Bioinformatics, 31(17), 2015, 2801–2807
doi: 10.1093/bioinformatics/btv270
Advance Access Publication Date: 25 April 2015
Original Paper

OXFORD

Structural bioinformatics

DockStar: a novel ILP-based integrative method for structural modeling of multimolecular protein complexes

Naama Amir*, Dan Cohen and Haim J. Wolfson*

Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

*To whom correspondence should be addressed.
Associate Editor: Anna Tramontano

Received on January 29, 2015; revised on April 9, 2015; accepted on April 19, 2015

Abstract

Motivation: Atomic resolution modeling of large multimolecular assemblies is a key task in Structural Cell Biology. Experimental techniques can provide atomic resolution structures of single proteins and small complexes, or low resolution data of large multimolecular complexes.
Results: We present a novel integrative computational modeling method, which integrates both low and high resolution experimental data. The algorithm accepts as input atomic resolution structures of the individual subunits obtained from X-ray, NMR or homology modeling, and interaction data between the subunits obtained from mass spectrometry. The optimal assembly of the individual subunits is formulated as an Integer Linear Programming task. The method was tested on several representative complexes, both in the bound and unbound cases. It placed correctly most of the subunits of multimolecular complexes of up to 16 subunits and significantly outperformed the CombDock and Haddock multimolecular docking methods.
Availability and implementation: http://bioinfo3d.cs.tau.ac.il/DockStar
Contact: naamaamir@mail.tau.ac.il or wolfson@tau.ac.il
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many cellular processes are performed by multimolecular protein complexes (Krogan *et al.*, 2006). A cell consists of hundreds of different functional complexes, such as the RNA exosome, the proteasome and the nuclear pore complex (Robinson *et al.*, 2007). Detailed structural characterization of such complexes is a primary goal of structural biology. However, structural characterization of large complexes by high resolution experimental techniques, such as X-ray crystallography or NMR spectroscopy, is challenging, whereas electron microscopy or mass spectrometry (MS) produce low resolution data (Alber *et al.*, 2007). It is becoming clear that integration of data derived from a variety of bio-physical techniques at multiple levels of resolution is essential for the structural analysis of large complexes (Thalassinos *et al.*, 2013).

In the last couple of decades significant progress has been achieved in computational modeling of binary protein–protein complexes as has been showcased in the CAPRI (Critical Assessment of Prediction of Interactions) challenge (Chen and Weng, 2002; Dominguez *et al.*, 2003; Duhovny *et al.*, 2002; Gray *et al.*, 2003; Janin, 2010; Kozakov *et al.*, 2006). However, only a handful of methods has been developed for the modeling of multimolecular complexes. The better performing ones have been limited to symmetric homo-oligomers (André *et al.*, 2007; Berchanski and Eisenstein, 2003; Comeau and Camacho, 2005; Pierce *et al.*, 2005; Schneidman-Duhovny *et al.*, 2005a). For the non-symmetric case, we have previously developed CombDock (Inbar *et al.*, 2005), which formulates the multimolecular complex detection task as a search for an optimally scoring spanning tree of the complex subunit interactions graph. The optimization is done by a branch and bound technique. The multimolecular version of Haddock (Karaca *et al.*, 2010) is driven by experimental and bioinformatics data, but limits the number of subunits to six, apparently due to computational

complexity constraints. Multi-LZerD (Esquivel-Rodríguez *et al.*, 2012) builds the multimolecular assembly applying a stochastic search driven by a genetic algorithm. Kuzu *et al.* (2014) construct the multimolecular complex iteratively, where at each iteration the subassembly is grown by one subunit. Hall *et al.* (2012) model the complexes at a coarse grained resolution, where proteins are represented by single spheres. Experimental MS-based data is translated into spatial restraints and integrated into the scoring function using the IMP platform (Russel *et al.*, 2012). The generation of candidate models is done by an exhaustive Monte Carlo search of the conformational space.

The large multimolecular assembly task can be divided into two subtasks: (a) Detection of the protein-protein interaction graph between the individual subunits; (b) Detection of a globally consistent pose of the subunits, so that there are no steric clashes between them and the binding energy of the whole complex is optimized.

Low resolution experimental data obtained from MS is especially suitable to assist in subtask (a). In particular, native MS of intact protein complexes and their subcomplexes can determine the stoichiometry of the complex subunits and deduce the interaction graph of the multimolecular complex (Taverner *et al.*, 2008). Chemical cross-linking combined with MS (XL-MS) provides distance constraints between surface residues both on the same and on neighboring subunits (Leitner *et al.*, 2010), thus providing information both for the detection of the interaction graph as well as constraints on the relative spatial poses of neighboring subunits. Such constraints have been successfully exploited in the modeling of the 26S proteasome (Lasker *et al.*, 2012), the proteasome lid (Politis *et al.*, 2014), the TRiC/CCT chaperonin (Kalisman *et al.*, 2012; Leitner *et al.*, 2012), the RNA polymerase II–TFIIF complex (Chen *et al.*, 2010) and more.

Here, we introduce DockStar, a novel protein assembly modeling algorithm which integrates MS-based data. We assume knowledge of the subunit interaction graph [subtask (a)] and focus on a global solution to the '3D-puzzle assembly' task [subtask (b)]. Our algorithm performs a search for assemblies, optimizing satisfaction of cross-linking restraints as well as a knowledge-based interaction potential between neighboring subunits. The optimization is performed by formulating the task as an Integer Linear Program (ILP), thus not depending neither on greedy type iterative algorithms, which tend to miss solutions with low scoring intermediates, nor on time consuming Monte Carlo type search. This enables efficient handling of relatively large assemblies.

To validate this method, DockStar was tested on a set of examples, where both high resolution structures of the full complex and cross-linking data between neighboring subunits was available. The modeling was performed both in bound and unbound scenarios and compared with the state of the art Haddock (Karaca *et al.*, 2010) and CombDock (Inbar *et al.*, 2005) multimolecular assembly algorithms. DockStar proved to be significantly more time efficient than the other methods, while exhibiting better performance.

2 Methods

DockStar receives as input a set of subunit structures, the subunit interaction graph and (optionally) a set of cross-linked residue pairs. It applies a pairwise (soft) docking algorithm to compute candidate docking poses for neighboring subunits, which are deduced from the interaction graph. The pairwise docking hypotheses are assembled into globally optimal multimolecular complex hypotheses by an ILP-based optimization algorithm. The final output is a ranked set of 3D complex structures, which are scored by the binding energy between the interacting subunits and satisfaction of the cross-links induced distance restraints. More specifically, the method is composed of the following steps (Supplementary Fig. S1 in the Supplementary Materials):

1. **Generation of transformation sets.** The aim of this step is to produce an initial set of candidate 3D poses (rigid transformations) for each subunit.
2. **Pairwise scoring.** A (docking) pair of candidate transformations belonging to different subunits is scored according to the quality of the resulting interface between the transformed subunits. The scoring function takes into account the number of satisfied cross-link restraints and a knowledge-based potential.
3. **Choice of globally optimal solutions.** A solution includes one rigid transformation for each subunit. The transformations are chosen to optimize the sum of the resulting interfaces scores. This optimization problem is solved by formulating the task as an ILP. The algorithm is tuned to produce a ranked set of K (user predefined) best solutions.
4. **Integration of partial solutions.** In some cases, the method of choice for generating the transformation sets limits the algorithm to deal only with complexes in which one subunit interacts with all the other subunits (interaction graph with a star shaped spanning tree). In such a case, the interaction graph of the whole complex is partitioned into (overlapping) subcomplexes with star shaped spanning trees, and the top scoring solutions for the subcomplexes are reintegrated to detect global solutions to cover the whole complex (Supplementary Fig. S2).

2.1 Generation of transformation sets

In the first stage of the algorithm for each subunit a set of candidate rigid transformations is generated. These transformations are obtained from the top scoring docking poses of neighboring subunits. However, when modeling multimolecular complexes, it is important to ensure that all these transformations are consistent, namely, refer to the same 3D reference frame. Therefore, one subunit is chosen as an **anchor** subunit, and its coordinate system is used as the 3D reference frame for all the other subunits. Preferably, the anchor subunit should have the largest number of neighbors in the multimolecular assembly interaction graph. All other subunits which are known to interact with the anchor are then docked to it. Hence, in this case, transformation sets are generated only for the anchor and its neighbors requiring a star shaped spanning tree topology of the interaction graph. In cases where the interaction graph of the complex does not have a star shaped spanning tree, the complex is divided to sub-complexes with star shaped spanning trees, that are solved independently and then the top solutions are integrated to cover the whole complex. The docking step is carried out by the PatchDock (Duhovny *et al.*, 2002) algorithm, which optimizes shape complementarity, while satisfying distance constraints between residues of neighboring subunits, if such constraints are available. In our examples the cross-link induced maximal distance constraint, which includes the linker and extended residue lengths, was set to 30 Å (Rappsilber, 2011). The top 1000 PatchDock transformations are refined, rescored and reranked by the FiberDock (Mashiach *et al.*, 2010) algorithm. For each subunit, a predefined number of top ranked transformations is chosen as the subunit transformation set.

Using an anchor subunit as a 3D reference frame for a subcomplex with a star shaped spanning tree is the 'method of choice' in the difficult cases, where no additional data is available. However, if a

homologous complex or an intermediate resolution cryo-EM map of the complex is available, one can compute the transformations by aligning the atomic resolution subunit structures to the (homologous or lower resolution) template complex.

2.2 Pairwise scoring

For each of the n subunits, let P_i $(0 \leq i < n)$ be subunit i and $T(P_i)$ be the set of candidate transformations received from the previous stage for subunit P_i. Let $T_{i,r}$ be transformation r of subunit P_i and $S(T_{i,r}, T_{j,s})$ the pairwise interaction score of subunits P_i and P_j transformed by $T_{i,r}$ and $T_{j,s}$, respectively.

Each pair of transformations $T_{i,r}$ and $T_{j,s}$, where $i \neq j$, is scored by the number of satisfied cross-link constraints between the transformed subunits. In case of equality, the interfaces generated between the subunits are ranked according to the recently published statistically optimized atomic potentials (SOAP-PP) score (Dong *et al.*, 2013).

2.3 Choice of a globally optimal transformation set

The globally optimal solution *Sol* includes one transformation per subunit and maximizes *score*(*Sol*) defined as:

$$\text{score}(\text{Sol}) = \sum_{T_{i,r}, T_{j,s} \in \text{Sol} \cap i \neq j} S(T_{i,r}, T_{j,s}) \tag{1}$$

Graph theoretic formulation of the optimization task. This optimization task can be formulated as the following graph theoretic problem (Supplementary Fig. S9). Let $G = (V, E)$ be an undirected n-partite graph with vertex set $V = V_0 \cup \cdots \cup V_{n-1}$, so that for each transformation $T_{i,r} \in T(P_i)$ there is a vertex $u_{i,r} \in V_i$. Each pair of vertices is joined by an edge:

$$E = \{(u_{i,r}, v_{j,s}) | u_{i,r} \in V_i; v_{j,s} \in V_j; i \neq j\} \tag{2}$$

with the weight:

$$w(u_{i,r}, v_{j,s}) = S(T_{i,r}, T_{j,s}) \qquad \forall (u_{i,r}, v_{j,s}) \in E \tag{3}$$

The optimal solution is achieved by choosing one vertex per V_i that maximizes the edge-weight of the induced sub-graph.

ILP formulation. The above graph theoretic task can be formulated as an ILP (Nemhauser and Wolsey, 1988). Define a variable $X_{i,r}$ for each vertex $u_{i,r} \in V$ and a variable $Y_{i,r,j,s}$ for each edge $e = (u_{i,r}, v_{j,s}) \in E$ as follows:

$$X_{i,r} = \begin{cases} 1 & \text{if } u_{i,r} \text{ is chosen} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$Y_{i,r,j,s} = \begin{cases} 1 & \text{if both } u_{i,r} \text{ and } v_{j,s} \text{ are chosen} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The ILP objective function is:

$$\text{Maximize} \qquad \text{score}(\text{Sol}) = \sum_{(u_{i,r}, v_{j,s}) \in E} w(u_{i,r}, v_{j,s}) Y_{i,r,j,s} \tag{6}$$

Subject to the constraints:

$$\sum_{u_{i,r} \in V_i} X_{i,r} = 1 \qquad \forall i, 0 \leq i < n \tag{7}$$

$$\sum_{u_{i,r} \in V_i} Y_{i,r,j,s} = X_{j,s} \qquad \forall j, s, i, \quad j \neq i \tag{8}$$

The objective function (Eq. 6) is exactly the edge-weight of the chosen sub-graph. The first constraint (Eq. 7) ensures that exactly one transformation is chosen for each subunit. The second constraint (Eq. 8) ensures that an edge is chosen if and only if both vertices that it connects are chosen as well.

The ILP step was solved by the CPLEX 12.5 package (http://www.ilog.com/products/cplex/).

Missing subunit. Although the previous formulation ensures the choice of exactly one transformation per subunit, in a practical setting it might occur that no reasonably scoring candidate transformation exists at all for some of the subunits, due, for example, to poor homology modeling. In such a case, one would like to return a partial solution. To handle this event an extra transformation is added to each subunit transformation set—the blank transformation. This transformation represents a situation where a solution for this subunit is 'missing'. The pairwise score of this transformation with any transformation of another subunit is set to zero. Therefore, a solution with missing subunits will be chosen if there is no full solution (with all subunits) with a higher score.

Alternative solutions. The method presented above outputs one single highest scoring global solution. To retrieve additional high scoring solutions, the ILP step is applied iteratively to find a solution that maximizes the objective function and was not chosen before. To achieve it a linear constraint is introduced for each previously obtained solution F. Define: $\forall X_{i,r}\ F(X_{i,r}) = 1 \Leftrightarrow X_{i,r}$ was chosen in solution F. Then add the following constraint to the ILP constraint set:

$$\sum_{F(X_{i,r})=1} X_{i,r} < n \tag{9}$$

This constraint promises that solution F will not be chosen again.

2.4 Integration of partial solutions

The presentation above deals with complexes having a star shaped spanning tree, where an anchor subunit, which interacts with all the other subunits, can be chosen. However, this is a special case. To apply the method for arbitrary complexes, they are divided into overlapping sub-complexes, each with a star shaped spanning tree, which are solved separately as above. Then, top solutions of subcomplexes that share a subunit are merged, while defining the shared subunit as the new 'anchor'. All the transformations in the merged (new) subcomplex are recalculated vis-a-vis the reference frame of the new 'anchor'. These new transformation sets are used as input for steps 2–4 of the algorithm in order to solve the larger sub-complex. In several such iterations one can cover all the subunits of the assembly.

3 Results

DockStar is a multimolecular docking method which integrates experimental cross-links and graph topology data. The method was tested on a diverse dataset, which included bound and unbound examples, different numbers of complex subunits (from 3 to 16), and diverse additional experimental data (see summary of results in Table 1). The method was compared with the state of the art CombDock and Haddock multimolecular docking methods. All running times were measured on a 12 core XEON 3.06 GHz server.

Below we present in detail the results for several large complexes. Additional data, including figures, can be found in the Supplementary Materials.

3.1 Modeling of individual subunits

In order to simulate **unbound** modeling, models of the different subunits were created using homology modeling. Homologous proteins

Table 1. Summary of the DockStar's results

Target complex	Bound/ unbound	Subunits number	Rank	Global Cα-RMSD[a]	Number of contacts[b]	Quality of predicted contacts[c]				Run time HH:MM
						high	medium	acceptable	lenient	
PP2A	Bound	3	1	0.68	2	2	0	0	0	00:35
	Unbound	3	1	6.9	2	0	0	0	2	00:43
Beef liver	Bound	4	1	0.85	3	3	0	0	0	02:51
Catalase	Unbound	4	1	2.7	3	0	3	0	0	03:53
RNA polII	Bound	11	1	7.9	10	4	3	2	0	04:53
	Unbound	11	3	4.8	10	0	3	4	1	04:56
Yeast exosome	Bound	10	1	5.1	9	6	1	0	0	10:34
	Unbound	10	12	6.0	9	1	1	1	1	11:22

[a]Global Cα-RMSD between the predicted and the native assemblies including only predictions with lenient to high quality.
[b]Number of contacts in the spanning tree of the complex interaction graph.
[c]Predicted interfaces in the target complex that are of lenient to high quality.

were found using BLAST (Altschul *et al.*, 1990) and HHBLITS (Remmert *et al.*, 2012) and were modeled with MODELLER (Webb and Sali, 2014).

3.2 Assessment criteria

The assessment of the quality of the resulting solutions was determined by three measures: (i) global Cα-RMSD between the predicted complex and the (known) native complex, (ii) i-RMSDbb (interface RMSD of the backbone) and (iii) Fnat (fraction of native contacts) between interacting subunits of the complex interaction graph spanning tree (Supplementary Materials for comprehensive Fnat and i-RMSDbb definitions). The interacting subunit interface prediction quality was divided into four categories (the three first follow the CAPRI challenge definition): high quality predictions (Fnat $\geq$ 0.5 and i-RMSDbb $\leq$ 1.0 Å), medium predictions (Fnat $\geq$ 0.3 and i-RMSDbb $\leq$ 2.0 Å), acceptable predictions (Fnat $\geq$ 0.1 and i-RMSDbb $\leq$ 4.0 Å) and lenient predictions (i-RMSDbb $\leq$ 8.0 Å). The lenient criterion was introduced, since analysis of docking funnels suggested that predictions with i-RMSDbb up to 8–10 Å can be locally minimized and refined to near native structures (Hunjan *et al.*, 2008; Kundrotas *et al.*, 2010).

3.3 The complexes modeled

PP2A. The protein phosphatase 2A has a key role in regulating diverse signaling pathways including cell growth, differentiation, apoptosis, cell motility, the DNA damage response and cell cycle progression (Janssens and Goris, 2001; Wurzenberger and Gerlich, 2011). The complex in its active form is composed of a catalytic subunit C, a scaffold subunit A and a regulatory subunit B (Wurzenberger and Gerlich, 2011). Cross-link data of this complex is available (Herzog *et al.*, 2012; Kahraman *et al.*, 2013) as well as the 3D atomic resolution of the subunits (Xu *et al.*, 2009). The cross-link data includes three inter cross-links. As all inter cross-links included subunit A. This subunit was used as the anchor subunit and subunits B and C were docked to it using PatchDock with the inter cross-links as distance constraints. The top scored 50 PatchDock+FiberDock solutions were used as candidate transformations for DockStar.

In the bound case, out of $\sim 2.6 \times 10^3$ possible solutions (the number of all possible solutions is the product of the size of the transformations set of each subunit), the first solution was correct with only 0.6 Å average i-RMSDbb of the predicted interfaces and low global Cα-RMSD of 0.68 Å [Fig. 1a, Table 1, Supplementary Table S1 in the Supplementary Materials]. The runtime of DockStar on this example was <32 min for the pairwise scoring step and 1:16 min for the ILP step.

In the unbound case, MODELLER was used to calculate the structures of the individual subunits according to templates with sequence identity of 40 to 100% (Supplementary Table S2). These calculated models had 1.9–5.3 Å RMSD to the bound native subunits. Sets of 50 alternative transformations for B and C were created by taking the best 50 scored docking results of PatchDock+FiberDock for the pairs AB and AC as in the bound case. In these sets, the smallest i-RMSDbb of subunits AB and AC were 5.6 and 5.8 Å, respectively. The first solution of DockStar included these best transformations (Supplementary Table S3 and Fig. S3B) and the global Cα-RMSD of this solution was 6.9 Å (Table 1). The major reason for this relatively high i-RMSDbb is due to the limited accuracy of the homology modeling. The homologous models of subunits B and C have 5.3 and 3.2 Å RMSD to the native, respectively. The run time was 40 min for the pairwise scoring step and 2 min for the ILP step.

Beef Liver Catalase. The beef liver catalase is a common homotetramer, which catalyzes the decomposition of hydrogen pyroxide to water and oxygen (Ko *et al.*, 1999). Its subunits contain a hem group and a NADPH molecule. Cross-link data of this complex is available (Kahraman *et al.*, 2013; Lauber and Reilly, 2010) and includes three inter cross-links that connect subunits: B-D, B-C, A-D. Therefore, the complex was solved in two steps. First, subunits B and D were selected as anchor subunits and the complexes D-B-C and A-D-B were solved. Then, the top 100 results from the previous run were merged by applying a second run.

In the bound case, the first solution was correct with three out of three high quality interface predictions (with average i-RMSDbb of 0.54 Å) and global Cα-RMSD of only 0.85 Å to the native (Fig. 1b, Table 1, Supplementary Table S2). The run time in this case was 35 min and 1:48 min for the pairwise scoring step and the ILP step of the first run, respectively; 2:13 h and 1:14 min for the pairwise scoring step and the ILP step of the second run.

For the unbound case, each of the subunits was modeled according to the human erythrocyte catalase template with sequence identity of 91% to the beef liver catalase and the obtained models had average RMSD of 0.73 Å to the native bound structures (Supplementary Table S2). The first solution had three medium quality interfaces (with average i-RMSDbb of 1.57 Å) and global Cα-RMSD of only 2.7 Å to the native (Table 1, Supplementary Table S3, Supplementary Fig. S4B). The run time in this case was 56 min and 1:43 min for the pairwise scoring step and the ILP step of the first run respectively; 2:51 h and 3:39 min for the pairwise scoring step and the ILP step of the second run.

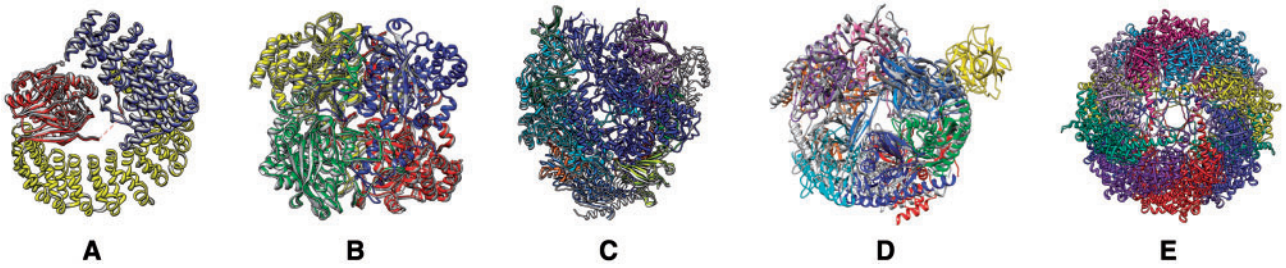

Fig. 1. The predicted models of the bound cases (coloured by chains) superimposed on the correct complex structures taken from the PDB (grey). (**A**) PP2A (A(yellow), B(blue), C(red)), (**B**) The Beef Liver Catalase [A(yellow), B(blue), C(red), D(green)], (**C**) RNA polymerase II [Rbp1(blue), Rbp2(cyan), Rbp3(light blue), Rbp5(purple), Rbp6(green), Rbp7(pink), Rbp8(yellow), Rbp9(dark green), Rbp10(orange), Rbp11(brown), Rbp12(red)], (**D**) The Yeast Exosome [Rrp45(blue), Rrp41(cyan), Rrp43(light blue), Rrp46(green), Rrp42(purple), Mtr3(pink), Rrp40(red), Rrp4(orange), Csl4(yellow), Dis3(dark green)]. (**E**) The predicted order of chains in the model of the TRiC/CCT Chaperonin: Z(red) Q(blue) H(yellow) E(light blue) B(pink) D(grey) A(green) G(purple)

RNA polymerase II. In eukaryotic cells, RNA pol II catalyses mRNA synthesis during transcription of protein-coding genes. This large asymmetric complex is composed of 12 subunits: a 10-subunit catalytic core (Rpb1-3,5,6,8-12) and a complex of two subunits (Rpb4/7) (Cramer, 2004). The two big core subunits Rpb1/2 are connected to all other subunits except of Rpb4, which is connected to the complex through Rpb7. Data which includes 34 inter cross-links is available (Chen *et al.*, 2010; Kahraman *et al.*, 2013).

We have solved the complex both in the 'bound' and 'unbound' scenarios. In both cases, first the two big core subunits (Rpb1/2) were docked to each other using PatchDock+FiberDock (Duhovny *et al.*, 2002; Mashiach *et al.*, 2010; Schneidman-Duhovny *et al.*, 2005b). Then, all other nine subunits were docked to the Rpb1/2 complex. Finally, DockStar was executed with sets of the top 200 FiberDock candidate transformations and the inter cross-link data.

For the bound case, the complex subunits from PDB 1WCM (Armache *et al.*, 2005) were used. DockStar's first solution predicted correctly 9 out of the 10 interfaces and had 7.9 Å global Cα-RMSD of the 10 subunits subcomplex relative to the native. The average i-RMSDbb was 1.85 Å (Fig. 1c, Table 1, Supplementary Fig. S5 and Table S1 in the Supplementary Materials). This solution satisfied 32 out of the 34 inter cross-links. The run time of this example was 4:35 h for the pairwise scoring step and 18:22 min for the ILP step.

In the unbound case, models of the different subunits were generated using homology modeling templates ranging from 33 to 100% sequence identity to the native subunits. Unbound structures were modeled with an average RMSD of 2.8 Å (Supplementary Table S2). The third ranked DockStar solution predicted correctly eight interfaces with an average of 3.0 Å i-RMSDbb relative to the native structure (Supplementary Fig. S6 and Table S3). The global Cα-RMSD of the solution subcomplex of these nine correctly transformed subunits is only 4.8 Å (Table 1) satisfying 31 out of the 34 inter cross-links. The runtime in this example was 4:47 h for the pairwise scoring step and <4 min for the ILP step.

Yeast Exosome. The exosome complex is the main cellular machinery responsible for degrading RNA molecules (Mitchell *et al.*, 1997). The interaction graph of this 10 subunits complex was deduced by the SUMMIT algorithm based on native MS experimental data (Taverner *et al.*, 2008). The core of the complex consists of a six subunit ring (Rrp41-Rrp42-Mtr3-Rrp43-Rrp46-Rrp45) and the remaining four subunits (Rrp4, Rrp40, Csl4, Dis3) interact with the ring subunits (Supplementary Fig. S8A). Later, an X-Ray structure was published (Makino *et al.*, 2013). Because this complex spanning tree is *non-star shaped*, it was modeled in three rounds with different anchor units for each round (see Supplementary Fig. S8 for a detailed explanation).

In the bound case, the top ranked solution by DockStar correctly predicted seven out of the nine interfaces of the complex interaction graph spanning tree, (Fig. 1d) with average i-RMSDbb of 1.72 Å (Supplementary Table S1) and 5.1 Å global Cα-RMSD of the eight subunits subcomplex that includes only the correct interactions (Table 1). The total run time of this example was 10:34 h.

In the unbound case, for 3 out of the 10 subunits the structure similarity between the homology modeled and the native structures was large (above 15 Å RMSD). In these cases (subunits: Rrp43, Mtr3, Dis3), the bound structures were used. The other homologous templates had 21–100% sequence identity to the native subunits with an average of 36%. The modeled structures had RMSD range of 1.9–10.5 Å to the native bound structures (Supplementary Table S2). This is a difficult example for which there is no cross-links data and the unbound subunits have large RMSDs to the native. Despite that, the 12th solution of DockStar predicted correctly four interfaces with 3.5 Å average i-RMSDbb (Supplementary Fig. S7D and Table S3). The global Cα-RMSD of the subcomplex that includes only these interactions was 6.0 Å (Table 1). The total run time was 11:22 h.

TRiC/CCT. The eukaryotic TRiC chaperonin plays a central role in assisting the folding of polypeptide chains (Spiess *et al.*, 2004). In its open state the chaperonin binds the substrate, while in its closed state, the substrate is inserted into a large cavity where folding occurs (Reissmann *et al.*, 2007). The eukaryotic TRiC chaperonin is composed of 16 subunits, two copies of eight different subunits, that are arranged in two octameric rings (Ditzel *et al.*, 1998). Although each ring consists of eight different subunits the order of which allows specific substrate binding modes (Martín-Benito *et al.*, 2007), the sequence identity between the different subunits in the same organism is 30% resulting in high structural identity, which makes reliable identification of the subunit order difficult as the subunits look almost the same in low resolution maps (Booth *et al.*, 2008; Cong *et al.*, 2010; Dekker *et al.*, 2011; Spiess *et al.*, 2004).

Both Kalisman *et al.* (2012) and Leitner *et al.* (2012) have used cross-link data to determine the correct order of the subunits in the eukaryotic bos taurus chaperonin, yet their methods are highly time consuming. Here, DockStar was applied to the cross-link data published by these two groups to detect the same structure in 10 and 14 min of CPU time, respectively. The published experimental datasets were used independently. First, all the individual subunits were modeled by homology modeling according to the homo-oligomer Thermococcus chaperonin closed state (Shomura *et al.*, 2004) that has 34–40% sequence identity with the subunits (Supplementary Table S2). Then, 3D rigid transformation sets were generated by

structurally aligning each of the different modeled subunits with each of the Thermococcus chaperonin subunits. Sixteen subunits were used as input—two copies for each of the eight different subunits. Each copy has a set of eight 3D rigid transformations for the different locations in the octameric ring.

The first cross-link set was taken from Kalisman *et al.* (2012). Out of 63 high confidence cross-links, 17 are cross-links between different subunits. The top ranked solution (Fig. 1e) was identical to the solution obtained by Kalisman *et al.* (2012) and Leitner *et al.* (2012). This solution satisfied 15 cross-links out of the 17. The total runtime of the algorithm in this example was <10 min.

The second cross-link set (Leitner *et al.*, 2012) included 87 cross-links between different subunits. As in the first case, the top solution was identical to the solution obtained by Kalisman *et al.* (2012) and Leitner *et al.* (2012), and it satisfied 67 cross-links out of the 87. The total run time was 14 min.

3.4 Comparison with other methods

DockStar was compared with the state of the art Haddock (Karaca *et al.*, 2010) and CombDock (Inbar *et al.*, 2005). Here, the results of the three methods were compared on the large complexes RNA polII, Yeast Exosome, and the small complexes PP2A and Beef liver catalase.

Comparison with CombDock. CombDock is designed to model multimolecular complexes from subunit data alone, without receiving interaction graph information. It gets as input 3D structures of the subunits and a predefined number of top docking results of all subunit pairs in the complex. In each iteration, CombDock produces subcomplexes of size i from smaller subcomplexes produced in previous iterations and saves the K best scored structures. The pairwise docking results were calculated by PatchDock with cross-links as distance constraints. To perform a 'fair' comparison between CombDock and DockStar, we have run CombDock twice on each example. In the first run, the regular CombDock procedure was performed. In the second run, we artificially provided CombDock the interaction graph information by supplying PatchDock docking results for subunit pairs that are known to be in contact, while for the non-contacting subunits random transformations were introduced. The results are summarized in Supplementary Table S4 of the Supplementary Materials. DockStar outperformed CombDock in all the above mentioned examples. Although CombDock succeeds to predict good solutions in most of the bound cases, this is not the case in the unbound examples. Although the knowledge of the interaction graph improves performance, still in all of the unbound cases CombDock succeeds to predict correctly only 0–2 of the interacting pairs. In all cases, DockStar was able to rank good solutions higher than CombDock. DockStar's better ranking ability can be attributed to its scoring function, which ranks first cross-link satisfaction, followed by a SOAPs score (Dong *et al.*, 2013), while CombDock uses a purely shape complementarity-based scoring function. Also, DockStar utilizes the ILP technique which returns an optimal scoring solution in the majority of the cases.

Although in the small examples (3–4 subunits) Combdock's runs were extremely fast and lasted <2 min, in some of the large examples (10–11 subunits), due to its exponential complexity, the runtime was prohibitive. The efficiency of the Dockstar method is due to the ILP technique which allows tackling problems with extremely large search space as multimolecular modeling, in relatively short running time.

Comparison with HADDOCK. HADDOCK uses experimental and/or bioinformatics data to predict the complex structure. This data is translated to ambiguous and/or unambiguous restraints which are integrated into the score function. Here, CPORT, a Consensus Interface Predictor (de Vries and Bonvin, 2011), was used to predict the interface residues of each subunit. Ambiguous distance constraints between predicted interface residues of subunit pairs that are known to interact (according to the interaction graph) were generated using GenTBL. The cross-links data was used as unambiguous constraints. As HADDOCK is limited to six subunits, for the RNA polII and the Yeast Exosome examples (composed of 11 and 10 subunits, respectively) a subcomplex of six connected subunits was chosen to be remodeled.

In the bound case, out of the four examples only one example has completed its run, the PP2A. None of the PP2A solutions contained a correct interface prediction. As there were no correct solution or sub-solution in the bound cases, HADDOCK was not tested for the unbound cases.

4 Discussion

DockStar is a novel integrative multimolecular docking algorithm. The method integrates experimental or modeled atomic resolution individual protein subunit structure data with MS-based low resolution data. The search for a globally optimal multimolecular assembly is formulated as an ILP and, thus, efficiently solved in the majority of the cases. The method can be applied in additional situations where (low resolution) complex data is obtained from electron microscopy or homologous complexes.

Future challenges include handling significant conformational flexibility of the modeled individual subunits.

Acknowledgements

We are grateful to Alexandre Bonvin for his help in running the Haddock experiments and to Dina Schneidman on her advise in the application of SOAP-PP.

Funding

This research was supported by the Israel Science Foundation (grant No. 1112/12), the I-CORE program of the Budgeting and Planning Committee and the Israel Science Foundation (center No. 1775/12) and by the Hermann Minkowski Minerva Geometry Center. N.A. acknowledges the Edmond J. Safra Bioinformatics Center fellowship.

Conflict of Interest: none declared.

References

Alber,F. *et al.* (2007) Determining the architectures of macromolecular assemblies. *Nature*, **450**, 683–694.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

André,I. *et al.* (2007) Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci.*, **104**, 17656–17661.

Armache,K.-J. *et al.* (2005) Structures of complete RNA polymerase ii and its subcomplex, rpb4/7. *J. Biol. Chem.*, **280**, 7131–7134.

Berchanski,A. and Eisenstein,M. (2003) Construction of molecular assemblies via docking: modeling of tetramers with d2 symmetry. *Proteins*, **53**, 817–829.

Booth,C.R. *et al.* (2008) Mechanism of lid closure in the eukaryotic chaperonin TRiC/CCT. *Nat. Struct. Mol. Biol.*, **15**, 746–753.

Chen,R. and Weng,Z. (2002) Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, **47**, 281–294.

Chen,Z.A. *et al.* (2010) Architecture of the RNA polymerase ii–tfiif complex revealed by cross-linking and mass spectrometry. *EMBO J.*, **29**, 717–726.

Comeau,S.R. and Camacho,C.J. (2005) Predicting oligomeric assemblies: N-mers a primer. *J. Struct. Biol.*, **150**, 233–244.

Cong,Y. *et al.* (2010) 4.0-å resolution cryo-em structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 4967–4972.

Cramer,P. (2004) Structure and function of RNA polymerase ii. *Adv. Protein Chem.*, **67**, 1–42.

de Vries,S.J. and Bonvin,A.M. (2011) Cport: a consensus interface predictor and its performance in prediction-driven docking with haddock. *PLoS One*, **6**, e17695.

Dekker,C. *et al.* (2011) The crystal structure of yeast CCT reveals intrinsic asymmetry of eukaryotic cytosolic chaperonins. *EMBO J.*, **30**, 3078–3090.

Ditzel,L. *et al.* (1998) Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. *Cell*, **93**, 125–138.

Dominguez,C. *et al.* (2003) Haddock: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.

Dong,G.Q. *et al.* (2013) Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics*, **29**, 3158–3166.

Duhovny,D. *et al.* (2002) Efficient unbound docking of rigid molecules. In: *Algorithms in Bioinformatics*. Springer, Berlin Heidelberg, pp. 185–200.

Esquivel-Rodríguez,J. *et al.* (2012) Multi-lzerd: multiple protein docking for asymmetric complexes. *Proteins*, **80**, 1818–1833.

Gray,J.J. *et al.* (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.

Hall,Z. *et al.* (2012) Structural modeling of heteromeric protein complexes from disassembly pathways and ion mobility-mass spectrometry. *Structure*, **20**, 1596–1609.

Herzog,F. *et al.* (2012) Structural probing of a protein phosphatase 2a network by chemical cross-linking and mass spectrometry. *Science*, **337**, 1348–1352.

Hunjan,J. *et al.* (2008) The size of the intermolecular energy funnel in protein–protein interactions. *Proteins*, **72**, 344–352.

Inbar,Y. *et al.* (2005) Prediction of multimolecular assemblies by multiple docking. *J. Mol. Biol.*, **349**, 435–447.

Janin,J. (2010) Protein–protein docking tested in blind predictions: the capri experiment. *Mol. BioSyst.*, **6**, 2351–2362.

Janssens,V. and Goris,J. (2001) Protein phosphatase 2a: a highly regulated family of serine/threonine phosphatases implicated in cell growth and signalling. *Biochem. J.*, **353**, 417–439.

Kahraman,A. *et al.* (2013) Cross-link guided molecular modeling with rosetta. *PloS One*, **8**, e73411.

Kalisman,N. *et al.* (2012) Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 2884–2889.

Karaca,E. *et al.* (2010) Building macromolecular assemblies by information-driven docking introducing the haddock multibody docking server. *Mol. Cell. Proteomics*, **9**, 1784–1794.

Ko,T.-P. *et al.* (1999) Structure of orthorhombic crystals of beef liver catalase. *Acta Crystallogr. Sec. D Biol. Crystallogr.*, **55**, 1383–1394.

Kozakov,D. *et al.* (2006) Piper: An FFT-based protein docking program with pairwise potentials. *Proteins*, **65**, 392–406.

Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Kundrotas,P. *et al.* (2010) Docking by structural similarity at protein-protein interfaces. *Biophys. J.*, **98**, 196a.

Kuzu,G. *et al.* (2014) Modeling protein assemblies in the proteome. *Mol. Cell. Proteomics*, **13**, 887–896.

Lasker,K. *et al.* (2012) Molecular architecture of the 26s proteasome holocomplex determined by an integrative approach. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 1380–1387.

Lauber,M.A. and Reilly,J.P. (2010) Novel amidinating cross-linker for facilitating analyses of protein structures and interactions. *Anal. Chem.*, **82**, 7736–7743.

Leitner,A. *et al.* (2010) Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell. Proteomics*, **9**, 1634–1649.

Leitner,A. *et al.* (2012) The molecular architecture of the eukaryotic chaperonin TRiC/CCT. *Structure*, **20**, 814–825.

Makino,D.L. *et al.* (2013) Crystal structure of an RNA-bound 11-subunit eukaryotic exosome complex. *Nature*, **495**, 70–75.

Martín-Benito,J. *et al.* (2007) The inter-ring arrangement of the cytosolic chaperonin cct. *EMBO Rep.*, **8**, 252–257.

Mashiach,E. *et al.* (2010) Fiberdock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic Acids Res.*, **38** (Suppl. 2), W457–W461.

Mitchell,P. *et al.* (1997) The exosome: a conserved eukaryotic RNA processing complex containing multiple 3 5 exoribonucleases. *Cell*, **91**, 457–466.

Nemhauser,G.L. and Wolsey,L.A. (1988) *Integer and Combinatorial Optimization*. Vol. 18. Wiley, New York.

Pierce,B. *et al.* (2005) M-zdock: a grid-based approach for CN symmetric multimer docking. *Bioinformatics*, **21**, 1472–1478.

Politis,A. *et al.* (2014) A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat. Methods*, **11**, 403–406.

Rappsilber,J. (2011) The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.*, **173**, 530–540.

Reissmann,S. *et al.* (2007) Essential function of the built-in lid in the allosteric regulation of eukaryotic and archaeal chaperonins. *Nat. Struct. Mol. Biol.*, **14**, 432–440.

Remmert,M. *et al.* (2012) Hhblits: lightning-fast iterative protein sequence searching by HMM-hmm alignment. *Nat. Methods*, **9**, 173–175.

Robinson,C.V. *et al.* (2007) The molecular sociology of the cell. *Nature*, **450**, 973–982.

Russel,D. *et al.* (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.*, **10**, e1001244.

Schneidman-Duhovny,D. *et al.* (2005a) Geometry-based flexible and symmetric protein docking. *Proteins*, **60**, 224–231.

Schneidman-Duhovny,D. *et al.* (2005b) Patchdock and symmdock: servers for rigid and symmetric docking. *Nucleic Acids Res.*, **33** (Suppl. 2), W363–W367.

Shomura,Y. *et al.* (2004) Crystal structures of the group ii chaperonin from thermococcus strain ks-1: Steric hindrance by the substituted amino acid, and inter-subunit rearrangement between two crystal forms. *J. Mol. Biol.*, **335**, 1265–1278.

Spiess,C. *et al.* (2004) Mechanism of the eukaryotic chaperonin: protein folding in the chamber of secrets. *Trends Cell Biol.*, **14**, 598–604.

Taverner,T. *et al.* (2008) Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. *Acc. Chem. Res.*, **41**, 617–627.

Thalassinos,K. *et al.* (2013) Conformational states of macromolecular assemblies explored by integrative structure calculation. *Structure*, **21**, 1500–1508.

Webb,B. and Sali,A. (2014) Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics*, supplement 47, unit 5.6, Wiley Online Library.

Wurzenberger,C. and Gerlich,D.W. (2011) Phosphatases: providing safe passage through mitotic exit. *Nat. Rev. Mol. Cell Biol.*, **12**, 469–482.

Xu,Z. *et al.* (2009) Structure and function of the pp2a-shugoshin interaction. *Mol. Cell*, **35**, 426–441.