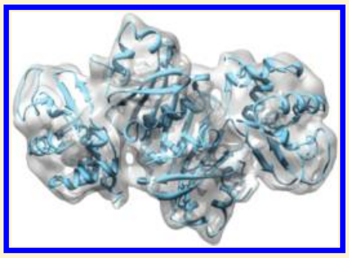
Fitting Multimeric Protein Complexes into Electron Microscopy Maps Using 3D Zernike Descriptors

Juan Esquivel-Rodríguez[‡] and Daisuke Kihara*[,†,‡,§]

[†]Department of Biological Sciences, [‡]Department of Computer Science, and [§]Markey Center for Structural Biology, Purdue University, West Lafayette, Indiana 47907, United States

ABSTRACT: A novel computational method for fitting high-resolution structures of multiple proteins into a cryoelectron microscopy map is presented. The method named EMLZerD generates a pool of candidate multiple protein docking conformations of component proteins, which are later compared with a provided electron microscopy (EM) density map to select the ones that fit well into the EM map. The comparison of docking conformations and the EM map is performed using the 3D Zernike descriptor (3DZD), a mathematical series expansion of three-dimensional functions. The 3DZD provides a unified representation of the surface shape of multimeric protein complex models and EM maps, which allows a convenient, fast quantitative comparison of the three-dimensional structural data. Out of 19 multimeric complexes tested, near native complex structures with a root-mean-square deviation of less than 2.5 Å were obtained for 14 cases while medium range resolution structures with correct topology were computed for the additional 5 cases.



INTRODUCTION

Multimeric protein complexes are at the center of many important biological functions, such as transport, gene regulation, translation, and enzymatic reactions. Although an increasing number of high-resolution structures of single proteins have been solved by X-ray crystallography and nuclear magnetic resonance (NMR), solving protein complex structures is still a challenging task by these methods. In recent years, cryo-electron microscopy (EM) has made significant advances to successfully determine macromolecular complex structures.[1,2] However, structures determined by EM are at relatively low resolution, ranging from about 4 Å to over 30 Å. Since atomic resolution structures of component proteins in a protein complex are available in many cases either from the experimental methods or computational modeling, it is of practical importance to establish efficient and accurate computational methods that can fit high-resolution structures into an EM density map of a protein complex.[3−5]

EMfit by Rossmann et al. optimizes the position of an atomic-resolution structure in an EM map by performing a six-dimensional local search starting from a predetermined position.[6] SITUS[7,8] employs a method called the codebook vector that allows fast density comparisons, without the need for explicit superimposition of maps. BCL::EM-Fit[9] uses geometric hashing combined with Monte Carlo refinement to perform rigid body fitting. Semiautomatic protocols using the EMAN[10] package have also been devised,[2,11] which combine computational programs with manual manipulation to generate fitted structures.

Since cryoEM has been often applied to large symmetric protein complexes, several methods construct symmetrical structures from a single protein and examine atomic clashes and the deviation from a structure with an idealized symmetry.[6,12−15] To account for conformational changes of proteins in multimeric forms, several methods focus on the application of molecular dynamics,[1,16−20] normal-mode analysis,[17,18] and elastic network models.[21]

The aforementioned methods mainly focus on optimizing the conformation and the orientation of high-resolution structures around their initially assigned positions, or mapping a component and its symmetric assembly into an EM map, one at a time. In conjunction with such methods, a couple of approaches were proposed that are aimed toward determining the positions and orientations of multiple high-resolution structures of component proteins simultaneously in an EM density map. Kawabata[22] and Lasker et al.[23] used Gaussian mixture models as a reduced representation of molecule shapes. In the former study, initial random positions for subunits are used as starting points, which are iteratively improved. In the latter method, given a set of anchor points in an EM map, the placements of high-resolution component structures are optimized.

Here, we present a novel computational method (EMLZerD) for fitting multiple high-resolution structures into an EM map, which combines a multiple protein docking procedure and an assessment for fitness of the protein complex structures and the EM map using the 3D Zernike descriptor (3DZD).[24−26] The multiple protein docking procedure generates a couple hundred plausible protein complex structures assembled from the component proteins.[27,28] Then, the overall surface shape of each candidate protein complex structure is compared with the EM map using the 3DZD. The 3DZD is a mathematical series expansion of a 3D function, which is a compact and rotationally invariant representation of a 3D object (i.e., the surface shape of the protein complexes and the EM map). The similarity of

Special Issue: Harold A. Scheraga Festschrift

Received: December 30, 2011
Revised: March 10, 2012
Published: March 14, 2012

ACS Publications © 2012 American Chemical Society

dx.doi.org/10.1021/jp212612t | *J. Phys. Chem. B* 2012, 116, 6854−6861

two objects can be quantified by the Euclidean distance of coefficients assigned to each term in the 3DZD series expansion of the two objects. Our approach has the following advantages: first, multiple component proteins can be fit into an EM map without providing initial positions of the proteins in the map, which is not addressed by the majority of the existing methods. The complex structure to be fit does not need to be symmetric. By using the compact and rotation invariant shape representation by the 3DZD, expensive computations required for superimposition of density maps or correlation calculations are avoided. Moreover, fitness of the complex structures to the EM map is quantitatively evaluated as the Euclidean distance of the 3DZDs. Therefore, the distance can be considered as the confidence score of the solution.

We tested our method on a data set of 19 multimeric asymmetric protein complexes, for which EM maps were simulated at 10 and 15 Å resolutions. Among them, near native complex structures within a root-mean-square deviation of 2.5 Å were obtained for 14 cases while medium range resolution structures with correct topology were computed for the remaining 5 cases.

■ EXPERIMENTAL METHODS

Our method, EMLZerD, takes a set of atomic resolution structures of component proteins and an EM map of the protein complex structure as input and determines the positions and the orientations of the component proteins within the EM map. The method consists of two logical steps: given a set of atomic resolution structures of proteins, a multiple docking program called Multi-LZerD, developed by our group, is employed to generate a couple hundred protein complex models. In the following step, the fitness of each complex model is evaluated against the target EM map using the 3DZD. Thus, the output of the EMLZerD is the ranked list of protein complex models with the confidence score. The overview of EMLZerD is illustrated in Figure 1. The two steps are described below in more detail.

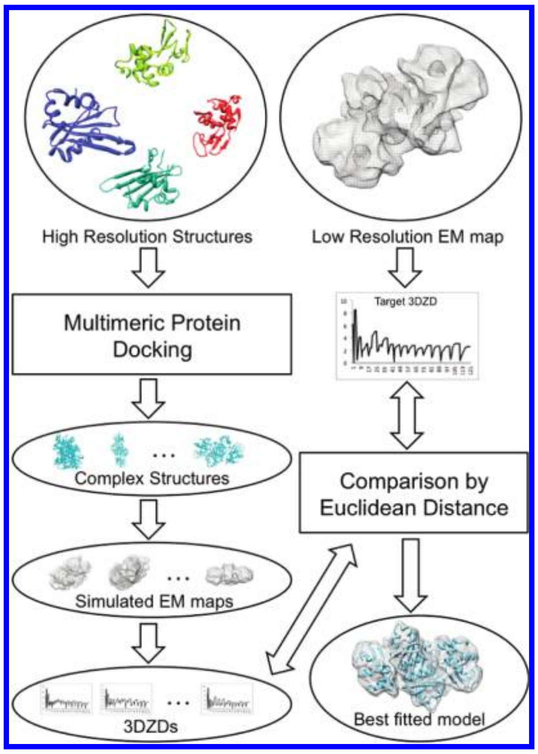


Figure 1. Overall flowchart of EMLZerD.

Multimeric Protein Docking with Multi-LZerD. Given a set of protein structures, Multi-LZerD assembles them into complex structures by combining pairwise docking solutions of pairs of proteins. The detailed algorithm and the benchmark results of multiple protein docking are provided elsewhere.[28,29] Here, we provide a brief explanation of the algorithm and the pseudocode of Multi-LZerD (Figure 2).

```
// Multi-LZerD Configuration
subunits = subunits in multimeric complex
population_size = complexes kept per iteration
generations = total iterations
threshold = acceptable number of atom clashes
// Pairwise prediction computations
predictions = ()
// Compute pairwise docking predictions
// for each pair of units
foreach 1 < i < subunits do
  foreach i+1 < j ≤ subunits do
    // The top 54,000 predictions are kept
    predictions.add(LZerD(i,j))
  end foreach
end foreach
// Genetic Algorithm initial population
population = ()
// 200 pop. size in this study
for 1 to population_size do
  // randomly select an edge and
  // a pairwise prediction
  // (e.g. from 1 to 54,000)
  t = random_spanning_tree(predictions)
  population.add(t)
end for
// Iterative phase
// Up to 5,000 iterations
for 1 to generations do
  // randomly replace one edge
  population.add(mutate(population))
  // new elements based on parent edges
  population.add(crossover(population))
  foreach individual in population do
    if individual.clashes > threshold then
      population.remove(individual)
    end if
  end foreach
  // cluster distance: <10Å RMSD
  population.cluster()
  population.select_top(population_size)
end for
// Final refinement (Monte Carlo based)
population.refinement()
```

Figure 2. Pseudocode of Multi-LZerD multimeric protein docking procedure.

Multi-LZerD takes atomic resolution structures of subunits from a multimeric protein complex and starts by generating over 54 000 docking poses for every pair of proteins using a pairwise docking program, LZerD (Local 3D Zernike descriptor-based docking program).[27] Note that pairwise docking is performed for every protein pair, including pairs that do not interact in the complex, since the correct complex structure is not known beforehand. Using the pairwise solutions for each protein pair, a whole complex structure can be uniquely specified as a graph, more precisely, a spanning tree, where nodes represent proteins and an edge between a pair

Table 1. Fitting Results Obtained for 19 Multimeric Complexes[a]

PDB	chains	number of residues[b]	rmsd (Å)[c]	rank by energy[d]	10 Å EM rank (euc. distance)[e]	15 Å EM rank (euc. distance)[e]
1A0R	3	650	0.85	1	1 (1.99)	1 (0.48)
1B9X	3	654	0.62	1	1 (0.95)	1 (0.18)
1K6N	3	855	1.11	1	1 (1.54)	1 (1.13)
1VCB	3	390	1.16	113	1 (1.36)	1 (0.53)
2AZE	3	307	1.00	1	1 (1.37)	1 (0.42)
2PRG	3	630	1.38	141	1 (1.50)	1 (0.57)
1ES7	4	410	1.85	4	1 (2.48)	1 (1.25)
1GPQ	4	528	1.74	1	1 (0.55)	1 (0.35)
1K2X	4	640	7.53	68	170 (9.62)	168 (4.95)
1LOG	4	466	1.90	63	1 (2.01)	1 (0.95)
1NNU	4	578	1.12	4	1 (1.32)	1 (0.64)
1QGW	4	497	3.24	4	1 (3.80)	1 (1.40)
1RHM	4	498	1.07	1	2 (2.57)	1 (0.27)
1WWW	4	442	2.48	54	17 (2.68)	17 (2.14)
2BBK	4	960	2.04	1	1 (1.67)	1 (0.38)
6RLX	4	104	4.49	171	84 (13.39)	23 (2.57)
1W88	5	1433	4.80	1	1 (1.62)	1 (0.70)
1I3O	6	812	2.07	57	7 (2.27)	3 (0.92)
1JYO	6	730	6.43	55	51 (5.38)	73 (4.31)

[a] The best fit structure after the final generation of Multi-LZerD is analyzed in terms of its RMSD to the native, as well as the ranking improvement obtained due to the EM fitting process. [b] The total number residues of all the subunits in the complex. [c] The best global Cα root-mean-square deviation (rmsd) to the native structure in the last GA generation. [d] The rank given by Multi-LZerD to this prediction, before testing the agreement with the EM map. [e] The original predictions are reranked according to the Euclidean distances of 3D Zernike descriptors for the native EM density map and the one from the prediction, using the two different resolutions tested.

specifies one of the precomputed docking poses of the two proteins. The spanning tree representation is very suitable for constructing a multiple docking complex from pairwise decoys because not all pairs of nodes need to be connected. Multi-LZerD explores the conformation space of spanning trees using a genetic algorithm (GA), which changes connections between nodes and exchanges docking poses between pairs of proteins. Combinatorial optimizations are performed starting from a population of 200 randomly generated spanning trees. Complex structures with too many atom clashes are removed. The complex structures are subject to clustering, and finally, the 200 lowest energy structures are selected. After each round of optimization, the best 200 conformations are kept for the next iteration. This is repeated up to 5000 iterations.

Multiple docking structures are evaluated with a physics-based score that linearly combines terms for the van der Waals potential,[30] the electrostatic potential,[31] hydrogen bond and disulfide bond terms,[32] a solvation term,[33,34] and a statistical atomic contact potential.[35] Clustering is performed at the end of each generation to promote structural variability in the population. 200 complex structures from the final generation are taken as input for the next step, where each of the candidate structures is compared with the EM map using 3DZD.

Evaluation of Fitness of Complex Structures with the EM Map Using 3D Zernike Descriptor. Previous studies[4,7,8] have emphasized the importance of having a concise yet precise representation of EM density maps that allows efficient comparison between proposed atomic models and low resolution EM data. The 3DZD[24−26,36] complies with these requirements for the purposes of EM-fitting. A 3DZD is a series expansion of a mathematical 3D function that, in this case, encodes the EM-maps and the surface shape of protein complexes. In our previous studies, the 3DZD was successfully applied to fast protein global shape comparison,[37,38] ligand binding pocket comparison,[39−41] and small ligand molecule search.[42] Here we provide a brief description of the 3DZD. For more mathematical details of the derivation, please refer to the previous papers.[24−26]

The 3DZD is a series expansion of a 3D function in terms of the Zernike−Canterakis basis:

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

where $-l < m < l$, $0 \leq l \leq n$, and $(n - l)$ is even. $Y_l^m(\vartheta, \varphi)$ are the spherical harmonics, and $R_{nl}(r)$ are radial functions, which are constructed so that $Z_{nl}^m(r, \vartheta, \varphi)$ can be converted to polynomials, $Z_{nl}^m(\mathbf{X})$, in the Cartesian coordinates as follows:

The Cartesian and spherical coordinates are converted between each other as

$$\begin{aligned} \mathbf{x} &= |\mathbf{x}|(\sin\vartheta\sin\phi, \sin\vartheta\cos\phi, \cos\phi)^{\mathrm{T}} \\ &= r(\sin\vartheta\sin\phi, \sin\vartheta\cos\phi, \cos\phi)^{\mathrm{T}} \end{aligned} \tag{2}$$

The spherical harmonics $Y_l{}^m(\vartheta, \varphi)$ are shown to be convertible to the Cartesian coordinates as

$$Y_{l}^{m}(\vartheta, \phi) = \frac{1}{r^{l}} e_{l}^{m}(\mathbf{x}) \tag{3}$$

Using eq 3, the 3D Zernike moments (eq 1) are converted to the Cartesian coordinates as

$$\begin{aligned} Z_{nl}^{m}(\mathbf{x}) &= R_{nl}(r)Y_{l}^{m}(\vartheta, \phi) \\ &= R_{nl}(r)\frac{1}{r^{l}}e_{l}^{m}(\mathbf{x}) \\ &= \sum_{\nu=0}^{k} q_{kl}^{\nu}|\mathbf{x}|^{2\nu} r^{l}\frac{1}{r^{l}}e_{l}^{m}(\mathbf{x}) \end{aligned} \tag{4}$$

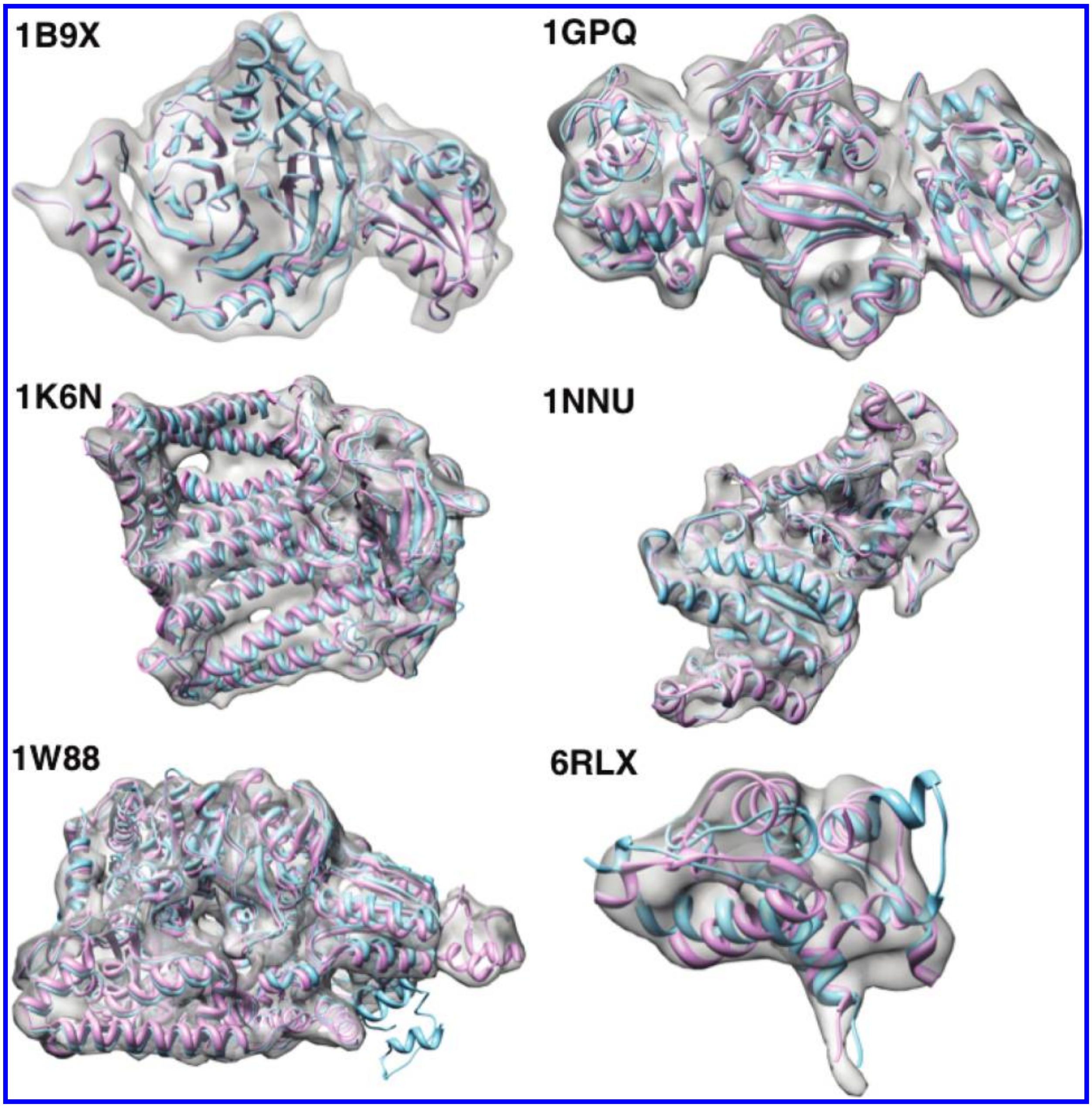


Figure 3. Examples of structures fitted into EM maps. The resolution of the simulated EM maps is 10 Å. Predicted subunit arrangements (cyan) fitted into the EM maps are superimposed on the crystal structures of the native complexes (magenta). The global rmsd of the fitted structures are the following: 1B9X, 0.62 Å; 1GPQ, 1.74 Å; 1K6N, 1.11 Å; 1NNU, 1.12 Å; 1W88, 4.80 Å; and 6RLX, 4.49 Å (Table 1).

Thus,

$$R_{nl}(r) = \sum_{\nu=0}^{k} q_{kl}^{\nu} |\mathbf{x}|^{2\nu} r^{l} = \sum_{\nu=0}^{k} q_{kl}^{\nu} r^{2\nu+l} \tag{5}$$

Here

$$e_l^m(\mathbf{x}) = r^l c_l^m \left(\frac{ix - y}{2}\right)^m z^{l-m} \sum_{\mu=0}^{\lfloor l-m/2 \rfloor} \binom{l}{\mu}\binom{l-\mu}{m+\mu} \left(-\frac{x^2+y^2}{4z^2}\right)^{\mu} \tag{6}$$

where

$$c_l^m = c_l^{-m} = \frac{\sqrt{(2l+1)(l+m)!(l-m)!}}{l!} \tag{7}$$

$2k = n - l$ and the coefficient q_{kl}^{ν} are determined as follows to guarantee the orthonormality of the functions within the unit sphere,

$$q_{kl}^{\nu} = \frac{(-1)^k}{2^{2k}} \sqrt{\frac{2l+4k+3}{3}} \binom{2k}{k} (-1)^{\nu} \frac{\binom{k}{\nu}\binom{2(k+l+\nu)+1}{2k}}{\binom{k+l+\nu}{k}} \tag{8}$$

To represent 3D structural data in the 3DZD, the 3D structure is mapped to a 3D grid function, $f(x)$. Now 3D Zernike moments of $f(x)$ are defined by the expansion in this orthonormal basis, as follows

$$\Omega_{nl}^{m} = \frac{3}{4\pi}\int_{|x|\le 1} f(\mathbf{x})\bar{Z}_{nl}^{m}(\mathbf{x})\,\mathbf{dx} \tag{9}$$

The rotational invariance is obtained by defining the 3DZD series F_{nl} as norms of vectors Ω_{nl}:

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^{m})^2} \tag{10}$$

The parameter n is called the order of 3DZD, which determines the resolution of the descriptor. An order of $n = 20$, which gives a total of 121 terms (the ranges of m and l are determined by n), is used in our current work, unless otherwise noted. The order of 20 is shown to be effective in global protein surface shape comparison in our previous work.[43] As the result, a 3D structure is represented by a vector of coefficients assigned to each term of the 3DZD. An additional comparison with $n = 24$ and $n = 28$ (169 and 225 coefficients, respectively) is provided in the Results section to show the impact of varying this parameter.

The fitness of generated protein complex structures with respect to a target EM map (i.e., similarity of the two 3D structures) is assessed by first simulating the EM density for the 200 complexes; then, 3DZD coefficients are generated for these. The similarity is quantified as the Euclidean distance between the vectors of 3DZD coefficients. Note that prealignment of a complex structure to the EM map is not necessary since the 3DZD is rotation invariant. A smaller Euclidean distance between a complex structure and an EM map indicates a considerable shape agreement, while high values mean significant differences in the overall shape between the two. As we will see in the Results section, the multimeric structure model with the smallest Euclidean distance represents the best fitting model into the target EM map for the majority of the test cases.

Data Set. To benchmark the proposed method, we prepared a data set of 19 multimeric protein complex structures. EM maps for these complex structures were simulated using the pdb2mrc program in EMAN2.[10] Simulated EM maps have been commonly used for studying structure fitting into EM maps.[9,17,21,22,44] For a protein complex, EM maps of two resolutions, 10 and 15 Å, were prepared. The 3DZDs of the EM maps were computed for an isosurface using a density range from 5 to 8 for the EM map at 10 Å resolution, while from 7 to 11 for the case of the 15 Å resolution. These density ranges were shown to be effective in capturing characteristic shape features of EM maps in our previous study.[45]

■ RESULTS

The modeling results for the 19 multimeric protein complexes are summarized in Table 1. Along with the results of high-resolution structure fitting into the EM maps, we also show multimeric protein docking results by Multi-LZerD without using the EM maps.

Multimeric Protein Docking Results. To begin with, we examine the accuracy of the multimeric protein complex structures generated by Multi-LZerD. 200 plausible complex models were generated and ranked by the physics-based score for a target protein complex. The fourth and fifth columns from the left in Table 1 show the root-mean-square deviation (rmsd) of the model that is closest to the native structure and the model's rank by the physics-based score.

Table 2. Fitting Results Using Different Order (n) to Generate the 3D Zernike Descriptors

		10 Å EM rank[a]			15 Å EM rank		
PDB	rank by energy[a]	$n = 20$	$n = 24$	$n = 28$	$n = 20$	$n = 24$	$n = 28$
1A0R	1	1	1	1	1	1	1
1B9X	1	1	1	1	1	1	1
1K6N	1	1	1	1	1	1	1
1VCB	113	1	1	1	1	1	1
2AZE	1	1	1	1	1	1	1
2PRG	141	1	1	1	1	1	1
1ES7	4	1	1	1	1	1	1
1GPQ	1	1	1	1	1	1	1
1K2X	68	170	168	163	168	168	173
1LOG	63	1	1	1	1	1	1
1NNU	4	1	1	1	1	1	1
1QGW	4	1	2	2	1	1	1
1RHM	1	2	1	1	1	1	1
1WWW	54	17	18	18	17	16	15
2BBK	1	1	1	1	1	1	1
6RLX	171	84	87	88	23	31	33
1W88	1	1	1	1	1	1	1
1I3O	57	7	12	8	3	2	2
1JYO	55	51	60	63	73	73	71

[a] The rank by energy and results of ranking for EM maps with $n = 20$ are the same as those shown in Table 1.

For 15 out of 19 cases, near native structures with an rmsd of less than 4.0 Å to the native were obtained. The 15 cases include 14 highly accurate structures whose rmsd is less than 2.5 Å. The physics-based score ranked a near native structure (rmsd of <4.0 Å) at the top in 7 out of the 15 cases, and within the fifth rank in 10 cases. Although the best (i.e., closest to native) models for the other four cases, 1K2X, 6RLX, 1W88, and 1JYO, have an rmsd of over 4.0 Å, the topologies (interactions of subunits) are almost correct. In the cases of 1W88 and 6RLX, their subcomplexes excluding one subunit were predicted within 4.0 Å rmsd. The subcomplex composed of chains A, B, C, and D in 1W88 was predicted at 1.28 Å rmsd and a slight displacement of chain I increased the overall rmsd. The second case, 6RLX, shows that the subcomplex with chains B, C, and D was predicted at 3.22 Å rmsd, and the displacement of chain A was the cause for a higher overall rmsd of 4.49 Å. Similar results were observed for 1JYO (six chain complex) and 1K2X (four chain complex). For 1JYO, an rmsd of 3.83 Å was observed for the subunit with four chains, A, B, D, and E. The rmsd increased to 4.75 Å when chain C is also considered and an overall rmsd ended up with 6.43 Å when all chains are taken into account. For the last case, 1K2X, the pairwise pose with chains A and B was predicted at 1.13 Å while C and D at 1.55 Å, and the union of these two subcomplexes yielded the medium quality overall rmsd of 7.53 Å. Therefore, overall Multi-LZerD managed to generate a near-native complex structure for the majority of the cases; however, not all of them were ranked high by the physics-based score among the 200 candidate structures. In the following section, we discuss how well the near-native structure models are ranked by using the target EM map and the 3DZD.

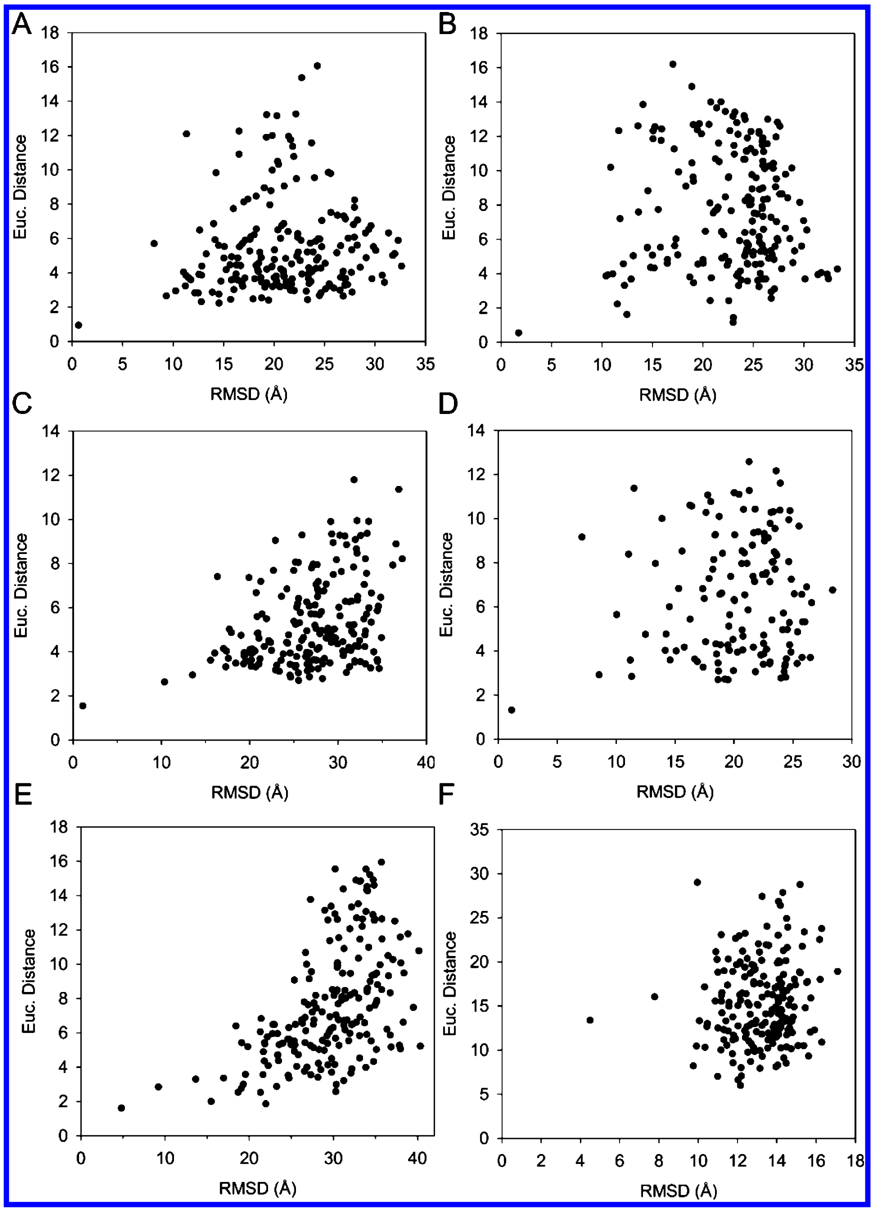


Figure 4. Euclidean distance between the 3DZD of the complex models and the EM map relative to the rmsd of the complex models: (**A**) 1B9X; (**B**) 1GPQ; (**C**) 1K6N; (**D**) 1NNU; (**E**) 1W88; (**F**) 6RLX. The complex structures from the final generation of the GA optimization are plotted. EM maps of 10 Å resolution were used.

Fitting Complex Structure Models to EM Maps. For each test case, 200 multimeric complex structure models were ranked according to their fitness to the EM map. The fitness of the models was quantified with the Euclidean distance between the 3DZD coefficients computed for the overall surface shape of the models and the isosurface of the EM map. The two rightmost columns in Table 1 show the ranks of the near native structures using the EM maps at 10 and 15 Å resolutions.

For the results using EM maps at 10 Å resolution, remarkably, the model closest to native was selected at the top rank in 14 out of the 19 cases. In addition, the best model for 1RHM and 1I3O, with rmsd values of 1.07 and 2.07 Å, was selected at the second and the seventh rank, respectively. However, the other four cases show a less successful ranking for the best structure model. This is caused mainly because the best available structures are not very accurate; their rmsd was in the ∼5 Å+ range for three out of the four cases (1K2X, 6RLX, 1JYO). Thus, the problem lies rather in the generation of accurate multimeric complex structures and not in the effectiveness of the 3DZD-based comparison. Indeed, the best docking model was selected as the top rank in all cases except one (1WWW, ranked 17) when the model's rmsd to native is closer than 2.5 Å. These ranks for the best structure models are, needless to say, better than the ranks obtained by the physics-based score without using the EM maps. Encouragingly, even slightly better results were obtained when 15 Å resolution EM maps were used. This indicates that EMLZerD can be applied even when the resolution of available EM maps is as low as 15 Å.

In Figure 3, six examples of fitted subunit structures into EM maps are shown. The global rmsd of the fitted structure to the crystal structure is shown in Table 1. The first four are successful cases, where structure fitting was made within 0.62 to 1.74 Å. As mentioned before, the case of 1W88 suffers from a slightly misplaced chain I (shown outside the EM map), while 6RLX shows the same with misplaced chain A in this case.

To analyze the impact of different values of the order *n*, in Table 2 we show the ranks obtained using *n* = 24 and 28 in comparison with the results using *n* = 20 (as in Table 1). Using a larger order provides finer representation of the input surface shape. It is shown that results of the ranks are almost the same. Given that an order of 20 gives a more compact representation than a larger order, and that using a higher order yields similar results, the use of *n* = 20 provides better efficiency while also being accurate.

3DZD Euclidean Distance between Complex Models and EM Maps. We show the distribution of the Euclidean distance between the 3DZDs of complex models and the EM map (*y*-axis) relative to the rmsd of the models (*x*-axis) in Figure 4. Six representative cases are shown: Figure 4A–D are cases where complex structures were predicted at an rmsd of less than 2.5 Å and successfully selected with the smallest Euclidean distance. Figure 4E is an example where the multimeric docking procedure did not yield a globally accurate model but the best model of a medium accuracy (rmsd: 4.80 Å) among the structure candidate pool was ranked as the top prediction by the Euclidean distance. The last example, Figure 4F, is a case where the docking procedure obtained only a medium accuracy model (rmsd: 4.49 Å), which was ranked low by the Euclidean distance. Note that there is only one near-native model at a low rmsd range among all the generated models because similar structures were clustered at the end of the multimeric docking procedure. The Euclidean distance of the best (i.e., closest to native) complex structure models are shown in parentheses within the "EM rank" columns in Table 1.

It is shown that highly accurate models with an rmsd <2.0 Å (Figure 4A–D) are well distinguished from the rest of the structure models by the Euclidean distance of 3DZDs. Significant correlation coefficients were not observed between the rmsd and the 3DZD Euclidean distance (0.0–0.53); however, this is reasonable because most of the structures have an rmsd of 10 Å or higher, where the overall conformations are very different from native and also because there are not many near-native structures in the pool due to the clustering.

To further investigate the relationship between the 3DZD Euclidean distance and the accuracy of the fitting, we plotted the 3DZD Euclidean distance and the rmsd of the top ranked model from the 19 test cases (i.e., the complex structure model that has the smallest 3DZD Euclidean distance to the EM map) in Figure 5. When the top ranked structure model has a significantly small Euclidean distance, they are highly accurate in the majority of cases. With the Euclidean distance threshold of 1.5, the models' rmsd values are within 2.0 Å in 7 out of 9 cases (77.8%), while models at the Euclidean distance of 2.0 or smaller are predicted at better than 2.5 Å rmsd for 11 out of 15 cases (73.3%). On the other hand, the plot shows that a highly accurate model may not be expected when even the top ranked structure model has the 3DZD Euclidean distance of 2.5 or higher. There are four structures with a Euclidean distance of over 2.5, whose rmsd to native are all above 3.0 Å, namely, 3.23, 19.3, 12.2, and 21.4 Å. Thus, an advantage of the proposed method is that the 3DZD Euclidean distance can indicate the quality of the generated complex model. Figure 4 also shows that almost all the structure models with a 3DZD Euclidean distance of 2.0 or higher have a very large rmsd to native.

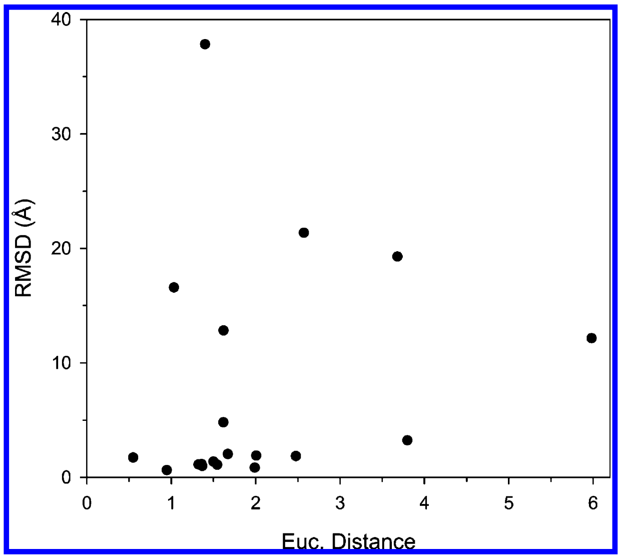


Figure 5. Models with the smallest 3DZD Euclidean distance from the 19 test cases are plotted. The EM maps at 10 Å resolution were used.

DISCUSSION

In this study we have shown that the 3DZD can effectively assess the fitness of high-resolution multimeric complex models for a given EM map. The combination of the 3DZD with a multimeric protein docking was implemented in the new method, EMLZerD, which was shown to be successful for the majority of the test cases in generating models that fit into given EM maps. Among the test cases, there are few where the method did not yield a near-native model. However, the 3DZD can indicate if a sufficiently accurate model is included in the pool of generated candidate structures or not. If the best fitting structure model (i.e., one with the smallest Euclidean distance to the EM map) still has a large distance of over 2.5, it is highly likely that an accurate model was not generated yet, thus it is worthwhile to continue to run the GA optimization for more generations, to explore further the conformational space.

Unlike existing methods that focus on local optimization of the placement of component proteins starting from preassigned anchoring positions in an EM map, our approach focuses on automatic placement of multimeric asymmetric complex simultaneously to an EM map. If additional biological or structural information of the protein complex is known beforehand, for example, symmetrical units or binding interface residues,[46] such information can be used in the multimeric docking stage to restrict the conformational space. Although the current method does not explicitly consider flexibility of protein subunits, existing methods for handling flexibility[1,16−20] can be employed for refinement of the fitted model provided by EMLZerD, especially for the cases where our final predicted model does not have a small distance to the EM map.

As more protein complexes are being solved by EM,[47,48] it is crucial to provide computational methods that aid analyzing low-resolution structures from EM by efficiently bridging available high-resolution structures and low-resolution EM data. EMLZerD, together with the other existing methods, will be a valuable tool for EM structural biology.

AUTHOR INFORMATION

Corresponding Author

*E-mail: dkihara@purdue.edu. Phone: 1-765-496-2284. Fax: 1-765-496-1189.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM075004, R01GM097528), the National Science Foundation (DMS0800568, EF0850009, IIS0915801), and a National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004). J.E.-R. is a Fulbright Science and Technology Fellow.

■ REFERENCES

(1) Falke, S.; Tama, F.; Brooks, C. L.; Gogol, E. P.; Fisher, M. T. *J. Mol. Biol.* **2005**, *348*, 219−30.

(2) Ludtke, S. J.; Chen, D.-H.; Song, J.-L.; Chuang, D. T.; Chiu, W. *Structure* **2004**, *12*, 1129−36.

(3) Rossmann, M. G.; Morais, M. C.; Leiman, P. G.; Zhang, W. *Structure* **2005**, *13*, 355−62.

(4) Fabiola, F.; Chapman, M. S. *Structure* **2005**, *13*, 389−400.

(5) Lindert, S.; Stewart, P. L.; Meiler, J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 218−25.

(6) Rossmann, M. G. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 1341−9.

(7) Wriggers, W.; Birmanns, S. *J. Struct. Biol.* **2001**, *133*, 193−202.

(8) Wriggers, W.; Milligan, R. A.; McCammon, J. A. *J. Struct. Biol.* **1999**, *125*, 185−95.

(9) Woetzel, N.; Lindert, S.; Stewart, P. L.; Meiler, J. *J. Struct. Biol.* **2011**, *175*, 264−76.

(10) Tang, G.; Peng, L.; Baldwin, P. R.; Mann, D. S.; Jiang, W.; Rees, I.; Ludtke, S. J. *J. Struct. Biol.* **2007**, *157*, 38−46.

(11) Baker, M. L.; Zhang, J.; Ludtke, S. J.; Chiu, W. *Nat. Protoc.* **2010**, *5*, 1697−708.

(12) André, I.; Bradley, P.; Wang, C.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17656−61.

(13) Berchanski, A.; Eisenstein, M. *Proteins* **2003**, *53*, 817−29.

(14) Chan, K.-Y.; Gumbart, J.; McGreevy, R.; Watermeyer, J. M.; Sewell, B. T.; Schulten, K. *Structure* **2011**, *19*, 1211−8.

(15) Rossmann, M. G.; Bernal, R.; Pletnev, S. V. *J. Struct. Biol.* **2001**, *136*, 190−200.

(16) Ahmed, A.; Whitford, P. C.; Sanbonmatsu, K. Y.; Tama, F. *J. Struct. Biol.* **2012**, *177*, 561−570.

(17) Tama, F.; Miyashita, O.; Brooks, C. L. *J. Mol. Biol.* **2004**, *337*, 985−99.

(18) Tama, F.; Miyashita, O.; Brooks, C. L. *J. Struct. Biol.* **2004**, *147*, 315−26.

(19) Grubisic, I.; Shokhirev, M. N.; Orzechowski, M.; Miyashita, O.; Tama, F. *J. Struct. Biol.* **2010**, *169*, 95−105.

(20) Tan, R. K.-Z.; Devkota, B.; Harvey, S. C. *J. Struct. Biol.* **2008**, *163*, 163−74.

(21) Zheng, W. *Biophys. J.* **2011**, *100*, 478−88.

(22) Kawabata, T. *Biophys. J.* **2008**, *95*, 4643−58.

(23) Lasker, K.; Sali, A.; Wolfson, H. J. *Proteins* **2010**, *78*, 3205−11.

(24) Canterakis, N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. *Proceedings of the 11th Scandanavian Conference on Image Analysis*; 1999; pp 85−93.

(25) Novotni, M.; Klein, R. 3D Zernike descriptors for content based shape retrieval. *Proceedings of the 8th ACM Symposium on Solid Modeling and Applications*; ACM Press: New York, 2003; pp 216−225.

(26) Sael, L.; Kihara, D. In *Biological Data Mining*; Chen, J. Y., Lonardi, S., Eds.; Chapman & Hall/CRC: Boca Raton, FL, 2009; pp 89−109.

(27) Venkatraman, V.; Yang, Y. D.; Sael, L.; Kihara, D. *BMC Bioinf.* **2009**, *10*, 407.

(28) Esquivel-Rodríguez, J.; Kihara, D. *BMC Bioinf.* **2012**, *13*, S6.

(29) Esquivel-Rodríguez, J.; Yang, Y. D.; Kihara, D. *Proteins* **2012**, in press.

(30) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. *J. Mol. Biol.* **2003**, *331*, 281−99.

(31) Andrusier, N.; Nussinov, R.; Wolfson, H. J. *Proteins* **2007**, *69*, 139−59.

(32) Meyer, M.; Wilson, P.; Schomburg, D. *J. Mol. Biol.* **1996**, *264*, 199−210.

(33) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133−52.

(34) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199−203.

(35) Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. *J. Mol. Biol.* **1997**, *267*, 707−26.

(36) Kihara, D.; Sael, L.; Chikhi, R.; Esquivel-Rodríguez, J. *Curr. Protein Pept. Sci.* **2011**, *12*, 520−30.

(37) Hawkins, T.; Chitale, M.; Kihara, D. *Mol. Biosyst.* **2008**, *4*, 223−31.

(38) La, D.; Esquivel-Rodríguez, J.; Venkatraman, V.; Li, B.; Sael, L.; Ueng, S.; Ahrendt, S.; Kihara, D. *Bioinformatics* **2009**, *25*, 2843−4.

(39) Sael, L.; Kihara, D. *Proteins* **2012**, *80*, 1177−95.

(40) Sael, L.; Kihara, D. *BMC Bioinf.* **2012**, *13*, S7.

(41) Chikhi, R.; Sael, L.; Kihara, D. In *Protein function prediction for omics era*; Kihara, D., Ed.; Springer: New York, 2011; pp 145−163.

(42) Venkatraman, V.; Chakravarthy, P. R.; Kihara, D. *J. Cheminform.* **2009**, *1*, 19.

(43) Sael, L.; Li, B.; La, D.; Fang, Y.; Ramani, K.; Rustamov, R.; Kihara, D. *Proteins* **2008**, *72*, 1259−73.

(44) Ceulemans, H.; Russell, R. B. *J. Mol. Biol.* **2004**, *338*, 783−93.

(45) Sael, L.; Kihara, D. *BMC Bioinf.* **2010**, *11* (Suppl 1), S2.

(46) Li, B.; Kihara, D. *BMC Bioinf.* **2012**, *13*, 7.

(47) Tagari, M.; Newman, R.; Chagoyen, M.; Carazo, J. M.; Henrick, K. *Trends Biochem. Sci.* **2002**, *27*, 589.

(48) Henrick, K.; Newman, R.; Tagari, M.; Chagoyen, M. *J. Struct. Biol.* **2003**, *144*, 228−37.