PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Fingerprint-based structure retrieval using electron density

Shuangye Yin and Nikolay V. Dokholyan*

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7260

ABSTRACT

We present a computational approach that can quickly search a large protein structural database to identify structures that fit a given electron density, such as determined by cryo-electron microscopy. We use geometric invariants (fingerprints) constructed using 3D Zernike moments to describe the electron density, and reduce the problem of fitting of the structure to the electron density to simple fingerprint comparison. Using this approach, we are able to screen the entire Protein Data Bank and identify structures that fit two experimental electron densities determined by cryo-electron microscopy.

Proteins 2011; 79:1002–1009.

Key words: cryo-EM; density fitting; structural genome; Zernike; geometric invariants.

INTRODUCTION

Assignment of protein structures using experimentally measured electron density maps is the crucial step in structure biology. If the electron density is obtained at high resolution (<3 Å), direct structure assignment is possible. However, for some proteins, only low-resolution (typically >5 Å) electron densities can be obtained, such as from using Cryo-electron microscopy (cryo-EM).[1–7] Practically, cryo-EM is commonly applied to large molecular complexes, for which the high-resolution structures are often already solved using X-ray crystallography.[3,8,9] In this case, the major challenge is the assembly of the existing crystal structures to fit the electron density of the complex. However, other challenges still exist, especially in the case when the structure of a single domain is unavailable,[6,7] or an individual protein undergoes large conformational change upon assembly of the complex. At ~5 Å resolution it is possible to identify single protein domain boundaries and separate the electron densities of the individual proteins. At this resolution, large secondary segments can already be assigned, but it is still impossible to directly determine the structure in atomic detail.[5] To determine the structure of a single protein within the complex, existing methods often use high-resolution x-ray structures and build homology models to fit the low-resolution density map,[5,7] or utilize *ab initio* folding guided by the density.[6]

We propose that it is possible to directly use the single-domain cryo-EM density to search a large structural database, such as the Protein Data Bank (PDB) or other database of protein models, to identify existing structures that best fit the electron density. These matching structures can be subsequently selected to build atomic models using existing methods.[5–9] Toward this goal, it is necessary to have a computational method that can rapidly compare the electron density with a large number of structural models (hundreds of thousands of structures per second of CPU time) independent of sequence similarity. Since it is often straightforward to derive the electron density map from a structural model, the essential challenge is to find an algorithm that can quickly compare two electron density distributions.

Electron density may be considered as a type of 3D object. Fast comparison of 3D objects has been a long-standing problem in computer graphics, modeling, and vision. To avoid explicit sampling in the rigid body degrees of freedom, one tactic is to use a vector of invariant descriptors (fingerprints) to describe the unique 3D features of the object.[10–12] In this way, the comparison of 3D objects is reduced to a comparison of two vectors, which is extremely fast. For example, we have previously used curvature distributions as fingerprints to describe local surface patches of a protein, and

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: NIH
*Correspondence to: Nikolay V. Dokholyan, Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260. E-mail: dokh@med.unc.edu.
Received 13 April 2010; Revised 8 October 2010; Accepted 5 November 2010
Published online 11 November 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.22941

successfully identified local surface similarity between proteins, independent of sequence and fold.[13]

The specified fingerprint we use in this study for the comparison of electron densities is made up of the 3D Zernike invariants,[10,12] which are constructed based on expansion of the 3D Zernike functions. Although other forms of expansion, such as the spherical harmonic function[11,14,15] and Hermite function,[16] have also been used, we choose 3D Zernike expansion due to its advantage of polynomial expansion in Cartesian coordinates.[12] Because of its speed and accuracy, 3D Zernike invariants have recently gained increasing popularity in the field of shape retrieval. Novotni and Klein first used 3D Zernike invariants for shape retrieval, and find better performance than spherical harmonic descriptors.[12] Sael *et al.* used 3D Zernike invariants to compare the geometry[17] and electrostatic properties[18] of protein surfaces. Venkatraman *et al.* further apply Zernike invariants to represent local surface shape for protein–protein docking.[19] 3D Zernike invariants were also adapted by Mek *et al.* to describe the molecular shape of ligands and proteins,[20] which was later extended by Grandison *et al.* to include flexibility to describe structural motion.[21] More examples of applications of spherical harmonic and 3D Zernike invariants can be found in a recent review by Venkatraman *et al.*[22] Despite the increasing research interest, application of Zernike invariants to the comparison of electron densities of proteins has not been reported yet.

In this study, we demonstrate the feasibility of using electron density maps of proteins to search a large protein structure database for matching structures. We benchmark this approach in a constructed test set, and also search the entire PDB for structures that fit two experimental cryo-EM electron densities, bovine metarhodopsin I[4] (5.5 Å resolution) and GroEL[4] (6 Å resolution). By ranking the protein structures based on their fingerprint similarity to the query electron density, we successfully identify matching structures among the top hits.

RESULTS AND DISCUSSIONS

Zernike invariants as fingerprints for electron density

Our fingerprint of any 3D electron density is constructed using 3D Zernike moments.[10,12,23,24] A detailed introduction to the 3D Zernike function was previously described.[10,12] Here we provide a brief description. The 3D Zernike polynomial function,

$$Z_{nl}^{m}(\vec{r}) = R_{nl}(r) \bullet Y_{l}^{m}(\theta, \phi) \tag{1}$$

is defined within the unit sphere ($r \leq 1$). Note that we use $(\vec{r})$ to denote the vector in 3D space and $r = |\vec{r}|$ for radial distance from the origin. By definition $Z_{nl}^{m}(\vec{r}) = 0$ if $n - l$ is an odd integer. For any function $f(\vec{r})$ defined within the unit sphere, the Zernike moments can be calculated as

$$\Omega_{nl}^{m} = \frac{3}{4\pi} \int_{r \leq 1} f(\vec{r}) \overline{Z}_{nl}^{m}(\vec{r}) d(\vec{r}) \tag{2}$$

More importantly, the combination of the Zernike moments

$$F_{nl} = \sqrt{\sum_{m=-l}^{l} \left| \Omega_{nl}^{m} \right|^2} \tag{3}$$

is shown to be invariant under arbitrary rotation about the origin,[10,12] which allows its usage as fingerprints. The implementation of the Zernike moment calculation is accomplished by expansion of Eq. (1) in Cartesian coordinates following Novotni and Klein.[12]

We constructed the fingerprint using the invariants as $\boldsymbol{F} = \{F_{nl}\}$, where $n = 0, 1, \ldots, 20$, and $l = 0, 1, \ldots, n$ and $n - l$ is even. Therefore, each fingerprint is a 121-dimensional vector. We choose $n = 20$ because previous works have shown that $n = 20$ gives a reasonable balance between accuracy and speed for object recognition.[12,17] We also perform reconstruction of some electron densities using the Zernike function with up to $n = 20$ and find sufficient resolution at this level (Fig. 1).

Fingerprint comparison

There are multiple ways to measure the difference between two fingerprints.[17,18,21] If we treat the two fingerprints as two vectors, the difference between the fingerprints can be defined as the distance between the two vectors. Commonly used distance measures include cosine angle distance, correlation coefficients distance, Manhattan distance, Euclidean distance, and Canberra distance. The definitions of the above distance measures are listed in Supporting Information Table 1.

To decide the most appropriate methods for fingerprint comparison, we apply the various fingerprint comparison methods to a benchmark set and compare the retrieval performances (Table I). We construct the benchmark set so that it contains protein structures that all have similar size and density, thereby representing challenging cases for structure retrieval experiments (See Methods section). We measure the performances quantitatively using both the receptor-operation curve (ROC) and the area under this curve (AUC). As evidenced from the ROC curves shown in Supporting Information Figure 1, we find that the Canberra distance has the worst performance in retrieving the matching structures. Use of the Canberra distance results in significantly lower AUC in five SCOP families except in the comparatively easy case of c.69.1.30, for which all of the methods give a similar high

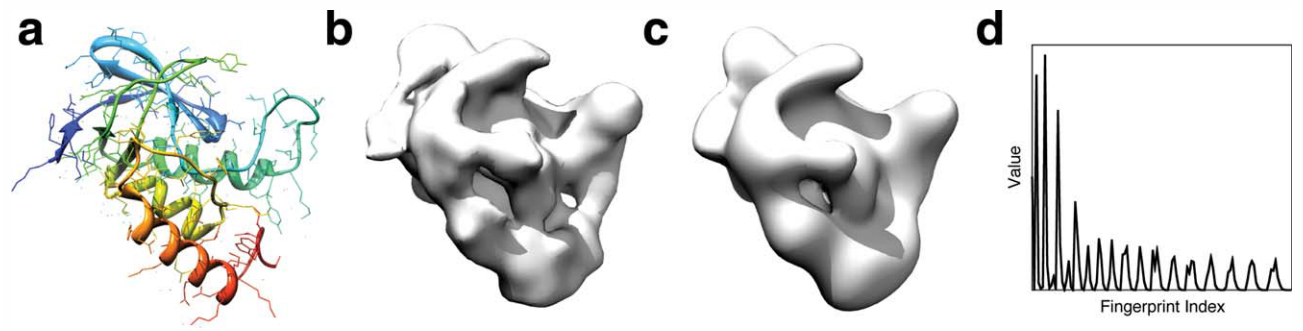

Figure 1

Fingerprint generation from protein structures. (**a**) The protein structure of staphylococcal nuclease (PDBID: 1stn). (**b**) Simulated electron density using EMAN.[9] (**c**) Reconstruction of the electron density using Zernike polynomials up to $n = 20$. (**d**) Calculated Zernike invariants up to $n = 20$ from the protein structure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

performance. Although the other methods have roughly similar performance across the six test cases, using the Manhattan distance seems to generate slightly more false positives at the early stage.

We note that for correlation and cosine distance, the absolute magnitudes of the vectors are ignored. In other words, the measured fingerprint difference will not change if the vectors are scaled by a constant factor. Nevertheless, these two measures still have similar accuracy to the Euclidean distance, making them the ideal choice for fingerprint comparison. In addition, cosine and correlation distances have the advantage of being insensitive to the total density fluctuation that may possibly occur from unresolved residues, mutations, or noise from experimental measurements. We choose the cosine distance over correlation distance in consideration of the slightly better computational efficiency.

We further study the statistical significance of fingerprint difference. Using the benchmark set, we constructed a decoy dataset for each query structure by removing all other structures that belong to the same SCOP family. We then calculate the fingerprint differences (using cosine angle distance) of the decoy datasets from the query structures. Since these decoys belong to different structure family, the comparison provides a statistical background of fingerprint differences. At the 5% statistical significance level, we find the cutoffs of significant fingerprint differences vary from 0.011 to 0.018, with an average value about 0.015 (Supporting Information Figure 2). Therefore, the fingerprint difference has to be 0.015 or smaller to be considered significantly similar.

Effect of Zernike cutoff

Next, we study the effect of Zernike function cutoff on retrieval performance. Using the same benchmark set, we apply different order cutoff values of $n = 5$, 15, 20, 25, and 30 and calculate the Zernike invariants accordingly. We find increasing overall retrieval performance as measured by AUC as a function of increasing n (Supporting Information Figure 3). The increase of performance seems to saturate at $n = 20$ even for the difficult benchmark set of c.37.1.1. Although larger values of n in general give better performance, the number of descriptors also increases quadratically with n. Based on the benchmark results, the choice of $n = 20$ seems justified, which value was also adapted by previous studies.[12,17]

Structure variation

Since the matching structures may be slightly different from the query structure as a result of conformational

Table I

The Summary of the Six SCOP Families used for Structure Retrieval in the Benchmark Set of 1597 Structures

SCOP family	N	N_{res}	R_{max}(Å)	R_g (Å)	Electron Density (e)	Centroid	RMSD (Å)	Center Shift (Å)
b.47.1.1	68	193.16	25.16	14.91	9029	1hpg-A	1.34	0.28
c.37.1.1	66	201.04	29.31	16.81	10316	3tmk-A	2.74	0.88
c.62.1.1	38	189.82	27.63	15.49	9915	1mf7-A	1.68	0.39
c.69.1.30	43	197.84	26.33	15.21	9518	1xzg-A	0.49	0.18
c.71.1.1	51	191.51	28.35	16.42	10181	1boz-A	1.20	0.26
d.3.1.1	60	213.68	28.46	16.30	10828	1khq-A	0.89	0.17

For each family, N is the number of structures in the family, N_{res} is the average number of residues, R_{max} and R_g are the average maximum distance and radius of gyration from the density center, RMSD and Center Shift are the average deviation of backbone C_α atom and shift of the density center from the centroid after structure alignment.

variation or limited experimental resolution, it is important to know how much structure difference can be tolerated using this method. For each of the six SCOP families, we compared the fingerprint difference as a function of the conformational difference from the query structure, which is measured using the root-mean-square deviation (RMSD) of the backbone C_α atoms after structural alignment (Supporting Information Figure 4). For the six SCOP families, we find that the fingerprint difference increases with the increasing RMSD. Moreover, although the six query structures belong to topologically different SCOP families, their dependencies on RMSD are quite similar. If we still select a 0.015 cutoff for fingerprint similarity, we find that the method is tolerant of conformational variations to up to 2.5 Å. This observation also explains the poor performance in retrieving the c.37.1.1 SCOP family, since many of the structures have RMSD in the 3 to 5 Å range, beyond the tolerance limit.

Rotation and center shift

Although mathematically the fingerprints are invariant upon any rotation around the center, the actual values can still vary because of numerical instability. To determine the effect of rotation, we select another six representative structures that have different overall shapes: protein structures 1a04-A and 1a06-A are compact and spherical, 16pk-A is dumbbell-shaped, 3ls0-A is cylindrical, 2xc8-A is beta-sandwich shaped, and 153l-A is asymmetric. For each protein, we randomly rotate the structure 20 times around the density center and calculate the fingerprints. We find minimal variation in the resulting fingerprints. As shown in Supporting Information Figure 5, the variation of the fingerprints is negligible. The fingerprint difference as measured using cosine angle distance is on the order of 1e-10, far below the significance level (~0.015).

The Zernike moment calculation depends on the choice of the origin, for which we take the center of mass of the electron density. In practice, the center of the electron density can be slightly different even for matching structures, as a result of, for example, unresolved residues, conformational change, or mutations. To study the dependence of the fingerprint on the choice of the center, we use the above representative protein structures, and randomly change the origin artificially and calculate the fingerprint differences (Supporting Information Figure 6a). Although the fingerprint differences in general all increase with increasing center shift, the exact response behaviors are slightly different. If we maintain 0.015 as the fingerprint difference cutoff, the tolerable center shift is on the order of 2–3.5 Å. On the other hand, we also study the shift of centers in the six SCOP families and find that the density center shifts are well below 2 Å (Supporting Information Figure 6b). Therefore, the variation of density centers is not a significant limitation for our fingerprint comparison method.

PDB screening

We demonstrate the application of the method on two experimental cryo-EM electron densities for bovine metarhodopsin I[4] (5.5 Å resolution) and GroEL[4] (6 Å resolution). We extract single protein domain densities from the experimental EM map, calculate the corresponding fingerprints, and compare with precalculated fingerprints for all protein structures in the PDB. Since the correlation of fingerprints is insensitive to the total density fluctuation, we applied an additional density filter and selected those structures whose total densities are within 20% fluctuation of the target protein. In this way, we allow small fluctuations in the density, but will not consider proteins with completely mismatching densities.

For GroEL, after applying the total density filter, there are 10,554 protein structures remaining in the dataset. We compare the fingerprints of these proteins with the target and rank them using ascending fingerprint difference. The best scoring structures (PDBID 1oel-G) is manually fit to the electron density using Chimera[25] and excellent fitting is found [Fig. 2(c)]. We find that all of the top 91 structures with the most similar fingerprints are GroEL structures. Furthermore, these 91 hits include all structures in the database that have the matching confirmations. Therefore, in this case, the fingerprint comparison clearly separates all the GroEL proteins from the rest in the database [Fig. 3(a)].

For bovine metarhodopsin protein, there are 20,962 protein structures that are within 20% total electron density variation of the target protein. We rank the structure using the fingerprint differences and find that 9 out of the top 10 hits are indeed rhodopsin structures. We manually fit the best scoring structure (PDBID 1hzx-B) to the electron density using Chimera[25] and find excellent fitting [Fig. 2(d)]. We plot the distribution of fingerprint difference score and find that all 21 rhodopsin structures are among the top 334 hits among the 20,962 entries in the dataset, with one exception. The exception (PDBID: 1ln6-A) has quite a large conformational change compared with the other 20 rhodopsin structures; The RMSD of the backbone C_α atoms of 1ln6-A from the other rhodopsin structures is 6.5 Å in average. Although the ranking is not perfect for the bovine metarhodopsin I test case, our method can still put all matching rhodopsin protein among the top ranks [Fig. 3(b)]. And the existence of multiple conformations of the same rhodopsin protein highlights the need to utilize more structural models to fit the experimental electron density.

We also examine the false positive hits for rhodopsin structure matching. We pick the top four distinct false-positive structures and compare them with the true positive structure of rhodopsin. As shown in Supporting

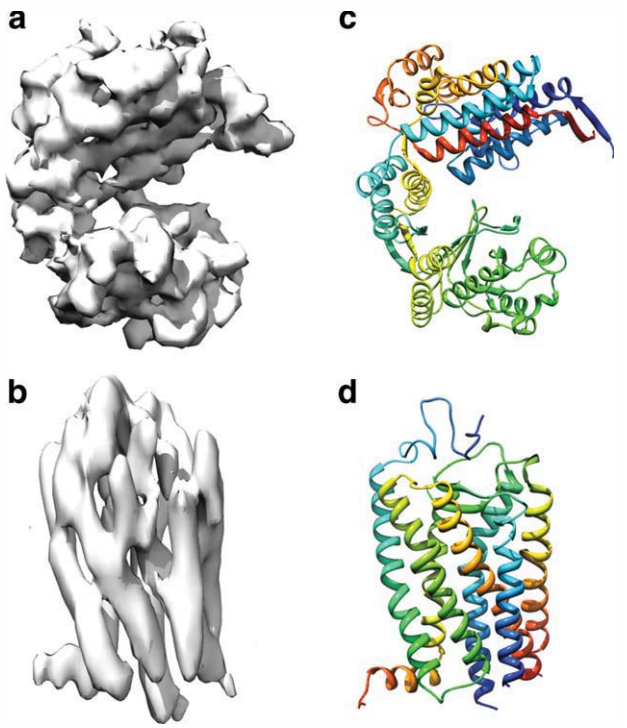

Figure 2

The experimental cryo-EM electron density for GroEL (**a**) and bovine metarhodopsin I (**b**) and their best matching structures (**c**, **d**) from searching the PDB. (The figures are prepared using USCF Chimera). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Information Figure 7, these false positive structures all have very different topology and secondary structure from the rhodopsin structure, although the overall shapes are somewhat similar. We note that compared with GroEL, the overall shape of the rhodopsin structure is not as unique, which may explain the high false positive rate in PDB screening. In addition, the fingerprint difference even for the first hit (0.024 for 1hzx-B) is quite large, compared with that for the GroEL case (the smallest fingerprint difference is 0.014 for 1oel-G), suggesting that other variations, such as the noise or conformational difference in the density map, may contribute to the low sensitivity in retrieving rhodopsin structures.

To test the performance of the fingerprint comparison, we also apply the fingerprint comparison without application of the total density filter. For the rescaling radius applied to both rhodopsin and GroEL (50 Å), there are 102,187 structures having fingerprints calculated at this radius (See Methods). Applying the same cosine angle distance, we find that all 91 GroEL structures are still correctly ranked highest among the 102,187 structures. For rhodopsin, the retrieval performance is slightly decreased, as shown by the ROC curves in Figure 4. In this case, removal of the density filter only slightly decreases the AUC value from 0.992 to 0.985. Therefore, we conclude that application of the density filter does not significantly affect the retrieval performance of the method. However, without the density filter, the absolute number of false positives increases as a result of the increasing dataset size. Applying the density filter can

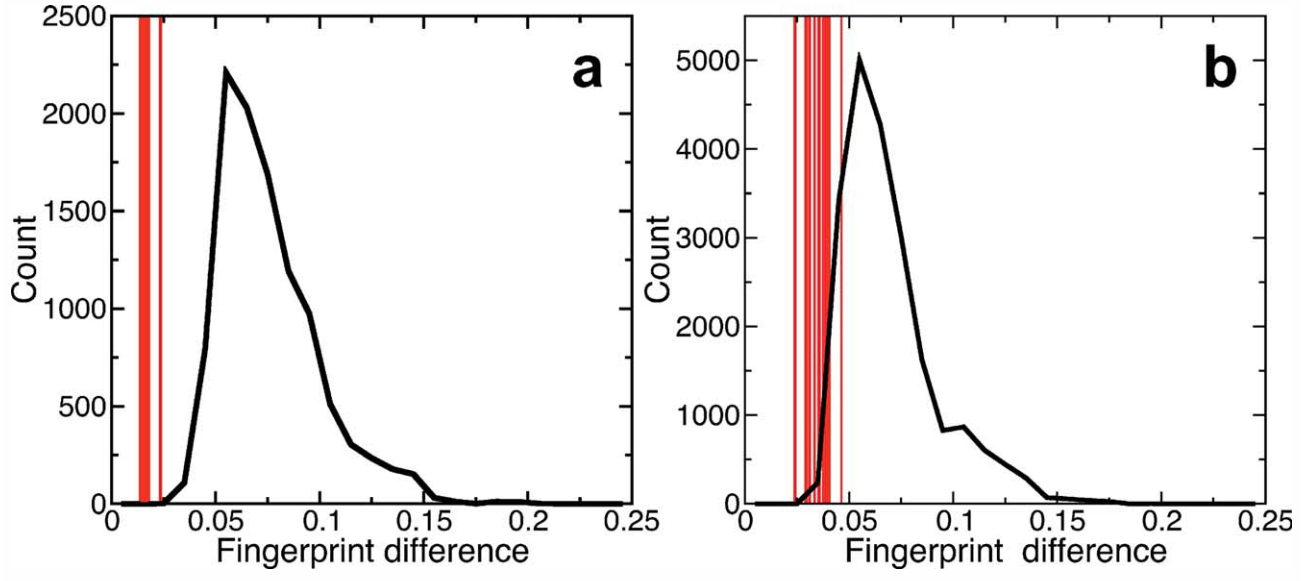

Figure 3

Fingerprint difference scores of all the GroEL (**a**) and rhodopsin (**b**) protein structures. The distribution of fingerprint differences between the experimental density and all decoy proteins in the datasets are plotted as black solid lines. The fingerprint difference scores for all GroEL and rhodopsin structures are marked using vertical lines. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

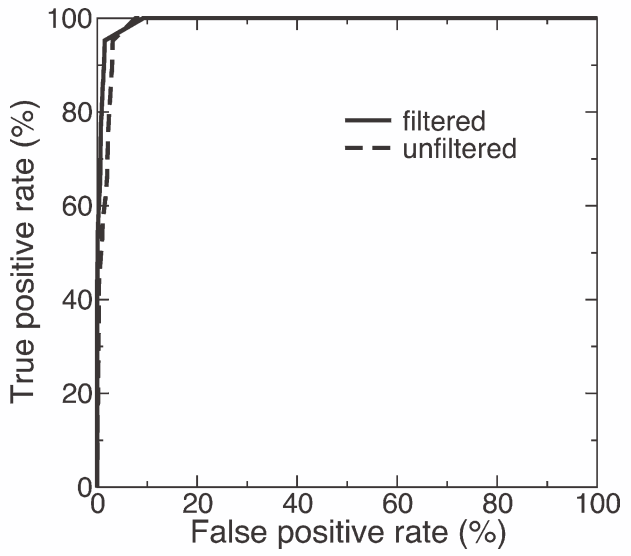

Figure 4

ROC curves of retrieving rhodopsin structures using the experimental cryo-EM density with and without applying the density filter.

prevent taking into account irrelevant structures and reduce the absolute number of false positives.

Performance

Using fingerprints, the fitting of structures to the electron density can be performed at very high speed. In our benchmark test, we can effectively screen 800,000 structures (without density filter) per second using a single 3 GHz Intel Xeon processor. Given this performance, it is possible to further expand the structure library by incorporating *ab initio* folding models and homology models.[5–7] The expanded library shall improve the screening accuracy and include the allosteric and dynamic properties of biomolecules. We expect the fingerprint-based screening approach to be a valuable tool in assisting cryo-EM for structure determination.

MATERIALS AND METHODS

Implementation of Zernike moment calculation

The implementation of the Zernike moment calculation is accomplished by expansion of Eq. (1) in Cartesian coordinates following Novotni and Klein.[12] We validate the implementation by performing Zernike transformation of a 3D density map and the reconstruction of the density using Zernike moments up to $n = 20$ [Fig. 1(c)]. The reconstructed density also verifies that $n = 20$ provides a sufficient resolution in describing the electron density.

Since the Zernike function is defined within a unit sphere, we transform the original electron density function (f_0) using $f(\boldsymbol{r}) = f_0((\boldsymbol{r}\text{-}\boldsymbol{r}_0)/r_m)$ before performing the Zernike moment calculations, where $\boldsymbol{r}_0$ is the center-of-mass of the electron density and r_m is the rescaling factor. Any density points with $r > r_m$ are discarded during the calculation.

Although previous studies[21,24] often choose r_m according to the maximum distance from the center, this approach is not suitable for our application to density comparison. First, the maximum distance is sensitive to small structure variations in the loops and end regions of the protein structures, which may lead to different r_m even when the two structures are similar. Second, the maximum distance is not well defined for experimental density maps because of the background noise. Therefore, in our implementation, we choose a set of fixed r_m values (20 Å, 30 Å, ... 100 Å) to rescale the density and generate multiple fingerprints accordingly (See the PDB screening section for details).

For the experimental electron density, the Zernike moments are calculated by direct integration of Eq. (2) over the 3D grid where the electron densities are assigned. For protein structure models, the Zernike moments are calculated by summation over all atoms using:

$$\Omega_{nl}^{m} = \frac{3}{4\pi}\sum_{i} \rho_i \overline{Z}_{nl}^{m}(\vec{r}) \qquad (4)$$

where ρ_i is the electron density of atom i. This approach is equivalent to the sampling of the density function at the atom centers and is much faster compared with integration over the grid.

Extraction of experimental density of single domain proteins

We retrieved the cryo-EM map from Electron Microscopy Data Bank (EMDB).[26] The entry IDs for bovine metarhodopsin I and GroEL are 1079[4] and 1081,[3] respectively. We separate the single domain densities using the density map tools available in UCSF Chimera volume viewer.[25] To eliminate the effect of background noise, we calculate the fingerprint only for voxels with densities above a certain threshold. The threshold is determined by visual inspection of the density map in Chimera so that the topology of the protein is best represented. The thresholds for bovine metarhodopsin I and GroEL are 25 and 0.3, respectively. The same thresholds are used for preparation of the surface representations shown in Figure 2(a,b).

PDB screening

The screening is performed on a snapshot of the Protein Data Bank (PDB)[27] created on September 11, 2008. Every PDB entry is parsed into separate chains, and all small molecules and nucleic acid chains are discarded. We keep only high quality structures with resolution better than 3 Å and having fewer than 20% of residues with missing

Table II
The Structure Retrieval Performance for the Six SCOP Families Using Different Fingerprint Comparison Methods of Cosine Angle, Correlation, Euclidean, Manhattan, and Canberra Distances

	AUC				
SCOP Family	Cosine	Correlation	Euclidean	Manhattan	Canberra
b.47.1.1	0.997	0.997	0.996	0.996	0.973
c.37.1.1	0.806	0.794	0.791	0.811	0.627
c.62.1.1	0.983	0.984	0.984	0.973	0.889
c.69.1.30	0.993	0.993	0.993	0.998	0.996
c.71.1.1	0.990	0.990	0.991	0.977	0.917
d.3.1.1	0.988	0.988	0.988	0.988	0.934

heavy atoms. After applying the filters, the total number of chains remaining is 110,109. We then calculate the fingerprints for every chain using various scaling factors (r_m) from 20 to 100 Å in increments of 10 Å. We also make sure that the entire protein density can be rescaled within the unit sphere and occupy a significant volume by enforcing $2r_g < r_m < 4r_g$, where r_g is the radius of gyration of the electron density. Statistics of all protein chains show that the r_{max}/r_g ratio is between 1.5 and 3.5 for the majority of the protein structures in the PDB (Supporting Information Figure 8), where r_{max} is the maximum distance of any atom from the density center. Therefore, selection of the rescaling radius r_m between $2r_g$ and $4r_g$ will ensure that the structure is enclosed within the unit sphere and covers a significant fraction of volume. For the experimental density map, we select the minimum r_m value from 20 to 100 Å range and conduct a visual inspection to ensure that this radius covers the complete density. After r_m is chosen, only the PDB chains whose fingerprints have been calculated with selected r_m will be compared. For the two examples presented here, we use $r_m = 50$ Å, and there are 102,187 PDB chains that have their fingerprints calculated at this rescaling level. The other 2078 PDB chains are simply ignored because they are too small or too large for the 50 Å rescaling radius.

To construct the validation set for the GroEL and rhodopsin hits, we use the SCOP and CATH classification databases to find all PDB chains that belong to the same family. We also visually inspect the structures to ensure that they have similar structures, and also to avoid possible errors in SCOP or CATH classification. For example, for rhodopsin, we use the SCOP family of f.13.1.2 and the CATH family of 1.20.1070.10., and find three structures (PDB code: 1xme-B, 2qpe-B, and 1ehk-B) incorrectly classified in the same SCOP family. Structural alignment calculation and visual inspection reveal that these three structures have completely different topology, and should not be classified into the rhodopsin family (Supporting Information Figure 9). In addition, there is a group of structures (mostly bacteriorhodopsin) classified as belonging to the same family, although they have quite large conformational change (RMSD in the 3.5–4.5 Å range) from the bovine rhodopsin. We exclude these structures from true positives, since our method can only tolerate structural differences of ~2 Å of RMSD (See Results). For GroEL, we use the SCOP family of a.129.1.1, c.8.5.1, and d.56.1.1, and CATH family of 1.10.560.10, 3.30.260.10, and 3.50.7.10. After structural alignment and visual inspection, we find that there is a cluster of protein structures (mostly thermosome alpha subunit) that belong to the same GroEL fold family but have very different conformations (RMSD in the range of 4-5 Å) and do not fit the experimental density well (Supporting Information Figure 10). At the end, we have a structurally consistent validation set of 91 GroEL structure and 21 rhodopsin structures.

Benchmark set

To assess the structure retrieval performance, we construct a benchmark set that contains 1597 protein structures. These are the nonredundant structures selected from the PDB snapshot that contain between 180 and 220 residues and have been assigned a SCOP classification. From these 1597 structure, we further select six SCOP families that have the most structures for the structure retrieval. We visually inspect all proteins in each family to ensure that they are indeed structurally similar. We then cluster the structures to pick the centroid structure as the query structure for each family, and use electron density calculated for this centroid to retrieve the other structures in the same SCOP family out of the 1597 structures. The statistic details of the six SCOP families are listed in Table I.

ACKNOWLEDGMENTS

The authors thank Dr. Steven Ludtke for informative discussions about cryo-EM resolution. They also thank Dr. Feng Ding, Srinivas Ramachandran, and Elizabeth A. Proctor during the development of the method. All structural alignments are calculated using the TMalign program.[28] The figures are prepared using UCSF Chimera[25] and Pymol.[29] The datasets are available for downloading at http://faust.med.unc.edu/pshape/.

REFERENCES

1. Zhou ZH, Baker ML, Jiang W, Dougherty M, Jakana J, Dong G, Lu G, Chiu W. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. Nat Struct Mol Biol 2001;8:868–873.
2. Chiu W, Baker ML, Jiang W, Zhou ZH. Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. Curr Opin Struct Biol 2002;12:263–269.
3. Ludtke SJ, Chen D-H, Song J-L, Chuang DT, Chiu W. Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. Structure 2004;12:1129–1136.
4. Ruprecht JJ, Mielke T, Vogel R, Villa C, Schertler GFX. Electron crystallography reveals the structure of metarhodopsin I. EMBO J 2004;23:3609–3620.
5. Topf M, Sali A. Combining electron microscopy and comparative protein structure modeling. Curr Opin Struct Biol 2005;15:578–585.

6. Baker ML, Jiang W, Wedemeyer WJ, Rixon FJ, Baker D, Chiu W. Ab Initio modeling of the herpesvirus VP26 core domain assessed by cryoEM density. PLoS Comput Biol 2006;2:e146–e146.
7. Topf M, Baker ML, Marti-Renom MA, Chiu W, Sali A. Refinement of protein structures by iterative comparative modeling and cryoEM density fitting. J Mol Biol 2006;357:1655–1668.
8. Wriggers W, Milligan RA, McCammon JA. Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. J Struct Biol 1999;125:185–195.
9. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, Ludtke SJ. EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol 2007;157:38–46.
10. Canterakis N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. Proceedings of 11th Scandinavian Conference on Image Analysis, Kangerlussuaq, Greenland, 1999. pp 85–93.
11. Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3D shape descriptors. Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing, Aachen, Germany, 2003. pp 156–164.
12. Novotni M, Klein R. 3D zernike descriptors for content based shape retrieval. Proceedings of the eighth ACM symposium on Solid modeling and applications, Seattle, Washington, USA, 2003. pp 216–225.
13. Yin SY, Proctor EA, Lugovskoy AA, Dokholyan NV. Fast screening of protein surfaces using geometric invariant fingerprints. Proc Natl Acad Sci USA 2009;106:16622–16626.
14. Kovacs JA, Wriggers W. Fast rotational matching. Acta Crystallogr Sect D 2002;58:1282–1286.
15. Kovacs JA, Chacón P, Cong Y, Metwally E, Wriggers W. Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. Acta Crystallogr Sect D 2003;59:1371–1376.
16. Leibon G, Rockmore DN, Park W, Taintor R, Chirikjian GS. A fast Hermite transform. Theor Comput Sci 2008;409:211–228.
17. Sael L, La D, Li B, Rustamov R, Kihara D. Rapid comparison of properties on protein surface. Proteins 2008;73:1–10.
18. Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, Kihara D. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins 2008;72:1259–1273.
19. Venkatraman V, Yang Y, Sael L, Kihara D. Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics 2009;10:407–407.
20. Mak L, Grandison S, Morris RJ. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. J Mol Graph Model 2008;26:1035–1045.
21. Grandison S, Roberts C, Morris RJ. The application of 3D Zernike moments for the description of "Model-Free" molecular structure, functional motion, and structural reliability. J Comput Biol 2009;16:487–500.
22. Venkatraman V, Sael L, Kihara D. Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. Cell Biochem Biophys 2009;54:23–32.
23. Zernike F. Inflection theory of the cutting method and its improved form, the phase contrast method. Physica 1934;1:689–704.
24. Novotni M, Klein R. Shape retrieval using 3D Zernike descriptors. Computer-Aided Design 2004;36:1047–1062.
25. Goddard TD, Huang CC, Ferrin TE. Visualizing density maps with UCSF Chimera. J Struct Biol 2007;157:281–287.
26. Tagari M, Newman R, Chagoyen M, Carazo J-M, Henrick K. New electron microscopy database and deposition system. Trends Biochem Sci 2002;27:589–589.
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
28. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
29. The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.