used to build the core segments (Fig. 1b). The remaining areas of the protein region are processed using the core-tracing algorithm of Swanson (1994). Specifically, the core segments are extended by associating successively lower nearby density points. The extension is completed when segments join (Fig. 1c).

2.3. Pattern recognition of the content of the segments

The segments obtained are analysed using the set of 11 third-order moment invariants (Lo & Don, 1989) as spatial descriptors (a feature vector) following the procedure described in Hattne & Lamzin (2008). In addition to the 11 scale-invariant moments, we use the radius of gyration (R_g). This combined feature vector provides a compact description of the shape of an object and is straightforward to compute. For the calculation of these features not only the shape of the segments but also the varying electron density inside them is considered. A 20 Å density map at different contour levels is shown in Fig. 2.

The same spatial descriptors are calculated for the unique domains present in the PDB structures. This database has been described for the application of automated molecular replacement with *BALBES* (Long *et al.*, 2008) and contained 30 146 unique domains as of August 2008. For the evaluation and development of the method presented here, we randomly selected 5000 domains from the database. For each domain we computed a 20 Å density map in the same way as for the test set described above and derived the feature vector. For all domains the values of the shape descriptors were averaged and their standard deviation (σ) was stored. Outliers, which are defined as domains in which at least one descriptor deviates by more than three standard deviations from the average, were excluded, which accounted for 5% of the data. The standard deviations subsequently used in the calculation of the scores (see below) were then recomputed. Otherwise, the presence of the outliers would disturb the scoring procedure.

The shape descriptors for the segment in question are slid through those computed from the domain database. The score is obtained by calculating the squared deviation among all descriptors, weighted by the inverse of the squared standard deviation (1), where f_i and $f_{i,j}$ are the shape descriptors for the search segment and the jth domain from the database, respectively.
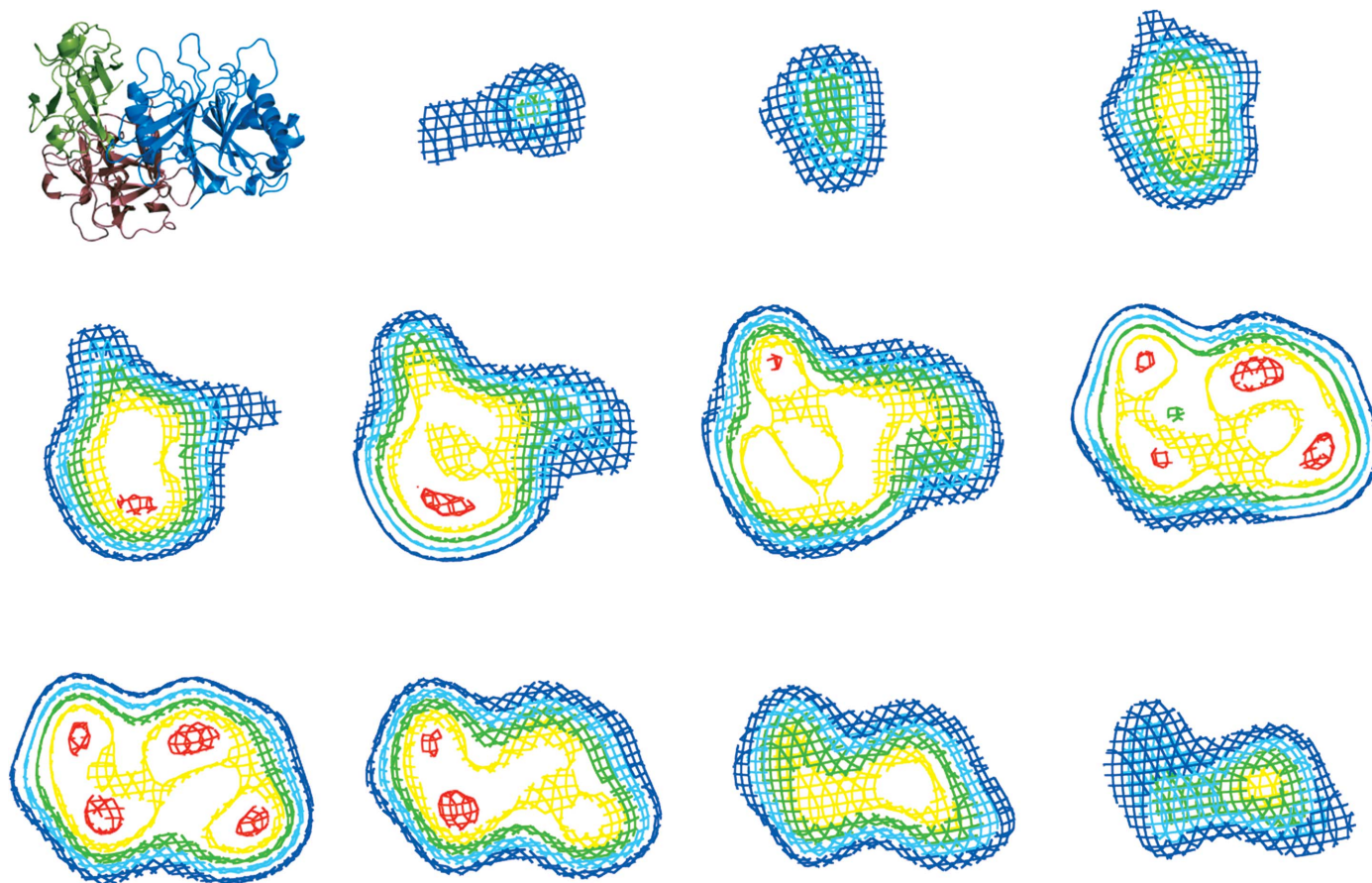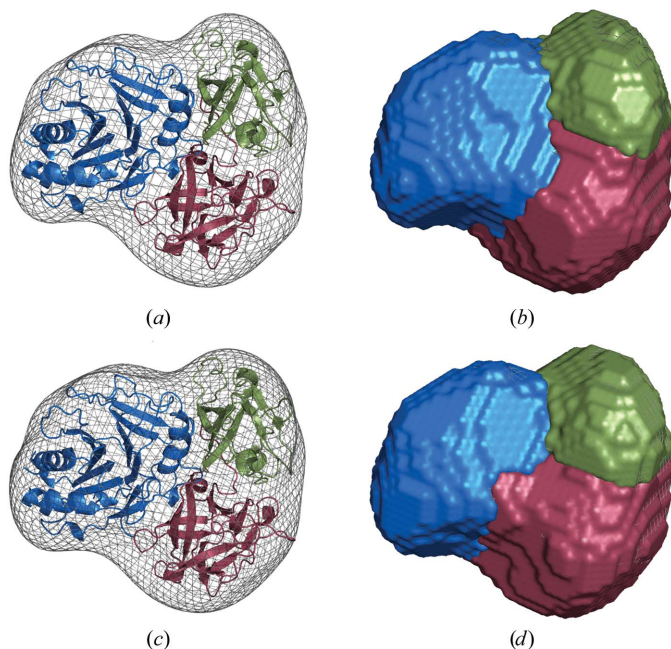

Figure 2
Slices through the bacterial genotoxin (PDB code 1sr4) 20 Å resolution density map showing different contour levels (blue, 1σ; turquoise, 2σ; green, 3σ; yellow, 4σ; red, 5σ).
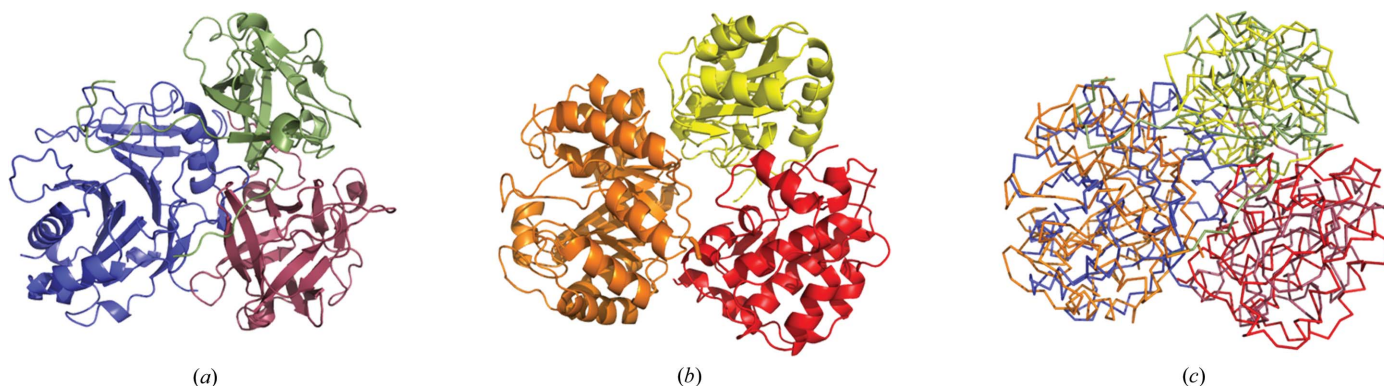
$$\text{Score}_j = \sum \frac{(f_i - f_{i,j})^2}{\sigma_i^2} \qquad (1)$$

Since the shape descriptors for the domain database are pre-calculated, the actual step of the recognition process for each segment is reasonably fast: about 1 s on a modern desktop computer.

Finally, the domains with the best (lowest) score are placed into the corresponding map segments using a relatively simple procedure. The centres of gravity for each map segment and the shape of the corresponding best-matched domain are aligned. Their orientation is deduced from the principal components of the 3×3 xyz variance–covariance matrices. Of the four possible superpositions (xyz, $\bar{x}\bar{y}z$, $\bar{x}y\bar{z}$ and $x\bar{y}\bar{z}$) the one that results in the highest overlap between the domain and the segment is applied to the domain structure. Per-

Table 1
Centres of gravity and radii of gyration for the three chains of 1sr4 and for the corresponding segments derived from the map at 20 Å resolution.

	Displacement of the centre of gravity (Å)		Radius of gyration (Å)		
	No phase error	20° phase error	Individual structures	Map segments, no phase error	Map segments, 20° phase error
Chain A	2.2	3.7	14.2	15.9	16.1
Chain B	3.8	4.4	16.5	17.3	16.9
Chain C	6.6	6.2	13.4	16.3	15.9
Average	4.2	4.8			

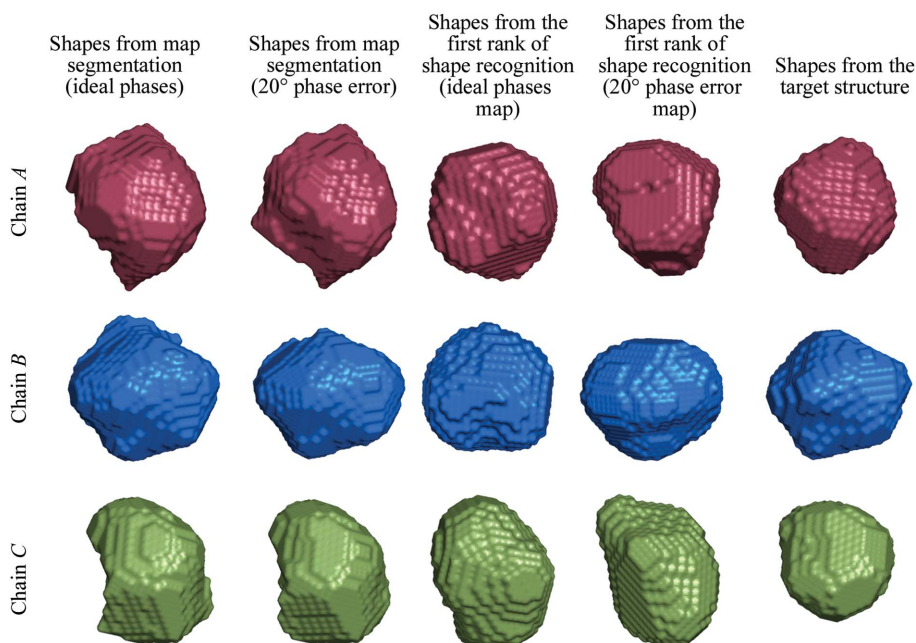forming this for all segments produces a 'structural' representation of the low-resolution density map.

For the comparative calculation of the phases from the placed domains the space group and the unit-cell parameters of the 1sr4 structure were used. Since map correlations can be formulated as an F^2-weighted average of the values of the cosine of the phase error (Lunin & Woolfson, 1993), they were computed in reciprocal space. Hereafter, we use the following notation: map correlation at a given resolution is the map correlation computed from the reflections within that narrow resolution shell, while overall map correlation at a given resolution is that computed using all reflections from infinity down to that resolution limit.

Figure 3
The segmentation of the 1sr4 map. (a) The true heterotrimeric structure and its density map calculated at 20 Å resolution; (b) results of the segmentation of the map into three core objects. (c) and (d) present the results for a map computed from the data with 20° phase error.

3. Results

The preliminary tests were very encouraging and indicate that a density map for a complex protein structure can be segmented reliably even with 20° phase error (Fig. 3). The matched domains from the database are placed into the map quite accurately; their centres of gravity are on average only about 4.2 Å from the centres of the domains in the 1sr4 structure (Table 1). However, these are not necessarily the correct domains in terms of their secondary and detailed atomic structure (see Fig. 4 and the text below). The displacement of the centres of the located domains is about 25% of the value of the segments' radii of gyration. The R_g values of the map-derived segments are about 10% higher than the R_g values of the individual domains from the database (Table 1), which we attribute to the properties of the

Figure 4
(a) 1sr4 in cartoon representation; (b) the fragments best matching the shape descriptors; (c) superposition of 1sr4 and the best matching fragments.

segmentation procedure and the density overlap caused by the presence of the other domains in the complex. For the map computed from the data with the 20° phase error, the centres of the placed domains match those in the 1sr4 structure with similar accuracy (4.8 Å). Also, the map-derived radii of gyration are very similar to the phase error-free case.

Figure 5
Comparison of the segments derived from the map with the shapes derived from the target structure and with the shapes derived from the first hit of the score-based ranking.
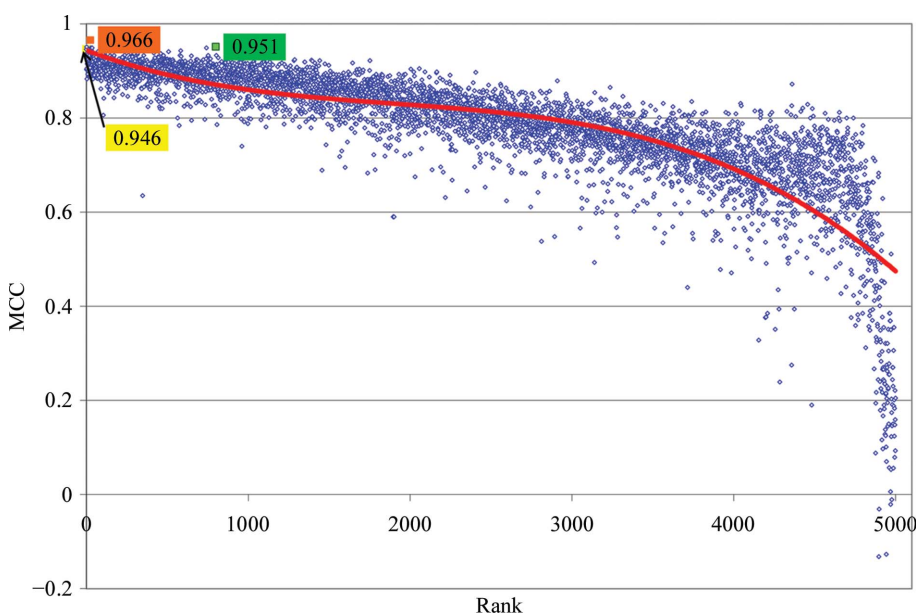
Figure 6
Each point on the graph represents one artifical complex structure composed of the three identified domains having the same rank in the scoring (A1–B1–C1, A2–B2–C2 etc.) using the error-free 1sr4 map. The overall map correlation to 20 Å resolution obtained by placing the target (true) domains in the three core segments is shown in green; this solution has an 'average' rank of about 800. The yellow box corresponds to the single solution with the best score, while the brown box represents the top 50 averaged domain structures. The red line shows the overall trend.

Use of the three-dimensional moment invariants and radii of gyration as shape descriptors works extremely well. The scan of the search targets against 5000 domains places the correct solution within the top 25% or even better. Chain A of 1sr4 can be found with rank 1150, chain B with rank 14 and chain C with rank 1238 (top 23%, 1% and 25%, respectively). The ranking for the map computed with the 20° phase error are 880 for chain A, 6 for chain B and 1038 for chain C (top 18%, 1% and 21%, respectively).

For real cases, one does not necessarily know the target domains and their conformations. We thus pick out the solutions whose shapes are best matched to the map segments. We see that although the selected solutions have quite different structures in comparison to the true targets (Fig. 4), their shapes at 20 Å resolution are very similar to the shapes of the map-derived core objects (Fig. 5).

Also, about the first 1000 solutions of the ranking show very similar scores (Fig. 6). Therefore, we employed averaging of the top 50 of them after the domains were superimposed on each map segment and calculated the map and the phases from the 'averaged' structure.

The placed domain structures can furthermore be used to calculate the phases and thus provide a means for phase extension beyond the 20 Å limit used for map interpretation. Clearly, a map computed at 20 Å resolution cannot provide useful phases beyond that limit. However, the phases computed from the superimposed domains with the highest score (even without any real-space rigid-body refinement) provide some useful phasing signal down to about 10–14 Å resolution. A map calculated with these phases has an overall map correlation to the map from the refined 1sr4 structure of 95% for data to 20 Å resolution and 79% for data to 10 Å resolution (Fig. 7a). For the case where the domains were matched to the map containing a 20° phase error these overall correlation coefficients are 92% and 73%, respectively (Fig. 7b).

Owing to the fact that the initial 20 Å resolution map computed from the final 1sr4 structure has a limited information content, the domains are placed with a certain error, at least in the currently implemented method. This results in the

fact that even the (re)placement of three correct target domains yields phases with rapidly decreasing quality. Indeed, these phases are comparable to those obtained from the best scoring single domains (Fig. 7*a*).

On the other hand, the phases computed from the averaged top 50 solutions (blue lines in Fig. 7) produce an overall map correlation of 96% to 20 Å resolution and 88% to 10 Å resolution, which is significantly better than either the phases derived from single top-ranked domains or the correct target domains.

4. Discussion and outlook

Rather than follow a simplistic approach of assigning the known domain structures to each segment of a low-resolution map, in this work we target a more general unbiased shape identification in terms of core template objects (domains). This considerably reduces the amount of structural knowledge needed in advance of the data interpretation. The use of a small set of spatial descriptors for the comparison of the deduced core segments to the shapes of the structures in the domain database plays an important role in the recognition process. Since the segmentation itself only gives a rough idea about the identity of each domain, the recognition procedure can be of special interest for validation purposes. This also makes the method capable of tolerating a certain amount of inter-domain and intra-domain flexibility which other approaches address explicitly using, for example, normal-mode analysis or domain family wide flexibility analysis (Tama *et al.*, 2004; Jolley *et al.*, 2007; Velazquez-Muriel *et al.*, 2005; Delarue, 2008).

The presented method is able to successfully identify the shapes of the domains forming a larger protein complex from a 20 Å resolution X-ray density map so that candidate structures from the PDB can be used for further placement, fitting and phase extension. According to the Rayleigh criterion

(Stenkamp & Jensen, 1984), in an analysis of a 20 Å resolution map two points can be seen to be separate if they are at least 14 Å apart from each other. Thus, the placement of the bacterial genotoxin segments with a 4 Å deviation in their centres is, perhaps, as good as it can be.

The concept of placing the fragments from a database of known structures in an unknown map to obtain more information is reminiscent of the approaches used by molecular-replacement pipelines. However, molecular-replacement methods are usually suitable for data extending to at least ~4 Å resolution and require far more knowledge in advance, for example the sequence.

The current implementation of the method uses a few assumptions that may not be valid for each real case, *e.g.* the *a priori* known number of domains constituting the structure. Additionally, there may be practical challenges in collecting all the low-angle reflections in an X-ray diffraction experiment and obtaining initial phases of sufficient quality. Nevertheless, the results obtained show that initial structural interpretation of a phased 20 Å resolution X-ray density map is realistic in cases where not much is known about the structure in advance. This might even be beneficial for the interpretation of the structural content of a cell as proposed, for example, by Baumeister & Steven (2000) and Muller *et al.* (2008). Furthermore, low-resolution data will become available from upcoming X-ray facilities with a free-electron laser (FEL) beam of coherent radiation and fine time structure. When diffraction patterns from such sources are obtained from biological samples before they turn into plasma in an FEL beam, the interpretation of (most likely) very low-resolution data will become a challenge.

An intriguing follow-up will be extensive tests and, most likely, tuning of the method for the use of EM images in cases where phased low-resolution X-ray data are not available. Interpretation of low-resolution maps in MX additionally requires the development of algorithms that use the structural
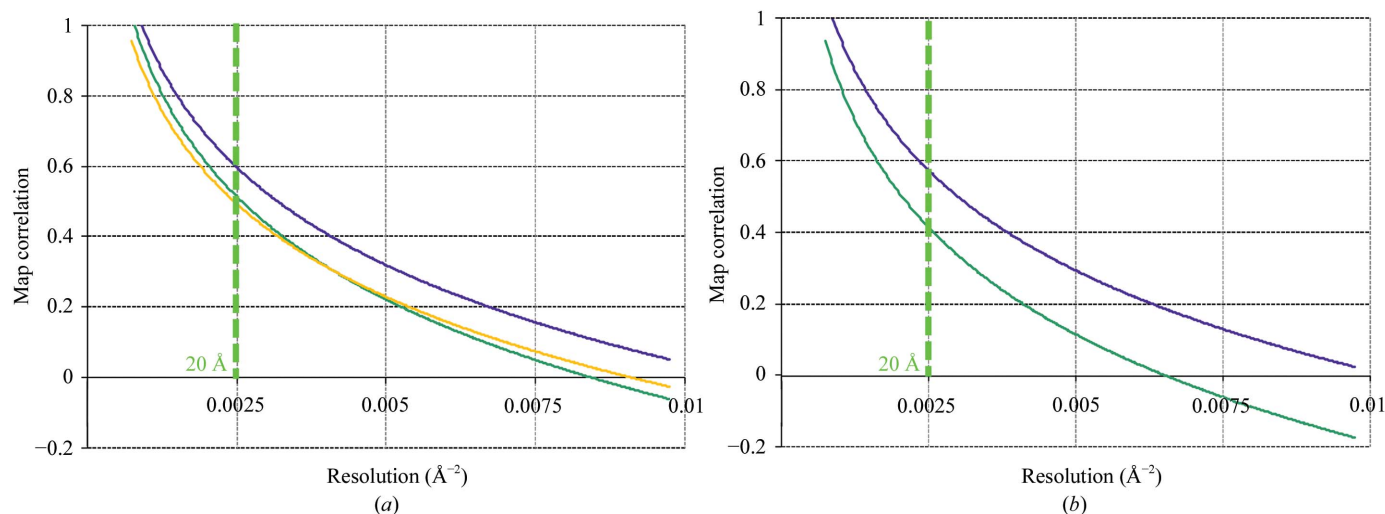
Figure 7
Map correlation at a given resolution computed from the placed domain fragments (green, using best scoring fragments; blue, using top 50 averaged domains; orange, using the target domains). (*a*) is for the map with no phase error and (*b*) is for the 20° phase error map. The high-resolution limit of 20 Å used for the computation of the initial density map is marked.

information from the interior of the molecule, since current EM analysis in essence provides only the envelope. In addition, the electron-density distribution in X-ray-derived maps differs from that in electron microscopy, which necessitates different pattern-recognition approaches and scoring functions. There are also method-specific challenges: the signal-to-noise ratio, data completeness (missing cone) and the three-dimensional reconstruction for EM and the phase problem and a general lack of correctly measured low-resolution reflections for MX.

The development of the presented method will benefit from an extension of the scoring function in order to improve the recognition process. For example, for a given domain in the database its interaction partners may be known and thus the interface regions of the domains can be derived. These regions should be compliant with those derived from the map segmentation. This information can either serve as a good validation criterion as to the correctness of the segmentation or can alternatively be used as additional information to post-refine the shapes of the identified segments during the segment-selection step or improve the ranking of the domain templates by means of likelihood.

After the selection of putative domains from the database a local real-space fit can be explored and this will certainly improve the capabilities of the method for phase extension. Plenty of good algorithms have been described for similar purposes (for reviews of some of them, see Cowtan, 2008; Roseman, 2000) and implementation of them or their variations would be a natural step to follow. Although identification of the corresponding structure is not yet performed unambiguously, it considerably limits the number of potential candidates which may be tried in the map and evaluated with more sophisticated methods.

Averaging of the identified suitable domains with shapes that give the best scores provided the best phasing information and proved very promising for phase extension. It is likely that performing the averaging after the real-space fit would provide even better results. Also, the best number of top-scored solutions to be averaged as well as their relative weighting in the averaging process remains to be investigated.

Another important task for future development is the implementation of an iterative procedure in which the recognized and placed models are used for map improvement and phase extension. Even if an incorrect domain structure but with a good shape match is placed in a 20 Å map during the first iteration, it may improve the phases for a subsequent density-modification step. This, in turn, may guide us to the correct domain structure in the next iterations.

References

Baker, M. L., Ju, T. & Chiu, W. (2007). *Structure*, **15**, 7–19.

Baker, M. L., Yu, Z., Chiu, W. & Bajaj, C. (2006). *J. Struct. Biol.* **156**, 432–441.

Baumeister, W. & Steven, A. C. (2000). *Trends Biochem. Sci.* **25**, 624–631.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Chacon, P. & Wriggers, W. (2002). *J. Mol. Biol.* **317**, 375–384.

Chiu, W., Baker, M. L., Jiang, W., Dougherty, M. & Schmid, M. F. (2005). *Structure*, **13**, 363–372.

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

Cowtan, K. (2008). *Acta Cryst.* D**64**, 83–89.

Delarue, M. (2008). *Acta Cryst.* D**64**, 40–48.

Dror, O., Lasker, K., Nussinov, R. & Wolfson, H. (2007). *Acta Cryst.* D**63**, 42–49.

Fabiola, F. & Chapman, M. S. (2005). *Structure*, **13**, 389–400.

Frangakis, A. S. & Hegerl, R. (2002). *J. Struct. Biol.* **138**, 105–113.

Garduno, E., Wong-Barnum, M., Volkmann, N. & Ellisman, M. H. (2008). *J. Struct. Biol.* **162**, 368–379.

Hattne, J. & Lamzin, V. S. (2008). *Acta Cryst.* D**64**, 834–842.

Holton, J. M. (2005). *Am. Crystallogr. Assoc. Ann. Meet.* Abstract W0308.

International Tables for Crystallography (2006). Vol. C, 1st online ed., Table 6.1.1.4, pp. 578–580. Chester: International Union of Crystallography. [doi:10.1107/97809553602060000600].

Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). *J. Mol. Biol.* **308**, 1033–1044.

Jolley, C. C., Wells, S. A., Fromme, P. & Thorpe, M. F. (2007). *Biophys. J.* **94**, 1613–1621.

Kong, Y. & Ma, J. (2003). *J. Mol. Biol.* **332**, 399–413.

Lo, C.-H. & Don, H.-S. (1989). *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 1053–1064.

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.

Lunin, V. Yu. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.

Muller, S. A., Aebi, U. & Engel, A. (2008). *J. Struct. Biol.* **163**, 235–245.

Navaza, J. (2008). *Acta Cryst.* D**64**, 70–75.

Nesic, D., Hsu, Y. & Stebbins, C. E. (2004). *Nature (London)*, **429**, 429–433.

Roseman, A. M. (2000). *Acta Cryst.* D**56**, 1332–1340.

Rossmann, M. G. (2000). *Acta Cryst.* D**56**, 1341–1349.

Stenkamp, R. E. & Jensen, L. H. (1984). *Acta Cryst.* A**40**, 251–254.

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

Tama, F., Miyashita, O. & Brooks, C. L. III (2004). *J. Struct. Biol.* **147**, 315–326.

Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. (2008). *Structure*, **16**, 673–683.

Velazquez-Muriel, J. A. & Carazo, J. M. (2007). *J. Struct. Biol.* **158**, 165–181.

Velazquez-Muriel, J. A., Sorzano, C. O., Scheres, S. H. & Carazo, J. M. (2005). *J. Mol. Biol.* **345**, 759–771.

Velazquez-Muriel, J. A., Valle, M., Santamaria-Pang, A., Kakadiaris, I. A. & Carazo, J. M. (2006). *Structure*, **14**, 1115–1126.

Wu, X., Milne, J. L., Borgnia, M. J., Rostapshov, A. V., Subramaniam, S. & Brooks, B. R. (2003). *J. Struct. Biol.* **141**, 63–76.

Xiong, Y. (2008). *Acta Cryst.* D**64**, 76–82.

Yu, Z. & Bajaj, C. (2005). *IEEE Trans. Image Process.* **14**, 1324–1337.