RESEARCH ARTICLE

# Structural modeling of histone methyltransferase complex Set1C from *Saccharomyces cerevisiae* using constraint-based docking

*Anne Tuukkanen*[1], *Bingding Huang*[1,2], *Andreas Henschel*[1], *Francis Stewart*[1] *and Michael Schroeder*[1]

[1] Biotechnology Center (BIOTEC), Technische Universität Dresden, Dresden, Germany
[2] Systems Biology Division, Zhejiang-California International NanoSystems Institute, Zhejiang University, Hangzhou, P. R. China

Received: April 27, 2010
Revised: August 11, 2010
Accepted: August 16, 2010

Set1C is a histone methyltransferase playing an important role in yeast gene regulation. Modeling the structure of this eight-subunit protein complex is an important open problem to further elucidate its functional mechanism. Recently, there has been progress in modeling of larger complexes using constraints to restrict the combinatorial explosion in binary docking of subunits. Here, we model the subunits of Set1C and develop a constraint-based docking approach, which uses high-quality protein interaction as well as functional data to guide and constrain the combinatorial assembly procedure. We obtained 22 final models. The core complex consisting of the subunits Set1, Bre2, Sdc1 and Swd2 is conformationally conserved in over half of the models, thus, giving high confidence. We characterize these high-confidence and the lower confidence interfaces and discuss implications for the function of Set1C.

**Keywords:**
Bioinformatics / Histone methyltransferase / Molecular docking / Multiprotein complex / Protein–protein interaction / Protein structure prediction

## 1 Introduction

### 1.1 The Set1 complex

Modification of chromatin structure is an essential aspect of the eukaryotic gene regulation [1, 2]. Histone H3 Lysine-4 (H3K4) methylation is a marker for transcriptionally active chromatin. The members of the SET1 family of histone methyltransferases (HMTs) are responsible for this modification in yeast and higher eukaryotes and all have a conserved SET-domain. The function of the Set1C complex from *Sacharomyces cerevisiae* is to mono-, di- or trimethylate histone 3 lysine 4 [3–5]. The Set1C complex consists of eight subunits (Set1, Bre2, Sdc1, Swd1, Swd2, Swd3, Spp1 and Shg1). Figure 1 shows a representation of the Set1C complex and its interactions derived from experimental data [6]. Set1 protein is the key catalytic component of the Set1C complex from which it derives its name. This subunit is a large protein with a C-terminal SET-domain containing the catalytic site for methylation and two N-terminal RNA recognition motifs. Other core subunits are Bre2, Swd1, Swd2 and Swd3. Swd1 and Swd3 subunits form a heteromer, which is essential for the integrity of the complex and absence of either of the subunits reduces the stability of the complex. Swd2 is also a subunit in another complex, cleavage and polyadenylation factor CPF, which is essential in yeast. In addition, the Swd2 subunit is part of the complex scaffold. Sdc1 and Bre2 subunits were observed to form a

**Correspondence:** Professor Michael Schroeder, Biotechnology Center (BIOTEC), Technische Universität Dresden, Tatzberg 47/49, 01309 Dresden, Germany
**E-mail:** ms@biotec.tu-dresden.de
**Fax:** +49-35146340061

**Abbreviations: CAPRI,** Critical Assessment of Predicted Interactions; **EM,** electron microscopy; **H3K4,** Histone H3 Lysine-4; **HMT,** histone methyltransferase; **L_RMSD,** Ligand Root-Mean-Square-Deviation; **MD,** molecular dynamics; **PDB,** protein data bank; **TAP,** tandem affinity purification

**Colour Online:** See the article online to view Figs. 1-4 in colour.

 

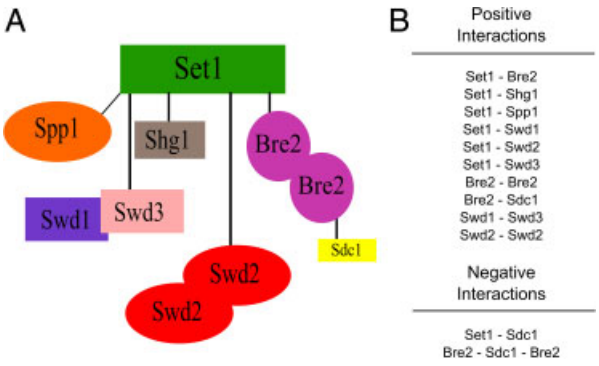


**Figure 1.** The experimental evidence on the protein interactions within the Set1C HMT complex [6]. (A) Schematic presentation of the interactions between the subunits. (B) Experimental data concerning the interactions. The positive interactions as well as the negative ones that are known not to take place in the complex are listed.

heteromer, which is required for trimethylation of histone 3 lysine 4 [3, 6–9]. It has been suggested that the Spp1 subunit regulates trimethylation efficiency.

There is an abundance of biochemical data available on the Set1C complex, but much less is known about its structural features. The used biochemical techniques only provide non-spatial information and do not describe anything about the nature of the interaction. This information can be obtained by 3-D structure determination of binary protein–protein interactions and multimeric complexes. It is technically demanding to produce high-resolution structures of large protein complexes. The number of structures in the protein data bank (PDB) is increasing, but the number of experimental structures of large protein assemblies remains still small relative to the total number of solved structures. About 50 000 structures are known, but only a few thousand of them have more than three subunits. There are no atomic-resolution or even low-resolution models of most protein complexes in the cell. However, understanding of many cellular processes requires structural knowledge about large protein assemblies such as HMTs. Generally, it is not sufficient to know the interacting components to understand the function of a large protein complex.

### 1.2 Protein interactions and docking

Computational approaches can be used in addition to existing experimental methods. A broad variety of sequence-based and structure-based methods exist. If available, one can use known interaction partners in orthologous proteins to infer interactions in another species. This concept of interologs has been described earlier and is particularly helpful when a structure of the interacting complex is available [10–12]. If one knows that A and B interact and there is a crystal structure in which homologous proteins A' and B' interact, one could use it to model the novel A–B interaction. Without such templates, one can employ docking methods, which aim to predict an atomic model of a protein complex by maximizing the shape and/or chemical complementarities between a given pair of interacting proteins. Docking methods predict the structure of a bound protein pair based on their coordinates in the unbound state and use sometimes background information [13]. Critical Assessment of Predicted Interactions (CAPRI) is a community-wide blind docking experiment where research groups can test their methods on new protein targets. ZDOCK [14], PIPER [15], ICM-DISCO [16], RosettaDock [17–19] and Patchdock [20] were the best performing docking methods in the CAPRI rounds 1–11 [21, 22]. Two important aspects in the docking problem are the search method to find possible binding conformations and the scoring function to discriminate the correct docked structure among thousands of possible ones. Fast Fourier Transform, Monte Carlo minimization and stochastic global minimization approaches have been used to search and refine docked structures. The identification of the correct docked structure from the list of false-positive possibilities using a scoring function is a difficult task. The used scoring functions approximate at different levels the binding energy between proteins. Another challenge is how to deal with protein flexibility and conformational changes that take place upon binding [23]. Approaches that have been proposed include implicit or explicit treatment of side-chain/backbone flexibility [24], principal component analysis-based techniques [25, 26] and use of structural ensembles in multiple docking simulations [27]. Computational docking is limited by low accuracy and even the best docking methods have had only a success rate of about 50% in the CAPRI experiment [22]. Thus, new approaches and addition of experimental data are required to improve the docking methods.

### 1.3 Modeling of large complexes

Recently, there have been several attempts to produce either atomic models or more coarse-grained architectural models of large protein complexes. Inbar and co-workers developed a docking algorithm for multimolecular assemblies [28]. Their CombDock approach employs docking methods typical for protein pair modeling for multiprotein complexes. They ranked the produced models using a scoring function based on subunit shape complementary and non-polar buried surface area. Aloy and co-workers started with a set of yeast protein complexes identified by tandem affinity purification (TAP) and selected the most promising ones for electron microscopy (EM) [29]. They modeled the interactions between the component proteins using their similarity (sequence or structural) to interacting proteins of known structure and the obtained EM maps. Alber *et al.* published

 

structural models of the nuclear pore complex, which were produced by combining together diverse biochemical and biophysical data at different levels of resolution [30]. Their structure determination was an iterative series of several steps where experimental data were translated into spatial restraints and an ensemble of structures fulfilling the restraints was calculated. These pieces of work show that it is feasible to predict and model the 3-D structures of large protein assemblies.

This work presents a novel constraint-based assembly approach for structural modeling of large protein complexes. Docking methods are still unreliable but can be improved by using constraints such as experimental knowledge or prediction of the binding site residues. In this work, the possibilities of producing 3-D models of protein complexes by combining computational methods and high-quality experimental data were explored. The experimentally derived high-quality data provided by protein tagging, MS and two-hybrid approaches forms the basis for 3-D modeling of HMT Set1C complex from *Saccharomyces cerevisiae* [6]. Our aim is to produce a 3-D model of the Set1C complex in order to shed light on the mechanism and regulation of histone methylation.

## 2 Materials and methods

Figure 2 presents the workflow of our complex assembly approach. The different methods used in the individual steps of the workflow were the following.

### 2.1 Comparative modeling and molecular dynamics (MD) simulations

Structural models were produced for each individual subunit with comparative modeling using MODELLER [31, 32] or I-TASSER structural modeling pipeline [33]. MD simulations were employed to study the structural stabilities of the proteins. MD simulations were carried out using the program NAMD [34]. CHARMM27 force field was used for description of the protein [35] and the TIP3P solvent model represented the water molecules [36]. The initial structures were routinely relaxed by 10 000 steps of conjugate gradient energy minimization prior to simulations. Simulations assumed constant particle number, constant pressure and constant temperature (NpT) ensembles. Langevin dynamics was used to maintain constant temperature and pressure was controlled using a hybrid Nose-Hoover Langevin piston method.

### 2.2 Modeling of binary interaction structures

To obtain structural models for each binary protein–protein interaction existing within the Set1C complex, three different techniques were employed. One approach to produce correct models of binary interactions was the use of structural interaction templates [29]. This is possible when there exists an experimental structure, which contains similar proteins in a complex with each other. SCOPPI database, which classifies domain–domain interactions from all known protein structures, was used to search interaction templates [37]. Often, there are no suitable structural interaction templates available and, hence, molecular docking has to be employed. RosettaDock [17–19] and ZDOCK (version 3.0.1) [14] were used for docking of the interacting protein pairs. Prediction of putative binding sites on the surfaces of the two interacting proteins can be used as a guide in the docking simulations. Here, we used metaPPI server to predict potential binding sites on the individual proteins. MetaPPI integrates the prediction results from five prediction servers and scores the predictions and the interface residues involved in binding [38].

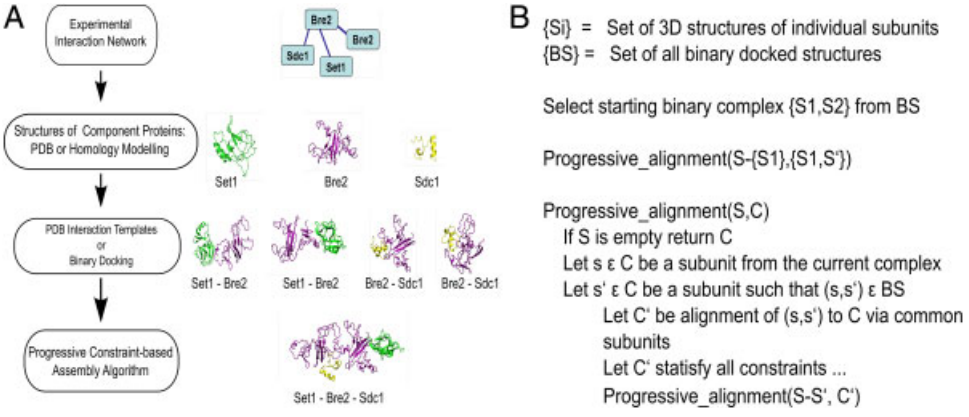


**Figure 2.** Schematic presentation of the modeling workflow. (A) Experimental protein–protein interaction data were the starting point of the modeling process. Progressive constraint-based assembly algorithm was used to produce the higher order structures after obtaining structures of the individual protein subunits and those of the all possible protein pairs. (B) Pseudo-code presenting the progressive constraint-based assembly algorithm. Binary docked structures *BS* are added one-by-one to the current complex *C*.

 

### 2.3 Progressive constraint-based assembly algorithm

The selected binary complexes were assembled together to form higher order structures. In our approach, all available information of the Set1C complex was used as constraints for producing the final structures. This includes experimental interaction data, binding site predictions, ensembles of binary-docked structures and structural interaction templates. The aim was to find a solution/structure which satisfies all the constraints. The basic idea in the complex assembly is combining of binary-docked structures by structurally aligning the common subunit of two binary-docked structures. In this way, a ternary complex is produced. Then, the next binary-docked subunit can be added to this three-subunit complex to form a four-subunit protein complex. In a similar manner, all subunits were added to the starting binary complex one-by-one. At this point of the assembly process, ten models (the ten best scoring clustered docked solutions) for each protein pair were considered instead of a single conformation. This leads to a huge number of possible combinations between the subunits. The complexity of this kind of progressive structural alignment of subunits is exponential (Fig. 3A). We have solved this problem by using different experimental and computational constraints (Fig. 3B). The first constraint check was whether subunits overlap in space after the addition of the new subunit. Combinations leading to spatial overlap were discarded from further analysis. Then, the fulfillment of all experimental constraints was checked to filter the solution space even further.

To demonstrate the general applicability of our assembly algorithm, RNA polymerase II complex (PDBid:1twf) was used as a benchmark. RNA polymerase II complex consists of ten different subunits (chains) and the interactions between them are shown in the Supporting Information Fig. S1. We used ZDOCK (version 3.0.1) to dock the interacting subunit pairs in RNA Polymerase II complex. For the pairwise docking by ZDOCK, the larger protein is kept static (called as the receptor) and the smaller one mobile (called as the ligand). The Ligand Root-Mean-Square-Deviation (L_RMSD) is calculated for ligand's CA atoms to assess the docking solutions after superimposing the receptor in docking solutions and the receptor in native complexes. If L_RMSD value is less than 10 Å, we define this docking solution as a near-native structure (hit). For each ZDOCK run, we kept 2000 solutions based on their pair-wise shape complementarity scores and only the top 10 are used for assembling. After docking all interaction pairs, the assembly was performed with the progressive-constraint-based approach. The order of assembly depended on the size of domains.

## 3 Results

The main research question of this work was to examine whether it is possible to produce atomic level 3-D models of

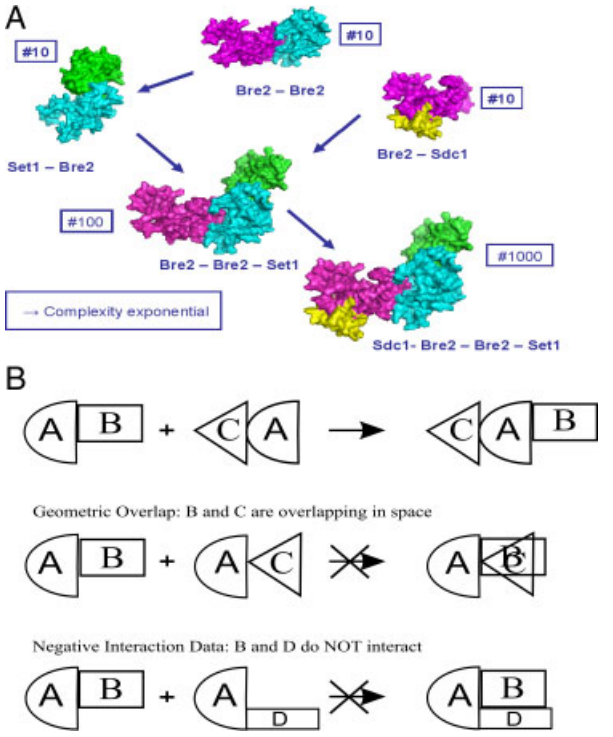


**Figure 3.** (A) The binary docked structures are combined by structurally aligning the common subunit. In this way, a ternary complex is produced. In a similar manner, all subunits were added to the starting binary complex one-by-one. Ten structural models for each pair were considered for each protein pair. This leads to a huge number of possible combinations between the subunits. (B) The combinatorial explosion is prevented by using different experimental and computational constraints. The first check was whether subunits overlap in space after the addition of the new subunit. Then all experimental constraints were checked.

the Set1C complex by combining computational techniques that are traditionally used only for pairs of proteins and high-quality experimental data. Figure 2 shows a schematic presentation of the work flow. We obtained 22 models of the eight-subunit Set1C complex (Fig. 4). This ensemble of structures satisfies the given experimental constraints and provides a starting point for experimental studies. The core part of the structural models consists of the subunits Set1, Bre2, Sdc1 and Swd2. This core part is conformationally conserved in over half of the models, thus, giving high confidence for interactions of these subunits. The sizes of the interaction interfaces varied from about 1500 to almost 2000 $Å^2$, which is typical for permanent interactions [39]. Our progressive assembly process takes into consideration the ten most probable (top 10) binding conformations for every protein subunit pair. During the assembly process, some of these binding conformations were discarded as they did not fit together with other subunits to fulfill the experimental constraints or led to spatial overlap with other subunits. The existing binary binding conformations in the

 

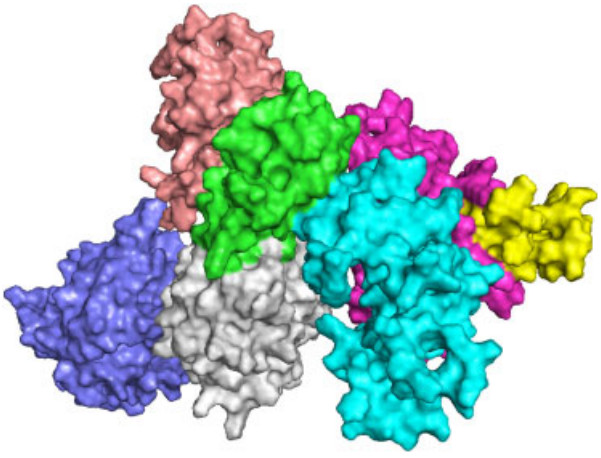

**Figure 4.** One of the final ten-subunit structural models of the Set1C HMT from *S. cerevisiae*.

final structural models were analyzed. The Set1–Bre2 and Set1–Swd2 subunit pairs had the most significant reduction in the number of possible binding orientations (see Section 4 below). All found final binary binding conformations can be considered as high-confidence interaction predictions as they were selected from the ensemble of possible conformations using constraints.

The assembly algorithm was run using top 5, top 10, top 20 and top 50 binary-docked structures for each protein pair in order to assess the stability of final results. The top 5 run ended after addition of the first three subunits as the addition of the fourth subunit lead in all cases to geometrical overlap. The number of final structures were 0, 22, 1456 and 52 180 for top 5, top 10, top 20 and top 50, respectively. The final number of models is quite moderate compared to the absolute number of possibilities, even in the case of top 50 docked structures, which indicates stability of the approach. In the case of top 20 docked structures, 37% of the final structures had the conformational conserved core of subunits. These runs were made using RosettaDock algorithm for the protein docking.

### 3.1 Experimental constraints

The experimental data used in this work come from specific small-scale experiments (bacterial two-hybrid approaches, TAPs and MS analysis on wild-type complex and domain deletion strains) reported by Dehe and colleagues [6]. The experiments provide information about positive and negative interactions within the complex (Fig. 1). The negative interactions are connections that are known not to exist in the complex. Figure 1 shows the experimentally observed physical interactions within the Set1C and the constraints used for structural modeling of their pairwise interaction. The catalytic subunit Set1 has been shown to interact directly with all subunits except Sdc1 [6]. It was observed that Sdw1–Swd3 heterodimer does not need the SET-domain of Set1 subunit for interactions. It was furthermore shown that association of Bre2 and Sdc1 requires the C-terminal SET-domain, but Sdc1 does not directly interact with Set1. These pieces of information form important constraints for our modeling problem. The interactions of Set1 subunit with the other subunits in our 3-D models are discussed in the following sections.

### 3.2 Comparative modeling and MD analysis of the subunits

The modeling process was started with prediction of the structures of the protein components with comparative modeling and threading [32, 33]. Comparative modeling requires high-quality multiple sequence alignment for the target protein and structural template with reasonable sequence similarity. Table 1 shows the subunits of Set1C complex and the sequence identity they have with the used structural template. It is well established that pairs of sequences sharing more than about 25% sequence identity can be confidently said to be similar in structure. Below this there is a twilight zone where dissimilar sequences might be mixed up with those sharing structural similarity. The structural stability of the models was tested by means of MD simulations. MD simulations provide detailed information on the fluctuations and the conformational changes of proteins. The fluctuations of the modeled structures were observed for 1 ns and the average RMSD of the backbone Cα atoms of each model was calculated (Table 2). The generated models were accepted if their overall structure was conserved during the simulation (RMSD $< 5$ Å). The RMSD values less than 3 Å during the simulation are generally

**Table 1.** The subunits of Set1C complex and the structural templates that were used in the modeling process

| Protein | PDB id | Seq. Id (%) | RMSD (Å) |
|---|---|---|---|
| Set1 | 1MVH | 36 | 3.7 |
| Bre2 | 1WOR, 1FLG, 2CN2, 1KV9, 1ZLG, 2YYO | 15 | 4 |
| Spp1 | 2FSA | 41 | 3.3 |
| Swd1 | 2HL4 | 28 | 4 |
| Swd2 | 2HL4 | 21 | 3.5 |
| Swd3 | 2HL4 | 37 | 3.7 |
| Shg1 | 1E5K | 18 | 4.8 |
| Sdc1 | 3JX8 & 3G36 | 40 | 3.5 |

The sequence identities between subunits and their templates and the average RMSD of the modeled structure Cα atoms are shown. The fluctuations of the modeled structures were observed for 1 ns and the average Root-Mean-Square-Deviation (RMSD) of the backbone Cα atoms of each model was calculated. The structural models were accepted if their overall structure was conserved during the simulation (average RMSD $< 5$Å).

 

**Table 2.** Interface properties derived from the final three-dimensional models of the full Set1C complex

| Protein pair | Interface1 (Å$^2$) | Interface2 (Å$^2$) | ♯ Int. residues | ♯ Hydrophobic | ♯ Charged/polar | Different conformations |
|---|---|---|---|---|---|---|
| Set1 – Bre2 | 1820 ± 11 | 1830 ± 7 | 15, 10 | 3, 1 | 10, 7 | 50%, 36%, 14% |
| Set1–Swd2 | 1600 ± 55 | 1900 ± 112 | 13, 16 | 3, 9 | 18, 16 | 79%, 21% |
| Set1 – Spp1 | 1610 ± 20 | 1780 ± 3 | 15, 18 | 3, 5 | 12, 8 | 27%, 18%, 18%, 14%, 14%, 9% |
| Set1 – Shg1 | 1550 ± 486 | 1470 ± 346 | 40, 3 | 15, 7 | 26, 23 | 18%, 18%, 16%, 16%, 13%, 13%, 6% |
| Set1–Swd3 | 1690 ± 390 | 2030 ± 309 | 25, 32 | 8, 12 | 15, 17 | 40%, 30%, 16% 10%, 4% |
| Swd1 – Swd3 | 1560 ± 308 | 1540 ± 340 | 38, 62 | 11, 18 | 25, 38 | 27%, 20%, 15%, 10%, 10%, 10%, 8% |
| Swd2 – Swd2 | 1496 ± 29 | 1447 ± 35 | 31, 27 | 11, 8 | 17, 17 | 37%, 35%, 28% |
| Bre2 – Sdc1 | 1900 | 1500 | 15, 17 | 2, 8 | 13, 9 | 100.00% |

Interface1 and Interface2 show the interface areas on the two binding partners in Å$^2$ and their standard deviations. The interface area was on average 1400 Å$^2$ which is typical for permanent interactions. The interaction between Bre2 and Sdc1 was modeled using a single structural template and, hence, the interface areas were constant in all final models. The percentage of occurrence of different binding conformations of each pair in the final models are listed.

considered to be a sign of a stable structure. Some of the models did not fulfill this criterion, which should be taken into account when further working on these proteins. On the other hand, large RMSD values of certain parts of proteins might indicate that they contain intrinsically disordered regions, which are possibly important for interactions with other proteins.

### 3.3 Binary protein–protein interaction between subunits

The next step was to structurally model all existing binary interactions within Set1C between the subunits. In the best case, there exists a structural interaction template, an experimental structure of two interacting proteins homologous to the proteins studied. Comparative modeling can be then used to build a model of a complex using the template. Henschel and co-workers extracted and classified all domain-domain interactions found in the PDB [40]. In 40% of the cases, the interacting domain families associate in multiple orientations, suggesting that all the possible binding orientations need to be explored for protein complexes. Hence, even homologous domain pairs can interact in geometrically different ways by employing different sets of residues to form interfaces. Often, proteins of a complex show good homology to a known single-protein structure, but there are no suitable templates on which to model its interactions. In this case, docking can be used to search the complex structure by maximizing shape complementary and physicochemical properties. In this way, it is possible to predict novel-binding configurations. We used RosettaDock, which provides simultaneous optimization of side-chain conformation and rigid body positions of the two docking partners. RosettaDock was chosen for the Set1 subunit docking, as it achieved good results in the CAPRI experiments on prediction of targets where one of the proteins was a homology model and side-chain repacking was essential to find correct binding conformation [22]. It also has a very accurate scoring function consisting of several energy terms. Ten thousand independent docking simulations were done for each protein–protein pair. Docking generated a large amount of data, which had to be processed and analyzed. First, the produced conformations were scored using Rosetta's energy function. Docking can produce complex conformations that are very similar. Thus, it is important to be able to select from the data set a smaller ''representative'' set of conformations for subsequent analysis. This was done with clustering analysis, which groups together similar structures based on their RMSD values. A representative binary-docked structure was chosen from each cluster and the ten best scoring cluster representatives were used for further analysis.

### 3.4 Constraint-based assembly algorithm

The chosen binary complexes were then assembled together to form higher order structures and finally the full model of Set1C complex. Computational docking is limited by low accuracy and one has to produce tens of thousands of structures to obtain some near the native structure. It is difficult to find the correct docked structure among all the candidates. To overcome this problem, we used several of the best ranking predicted binding conformations. The use of the docked structures to produce larger protein assemblies makes sense if experimental and computational information is used to restrict the solution space. We used an approach where all available information of the protein complex was taken as constraints for producing larger structures. These constraints include experimental data, binding site predictions, the structures of binary-docked complexes, interaction templates and binding free energy calculations. The catalytic subunit Set1 has been shown to interact directly with all subunits except Sdc1 ([6], Fig. 1). Dehe and co-workers analyzed direct interactions within the complex using a bacterial two-hybrid system and TAP-tagging with MS. The TAP-tagging

 

experiments were done in addition to the wild-type Set1 with the N-terminus of Set1 (residues 1–283), the Set1 RNA recognition motifs (residues 234–580) and Set1 C-terminal domain (residues 580–1080) and with yeast strains lacking Bre2, Sdc1, Spp1 or Shg1. The interactions of Set1 subunit with the other subunits in our 3-D models are discussed in the following sections.

### 3.5 Bre2, Sdc1 and their interactions

The Bre2 protein interacts directly with the Sdc1 and Sdc1 is known not to have direct contact with Set1 subunit. The structure of the human homolog of Sdc1, DPY-30L, is available and it exists as a homodimer (PDB identifier: 3G36) [9]. One DPY-30L monomer consists of two short helices linked by a short turn, which participate in dimer formation. The α-helices from the two monomers form an antiparallel bundle. The dimer interface is mainly hydrophobic and contains the highly conserved DPY-30 motif. Wang and co-workers proposed that ASH2L, the human homolog of Bre2, might interact with DPY-30L using the dimerization interface. In accordance, it has been suggested that Sdc1-dimerization is important for interaction of Sdc1 and Bre2 with Set1 [6]. The DPY-30L structure indicates that two conserved leucines in the Sdc1 (DPY-30 domain) protein are required for binding with Bre2 (or human ASH2L) [9]. Biochemical and mutational data together with this structural information also suggest that the C-terminus of Bre2 contains so-called SDI (Sdc1 Dpy-30 Interaction) domain which is essential to interaction with Sdc1 [8]. We modeled the Bre2–Sdc1 interaction structurally using the interaction template and found that the interface sizes between Bre2 and Sdc1 were 1900 and 1500 $Å^2$, respectively. These interfaces are relatively large and in accordance with the permanent nature of the interaction. The interaction between these subunits is mainly based on interactions between polar and charged residues.

### 3.6 Swd3 and Swd1 – a heterodimer within the complex

Swd1 and Swd3 subunits form a heteromer that is essential for the integrity of the whole complex [3]. Absence of either of the subunits reduces the stability of the complex [6]. Swd3 and its mammalian homolog Wdr5 have seven WD40 repeats that have a β-propeller fold and have been shown to be required for the complex assembly. No H3K4 methylation takes place in their absence [41, 42]. Structural evidence shows histone H3 binding in the central hole of the Wdr5 β-propeller (PDB identifier: 2H6N) [43]. Another recent crystal structure of MLL1 catalytic domain SET, a human homolog of Set1, shows that it binds dimethylated H3K4, which makes histone binding to Wdr5 impossible (PDB identifier: 2W5Z) [44]. In addition, a sequence motif named as the WIN-motif was identified in the N-terminal region of the catalytic SET domain and proposed to mediate interaction between MLL1 and Wdr5 [45, 46]. Hence, it seems likely that the Set1 subunit uses a binding motif similar to the histone tail to interact with Swd3. In our docking simulations, two prominent binding conformations which fulfill this interaction constraint were found for the Set1–Swd3 pair. The binding conformations suggest that interaction between these subunits is due to charged and polar residues. The interface sizes of these two conformations are 1770 and 2044 Å, both of which ensure stable interaction.

### 3.7 Swd2 interactions

Swd2 is probably a homodimer [3] and part of the complex scaffold [6]. It is also the subunit which forms a link between CPF (cleavage and polyadenylation factor) and Set1C by being part of the both complexes. Lack of Swd2 in yeast leads to global loss of H3K4 di- and trimethylation [47, 48]. The human homolog of Swd2, Wdr82, was shown to interact with the RNA recognition motif of the human Set1A histone methyltranferase complex (human homolog of Set1C) and to bind to phosphorylated RNA polymerase II (pol II) [49]. In yeast, the Set1C interaction with pol II takes place *via* the Paf1 complex [50]. Thus, the binding of Swd2 to Set1 subunit in yeast could take place using another structural region of Set1 than in the human complex. Initially, there was an ensemble of ten binary-docked structures for the Set1–Swd2 subunit pair. However, only two of the ten initial binary structures were present in the final structural models. These both models have several features in common: the interaction interfaces are rich in polar and charged residues in both interface sides. Furthermore, the interface region is between residues Set1 340–420, a region between the two Set1 RNA recognition motifs. This finding indicates that the binding between Set1 and Swd2 has similar features as binding between Wdr82 and Set1A, even though in yeast the association of Set1 with pol II involves the Paf1 complex. This piece of information about Swd2-binding site on Set1 subunit was not included in the assembly process and, thus, indicates correctness of the modeling process. The binary-docked structures of the Swd2–Swd2 showed three different patches of binding residues on the protein (Swd2 residues 16–17, 102–105 and 142–145), which are also conserved regions in the Swd2 proteins. Interface conservation has been shown to be a good indicative of correct binding orientation [51].

### 3.8 Shg1, Spp1 and their interactions

Association of Shg1 and Spp1 with the Set1 subunit does not require the C-terminal part of Set1 and they do not interact with any other subunit in the Set1C complex [6]. Spp1 was pulled down in the TAP-tagging experiments with

 

Set1 residues 580–900 [4, 6]. The produced Set1–Spp1 binary-docked structures satisfy this constraint and the best scoring representatives were enriched around two regions of Set1: 750–770 and 840–850. The used ensemble of binary-docked structures of Set1–Shg1 subunit pair contained very distinct binding conformations. The final structural models contained seven of these different binding conformations.

### 3.9 Benchmarking results: the RNA polymerase II complex

The pairwise docking results for all interacting subunit pairs of RNA Polymerase II complex are shown in Table 3. There was at least one binding conformation highly similar to the native interaction structure in the top 10 solutions for all pairs except the b-k subunit pair, which was not used in the final assembly (the c-k pair was used instead for adding the k subunit). The order of assembly was a-b, b-c, a-e, a-f, a-h, a-i, c-k, b-j and b-l. In the final assembled structures, there are six models whose average L_RMSD value is below 10 Å. Supporting Information Fig. S2 shows the best model after assembling (ranking: top 1, average L_RMSD: 1.63 Å), compared with the native structure.

## 4 Discussion

### 4.1 Mono-, di- and trimethylation efficiency depends on different subunits

The three different states of H3K4 methylation mark different transcription states of chromatin. H3K4 trimethylation is most prominent in the beginning of the actively transcribed regions of genes as dimethylated H3K4 is mostly found in the middle of the genes and monomethylated H3K4 is present the end of the genes [52]. The

**Table 3.** Detailed pair-wise docking results of RNA Polymerase II complex

| Pairs | Number of hits | Best L_RMSD | Ranking of hits |
|---|---|---|---|
| a-b | 1 | 2.06 | 1 |
| a-e | 3 | 1.26 | 1, 2, 5 |
| a-f | 2 | 1.00 | 1, 2 |
| a-h | 2 | 1.45 | 1, 2 |
| a-i | 1 | 1.48 | 1 |
| a-k | 2 | 1.64 | 3, 6 |
| b-c | 3 | 1.42 | 1, 6, 8 |
| b-i | 1 | 1.93 | 1 |
| b-j | 4 | 1.52 | 1, 4, 6, 9 |
| b-k | 0 | – | – |
| b-l | 1 | 1.45 | 1 |
| c-j | 2 | 1.05 | 3, 6 |
| c-k | 4 | 1.28 | 1, 2, 4, 6 |

ZDOCK generates more than 1 hit for all the pairs except b-k and most of the best hit is in the top 1.

SET-domain-containing proteins have a highly conserved tyrosine/phenylalanine residue in their lysine-binding pockets. This residue is essential for the product specificity, so-called Tyr/Phe switch, which controls whether the product is mono-, di- or trimethylated lysine [53, 54]. Set1 has a tyrosine residue (Tyr1052) on this position, which indicates pure monomethylation based on the Tyr/Phe switch theory. Hence, it was postulated that Tyr1052 has an important role in the modification of space available in the catalytic site of Set1 and regulates its transition from mono- to di- and trimethylation. Spp1 and Bre2 subunits were found to be essential for proper H3K4 trimethylation by Set1C [55]. It was suggested that they interact directly with the Set1 subunit and *via* this interaction they can alter the lysine-binding pocket and, hence, allow H3K4 trimethylation. In a Spp1 deletion mutant, the trimethylation was reduced by around three-fold, but dimethylation was not affected [6]. Binding of Spp1 subunit would alter the binding pocket and is a prerequisite for trimethylation. Mutation of residue Tyr1052 into a phenylalanine was shown to compensate the loss of Spp1 in the H3K4 trimethylation activity [56]. Removal of the tyrosine hydroxyl group would suppress the loss of Spp1 in trimethylation. In this study, two possible regions for the Set1–Spp1 interaction were found (Set1 residues 750–770 and 840–850). The latter of the predicted binding sites is in the vicinity of the catalytic site/Tyr1052 and is consistent with the experimental findings of Spp1 modification of the catalytic site.

### 4.2 Conclusions

We developed a progressive constraint-based assembly algorithm for 3-D modeling of protein complexes and applied it to the study of Set1C H3K4 methyltranferase from *S. cerevisiae*. The modeling constraints were taken from several experimental studies providing knowledge of existing interactions. There are only few existing methods for structural modeling of protein complexes with more than two subunits. The complex modeling approaches of Aloy and Alber rely on EM data which give overall shapes of complexes [29, 30]. However, production of an EM map of a complex is not always technically feasible and these data were not available for Set1C. Our protein complex assembly algorithm uses only biochemical protein–protein interaction and functional data as constraints for modeling. The EM-based approaches are scalable to our method, but the constraint-based assembly algorithm cannot yet be run fully automatically and at large scale. However, this work shows that the use of only non-spatial biochemical data is sufficient to produce reliable structural models. A problem in the use of structural interaction templates is that in 40% of the cases, the interacting domain families associate in multiple orientations. Our docking-based approach allows identification of these novel-binding orientations. Several docked structural alternatives were considered for each protein pair

 

in the complex. This gives a better sampling of the structural search space, but it also means that the search algorithm has a high branching factor. The progressive addition of subunits in the computational model could lead to exponential increase of possible models. However, the combinatorial explosion of possible structural models was prevented by data integration and the use of negative information, *i.e.* knowledge of interactions that do not exist in the complex. The generated structural models form a starting point for experimental testing and functional analysis. We tested our approach on a protein complex with a known 3-D structure, RNA polymerase II with ten subunits, and could produce structural models that were within RMSD of 10 Å from the native complex structure.

Overall, this work shows that constraint-based assembly of large complexes is feasible and can generate novel hypotheses. In the particular study in Set1, our approach demonstrates that Set1, Bre2, Sdc1 and Swd2 form the core structural backbone conserved in over half of the final models. Additionally, we find models where predicted binding sites are consistent with experimental evidence on the change in product specificity (mono-, di- or trimethylation) upon mutation.

*The authors have declared no conflict of interest.*

## 5 References


[1] Cheng, X., Collins, R. E., Zhang, X., Structural and sequence motifs of protein (histone) methylation enzymes. *Annu. Rev. Biophys. Biomol. Struct.* 2005, *34*, 267–294.

[2] Turner, B. M., Simplifying a complex code. *Nat. Struct. Mol. Biol.* 2008, *15*, 542–544.

[3] Roguev, A., Schaft, D., Shevchenko, A., Pijnappel, W. *et al.*, The *Saccharomyces cerevisiae* Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *EMBO J.* 2001, *20*, 7137–7148.

[4] Krogan, N. J., Dover, J., Khorrami, S., Greenblatt, J. F. *et al.*, COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression. *J. Biol. Chem.* 2002, *277*, 10753–10755.

[5] Nagy, P. L., Griesenbeck, J., Kornberg, R. D., Cleary, M. L., A trithorax-group complex purified from *Saccharomyces cerevisiae* is required for methylation of histone H3.*Proc. Natl. Acad. Sci. USA* 2002, *99*, 90–94.

[6] Dehe, P.-M., Dichtl, B., Schaft, D., Roguev, A. *et al.*, Protein interactions within the Set1 complex and their roles in the regulation of histone 3 lysine 4 methylation. *J. Biol. Chem.* 2006, *281*, 35404–35412.

[7] Dehe, P.-M., Geli, V., The multiple faces of Set1. *Biochem. Cell Biol.* 2006, *84*, 536–548.

[8] South, P. F., Fingerman, I. M., Mersman, D. P., Du, H.-N., Briggs, S. D., A conserved interaction between the SDI domain of Bre2 and the Dpy-30 domain of Sdc1 is required for histone methylation and gene expression. *J. Biol. Chem.* 2010, *285*, 595–607.

[9] Wang, X., Lou, Z., Dong, X., Yang, W. *et al.*, Crystal structure of the C-terminal domain of human DPY-30-like protein: A component of the histone methyltransferase complex. *J. Mol. Biol.* 2009, *390*, 530–537.

[10] Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L. *et al.*, Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 2000, *287*, 116–122.

[11] Gerstein, M., Lan, N., Jansen, R., Proteomics. Integrating interactomes. *Science* 2002, *295*, 284–287.

[12] Aloy, P., Russell, R. B., Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 5896–5901.

[13] Andrusier, N., Mashiach, E., Nussinov, R., Wolfson, H. J., Principles of flexible protein-protein docking. *Proteins* 2008, *73*, 271–289.

[14] Wiehe, K., Pierce, B., Tong, W. W., Hwang, H. *et al.*, The performance of ZDOCK and ZRANK in rounds 6–11 of CAPRI. *Proteins* 2007, *69*, 719–725.

[15] Kozakov, D., Brenke, R., Comeau, S. R., Vajda, S., PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* 2006, *65*, 392–406.

[16] Fernandez-Recio, J., Totrov, M., Abagyan, R., ICM-DISCO docking by globalenergy optimization with fully exible side-chains. *Proteins* 2003, *52*, 113–117.

[17] Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O. *et al.*, Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* 2003, *331*, 281–299.

[18] Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., Baker, D., Progress in modeling of protein structures and interactions. *Science* 2005, *310*, 638–642.

[19] Wang, C., Schueler-Furman, O., Andre, I., London, N. *et al.*, RosettaDock in CAPRI rounds 6–12. *Proteins* 2007, *69*, 758–763.

[20] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H. J., PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 2005, *33*(*Web Server issue*), W363-W367.

[21] Mendez, R., Leplae, R., Lensink, M. F., Wodak, S. J., Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins* 2005, *60*, 150–169.

[22] Vajda, S., Kozakov, D., Convergence and combination of methods in protein-protein docking. *Curr. Opin. Struct. Biol.* 2009, *19*, 164–170.

[23] Bonvin, A. M. J. J., Flexible protein-protein docking. *Curr. Opin. Struct. Biol.* 2006, *16*, 194–200.

[24] Ehrlich, L. P., Nilges, M., Wade, R. C., The impact of protein flexibility on protein-protein docking. *Proteins* 2005, *58*, 126–133.

[25] Hinsen, K., Reuter, N., Navaza, J., Stokes, D. L., Lacapre, J.-J., Normal mode-based _tting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys. J.* 2005, *88*, 818–827.

[26] Lindahl, E., Delarue, M., Refinement of docked protein-ligand and protein-DNA structures using low frequency


 

normal mode amplitude optimization. *Nucleic Acids Res.* 2005, *33*, 4496–4506.

[27] Krol, M., Tournier, A. L., Bates, P. A., Flexible relaxation of rigid-body docking solutions. *Proteins* 2007, *68*, 159–169.

[28] Inbar, Y., Benyamini, H., Nussinov, R., Wolfson, H. J., Prediction of multi-molecular *assemblies by multiple docking*. *J. Mol. Biol.* 2005, *349*, 435–447.

[29] Aloy, P., Russell, R. B., Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* 2004, *22*, 1317–1321.

[30] Alber, F., Dokudovskaya, S., Veenho, L. M., Zhang, W. *et al.*, Determining the architectures of macromolecular assemblies. *Nature* 2007, *450*, 683–694.

[31] Sali, A., Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 1993, *234*, 779–815.

[32] Marti-Renom, M., Stuart, A., Fiser, A., Sanchez, R. *et al.*, Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 2000, *29*, 291–325.

[33] Roy, A., Kucukural, A., Zhang, Y., I-TASSER: a unified platform for auto-mated protein structure and function prediction. *Nat. Protoc.* 2010, *5*, 725–738.

[34] Phillips, J. C., Braun, R., Wang, W., Gumbart, J. *et al.*, Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 2005, *26*, 1781–1802.

[35] MacKerell, J. A. D., Bashford, D., Bellott, M., Dunbrack, R. L. Jr., *et al.*, All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 1998, *102*, 3586–3616.

[36] Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R., Kein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 1983, *79*, 926–935.

[37] Winter, C., Henschel, A., Kim, W. K., Schroeder, M., SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.* 2006, *34*, D310–D314.

[38] Huang, B., Schroeder, M., Using protein binding site prediction to improve protein docking. *Gene* 2008, *422*, 14–21.

[39] Horton, N., Lewis, M., Calculation of the free energy of association for protein complexes. *Protein Sci.* 1992, *1*, 169–181.

[40] Henschel, A., Winter, C., Kim, W. K., Schroeder, M., Using structural motif descriptors for sequence-based binding site prediction. *BMC Bioinformatics* 2007, *8*, S5.

[41] Steward, M. M., Lee, J.-S., O'Donovan, A., Wyatt, M. *et al.*, Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nat. Struct. Mol. Biol.* 2006, *13*, 852–854.

[42] Dou, Y., Milne, T. A., Ruthenburg, A. J., Lee, S. *et al.*, Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nat. Struct. Mol. Biol.* 2006, *13*, 713–719.

[43] Ruthenburg, A. J., Wang, W., Graybosch, D. M., Li, H. *et al.*, Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nat. Struct. Mol. Biol.* 2006, *13*, 704–712.

[44] Southall, S. M., Wong, P.-S., Odho, Z., Roe, S. M., Wilson, J. R., Structural basis for the requirement of additional factors for MLL1 SET domain activity and recognition of epigenetic marks. *Mol. Cell* 2009, *33*, 181–191.

[45] Patel, A., Vought, V. E., Dharmarajan, V., Cosgrove, M. S., A conserved arginine-containing motif crucial for the assembly and enzymatic activity of the mixed lineage leukemia protein-1 core complex. *J. Biol. Chem.* 2008, *283*, 32162–32175.

[46] Song, J.-J., Kingston, R. E., WDR5 interacts with mixed lineage leukemia (MLL) protein *via* the histone H3-binding pocket. *J. Biol. Chem.* 2008, *283*, 35258–35264.

[47] Cheng, H., He, X., Moore, C., The essential WD repeat protein Swd2 has dual functions in RNA polymerase II transcription termination and lysine 4 methylation of histone H3. *Mol. Cell. Biol.* 2004, *24*, 2932–2943.

[48] Roguev, A., Schaft, D., Shevchenko, A., Aasland, R. *et al.*, High conservation of the Set1/Rad6 axis of histone 3 lysine 4 methylation in budding and fission yeasts. *J. Biol. Chem.* 2003, *278*, 8487–8493.

[49] Lee, J.-S., Shukla, A., Schneider, J., Swanson, S. K. *et al.*, Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell* 2007, *131*, 1084–1096.

[50] Ng, H. H., Robert, F., Young, R. A., Struhl, K., Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol. Cell* 2003, *11*, 709–719.

[51] Choi, Y. S., Yang, J.-S., Choi, Y., Ryu, S. H., Kim, S., Evolutionary conservation in multiple faces of protein interaction. *Proteins* 2009, *77*, 14–25.

[52] Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M. *et al.*, Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 2005, *122*, 517–527.

[53] Zhang, X., Tamaru, H., Khan, S. I., Horton, J. R. *et al.*, Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase. *Cell* 2002, *111*, 117–127.

[54] Collins, R. E., Tachibana, M., Tamaru, H., Smith, K. M. *et al.*, *In vitro* and *in vivo* analyses of a Phe/Tyr switch controlling product specificity of histone lysine methyltransferases. *J. Biol. Chem.* 2005, *280*, 5563–5570.

[55] Schneider, J., Wood, A., Lee, J.-S., Schuster, R. *et al.*, Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. *Mol. Cell* 2005, *19*, 849–856.

[56] Takahashi, Y.-H., Shilatifard, A., Structural basis for H3K4 trimethylation by yeast Set1/COMPASS. *Adv. Enzyme Regul.* 2010, *50*, 104–110.

 