Supplementary data

1 Methods

1.1 Incremental distance rank profile

We employ vector quantization to convert the atomic coordinates of 3D maps to a set of representative 3D points. The software package Situs is used for the conversion, and the programs qvol and qpdb are used for the 3D maps and the atomic coordinate data, respectively (Wriggers et al., 1998). The numbers of 3D points are set to 30 and 50, regardless of the resolution, size, and data type.

For a rapid comparison, we employ a rotation and translation invariant profile, which does not require any superimpositions of 3D shapes. We generate a distance rank (DR) profile, which is a one-dimensional array of the distance values of all of the 3D point pairs sorted in ascending order (Figure S1a). The DR profile expresses the overall shape of the structure data, as well as the size. For example, elongated thin structures have rather closer pairs than spherical structures. Repetitive molecules have multiple pairs with similar distances, reflecting their repeating units. Note that the DR profile looks similar to the pair-distance distribution function (PDF) (Serdyuk et al., 2007) and the D2 shape function (Osada et al., 2002), because they are also generated from pairwise distances between 3D points in the shape. However, they are clearly different from our DR profile. The vertical axis of the PDF or D2 shape function is the frequency, and the horizontal axis is the distance. However, the vertical axis of the DR profile is the distance, and its horizontal axis is the rank (fig. S1).

Structures with similar sizes tend to have similar DR profiles, even if their detailed shapes are different. This occurs because they all are monotonically increasing profiles, and the distance value is related to the size of the structure data. To overcome this problem, we introduce an incremental transformation to the DR profile, which calculates the sum of the increment value within a window size in the DR (Figure S1b). This modified profile is called the incremental distance rank (iDR) profile. The iDR profile is defined as follows:

$$P_r = \sum_{w=0}^{\min[W, L-r]} (D_{r+w} - D_r)$$

(1)
where $P_r$ is the $r$-th component of the iDR profile, $D_{r+w}$ is the $(r+w)$-th component of the DR profile, $L$ is the length of the DR profile and $W$ is the window size.

The iDR profile is more sensitive than the original DR profile. The window size is 30% of the number of 3D point pairs. For the case of the 50 point set, the total number of 3D point pairs is 1,225; therefore, the window size for the incremental transformation is $1,225 \times 0.3 = 368$.

Figure S1: (a) Schematic diagram of the generation of the DR profile. By vector quantization, a model (blue object) is converted into set of points (orange dots). While the number of points is four in this schematic diagram, 30 and 50 dots are generated in the actual system. The DR profile (red curve) is the array of distances (thick black lines) of all pairs of points sorted by length, from the shortest to the longest. (b) Example of the iDR profile of real data, EMDB 1003. The 3D map data with 50 points generated by vector quantization (left), the DR profile (right top), and the iDR profile (right bottom) for the model are shown. The iDR profile is the trace of the sum of incremental values (blue and red areas) of the DR profile.

For each structure, four types of profiles, $P_{30}$, $P_{50}$, $P_{o25}$ and $P_{PCA}$, are prepared. The $P_{30}$,
$P_{50}$, and $P_{0.25}$ profiles are iDR profiles, and the $P_{PCA}$ profile is derived by principal component analysis (PCA) of the 50 point set, and has three values, the standard deviations along the first, second, and third principle axes. These values are related to the approximate size of the 3D model along the three axes. The iDR profiles $P_{30}$ and $P_{50}$ are generated by 30 and 50 points, respectively. The iDR profile $P_{0.25}$ is generated by the outermost 25 points, which are the 25 furthest points from the center of the molecule, among the 50 points. This is used in order to decrease the effects of the high-pass filtering and the correction of the contrast transfer function applied during the EM image analyses, as they affect the density ratio between the inner and outer parts.

The lengths of the iDR profiles are 435, 1225, and 300 for $P_{30}$, $P_{50}$, and $P_{0.25}$, respectively. For faster computation, the profile lengths are reduced by an averaging operation. By this operation, every $m$ ranks of the profiles are averaged and merged into one rank. We prepare two sets of reduced profiles: short and long profiles. For the short profiles, the value $m$ is set to 50 for $P_{30}$, and to 8 for the others. For the long profiles, the value $m$ is set to 24 for $P_{30}$, and to 4 for the others.

1.2 Shape using the iDR profile

The similarity $S_k(i,j)$ between two profiles $P_k(i)$ and $P_k(j)$ is determined as follows,

$$S_k(i,j) = 1 - \frac{(P_k(i) - P_k(j))^2}{(P_k(i) + P_k(j))^2}$$

(2)

where $P_k(i)$ and $P_k(j)$ are the profiles of data $i$ and $j$, respectively. $K$ is the type of profile, which is 30, 50, 0.25, or $PCA$. This similarity $S_k(i,j)$ is related to the Sorensen-Dice coefficient and the Hodgkin index. The merging shape similarity $S_{RMS}$ is the root mean squared value of four profiles, $S_{30}, S_{50}, S_{0.25}$ and $S_{PCA}$:

$$S_{RMS}(i,j) = \sqrt{\frac{S_{30}^2(i,j) + S_{50}^2(i,j) + S_{0.25}^2(i,j) + S_{PCA}^2(i,j)}{4}}$$

(3)

Finally, the score $S_{RMS}$ is adjusted to modify its range from -1 to 1:

$$S(i,j) = 2[S_{RMS}(i,j)]^p - 1$$

(4)

where $\alpha$ is 2.17938653227261, which is determined from the mean of $S_{RMS}$ values of the comparison of 90,000 randomly chosen pairs (Figure S2). This function gives 1 for identical data, and 0 for the value expected for a randomly chosen pair. The $p$ values for scores of 0.9, 0.8 and 0.7 are 0.0044%, 0.6878% and 4.4089%, respectively.
**Figure S2.** Cumulative histogram of the score values (S) of 90,000 randomly chosen pairs.

1.3 Similarity search
To decrease the computational cost for more than 200,000 comparisons, the search is performed in three stages. In the first stage, only the PCA profiles are used to select the structures with roughly similar sizes. The structures in the dataset are excluded, if one of its three PCA standard deviations is either larger than 200% or smaller than 50% of the corresponding value of the query. In the second stage, the short profiles are used to exclude dissimilar structures with rapid computation. The second stage chooses the 3,000 best similar data with similarity scores greater than 0.4, which corresponds to the top 30% of the random distribution as shown in Figure S2. At the last stage, the similarities using the long profiles are calculated for the data chosen in the second stage, and the top 2,000 data are shown. The profiles of the data in the dataset are prepared beforehand. When the data in the dataset are used for the input, the computation for only these three stages is required, and it usually takes less than one minute. For the case of uploaded data, several extra minutes are required for uploading data, vector quantization, and profile generation.

1.4 Pairwise superimposition using the *gmfit* program
Pairwise superimposition is performed by the *gmfit* program for two Gaussian mixture models (GMMs). Its strategy is similar to that described in our previous paper (Kawabata, 2008), in
terms of the generation of initial configurations and the succeeding steepest descent local searches.

For a “many-to-one” superimposition, we employed a random position generator from the GMM density map for the initial configuration. For a pairwise superposition, we employed the principal component axes alignment method, which is popular among researchers (Lohmann, 1998). In this method, three principal component axes are calculated for the two assemblies, as shown in Figure S3a: \( \{ \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \} \) and \( \{ v_1, v_2, v_3 \} \). The indices of the axes are defined by the order of their eigen values. The eigen value of \( \mathbf{u}_1 \) is the largest among the three, and that of \( \mathbf{u}_3 \) is the smallest. The reference GMM is fixed, and the target GMM is translated to align their centers, and then rotated to align the three corresponding principal axes. If the axis matches are \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), then the rotation matrix is \( \mathbf{R}=\mathbf{V}\mathbf{U}^T \), where \( \mathbf{U}=\{ \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \} \) and \( \mathbf{V}=\{ v_1, v_2, v_3 \} \) (Figure S3b). As the directions of the principal axes are arbitrarily decided, four combinations of the signs of the axes must be tried for each matching of the axes, as shown in Figure S3c.

The matching of the axes by the order of their eigen values works well for most of the assemblies. However, we found that some symmetrical assemblies generate similar eigen values. For these cases, the subtle differences between these eigen values are not useful to determine the matching of the axes. An example is the cylinder-shaped chaperonin assemblies: EMDB 1397 and EMDB 5140. Finally, we decided to try all of the combinations \((3!)=6\) of the three axes while ignoring their eigen values: \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \), \( \{ \mathbf{u}_1; v_1, \mathbf{u}_2; v_2, \mathbf{u}_3; v_3 \} \). Considering the arbitrary nature of the signs (Figure S8c), we finally used \( 24 \) \((=3! \times 4)\) initial configurations.

For each initial configuration, the steepest descent method is performed to minimize the fitness energy \( E_{\beta s} \), defined in our previous paper. Finally, we chose the configuration with the lowest energy, among the 24 minimized configurations. The computation time for the pairwise superimposition of the two GMMs with the 20 Gaussian distribution function is less than one second.

Similarities between two GMMs are evaluated using the correlation coefficient \( CC \) of the two distribution functions of GMM over all space:

\[
CC = \frac{\int_{-\infty}^{\infty} f_a(r)f_b(r)dr}{\sqrt{\int_{-\infty}^{\infty} f_a^2(r)dr\int_{-\infty}^{\infty} f_b^2(r)dr}} \tag{5}
\]

where \( f_a(r) \) is the distribution function of one GMM, \( f_b(r) \) is that of another GMM. These integrals are analytically obtained as shown in our previous paper.
Figure S3. Outline of the principal axes alignment. (a) Definition of the eigenvector sets $U=\{u_1, u_2, u_3\}$ for the target assembly, and $V=\{v_1, v_2, v_3\}$ for the reference assembly. (b) Superimposition procedure. (c) Four combinations of the signs of the axes.
2 Case studies
The characteristic points of the Omokage search are as follows. 1) It is a search based on the global shape similarities of the structure data, and does not consider the similarity of the detailed atomic structures or the atomic compositions of the component molecules. 2) Atomic models and density maps can be used for the search query and the dataset. Omokage search is the first web-based tool that can perform a structure-based search using both PDB and EMDB data. 3) It is rapid and easy to use. Computation times range from tens of seconds to several minutes. The user can visually check the similarities and differences between the query and the found structures in a superimposed view with the interactive structure viewer, Jmol, embedded in the webpage. Based on these points, we introduce three types of case studies.

2.1 Structural comparison with different resolutions in both the EMDB and PDB
The first example is a search using the RNA polymerase (RNAP) structures from both the PDB and EMDB. For a query of the 25 Å resolution 3D map data of RNAP (EMDB 2190), more than 100 RNAP structures were found in both databanks, independent of their resolutions. For example, the 25 Å resolution 3D map (EMDB 2191; Score 0.8984) and the 3.5 Å resolution atomic model (PDB 4bbr; Score 0.8598) were found as the first and second ranks, respectively (Figure S4a). The search with the atomic model, PDB 4bbr, yielded a similar result. Pairwise superimposition by gmfit also supported the shape similarities of both types of data, density maps and atomic models (Figure S4b). The correlation coefficient values to EMDB 2190 were 0.9777 for EMDB 2991 and 0.883 for PDB 4bbr. Thus, Omokage search is borderless with regard to the data types, and can serve as a bridge between both databanks.
Figure S4. (a) The results of the search using the 20 Å resolution density map, EMDB 2190. Structure data ranked from 1 to 5 are shown. The first and third data, indicated by arrows, are the density maps data at 25 Å resolution from the EMDB (2191 and 2192), and the others are atomic models at 4.3 to 3.4 Å resolutions from the PDB (4bbr, 3k1f, and 2r92). (b) Superimposition of the maps and the models fitted by gmfit. Map vs. map (left, EMDB 1883 in magenta vs. EMDB 2190 in green), the model vs. the map (center, PDB 4bbr in magenta vs. EMDB 2190 in green), and the model vs. the model (right, PDB 3k1f in green vs. 4bbr in magenta). All images were captured from Omokage search and gmfit pages.

2.2 Different sequence orders or component arrangements
Using tools based on sequence similarities, such as BLAST, it is not easy to find similar structure data with different sequence orders or component arrangements. An example of circular permutation is shown by the results of a search using the lectin monomer, biological unit (BU) 2 of PDB 1rin, which yielded another circularly permuted lectin, PDB 3ax4, as the 16th similar shape data with the score of 0.8856. The gmfit results also clearly showed the similarity between them. Although the positions of the termini
are different due to their circular permutation, the entire parts of both structures superimposed very well, with a correlation coefficient of 0.897 (Figure S5).

**Figure S5.** Circular permutation found by the *Omokage search.* (a) Schematic diagram of the sequence homology of the lectins, based on the alignment by *MICAN* (Minami *et al.*, 2013), with the sequence identity of 43.2% between the N-terminal half (1-108) of 1rin-2 and the C-terminal half (123-237) of 3ax4, and 40.5% between the C-terminal half (112-234) of 1rin-2 and the N-terminal half (1-116) of 3ax4. The N and C terminal parts are colored blue and red, respectively. (b-d) Superimposition of the lectin structures, BU 2 of PDB 1rin (b), PDB 3ax4 (c), and superimposition (d), viewed from the same orientation. The colors in (b) and (c) correspond to those in (a). In (d), the 1rin-2 and 3ax4 structures are colored green and magenta, respectively.

As the different component arrangement we show the DNA clamps, which are processivity-promoting factors in DNA replication, encircling the DNA strand and acting as a "moving platform" for DNA polymerase and other proteins. They are found in the three domains of life (bacteria, archaea and eukarya), and all of them form similar ring structures to perform their common “platform” function. However, the number of components is different, despite the similar secondary structure arrangements. In bacteria, two DNA polymerase beta subunits form the clamp ring. In contrast, in archaea and eukaryotic nuclei, three proliferating cell nuclear antigen
(PCNA) molecules are assembled into a similar ring (Krishna et al., 1994). A search from the archaeal PCNA clamp in the trimer ring form (PDB 3IFV) yielded more than 100 clamps, including bacterial beta clamps in the dimer form, such as PDB 4k3l (Omokage score, 0.9019; gmfit correlation coefficient, 0.885). The shape similarity of both types can be seen in the gmfit results (Figure S6).

**Figure S6.** Superimposition of DNA clamps. (a, b) The trimer clamp (a, PDB 3ifv) and the dimer clamp (b, PDB 4k3l). The models are colored by the chains, and the arrows indicate the subunit boundaries. (c) Superimposition of both clamps, colored green and magenta, respectively.

### 2.3 Molecular mimicry

The third case is the shape similarity between RNA and protein. A famous example is the structural similarity between the complex of transfer RNA with elongation factor Tu (tRNA-EF-Tu complex) and the elongation factor G (EF-G). The former is an RNA-protein complex, while the other is a monomeric protein. Both factors bind to the A-site of the translating ribosome in similar manners. Their shape similarity is called "molecular mimicry" (Nissen et al., 1995 and Agrawal et al. 1998). The search using the tRNA-EF-Tu complex structure (PDB 1ob2) as a query yielded some EF-G structures, such as PDB 1efg (score, 0.8674) and PDB 1elo (score, 0.8588). The EF-G protein structure successfully fit not only with the protein part but also with the RNA part in the tRNA-EF-Tu complex structures (Figure S7), with a correlation coefficient of 0.694 for PDB 1elo and 0.684 for PDB 1efg.
Figure S7. Superimposition of elongation factors. The EF-Tu-tRNA complex (a, PDB 1ob2) and EF-G (b, PDB 1efg) are colored by their chains. The arrow indicates the RNA chain. (c) Superimposition of both factors, colored green and magenta, respectively.
3 Evaluation of searching performances of Omokage score, EM-SURFER distance, and gmfit score

3.1 Datasets and evaluation methods

It is important to compare the searching performance of our comparison method with the distances of 3D-Zernike descriptors, employed by the EM-SURFER server (Esquivel-Rodriguez et al., 2015). In order to evaluate the performances to discriminate similar structures, the “correct” standard datasets for similar maps must be prepared. This is not easy, because there is no classification database for EM maps, such as SCOP and CATH for protein atomic structures.

For this study, we prepared two datasets. For the first dataset “ClpB-ClpP”, we focused on nine maps (from EMDB 2555 to 2563), following Esquivel-Rodriguez et al. These nine maps were determined for mutants of ClpB and ClpP under different conditions, and were reported in the same paper (Carroni et al., 2014). We suppose that these nine maps have similar shapes and no other maps in EMDB have similar shapes to them. EMDB 2556 was chosen for the query.

The second dataset “70S ribosome” was generated by our manual checks. We chose EMDB 1003 (70S ribosome) for the query, and picked up all of the entries for prokaryotic 70S ribosome. We first downloaded all 160 EMDB entries classified as “ribosomes -70S & others” from the Gallery page of EM navigator. Among these 160 entries, we deleted the entries for organelle ribosome (6329, 2876, 2877, 2878, 2914, 2827), ribosomes with other molecules (1750, 5174, 1582, 2826), irregularly shaped ribosomes (1068, 1070, 1071, 1072) and entries with incorrect pixel spacing (1143, 1172, 1173, 1391, 2170, 2172, 2976). Two entries (6315 and 6316) were deleted because EM-SURFER server did not provide their descriptors. Finally, 137 entries were chosen for the 70S ribosome dataset. These 137 EMDB-IDs are summarized in Table S1.

The 3D-Zernickke descriptors for all of the EMDB entries were downloaded from the EM-SURFER web server. The Euclid distances among them were calculated by our in-house python script. The descriptors are based on the author-recommended density level. The volume filter and the resolution filter were not used. The Omokage score and the gmfit score were calculated using our WEB server. The score using the long profiles defined in Equation 4 was employed as the Omokage score. The gmfit score was obtained from the correlation coefficient between two GMMs, after optimal superimposition.

For the database search, we used 3,047 entries registered in the EMDB, downloaded on July 29, 2015. EM-SURFER failed to generate descriptors for 346 EMDB entries, and gmfit server also failed to generate GMMs for 29 EMDB entries. The number of
entries handled by all three servers was 2,690.

3.2 Correlation Analyses

For the preliminary analysis, we checked the correlations among these three scores (Figure S8). The Omokage score and the EM-SURFER distance did not correlate well, as their correlation coefficient was -0.342, as shown in Figure S8a. The correlation between the gmfit score and the EM-SURFER distance was similarly low (-0.376). However, the Omokage score and the gmfit score showed high correlation, as their correlation coefficient was 0.820. Considering the algorithmic difference between the iDR profiles of Omokage and the gmfit superimposition, we think their high correlation is quite surprising. In contrast, we suggest that the EM-SURFER score may be influenced by a factor that does not affect both Omokage and gmfit. We will discuss this unique factor of EM-SURFER later.

3.3 Evaluation using the ClpB-ClpP dataset

Figure S9 shows the accumulation curves for the ClpB-ClpP dataset. The scores and ranks for the three scores are summarized in Table S2. The AUC values of the accumulation curves indicate that the discrimination powers of these three scores are not very different; however, the EM-SURFER score is slightly inferior to the other two scores. Figure S10 shows examples of 3D maps in the ClpB-ClpP dataset, and the map of EMDB 2560 is shown in Figure S10b. All three methods judged that this is similar to EMDB 2556, although their details are different. The map of EMDB 2561, which looks larger than that of EMDB 2556, is shown in Figure S10c. Its query volume ratio is 1.381, although its PCA volume ratio is 0.921. Its ranks for Omokage and gmfit were 4 and 3, respectively; however, the rank for EM-SURFER was 368. Thus, EM-SURFER tended to fail to recognize a similar map with a larger volume. This tendency will be observed more clearly in the 70S ribosome dataset.

3.4 Evaluation using the 70S ribosome dataset

Figure S11 shows the accumulation curves for the 70S ribosome dataset. The scores and ranks for the best 30 maps for the Omokage score are summarized in Table S3. The curves in Figure S11 indicate that the discrimination by the EM-SURFER distance was much worse than those by Omokage and the gmfit score. When we assessed the failure of EM-SURFER shown in Table S3, we found that EM-SURFER tends to return large distances for maps with different volumes from the query’s volume. Figure S12b is an example for the large volume (EMDB 5362 with the volume ratio 1.907), and Figure S12c is that for the small volume (EMDB 5030 with the volume ratio 0.487). The ranks
of EM\textunderscore SURFER for these two maps are large (1015 and 1099), whereas those of Omokage (15 and 20) and gmfit (33 and 52) are small. The reason why EM\textunderscore SURFER showed poor performance for the 70S ribosome dataset may be that the 70S ribosome dataset has various maps with different volumes, and that EM\textunderscore SURFER is more sensitive to volume differences than Omokage and gmfit. The volume ratio is affected by many factors, such as resolution, density level for contour surface, and filtering operations. Both the Omokage and gmfit scores were calculated for low-resolution models. The Omokage score is based on vector quantization of 30\textasciitilde50 points, while gmfit used 20 Gaussian distribution functions. Therefore, the Omokage and gmfit scores may not be affected much by resolution differences. We concluded that our Omokage server is better than EM\textunderscore SURFER to detect biological similarities among various density maps with different resolutions and volumes.

Interestingly, the ratios of the PCA volumes shown in Table S3 do not vary less than the ratios of the volumes. The PCA volume is the volume of the ellipsoid obtained from the PCA analysis, which is defined as $V_{\text{pca}} = \frac{4}{3} \pi \text{sqrt}(v_1v_2v_3)$, where $v_1$, $v_2$, $v_3$ are the variances for the first, second and third principal axes, respectively. The ratios of the volumes of EMDB 5362 and EMDB 5030 were 1.907 and 0.487, whereas those of the PCA volumes were 1.143 and 1.069. This means that the PCA variances are more robust than the volumes, and would serve well for descriptors. That is why we employed the $S_{\text{pca}}$ score for the first stage of our three-stage search.

### 3.5 Computational time

From the viewpoint of computational time, the gmfit program should be much slower than the Omokage search and the EM\textunderscore SURFER server. It takes 0.1 to 1.0 second to superimpose one pair of GMMs. Therefore, gmfit is not suitable for searching a large database, but it has the ability to superimpose 3D structures, which cannot be done by Omokage and EM\textunderscore SURFER. The computational speeds of the Omokage and EM\textunderscore SURFER algorithms cannot be directly compared through their WEB servers. However, both methods employ 1D-vectors to describe 3D structural features, and thus we can expect their computational costs to be proportional to the length of their 1D vectors for descriptors. The length of the vector for EM\textunderscore SURFER is 121, that of the Omokage short profile is 153, and that of the Omokage long profile is 303. This means that Omokage with the long profiles may be about three times slower than EM\textunderscore SURFER, whereas Omokage with the short profiles may be as fast as EM\textunderscore SURFER. As we explained, the Omokage WEB server employs a three-stage search, and we expect
that the computation cost for our *Omokage* three-stage search will be similar to that for the *EM-SURFER* search.

![Figure S8](image-url)

**Figure S8.** Plots for the three scores for searches, using EMDB 1003 as the query. The red points are the maps in the 70S ribosome dataset. (a) *Omokage* score vs. *EM-SURFER* distance. (b) *gmfit* score vs. *EM-SURFER* distance. (c) *Omokage* score vs. *gmfit* score.
**Figure S9.** Accumulation curves for the ClpP-ClpB dataset, using EMDB 2556 as the query. The red, green, and blue lines represent the *Omokage* score, the *EM-SURFER* distance, and the *gmfit* score, respectively.
**Figure S10.** 3D density maps in the ClpP-ClpB dataset. All of them were superimposed onto the EMDB 2556 map using the *gmfit* program. Contour surfaces were drawn by the author-recommended density level, using *UCSF Chimera*. (a) EMDB 2556. This is the query. (b) EMDB 2560. The volume ratio is 0.947. (c) EMDB 2561. The volume ratio is 1.381.
**Figure S11.** Accumulation curves for the 70S ribosome dataset, using EMDB 1003 as the query. The red, green, and blue lines represent the *Omokage* score, the EM-SURFER distance, and the *gmfit* score, respectively.
Figure S12. 3D density maps in the 70S ribosome dataset and the corresponding atomic model. All of them were superimposed onto the EMDB 1003 map using the gmfit program. Contour surfaces were drawn by the author-recommended density level, using UCSF Chimera. (a) EMDB 1003. This is the query map. (b) BU 1 of PDB 4v5d. The 23S rRNA is shown in blue, and the 16S rRNA is shown in green. (c) EMDB 5362. The volume ratio is 1.907. (d) EMDB 5030. The volume ratio is 0.487.
Table S1. Two correct standard datasets for similar maps.

| Dataset       | EMDB-ID                                        |
|---------------|-----------------------------------------------|
| ClpB-ClpP     | 2555 2556 2557 2558 2559 2560 2561 2562 2563 |
| 70S ribosome  | 1003 1004 1005 1006 1007 1008 1045 1055 1056 1064 |
|               | 1065 1077 1110 1122 1128 1184 1185 1248 1250 1251 |
|               | 1261 1262 1263 1302 1310 1311 1312 1315 1323 1324 |
|               | 1362 1363 1365 1366 1370 1395 1417 1484 1499 1524 |
|               | 1540 1541 1551 1553 1554 1564 1565 1615 1616 1657 |
|               | 1716 1717 1718 1719 1720 1721 1722 1723 1724 1725 |
|               | 1726 1728 1762 1798 1799 1829 1830 1849 1850 1854 |
|               | 1858 1915 1917 1920 1921 2009 2183 2277 2316 2373 |
|               | 2446 2472 2473 2474 2475 2564 2565 2695 2696 2705 |
|               | 2773 2847 2917 2977 2978 5030 5036 5041 5113 5125 |
|               | 5126 5141 5188 5189 5234 5262 5265 5266 5267 5307 |
|               | 5359 5360 5361 5362 5363 5364 5386 5562 5691 5692 |
|               | 5693 5771 5775 5784 5785 5786 5796 5797 5798 5799 |
|               | 5800 5841 5842 5843 6211 6306 6311            |

20
Table S2. Scores and ranks for the ClpP-ClpB dataset, using EMDB 2556 as the query.

| ID    | Score (distance) | Rank | Ratio of volume | Resolution (Å) |
|-------|------------------|------|----------------|----------------|
|       | OM \(^a\) | ES \(^b\) | GM \(^c\) | OM | ES | GM | voxel \(^d\) | pca \(^e\) |
| 2556  | 1.000           | 0.0  | 1.000          | 1  | 1  | 1  | 1.000         | 1.000         |
| 2560  | 0.805           | 7.6  | 0.840          | 2  | 4  | 6  | 0.947         | 0.993         |
| 2555  | 0.763           | 7.2  | 0.886          | 3  | 3  | 2  | 1.194         | 1.002         |
| 2561  | 0.759           | 17.8 | 0.876          | 4  | 368| 3  | 1.381         | 0.921         |
| 2558  | 0.735           | 7.1  | 0.826          | 14 | 2  | 7  | 1.231         | 1.294         |
| 2557  | 0.704           | 16.3 | 0.844          | 41 | 220| 5  | 1.480         | 1.189         |
| 2559  | 0.688           | 12.3 | 0.856          | 50 | 34 | 4  | 1.034         | 0.854         |
| 2562  | 0.530           | 12.8 | 0.494          | 309| 42 | 529| 0.487         | 0.495         |
| 2563  | 0.214           | 22.9 | 0.465          | 928| 956| 705| 0.512         | 0.432         |

\(^a\) “OM” stands for “Omokage search”.

\(^b\) “ES” stands for “EM-SURFER”.

\(^c\) “GM” stanrds for “gmfit”.

\(^d\) “voxel” means that the volume is evaluated number of voxels over the given contour level.

\(^e\) “pca” means that the volume is evaluated by that of the PCA ellipsoid. The volume of PCA ellipsoid is defined as $V_{pca} = \frac{4}{3} \pi \sqrt{v_1 v_2 v_3}$, where $v_1, v_2, v_3$ are variances for the first, second and third principal axes, respectively.
Table S3. Scores and ranks for the 70S ribosome dataset, using EMDB 1003 as the query. The best 30 maps for the Omokage score among the 137 maps are shown.

| ID   | Score (distance) | Rank | Ratio of volume | Resolution (Å) |
|------|------------------|------|-----------------|----------------|
|      | OM   | ES | GM | OM | ES | GM | voxel | pca |                |
| 1003 | 1.000 | 0.0 | 1.000 | 1 | 1 | 1 | 1.000 | 1.000 | 11.5          |
| 5843 | 0.904 | 10.4 | 0.945 | 2 | 137 | 25 | 1.251 | 1.079 | 7.7           |
| 5785 | 0.904 | 3.3 | 0.952 | 3 | 11 | 10 | 1.004 | 1.065 | 9.1           |
| 5784 | 0.902 | 3.9 | 0.950 | 4 | 13 | 15 | 0.976 | 1.058 | 7.5           |
| 5841 | 0.901 | 7.6 | 0.952 | 5 | 71 | 7 | 1.173 | 1.045 | 7.5           |
| 1849 | 0.899 | 5.8 | 0.951 | 6 | 34 | 11 | 1.298 | 1.160 | 8.3           |
| 5265 | 0.896 | 4.1 | 0.930 | 7 | 16 | 58 | 1.157 | 1.124 | 13.2          |
| 5562 | 0.892 | 7.9 | 0.945 | 8 | 78 | 27 | 1.365 | 1.146 | 9.8           |
| 1829 | 0.892 | 8.1 | 0.835 | 9 | 84 | 160 | 0.833 | 1.027 | 5.6           |
| 1657 | 0.891 | 20.2 | 0.842 | 10 | 841 | 154 | 0.531 | 0.999 | 5.8           |
| 1799 | 0.890 | 12.9 | 0.922 | 11 | 191 | 73 | 0.746 | 1.117 | 7.6           |
| 5842 | 0.890 | 8.4 | 0.957 | 12 | 93 | 3 | 1.149 | 1.050 | 9.1           |
| 1850 | 0.887 | 9.4 | 0.947 | 13 | 113 | 22 | 1.409 | 1.137 | 13.2          |
| 2695 | 0.885 | 10.2 | 0.838 | 14 | 129 | 156 | 0.841 | 1.094 | 7.7           |
| 5362 | 0.884 | 21.5 | 0.943 | 15 | 1015 | 33 | 1.907 | 1.143 | 11.5          |
| 1417 | 0.884 | 5.0 | 0.943 | 16 | 24 | 32 | 0.993 | 1.119 | 9.4           |
| 1720 | 0.883 | 5.3 | 0.887 | 17 | 26 | 111 | 1.146 | 1.215 | 17.0          |
| 1541 | 0.881 | 5.9 | 0.948 | 18 | 36 | 17 | 1.308 | 1.143 | 8.9           |
| 1315 | 0.879 | 4.3 | 0.821 | 19 | 20 | 177 | 1.086 | 1.099 | 7.3           |
| 5030 | 0.879 | 22.2 | 0.934 | 20 | 1099 | 52 | 0.487 | 1.069 | 6.4           |
| 5359 | 0.879 | 21.4 | 0.960 | 21 | 992 | 2 | 1.928 | 1.161 | 14.7          |
| 5262 | 0.878 | 3.0 | 0.948 | 22 | 5 | 19 | 1.143 | 1.132 | 13.2          |
| 2446 | 0.878 | 3.1 | 0.950 | 23 | 7 | 12 | 1.135 | 1.134 | 7.3           |
| 1261 | 0.877 | 10.2 | 0.803 | 24 | 128 | 206 | 0.960 | 1.261 | 9.5           |
| 5361 | 0.877 | 20.9 | 0.952 | 25 | 944 | 9 | 1.897 | 1.165 | 12.1          |
| 5360 | 0.876 | 20.5 | 0.950 | 26 | 887 | 14 | 1.910 | 1.140 | 13.1          |
| 2705 | 0.876 | 3.3 | 0.944 | 27 | 8 | 29 | 1.000 | 0.996 | 8.0           |
| 1716 | 0.875 | 8.3 | 0.913 | 28 | 85 | 79 | 0.952 | 1.175 | 12.0          |
| 5267 | 0.875 | 6.7 | 0.946 | 29 | 54 | 24 | 1.209 | 1.125 | 13.2          |
| 1248 | 0.874 | 5.7 | 0.936 | 30 | 32 | 46 | 0.976 | 1.187 | 13.8          |

Abbreviations of methods are shown in Table S2.
References

Agrawal R. K. et al. (1998) Visualization of elongation factor G on the Escherichia coli 70S ribosome: the mechanism of translocation. Proc. Natl. Acad. Sci. U. S. A. 95, 6134-6138.

Carroni et al. (2014) Head-to-tail interactions of the coiled-coil domains regulate ClpB activity and cooperation with Hsp70 in protein disaggregation. Elife, 3, e02481.

Esquivel-Rodriguez, J. et al. (2015) Navigating 3D electron microscopy maps with EM-SURFER. BMC Bioinformatics, 16, 181.

Kawabata, T. (2008). Multiple subunit fitting into a low-resolution density map of a macromolecular complex using Gaussian mixture model. Biophys. J. 95, 4643-4658.

Krishna T. S. et al. (1994) Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. Cell, 79, 1233-1243.

Lohmann, G. (1998) Volumeric Image Analysis. John Wiley & Sons, West Sussex, 211-213.

Minami et al. (2013) MICAN: a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, Ca only models, Alternative alignments, and Non-sequential alignments. BMC Bioinformatics, 14, 24.

Nissen, P. et al. (1995) Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog. Science, 5241, 1464-1472.

Osada, R., et al. (2002) Shape distributions. ACM Transactions on Graphics, 21, 807-832.

Serdyuk, I.N., et al. (2007) Methods in Molecular Biophysics: Structure, Dynamics, Function. Cambridge University Press, Cambridge, 802-805.

Wriggers, W., et al. (1998) Self-organizing neural networks bridge the biomolecular resolution gap. J. Mol. Biol. 284, 1247-1254.