PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Large oligomeric complex structures can be computationally assembled by efficiently combining docked interfaces

Matthias Dietzen,[1] Olga V. Kalinina,[1]* Katerina Taškova,[2,3] Benny Kneissl,[2,4] Anna-Katharina Hildebrandt,[5] Elmar Jaenicke,[6] Heinz Decker,[6] Thomas Lengauer,[1] and Andreas Hildebrandt[2]

[1] Max Planck Institute for Informatics, Campus E1 4, Saarbrücken 66123, Germany

[2] Institute of Computer Science, Johannes Gutenberg University, Staudingerweg 9, Mainz 55128, Germany

[3] Institute for Molecular Biology, Johannes Gutenberg University, Ackermannweg 4, Mainz 55128, Germany

[4] Roche Pharma Research and Early Development, pRED Informatics, Roche Innovation Center Penzberg, Nonnenwald 2, Penzberg 82377, Germany

[5] Center for Bioinformatics, Saarland University, Campus E2 1, Saarbrücken 66123, Germany

[6] Institute of Molecular Biophysics, Johannes Gutenberg University, Jakob-Welder-Weg 26, Mainz 55128, Germany

ABSTRACT

Macromolecular oligomeric assemblies are involved in many biochemical processes of living organisms. The benefits of such assemblies in crowded cellular environments include increased reaction rates, efficient feedback regulation, cooperativity and protective functions. However, an atom-level structural determination of large assemblies is challenging due to the size of the complex and the difference in binding affinities of the involved proteins. In this study, we propose a novel combinatorial greedy algorithm for assembling large oligomeric complexes from information on the approximate position of interaction interfaces of pairs of monomers in the complex. Prior information on complex symmetry is not required but rather the symmetry is inferred during assembly. We implement an efficient geometric score, the transformation match score, that bypasses the model ranking problems of state-of-the-art scoring functions by scoring the similarity between the inferred dimers of the same monomer simultaneously with different binding partners in a (sub)complex with a set of pregenerated docking poses. We compiled a diverse benchmark set of 308 homo and heteromeric complexes containing 6 to 60 monomers. To explore the applicability of the method, we considered 48 sets of parameters and selected those three sets of parameters, for which the algorithm can correctly reconstruct the maximum number, namely 252 complexes (81.8%) in, at least one of the respective three runs. The crossvalidation coverage, that is, the mean fraction of correctly reconstructed benchmark complexes during crossvalidation, was 78.1%, which demonstrates the ability of the presented method to correctly reconstruct topology of a large variety of biological complexes.

Proteins 2015; 83:1887–1899.

Key words: macromolecular assembly; structural modeling; protein–protein interactions; transformation match score; complex match score; 3D-MOSAIC.

INTRODUCTION

Protein complexes mediate many essential processes in the cell. Often, very large multimeric protein complexes are formed for regulating the metabolic processes, nutrient delivery or defense mechanisms. Protein monomers can aggregate with assistance of molecular chaperones[1]

Additional Supporting Information may be found in the online version of this article.
Abbreviations: *cms*, complex match score; 3D-MOSAIC, 3-dimensional modeling of oligomeric structural assemblies based on pairwise interaction combination; RMSD, root-mean-square deviation; *tms*, transformation match score; tRMSD, topology RMSD

Grant sponsor: International Max Planck Research School for Computer Science (IMPRS-CS; M.D.), Saarbrücken, Germany.
Institution at which the work was performed: Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany.
*Correspondence to: Olga V. Kalinina, Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany. E-mail: kalinina@mpi-inf.mpg.de
Received 11 March 2015; Revised 20 July 2015; Accepted 29 July 2015
Published online 6 August 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24873

or in a self-assembly process in which they find their position within the protein complex without any further help.[2] However, neither the processes in chaperone-assisted assembly nor the molecular principles of self-recognition and the hierarchical order of the association process are well-understood. In the latter case, hydrophobic interactions are often involved[3–5]; in some cases electrostatic interactions guide the process.[6] To obtain a deeper understanding of the biophysical basis of the assembly processes, bioinformatics tools such as protein–protein docking programs are often used.

The assembly of oligomeric complexes from protein monomers resembles solving a three-dimensional jigsaw puzzle. Yet, in contrast to a real jigsaw puzzle, protein interfaces, that is, the surface patches that can interact with a binding partner, are not as well defined and only roughly complementary. If their location is known approximately, for example, from crosslinking,[7–9] correlated mutation studies,[10–13] clustering of global docking poses to identify potential binding modes,[14,15] or databases such as Interactome3D,[16] KBDOCK,[17] or PRISM,[18] the corresponding interface areas can be locally probed more extensively using docking methods.[19] However, although state-of-the-art algorithms usually yield near-native solutions, the employed scoring functions typically fail to appropriately rank these poses and distinguish them from decoy solutions.[19,20]

Hence, the computational assembly of large protein complexes is challenging and the development of algorithms for solving this problem has received attention only in the last decade. Few multi-body docking approaches exist, the most prominent being HADDOCK,[21] an information-driven docking algorithm that allows the simultaneous docking of up to six protein monomers. Many algorithms incorporate symmetry information to restrain the combinatorial space, for example, ClusPro,[22] SymmDock,[23] Rosetta's symmetry docking protocol,[24] M-ZDock,[25] or a particle swarm optimization-based method which predicts homo-oligomers of up to 24 monomers.[26] Methods not relying on symmetry information also exist: DockTrina[27] can compute asymmetric trimers by scanning combinations of protein dimers via an RMSD-based test; MDOCK_HEX[28] and CombDock[29] use pairwise dockings to compute clash-free minimum-weight spanning trees (based on docking scores) representing protein complexes. Other methods employ genetic algorithms and Monte-Carlo refinement during assembly from pairwise dockings,[30] or assemble complexes using interaction data predicted by structural matching of protein–protein interfaces.[31]

In this work, we present 3D-MOSAIC (**3**-dimensional modeling of oligomeric structural assemblies based on pairwise interaction combination), a novel, time-efficient combinatorial algorithm that employs a tree-based greedy scheme for assembling protein complexes from docked complexes of pairs of monomers. To deal with the ranking problem typical for commonly used scoring functions,[19,20] we introduce a novel measure, called transformation match score (*tms*), that scores (sub)complexes solely based on the compatibility of pairwise complexes produced for each pair of interacting monomers with a docking algorithm of the user's choice (RosettaDock[32] in this study).

A similar idea has been proposed in DockTrina,[27] where however the authors limit themselves to considering only trimers and thus evade the largest part of the combinatorial burden. DockTrina also exploits the idea to reward implicitly produced interfaces compatible with pregenerated docking poses, and does not rely on advance information on interaction interfaces. Unlike CombDock,[29] which does not require information on these interfaces either, we explicitly use such information to assemble a complex by successive attachment of monomers and perform a greedy search in order to find the correct complex topology. Our algorithm does not rely on a priori symmetry information, but rather infers symmetry during assembly and optimizes the complexes accordingly. The successful validation on a diverse benchmark set of 308 complexes with 6 to 60 monomers and up to 15 different protein types involved in complex formation shows that 3D-MOSAIC considerably extends the limitations of previous tools. 3D-MOSAIC is implemented in BALL[33] and is currently limited to PDB files (<63 chains and 100,000 atoms), but will be extended to other file formats. The algorithm requires the knowledge of the stoichiometry of the complex, of three-dimensional structures of all distinct monomers, and of the approximate location of the interaction interfaces for each pair of monomers in contact, from which it generates a set of candidate docking poses using an established docking algorithm. The availability of the latter information from experiment or prediction currently presents the major limitation of the proposed algorithm. On the one hand, the relevant experimental data may be hard to come by, on the other hand, our tests show that in the absence of such data, current pairwise docking algorithms often do not find near-native poses. With the progress in this area, we expect 3D-MOSAIC to become applicable in cases when no advance information on the location of interaction interfaces is available.

MATERIALS AND METHODS

Transformation match score

The central idea of 3D-MOSAIC is the transformation match score (*tms*). In a complex, each monomer typically interacts with multiple binding partners via different interfaces. 3D-MOSAIC uses a set of given pairwise complexes of all involved monomers, and if this set includes poses corresponding to the near-native interactions between the monomers, it is possible to find a rigid transformation that superimposes a pair of monomers in the complex onto a suitable docking pose. If we continue this process for other

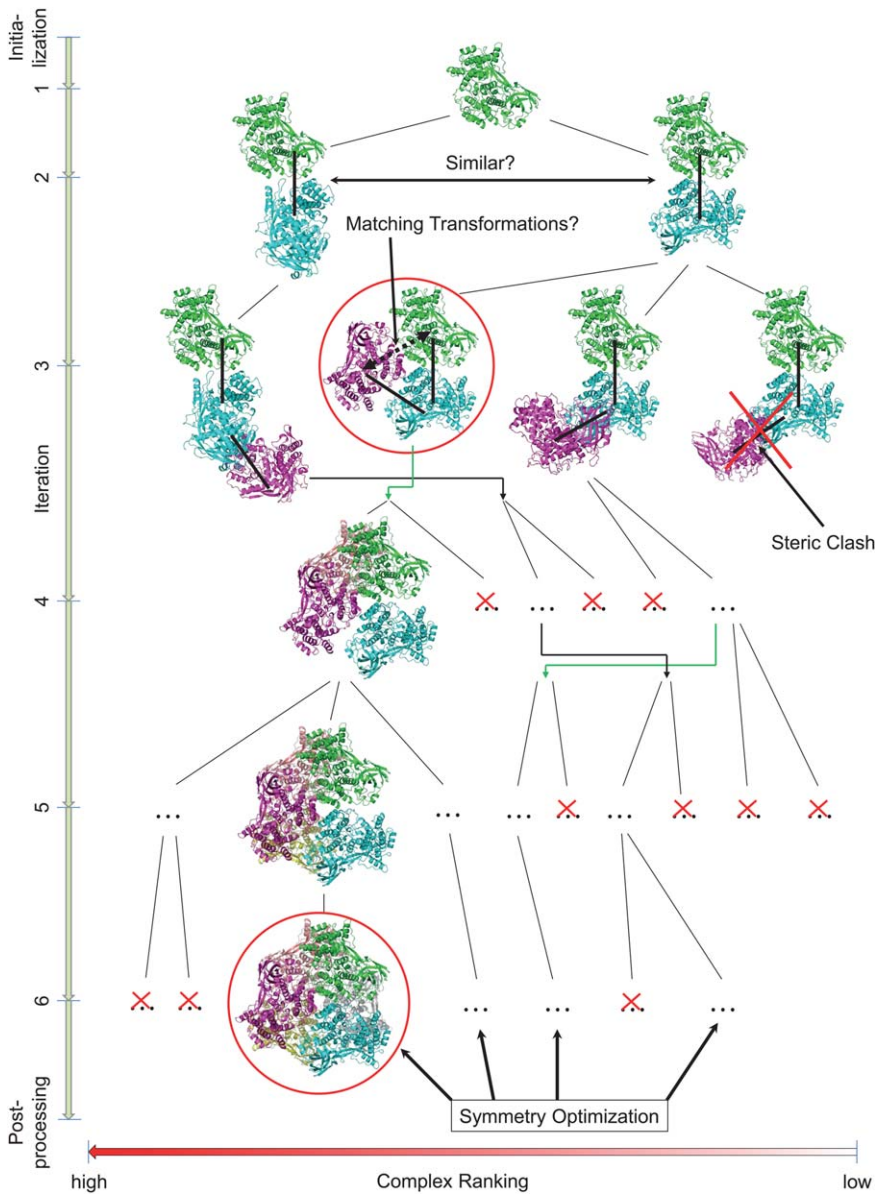

Figure 1

Exemplary assembly of the homo-hexameric hemocyanin from Panulirus interruptus (PDB code 1HCY) using 3D-MOSAIC. In each iteration, new monomers can be attached to all previously retained solutions. If a matching interface is found, the complex match score increases and the corresponding complex might be ranked further up in the list of solutions (green double-tilted arrows). Solutions similar to better-ranked ones or yielding severe steric clashes are discarded. After complex construction, a symmetry optimization can be performed. Complex images created with PyMOL.[41]

monomers, using one of the monomers of the preassembled partial complex as an anchor, and adding a new monomer consistent with a docking pose out of the set of pairwise dockings, each new monomer will have an associated transformation $\mathbf{T}^A$. However, if the added monomer forms an additional interface with another monomer, another rigid transformation $\mathbf{T}^B$ can be associated with it.

The core of 3D-MOSAIC comprises a novel and quickly computable score for measuring the pairwise similarity of any two such rigid transformations $\mathbf{T}^A$ and $\mathbf{T}^B$, *tms*. This score is exact in a very fundamental sense, since it is based on RMSD. Furthermore, it is efficiently computable and easy to interpret.

Let $\mathbf{t}^A, \mathbf{t}^B$ be the translations and $\mathbf{R}^A, \mathbf{R}^B$ the rotations associated with transformations $\mathbf{T}^A$ and $\mathbf{T}^B$, respectively. As we have previously shown,[34] the RMSD between two rigid transformations $\mathbf{T}^A$ and $\mathbf{T}^B$ associated with two docking poses of a protein P can be calculated in constant time, that is, the computing time does not depend on the number n of atoms involved. Using the protein's

covariance matrix $\mathrm{cov}(P)$ which has to be precomputed only once, the RMSD between two rigid transformations $\mathbf{T}^{\mathrm{A}}$ and $\mathbf{T}^{\mathrm{B}}$ involving P can be computed as:

$$\mathrm{RMSD}(\mathbf{T}^{\mathrm{A}},\mathbf{T}^{\mathrm{B}})=\sqrt{|\mathbf{t}^{\mathrm{AB}}|^2+\frac{1}{n}\mathrm{tr}(\mathbf{R}^{\mathrm{AB}}\cdot\mathrm{cov}(P))} \quad (1)$$

where $\mathbf{R}^{\mathrm{AB}} := \mathbf{R}^{\mathrm{B}}-\mathbf{R}^{\mathrm{A}}$. The elements of the covariance matrix are defined as $\mathrm{cov}(P)_{ij}=\sum_{k=1}^{n} x_{ki}x_{kj}$, where x_{ki} and x_{kj} are the i-th and j-th coordinates of atom k, and $i,j=1,2,3$.

Letting $\mathrm{rmsd}_{\max}$ be the threshold for the maximum RMSD under which $\mathbf{T}^{\mathrm{A}}$ and $\mathbf{T}^{\mathrm{B}}$ are considered similar, we define the *tms* as follows:

$$S^{\mathrm{rmsd}}(\mathbf{T}^{\mathrm{A}},\mathbf{T}^{\mathrm{B}}) := \left(\max\left(1-\frac{\mathrm{RMSD}(\mathbf{T}^{\mathrm{A}},\mathbf{T}^{\mathrm{B}})}{\mathrm{rmsd}_{\max}},0\right)\right)^2 \quad (2)$$

S^{rmsd} is 1 for identical transformations and decreases toward the value 0 obtained when the cutoff $\mathrm{rmsd}_{\max}$ is reached or exceeded. To speed up the calculations, we introduce an additional parameter $l_{\max}$ that represents the displacement of centers of masses resulting from transformations $\mathbf{T}^{r,d}$ and $\mathbf{T}^{a,d_a}$. If this parameter exceeds a threshold, there is no need to calculate S^{rmsd}, since the transformations are clearly too dissimilar.

An additional score that employs translational and rotational displacement instead of RMSD was developed.[35] It produced similar results on the benchmark that are not reported here.

Outline of 3D-MOSAIC

3D-MOSAIC requires high-resolution three-dimensional structures of a representative of each protein involved in forming the complex (hereafter, protein types), information on the stoichiometry of the complex, that is, the multiplicity of each protein type, and pairwise interfaces that provide the presumed binding modes in the complex. 3D-MOSAIC assembles the complex in an iterative tree-based greedy fashion with each node representing a monomer attached in a particular orientation (Fig. 1): starting from a seed monomer with the largest number of interfaces (identified from pairwise docking poses after clustering) of all protein types as the initial parent solution, in each iteration, the algorithm generates new child solutions, that is, partial or subcomplexes by attachment of an additional monomer to each of the parent solutions retained from the previous iteration.

For monomer attachment, each monomer r in the parent solution is considered a potential interaction partner. A new monomer l of a particular protein type p can be attached to r, if i) the number of occurrences of p in the parent solution has not yet reached its maximum multiplicity, and ii) r has unoccupied interfaces (that can be deduced from docking poses) for an interaction with a protein of type p. Each docking pose associated with such an interface is considered to be a new child solution, if the placement of the new monomer l of type p according to that pose does not lead to a severe steric clash of l with other monomers already present in the parent solution. Particularly, we consider the aggregate transformation $\mathbf{T}^{r,d} := \mathbf{T}^{r}\cdot\mathbf{T}^{d}$, consisting of the transformation $\mathbf{T}^{r}$ given by node r and the docking pose transformation $\mathbf{T}^{d}$. The new child monomer l is scored according to the number of interfaces it has with all ancestor monomers a already present in the complex. We investigate each a for a docking pose d_a, such that the aggregate transformation $\mathbf{T}^{a,d_a}$ is maximally similar with the transformation $\mathbf{T}^{r,d}$ of the newly attached monomer l, that is,

$$d_a := \underset{d'\in\text{all docking poses of all interfaces}}{\arg\max} S^{\mathrm{rmsd}}(\mathbf{T}^{r,d},\mathbf{T}^{a,d'}) \quad (3)$$

We define the complex match score (*cms*) $S(l)$ of a child node l as the sum of the *cms* of its parent node r and the obtained *tms* for all poses d_a over all ancestors a except the monomer r to which l was attached by docking pose d:

$$S(l) := 1+S(r)+\sum_{\text{all ancestral nodes } a\neq r} S^{\mathrm{rmsd}}(T^{r,d},T^{a,d_a}) \quad (4)$$

The additional summand 1 accounts for the attachment of l via r using d for which d itself already yields a "perfect" transformation similarity score. The complex match score of the root node is zero. In case no additional interfaces are established, and the *cms* is equal for all solutions (for example, first attachment step), the solutions are ranked according to the score produced by the external pairwise docking algorithm for the respective docking poses. The position of the monomers, to which the additional interfaces have been formed (if any), can be adjusted by interpolating between the ones obtained from transformations $\mathbf{T}^{r,d}$ and $\mathbf{T}^{a,d_a}$ for all ancestors a.

After each iteration, the generated child solutions are clustered based on C_α RMSD to ensure a diverse solution set in each iteration. Starting with the top-ranking solution as the first representative, each subsequent solution is compared to each previously retained representative: if the C_α RMSD of the RMSD-minimizing mapping (Supporting Information, Section 1.1) of the monomers of the new solution to the representative is below a threshold, the solution is discarded, otherwise it is added to the set of representatives. This procedure is iterated until a user-defined number of diverse representatives has been found.

After the final iteration, a symmetry optimization is attempted for each complex (Supporting Information, Section 1.2), provided that no steric clashes are thus

introduced. First, all possible nontrivial superimpositions of the complex onto itself, which map identical monomers onto each other and produce RMSD below a threshold, are identified. From all such superimpositions, the final placement of each monomer is then averaged, yielding a symmetry-optimized solution.

Benchmark data set

To validate 3D-MOSAIC, we established a diverse, representative, high-quality benchmark set of protein complexes obtained from the protein data bank (PDB)[36] containing 52,112 structures. For each type of protein in each complex, we determined the monomers with the same sequence. Chains with sequence differences between RESSEQ entry in the PDB file and actual structure, missing internal loops, mutations or nonstandard residues were excluded from the set of candidates. We also excluded structures with steric clashes, containing multiple connected subcomplexes, nucleic acids, or antibodies, split into several PDB entries and containing hetero groups that cannot be handled by RosettaDock. For structures containing complexes of several protein chains, we selected those with more than six chains, and then searched the rest of the PDB for sequence-identical monomers. The resulting dataset consists of 350 complexes, of which nine contain monomers also found in a structure with only one protein chain (unbound category), 10 contain monomers that are also found as dimers in the PDB (dimer category), 122 contain monomers that are found also in other multimeric complexes (foreign category), and 209 contain unique monomers not found in any other complex structure (same category).

The binding modes, that is, the representative mutual placement of each pair of interacting residues, in each complex were determined via structural alignment to the representative sequence-identical chains with best resolution and structural quality. If there were at least 20 pairwise distances below 10 Å between C_α atoms of residues of the two representative chains, a dimer corresponding to a potential binding mode was recorded. All such dimers per complex were subsequently clustered using a C_α RMSD of 5.0 Å to identify unique binding modes. From each cluster, the dimer with the smallest number of steric clashes was kept as the representative of a unique binding mode. Several complexes contained a few interfaces that were smaller than 20 contacting residue pairs, and such interfaces were removed. In 42 cases, this led to a complex falling apart into disjoint components of less than six monomers, and such complexes were discarded. The final set comprises 308 complexes (1044 unique binding modes) with 9, 8, 108, and 183 complexes in the unbound, dimer, foreign, and same categories, respectively. Complexes in the dataset are symmetric and asymmetric, contain from 6 to 60 monomers and up to 15 protein types. For each representative chain, we created a randomly rotated copy of the corresponding protein structure centered at the origin which will be used for assembly with 3D-MOSAIC.

Dimer preparation and docking

All representative dimers determined in the previous section were prepared using RosettaDock's[32] prepack protocol, and 10,000 docking poses per binding mode were generated using RosettaDock standard parameters (-dock_pert 3 8, -spin, -ex1, and -ex2aro) for local docking in low-resolution mode (side chains represented by centroid atoms). The all-atom refinement stage was skipped because optimal side-chain and rigid-body orientations of the dimers can be expected to differ from those in the actual complex.

3D-MOSAIC can then assemble protein complexes based on the docking transformations and interaction energies associated with these docking poses. To obtain the interaction energies, we rescored all poses using RosettaDock's cen_std weights and subtracted the internal energies of each binding partner. In doing so, we avoid multiple contributions from the internal monomer energies to the total complex energy upon assembly.

Each pairwise docking pose describes two relative placements which can be expressed as transformation matrices: the placement of the second monomer relative to the first one and vice versa, depending on which monomer is considered to act as the reference (hereafter receptor) to which the other one is bound (hereafter ligand). In the case of a heterodimer consisting of two monomers m_A and m_B with nonidentical protein types A and B, respectively, one docking pose describes a relative placement of (ligand) m_B with respect to (receptor) m_A and, analogously, a relative placement of (ligand) m_A with respect to (receptor) m_B. In the case of a homodimer, the docking pose describes two placements for the same protein type if the docking pose is asymmetric, otherwise one. 3D-MOSAIC identifies a binding mode as symmetric, if at least 1% of the poses associated with the binding mode were symmetric, that is, could be superimposed on themselves with RMSD less than 0.5 Å. In these cases, all nonsymmetric docking poses were disabled. In the course of the algorithm, 3D-MOSAIC was free to use any of these placements for a potential attachment of a new monomer to a particular subcomplex, depending on the protein types present and the sites of attachment available in this subcomplex. From each of these placements, the corresponding interaction interface and transformation were defined.

Assembly experiments on the benchmark

The algorithm contains a large number of parameters that were thoroughly explored. C_α RMSD in clustering

after the first, all intermediate and last iterations were set to either 1.0, 2.0, and 3.0 Å or 1.0, 3.0, and 5.0 Å, respectively. The number of tolerated clashes [37] was set to 10, 25, 50, or 150, interpolation of the monomer placement with respect to matching docking transformations was either disabled or enabled. For S^{rmsd} [Eq. (2)] thresholds for displacement-based prefiltering (l_{max}) and $rmsd_{max}$ were set to 1.0 Å/3.0 Å, 1.5 Å/4.5 Å, and 2.5 Å/7.5 Å (see Supporting Information Section 2.1 for details).

All runs employed a so-called solution reduction scheme, that is, they considered 2000 solutions in the first iteration, reduced by a factor of 2 in each subsequent iteration, with a threshold of at least 100 poses to consider per iteration. Parameters for large assemblies with 20 or 40 monomers were changed to reduce the required computational time: after adding the 20th monomer, the cluster parameters were reduced by a factor of 5 and the number of solutions to retain per iteration was reduced to 50. After placement of the 40th monomer, cluster parameters were reduced by 50%, the number of solutions to be kept per iteration was decreased to 25. In the first 20 iterations, all docking poses were enabled for attachment; after 20 (40) levels, only the 500 (250) poses of each interface yielding the highest *tms* were enabled.

In total, 48 combinations of parameters were explored. The total number of runs of 3D-MOSAIC on the 308 benchmark complexes thus amounts to 14,784.

Topology RMSD

Routinely, the quality of a modeled structure is assessed using C_α RMSD from a reference. However, this measure has two disadvantages in the context of our approach. First, when using only one representative monomer structure per type of protein involved in the complex, a certain amount of the measured C_α RMSD will be due to conformational differences in the representative structure and the corresponding monomers in the reference complex used for validation. Although this is already true when using a representative monomer from the reference complex itself, the effect becomes even more dominant if monomers from a different complex or an unbound structure are used for assembly. Second, due to the iterative nature of complex assembly from pairwise dockings, each docking pose which is not identical to the native binding mode will introduce a certain amount of error, depending on its deviation from the ideal interaction geometry. During iterative assembly such errors will accumulate, and can result in large C_α RMSDs, especially for complexes with many components, even though the complex topology is correct and the differences between the individual native dimers in the reference complex and their respective counterparts in the modeled complex are small. Also, this renders the measure incomparable between complexes of different size.

Here, we introduce another measure for comparison with a reference complex called topology RMSD (tRMSD), which is inspired by the iRMSD[38] for protein dimers. To compute iRMSD, a protein is represented by seven anchor points: its centroid and six points at ±5.0 Å in *x*-, *y*-, and *z*-direction. The RMSD values for this reduced representation (iRMSD) have been demonstrated to be more robust with respect to conformational differences between the compared structures, particularly if these changes are located in regions which do not contribute to the interaction.[38] We extend this measure to oligomers as follows: for each protein type, the relevant seven points are computed. For each pair of monomers interacting in the reference, the corresponding matched monomers from the complex are determined, and the RMSD between the anchor points of the respective dimers is computed. Finally, tRMSD is obtained as the mean RMSD between the anchor points of all dimers. The tRMSD thus assesses the correctness of the relative position of the interacting monomers compared to the reference, while ignoring conformational differences in those areas of the proteins that do not participate in any binding mode. We consider a complex to be correctly reconstructed if its tRMSD from the reference is at most 2.5 Å.

Computational resources and availability

All docking and assembly experiments were performed on the high-performance cluster MOGON (Johannes Gutenberg University, Mainz, Germany), consisting of 535 nodes, each with 4 CPUs and 16 cores per CPU, clocked with 2.1 GHz. 3D-MOSAIC will be available officially as part of the next release version of the open-source project BALL.[33] A pre-release version and the benchmark data set (1.2 GB) are available upon request.

RESULTS

Benchmark performance and crossvalidation

Of 308 benchmark complexes comprising 9, 8, 108, and 183 complexes in the unbound, dimer, foreign, and same categories, respectively, 267 (86.7%) could be reconstructed correctly. Specifically, for each of those 267 complexes there is a parameter combination such that the structure model generated with this combination deviated from the reference complex with a tRMSD score not greater than 2.5 Å, and ranked within the top 100 solutions. However, owing to limitations in computational resources, the number of parameter combinations that can be tested in a real application scenario is small. We thus performed an exhaustive search and determined the combination that provides the best coverage. It enables reconstruction of 71.8% of complexes in the benchmark. The respective parameter settings are: clustering C_α RMSD of 1 Å after the first, 5 Å after the last, and 3 Å

Table I
Performance of 3D-MOSAIC in the Benchmark

Parameter setting	*N* (cov)	cov_{cv}	*tms* disabled *N* (cov)	*tms* and clustering disabled *N* (cov)
Best one	221 (71.8)	69.1	110 (35.7)	60 (19.5)
Best two	245 (79.5)	76.6	125 (40.6)	69 (22.4)
Best three	252 (81.8)	78.1	128 (41.6)	73 (23.7)

Number *N* (and coverage cov [%]) of the benchmark complexes reconstructed using the best one, two or three combinations of parameters, with corresponding crossvalidation coverage (cov_{cv} [%]) rates.

after each intermediate iteration, 150 clashes (as defined in Ref. 37) for each pair of monomers, prefiltering displacement threshold 2.5 Å and RMSD threshold for *tms* calculation 7.5 Å thus being the most error-tolerant combination. In the case that no information is available on the nature of the interactions in a complex, it is also a common practice to run an algorithm several times with modified parameters. Thus, we have also identified three sets of parameters that jointly produce optimal results in terms of coverage (see Supporting Information, Section 2.1). In a practical scenario, one needs to run the assembly reconstruction three times using each of these settings to achieve best chances to obtain a correct solution. In addition, we carried out a 1000× 10-fold crossvalidation to investigate how well this performance generalizes to unseen data (Table I). In many cases, the correct solution is ranked among the top solutions (Table II).

The number of correctly reconstructed complexes increases to 81.8% (252) when one takes the best result from three independent runs with different combinations of parameters. The crossvalidation results show that the best-performing parameter settings determined on our benchmark are well-suited for assembly of unknown complexes: the best combination of parameters determined for randomly selected 90% of the benchmark complexes yielded mean coverages (Table I) almost as good as determined for the whole data set, with a maximum deviation of 3.7%. If we disable *tms*, the key feature of the presented algorithm, the number of correctly reconstructed complexes drops by half (Table I). Disabling clustering of solutions after each iteration leads to a further performance drop.

The generated solutions are ranked based on complex match score [Eq. (4)]. We have noticed that applying symmetrization to the generated complex and then re-ranking the solutions by favoring those with detected symmetry can often further improve the ranking (Table III). Indeed, symmetrization improves ranking of correct solution, and if 3D-MOSAIC can reconstruct a complex correctly at all, this reconstruction almost certainly is found within top 25 solutions, and in most cases appears at the top of the list.

Of all benchmark cases, complexes in unbound (nine complexes, 6–10 monomers) and dimer (eight complexes, 6–12 monomers) categories represent a setting closest to reality. Six of nine complexes in the unbound and five of eight in the dimer category could be correctly reconstructed using the best performing combination of parameters. In most cases the correct solution is ranked as first or second, the lowest rank of the first correct solution being five.

Docking results and determination of essential binding modes

Performance of 3D-MOSAIC critically depends on the quality of the generated docking poses, which we generate using RosettaDock[32] in this study. For each of the

Table II
Joint Performance of the Best Three Combinations of Parameters

Category	Number of complexes	Top 1	Top 10	Top 25	Top 100	All
Unbound	9	6	7	7	7	7
		1.00 (0.69)	1.19 (0.78)	1.19 (0.78)	1.19 (0.78)	1.19 (0.78)
Dimer	8	5	6	6	6	6
		1.00 (0.67)	1.77 (0.75)	1.77 (0.75)	1.77 (0.75)	1.77 (0.75)
Foreign	108	74	83	86	86	90
		1.00 (0.67)	1.29 (0.74)	1.88 (0.77)	2.01 (0.77)	8.06 (0.80)
Same	183	130	143	146	148	149
		1.00 (0.68)	1.36 (0.75)	1.47 (0.76)	1.87 (0.77)	3.18 (0.77)
Total	308	215	239	245	247	252
		1.00 (0.68)	1.34 (0.75)	1.62 (0.76)	1.90 (0.77)	4.84 (0.78)

Number of complexes with correct solution within top-ranked *N* solutions is reported. The mean rank of the first correct solution and the crossvalidation accuracy (in parentheses) are given in the next line. The ranks are computed by generating three ranked lists, one for each combination of parameters. Each list is ordered lexicographically with respect to symmetry, then *cms*. The resulting ranked lists are merged and items with equal rank are ordered with respect to the accumulated docking score.

Table III
Effect of Symmetry Optimization on Ranking of Solutions

Ranking by	Top 1	Top 10	Top 25	Mean Rank
cms	78.2 (3.8%)	92.9 (1.8%)	96.0 (1.0%)	4.00 ± 11.27
symmetry, *cms*	82.9 (5.4%)	94.9 (2.2%)	97.2 (1.3%)	3.18 ± 9.77

Mean percentage (and standard deviation) of correctly reconstructed benchmark complexes per parameter setting with a near-native solution among the top 1, 10, 25 ranks, as well as the mean rank of the first correctly reconstructed complex. Ranking is either based on *cms* or on a lexicographical ordering with respect to the extent of symmetry involved and then by *cms*.

1044 binding modes retained in the benchmark dataset, the minimum, median, and maximum C_α dimer RMSD of all corresponding 10,000 docking poses from the reference binding mode was determined (Fig. 2), which reveals that RosettaDock was able to find a near-native pose (C_α dimer RMSD at most 2.0 Å) for 1031 thereof, with a mean C_α dimer RMSD for the minimum-RMSD distribution of 0.654 Å and standard deviation 0.326 Å. The median and maximum RMSD distributions with mean RMSDs of 12.457 Å and 22.756 Å and standard deviations of 3.718 Å and 7.046 Å, respectively, demonstrate that the employed docking protocol generated a sufficient number of decoys to provide a reasonable test scenario for 3D-MOSAIC.

Comparison with Comeau and Camacho[22] and CombDock[29]

In addition to the crossvalidation, we evaluated the best-performing parameter settings determined in the benchmark experiment on an independent data set of 17 homo-hexamers obtained from Comeau and Camacho[22] Two of the 17 complexes (PDB codes 1I40, 1NSF) are also present in our benchmark data set, but are assembled using monomers from structures determined in independent X-ray crystallography experiments, whereas in the evaluation on the data set from Comeau and Camacho they are assembled using monomers from 1I40 and 1NSF, respectively, making the assembly problem easier. 3D-MOSAIC successfully reconstructs 12 of them.

We also applied CombDock[29] to our benchmark set. The program failed to finish in 118 cases, producing an assembly for 190 complexes. Only one complex was reconstructed correctly, that is, with tRMSD below 2.5 Å. This comparison is unfair, however, because CombDock does not employ information on the approximate location of interaction interfaces, that is essential for 3D-MOSAIC to generate docking poses. So we also refrained from using this information to conduct a fair experiment. Instead of generating docking poses from a representative dimer (cf. Section "Dimer Preparation and Docking"), we performed unconstrained global protein–protein docking of each two monomers with CombDock,[29] and used these docking poses as input for 3D-MOSAIC run with slightly relaxed parameters (see Supporting Information Section 3.1 for details). The interaction interfaces were thus not known in advance, so we consider the whole monomers as a single interface and allow multiple attachments. In this setting, 3D-MOSAIC produces no correct solutions. However, if we consider a slightly extended threshold of 5 Å for tRMSD, we find correct solutions with 3D-MOSAIC in 19 cases versus 2 with CombDock, and 3D-MOSAIC typically yields a smaller tRMSD per complex than CombDock (Fig. 3). These solutions were also usually ranked higher with 3D-MOSAIC. Generally, both methods perform poorly, which can mainly be attributed to the difficulties in generating near-native binding modes among the docking poses produced by the unconstrained pairwise docking of CombDock[29]: C_α dimer RMSD > 3.0 Å for 94% of the binding modes.

Experiments using single Residue-pair interaction constraints

As the above comparison shows, we cannot completely avoid using knowledge of approximate positions of interaction interfaces. Thus, we set out to find the minimal amount of information necessary. Literature and database searches did not provide enough data of this kind. Thus, we modeled it by assigning to each interface one pair of contacting residues: for each of the native binding

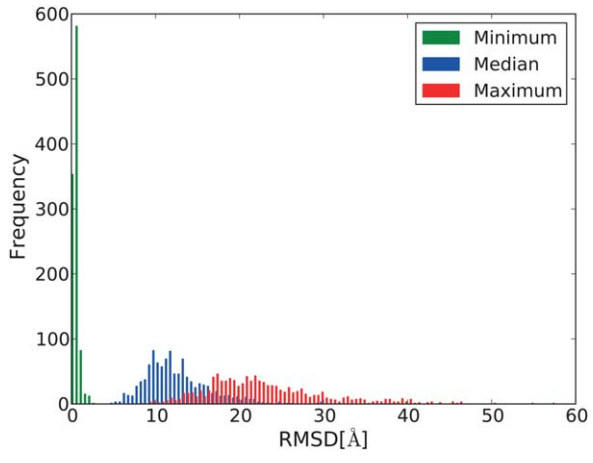

Figure 2
Histogram of docking performance over all 1044 reference binding modes: for each binding mode, the minimum, median and maximum C_α dimer RMSD from the reference mode over all 10,000 docking poses was determined.

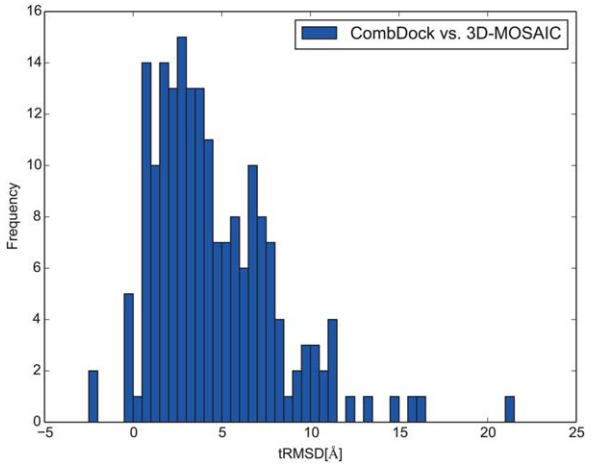

Figure 3

Distribution of difference of best tRMSDs per assembly between CombDock and 3D-MOSAIC. In only seven out of 190 cases, CombDock yielded a better tRMSD than 3D-MOSAIC (bars below zero). Images created with Matplotlib.[42]

modes, we randomly selected a pair of residues, one from each protein in the binding mode, whose C_α distance is at most 10 Å. Then we generated the start dimers by aligning the monomer centroids with the assigned contacting pair along a straight line and randomly rotating each monomer. The contacting residues were placed at 10 Å from each other. The docking poses were then generated with RosettaDock[32] using the same protocol as for the benchmark (see Section "Dimer Preparation and Docking"). Moreover, in addition to actually contacting pairs, three, six, or ten non-native contacting pairs were added to each complex to model experimental or prediction error. We performed this experiment on ten complexes, all of which could be reconstructed in the benchmark: 1HI9 (10), 1KW6 (8), 1PVV (12), 1QK1 (8), 1X1O (6), 1YNB (6), 2BJK (6), 2F1D (24), 2UYU (8), 3Q46 (6) (the number in parentheses represents the number of monomers in each).

We ran 3D-MOSAIC with slightly relaxed parameters (see Supporting Information Section 3.1 for details) and could reconstruct seven of them (Fig. 4). Three complexes: 2BJK, 2UYU, and 1PVV, could not be reconstructed, mainly due to the fact that at least one of the native binding modes was not found among pairwise docking poses for some of the interfaces. Additionally, the topology of 1PVV resembles a hollow sphere, which leads to a hardly restrained configurational space for the docking poses that can be validly attached without introducing severe steric clashes. When adding six non-native binding modes, three complexes (1KW6, 1QK1, and 3Q46) could be reconstructed, and two of them (1KW6 and 1QK1) could still be reconstructed correctly, when ten non-native binding modes were added.

Examples

Figure 5 shows some successfully reconstructed complexes whose properties emphasize some regards in which 3D-MOSAIC is superior to other methods: for example, the 20S proteasome in complex with activator PA26 (PDB code 1Z7Q, 15 protein types, 42 monomers), a protein complex that degrades proteins, could be reconstructed with a tRMSD of 0.93 Å from the reference complex using monomers from the reference complex. Similarly, the proteasome core complex (PDB code 1RYP, 14 protein types, 28 monomers, not shown) could be reconstructed with a tRMSD of 0.67 Å which is remarkable because monomers from five different sources have been used (PDB codes 1Z7Q, 1FNT, 3L5Q, 3UN4, 1VSY).

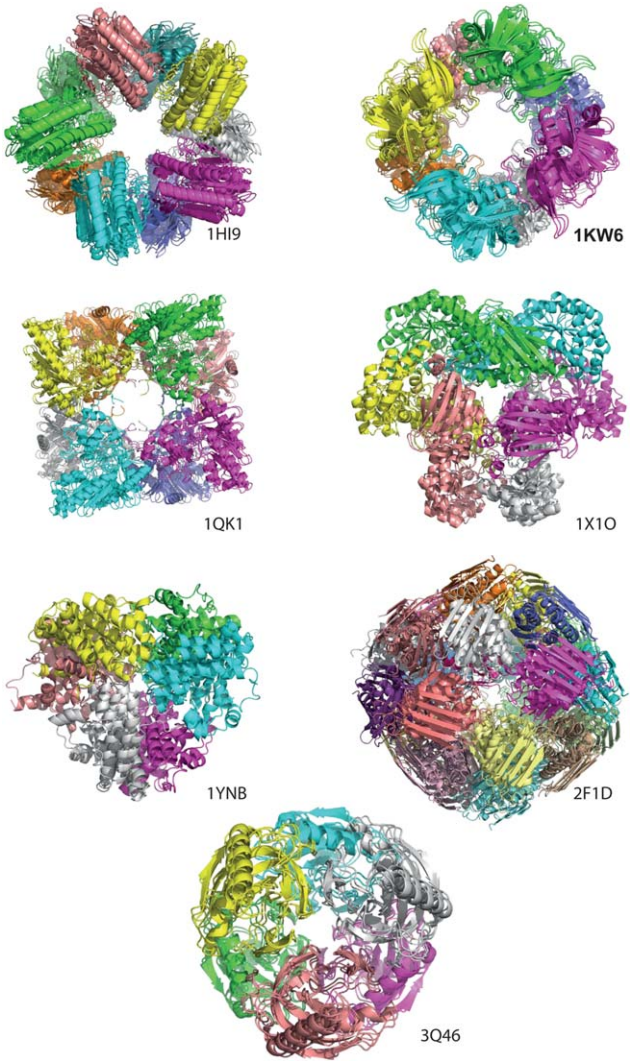

Figure 4

The seven complexes that could be reconstructed in the single residue-pair interaction constraints experiments. Each assembled complex is superimposed onto the respective reference. Complex images created with PyMOL.[41]

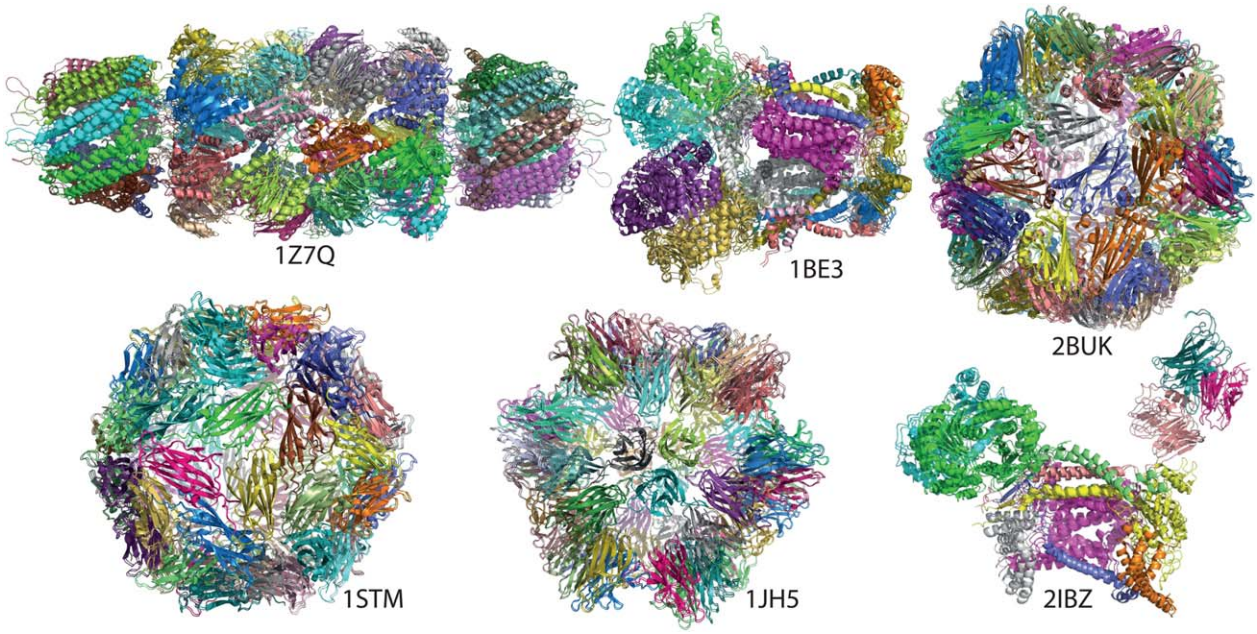

Figure 5

Examples of successfully reconstructed assemblies, superimposed onto the corresponding reference complex. Images created with PyMOL.[41]

Another interesting example is the bovine cytochrome BC1 trans-membrane complex (PDB code 1BE3, 22 monomers, 11 protein types) which is part of the final stages of energy conversion in the electron transport chain and could be reconstructed with a tRMSD of 0.91 Å. None of these complexes could be reconstructed with CombDock.[29]

3D-MOSAIC was even capable of assembling 60-mers, as, for example, the capsids of satellite panicum mosaic virus (PDB code 1STM) and satellite tobacco necrosis virus (PDB code 2BUK), reconstructed with tRMSDs of 0.23 Å and 0.31 Å, respectively, as well as a member of the tumor necrosis factor family (TNF), sTALL-1 (PDB code 1JH5), which also exhibits a capsid-like structure and could be reconstructed with an outstanding tRMSD of 0.06 Å. Asymmetric multimeric assemblies are rare in Nature, and hence depleted in our benchmark. An example of such a case is the complex of yeast cytochrome BC1 with stigmatellin (PDB code 2IBZ), which contains 11 monomers of 11 protein types and could be reconstructed with a tRMSD of 0.87 Å.

To visually demonstrate the characteristics of complexes which are more difficult to reconstruct with 3D-MOSAIC, Figure 6 shows some examples of complexes for which a successful reconstruction could not be achieved. For example, ring-like complexes with many monomers [Fig. 6 (a)] where each monomer provides two interfaces, one for each of its neighbors in the ring, can be considered to be an extreme case because of their low connectivity. In such cases, the search for similar transformations is only reasonable upon ring closure, that is, when the last monomer is attached. Here, an additional interaction with the initially placed monomer can be established, yielding a non-zero *tms*. In preceding iterations, the *cms* of all solutions are equal and 3D-MOSAIC must rely on the ranking based on the accumulated docking scores. Hence, near-native solutions must be ranked accurately in the set of docking poses for the assembly to be successful. Similarly, cage-like structures, for example, the pyruvate dehydrogenase complex (PDB code 1B5S), are hard to assemble. Here, two monomers of each of five well-connected trimers form decameric rings [Fig. 6 (b)] and, while 3D-MOSAIC easily correctly reproduces the involved trimers, their proper ring-like arrangement is hard to achieve. Complexes with monomers that are mostly helical [Fig.6 (c)] or heavily intertwined via β-sheets [Fig.6 (d)] are also difficult to assemble. Helical monomers exhibit almost no complementary surfaces and all docking poses are equally likely. In case of intertwining β-strands, the number of compatible docking poses is highly limited as the docking funnel is very narrow, and an assembly will likely lead to severe steric clashes.

DISCUSSION

3D-MOSAIC is a novel combinatorial greedy algorithm for assembling large oligomeric protein complexes from pairwise docking poses that uses a new function,

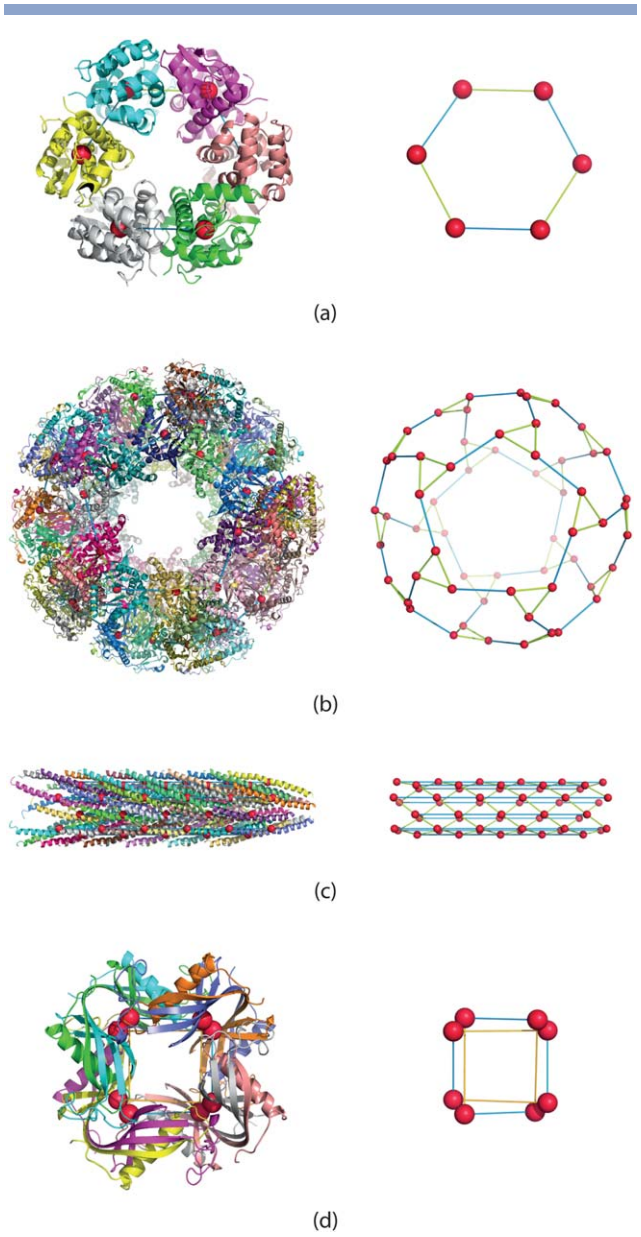

Figure 6

Examples of complexes and corresponding topology graphs for hard cases: (**a**) ring-like topology of T4 lysozyme hexamer (PDB code 3SBA), (**b**) cage-like topology of pyruvate dehydrogenase E2 60-mer core complex (PDB code 1B5S), (**c**) inovirus coat protein filament (PDB code 2C0W) composed of helical monomers, and (**d**) human cystatin C complex (PDB code 1R4C) forming interchain β-sheets. Different node colors correspond to different protein types, different edge colors to different binding modes. Images created with PyMOL.[41]

called transformation match score (*tms*), scoring the similarity between two rigid (docking) transformations of the same protein. The validation of 3D-MOSAIC on a diverse benchmark set of 308 complexes, with one single combination of parameters allowing for reconstructing 71.8% of all complexes, shows that the introduced scoring function is efficiently capable of selecting docking poses that represent native pairwise binding modes to form an oligomeric complex. 3D-MOSAIC extends capabilities of docking-based reconstruction of large complexes by increasing the number of complex components as well as the number of protein types that can be handled, with running times of a few hours for smaller complexes (up to ten monomers) to as little as approximately one day for complexes with 60 monomers, when using 10,000 docking poses per binding mode (for a detailed analysis of the running times see Supporting Information, Section 2.2).

3D-MOSAIC relies on prior knowledge of the approximate location of the binary protein interaction interfaces. This can be considered a limitation of the algorithm. 3D-MOSAIC favors complex topologies in which multiple binding modes are simultaneously established upon attachment of a new monomer. Complexes with a low degree of connectivity between the monomers can be expected to be more difficult to assemble. These complexes do not exhibit many binding modes, consequently, *tms* will rarely find matching docking poses, and thus, 3D-MOSAIC must resort to the ranking based on the sum of docking scores of the poses used for assembly. Here, the method suffers from the same problem that is common to all state-of-the-art assembly algorithms based on pairwise dockings in such a situation: common docking scoring functions are seldom able to effectively score and rank near-native solutions and to discriminate them from decoys and, thus, near-native solutions will rarely be found among the top ranks.[19,20] Due to the combinatorial nature of the complex assembly problem using pairwise dockings, viable solutions will be rapidly downranked. Although 3D-MOSAIC is often still able to assemble such complexes, it definitely performs better for well-connected complexes, for which the *tms* will, upon attachment of a new monomer, reward many surrounding monomers that form docking poses compatible with the position of the newly placed monomer. 3D-MOSAIC does not assume a high degree of connectivity between the monomers in the complexes, but benefits from it. In such cases, upon attachment of a new monomer to one already present in a given subcomplex, the *tms* can detect additionally established interactions with other monomers in the subcomplex. The benchmark set derived from the PDB in a semi-automated fashion with a minimum amount of manual intervention indicates that the majority of known complexes exhibit a sufficient amount of connectivity for the *tms* to be successfully applied. *tms* thus provides a valuable measure which significantly advances the field of assembling and ranking complexes based on pairwise dockings.

3D-MOSAIC can be used with any kind of pairwise interaction information that either (i) can be used to generate docking poses, for example, the aforementioned single residue pair interactions which can for example be obtained from crosslinking experiments[7–9] as well as

correlated mutation studies,[10–13] or (ii) can be described as transformations of the respective binding partners relative to each other, for example, dimers obtained from the PDB,[36] Interactome3D,[16] or PRISM.[18]

The availability and collection of such data form the basis for the field of integrative modeling approaches which combine a multitude of different information sources providing data on distances between components of the complex to generate medium-to-high resolution structural models of macromolecular assemblies. For example, Lasker *et al.*[39] have demonstrated that a sufficient amount of data can be collected (in this case cryo-EM maps and densities, residue-specific crosslinks, protein–protein interaction data from *in vitro* binding assays, crosslinking experiments and others, as well as structures for the individual monomers), to model complexes as large as the 26S proteasome (2.5 MDa, 33 protein types, 66 monomers).[40]

Depending on the application scenario and the available information, our approach can thus either provide assistance for or be used as a complement to such integrative approaches. Currently, additional features such as the incorporation of cryo-EM data to guide assembly and the step-wise generation of protein subcomplexes to facilitate reconstruction of weakly connected complexes and to improve runtime efficiency are being explored.

AUTHOR CONTRIBUTIONS

M.D., O.V.K., E.J., H.D., T.L., A.H. designed research; M.D. performed research, data analysis and developed 3D-MOSAIC; M.D., K.T., B.K., A.K.H. implemented 3D-MOSAIC; all authors wrote the article.

ACKNOWLEDGMENTS

The authors thank the BALL developer team for support, Prof. Dr. Bernd Gärtner (Institute of Theoretical Computer Science, ETH Zurich, Switzerland) for granting a GPL license to the miniball software employed in the hierarchical clash checking, as well as the Johannes-Gutenberg University, Mainz, Germany and the high-performance cluster MOGON for providing the computational resources.

REFERENCES

1. Makhnevych T, Houry WA. The role of Hsp90 in protein complex assembly. Biochim Biophys Acta, Mol Cell Res 2012;1823:674–682.
2. Villar G, Wilber AW, Williamson AJ, Thiara P, Doye JPK, Louis AA, Jochum MN, Lewis ACF, Levy ED. Self-assembly and evolution of homomeric protein complexes. Phys Rev Lett 2009;102:118106
3. Chothia C. Hydrophobic bonding and accessible surface area in proteins. Nature 1974;248:338–339.
4. Chothia C, Wodak S, Janin J. Role of subunit interfaces in the allosteric mechanism of hemoglobin. Proc Natl Acad Sci USA 1976;73:3793–3797.
5. Chothia C. The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–12.
6. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein–protein interactions. Curr Opin Struct Biol 2000;10:153–159.
7. Back JW, de Jong L, Muijsers AO, de Koster CG. Chemical cross-linking and mass spectrometry for protein structural modeling. J Mol Biol 2003;331:303–313.
8. Sinz A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. Mass Spectrom Rev 2006;25:663–682.
9. Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, Aebersold R. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. Mol Cell Proteomics 2010;9:1634–1649.
10. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012;30:1072–1080.
11. Sandler I, Medalia O, Aharoni A. Experimental analysis of co-evolution within protein complexes: the yeast exosome as a model. Proteins: Struct, Funct, Bioinf 2013;81:1997–2006.
12. Sprinzak E, Margalit H. Correlated sequence-signatures as markers of proteinprotein interaction. J Mol Biol 2001;311:681–692.
13. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28:184–190.
14. Lorenzen S, Zhang Y. Identification of near-native structures by clustering protein docking conformations. Proteins: Struct, Funct, Bioinf 2007;68:187–194.
15. O'Toole N, Vakser IA. Large-scale characteristics of the energy landscape in protein–protein interactions. Proteins: Struct, Funct, Bioinf 2008;71:144–152.
16. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods 2013;10:47–53.
17. Ghoorah AW, Devignes M-D, Smaïl-Tabbone M, Ritchie DW. KBDOCK 2013: a spatial classification of 3D protein domain family interactions. Nucleic Acids Res 2014;42:D389–D395.
18. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A. PRISM: web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res 2014;42:W285–W289.
19. Janin J. Docking predictions of protein-protein interactions and their assessment: The CAPRI experiment. In: Roterman-Konieczna I, editor. Identification of ligand binding site and protein-protein interaction area. Dordrecht, Netherlands: Springer; 2013. pp 87–104.
20. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins: Struct, Funct, Bioinf 2013;81:2082–2095.
21. Karaca E, Melquiond ASJ, de Vries SJ Kastritis PL., Bonvin AMJJ. Building macromolecular assemblies by information-driven docking introducing the HADDOCK multibody docking server. Mol Cell Proteomics 2010;9:1784–1794.
22. Comeau SR, Camacho CJ. Predicting oligomeric assemblies: N-mers a primer. J Struct Biol 2005;150:233–244.
23. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and Symm-Dock: servers for rigid and symmetric docking. Nucleic Acids Res 2005;33:W363–W367.
24. André I, Bradley P, Wang C, Baker D. Prediction of the structure of symmetrical protein assemblies. Proc Natl Acad Sci USA. 2007;104:17656–17661.
25. Pierce B, Tong W, Weng Z. MZDOCK: a grid-based approach for Cn symmetric multimer docking. Bioinformatics 2005;21:1472–1478.
26. Degiacomi MT, Dal Peraro M. Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling. Structure 2013;21:1097–1106.
27. Popov P, Ritchie DW, Grudinin S. DockTrina: docking triangular protein trimers. Proteins: Struct, Funct, Bioinf 2014;82:34–44.
28. Venkatraman V, Ritchie DW. Predicting multi-component protein assemblies using an ant colony approach. Ijsir 2012;3:19–31.

29. Inbar Y, Benyamini H, Nussinov R, Wolfson HJ. Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. Phys Biol 2005;2:S156
30. Esquivel-Rodrguez J, Yang YD, Kihara D. Multi-LZerD: multiple protein docking for asymmetric complexes. Proteins: Struct, Funct, Bioinf 2012;80:1818–1833.
31. Kuzu G, Keskin O, Nussinov R, Gursoy A. Modeling protein assemblies in the proteome. Mol Cell Proteomics 2014;13:887–896.
32. Chaudhury S, Berrondo M, Weitzner BD, Muthu P, Bergman H, Gray JJ. Benchmarking and analysis of protein docking performance in Rosetta v3.2. PLoS One 2011;6:e22477
33. Hildebrandt A, Dehof AK, Rurainski A, Bertsch A, Schumann M, Toussaint NC, Moll A, Stöckel D, Nickels S, Mueller SC, Lenhof H-P, Kohlbacher O. BALLbiochemical algorithms library 1.3. BMC Bioinformatics 2010;11:531
34. Hildebrandt AK, Dietzen M, Lengauer T, Lenhof H-P, Althaus E, Hildebrandt A. Efficient computation of root mean square deviations under rigid transformations. J Comput Chem. 2014;35:765–771.
35. Dietzen MM. Modeling protein interactions in protein binding sites and oligomeric protein complexes. 2014. Available at: url: http://scidok.sulb.uni-saarland.de/volltexte/2014/5940/.
36. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computerbased archival file for macromolecular structures. Arch Biochem Biophys 1978;185:584–591.
37. Bugalho MMF, Oliveira AL. Constant time clash detection in protein folding. J Bioinform Comput Biol 2009;7:55–74.
38. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003; 332:989–998.
39. Lasker K, Förster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proc Natl Acad Sci USA 2012;109:1380–1387.
40. Beck F, Unverdorben P, Bohn S, Schweitzer A, Pfeifer G, Sakata E, Nickell S, Plitzko JM, Villa E, Baumeister W, Förster F. Near-atomic resolution structural model of the yeast 26s proteasome. Proc Natl Acad Sci USA 2012;109:14870–14875.
41. Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.5.0.1". 2012.
42. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9:90–95.