**Supplemental Information**

**Integrative modeling of biomolecular complexes: HADDOCKing with Cryo-EM data**

Gydo C.P. van Zundert1, Adrien S.J. Melquiond1,2, Alexandre M.J.J. Bonvin1\*

1Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, 3584CH, the Netherlands.

2Current address: Hubrecht Institute - KNAW & University Medical Center Utrecht

Uppsalalaan 8, 3584 CT Utrecht, the Netherlands

\*Correspondence: [a.m.j.j.bonvin@uu.nl](mailto:a.m.j.j.bonvin@uu.nl) (A.M.J.J.B.)

**Modeling a virus-antibody complex with 21Å resolution cryo-EM data**

In order to show how a HADDOCK run is performed with cryo-EM data, we will go through an example step by step to fully show all parameter settings and stages involved. The immature Dengue virus will be used for demonstration purposes for which 21Å resolution cryo-EM data is available, in addition to high-resolution models. Also, the biology of the system can be leveraged during the docking, as the approximate interface of both the antibody and the prM chain are known.

*Preparing the input data for docking*

First, cryo-EM data were downloaded from the EMDB (<http://www.emdatabank.org>, EMD-5675). The virus-antibody map is sizeable ( voxels) and since we are only interested in a small subset of the map, the cryo-EM data were cropped such that only one subunit is present in the density. This will significantly accelerate the search, as the cross-correlation potential and its derivative are calculated via FFTs and thus benefits greatly from the reduced map size. The data can be manually cropped using UCSF Chimera (<http://www.cgl.ucsf.edu/chimera/download.html>) by selecting a sub-region in space and saving the density to file (*subunit.mrc*). Next, we adjusted the size of the map further, such that the number of voxels in each dimension is a multiple of 2, 3 and 5 as this is a requisite for the FFT calculations in CNS, ultimately resulting in a significantly smaller map size ( voxels). The cryo-EM data format was subsequently converted to XPLOR/CNS format. For both tasks a Python script was made, which are included in the HADDOCK-EM distribution, and were executed with the commands (after setting up the HADDOCK-EM environment):

>>> $HADDOCKEMTOOLS/resize\_to\_235multiple subunit.mrc subunit\_resized.mrc

>>> $HADDOCKEMTOOLS/convert\_em subunit\_resized.mrc subunit.xplor

Next, high-resolution models were downloaded from the PDB for the envelope protein-prM complex and the antibody. The latter was extracted from the current model (PDB-ID: 3J42, chains K and L) and saved to file (*antibody.pdb*), while the former was taken from the original model (PDB-ID: 3C5X), since the 3J42-model lacked side-chains. As in standard HADDOCK protocol, the chainID and segID fields were set to “A” for the envelope protein-prM complex and to “B” for the antibody using standard HADDOCK tools. The commands were as follows:

>>> $HADDOCKTOOLS/pdb\_setchain –v CHAIN=A 3c5x.pdb > 3c5x\_chainA.pdb

>>> $HADDOCKTOOLS/pdb\_chain-to-segid 3c5x\_chainA.pdb > 3c5x\_segidA.pdb

>>> $HADDOCKTOOLS/pdb\_setchain –v CHAIN=B antibody.pdb > antibody\_chainB.pdb

>>> $HADDOCKTOOLS/pdb\_chain-to-segid antibody\_chainB > antibody\_segidB.pdb

Furthermore, the waters were removed from the 3C5X-model. Finally, we renumbered the residues of the prM chain using a script from the HADDOCK pdb-tools repository (<https://github.com/haddocking/pdb-tools>). The command was

>>> python pdb\_reres –i -3 3c5x\_segidA.pdb > 3c5x\_segidA\_renum.pdb

This guarantees that each residue of segID “A” has their unique residue identifier.

To determine the centroids in an objective method, we performed an exhaustive cross-correlation search of both subunits in the cropped density map using PowerFit (<https://github.com/haddocking/powerfit>) (van Zundert and Bonvin, to be published). We ran the command line tool *powerfit* with 5.27° rotational sampling density, Laplace prefilter and no resampling, as thus

>>> powerfit 3c5x\_segidA\_renum.pdb subunit.xplor 21 –a 5.27 –d results/3c5x –l -nr

>>> powerfit antibody\_segidB subunit.xplor 21 –a 5.27 –d results/antibody –l -nr

The Laplace pre-filter was used as it performs better on low-resolution cryo-EM data, and we refrained from using the resampling option since this results in a more accurate placement of the subunits. PowerFit outputs a file (*solutions.out*) that includes the center of mass position of each chain for a found fit. For both subunits the top fit was taken with its corresponding centroid position. The density map and subunits are included as a test-case in the PowerFit source code.

Lastly, the ambiguous interaction restraints (AIRs) need to be determined. Since the interface of both subunits is approximately known, active residues were extracted from the 3J42-model. Residues that were within 5Å distance of a residue of the other chain were designated active. We used the *contact* script in the HADDOCK tools directory together with other standard UNIX programs. The precise commands were:

>>> $HADDOCKTOOLS/contact 3J42.pdb 5 | awk ‘{print $1}’ | sort –n \

| uniq | tr ‘\n’ ‘ ‘ > res-3c5x.out

>>> $HADDOCKTOOLS/contact 3J42.pdb 5 | awk ‘{print $4}’ | sort –n \

| uniq | tr ‘\n’ ‘ ‘ > res-antibody.out

AIRs were created by inputting the resulting residue numbers into the generate-AIR webserver (<http://www.bonvinlab.org/software/haddock2.2/generate_air.html>) as active residues and the output was saved to file (*ambig.tbl*). This wraps up the first part of preparing the input data.

*Setting up and performing the docking run*

At the start of the run, a *new.html* file was created containing all the paths to the data (<http://www.bonvinlab.org/software/haddock2.2/start_new.html>), including the prepared PDB files and the file containing the AIRs. HADDOCK-EM requires in addition also the path to the density data, described by the new CRYO-EM\_MAPFILE tag in the *new.html* file, and was manually added. The content of this *new.html* file is:

<html>

<head>

<title>HADDOCK - start</title>

</head>

<body bgcolor=#ffffff>

<h2>Parameters for the start:</h2>

<BR>

<h4><!-- HADDOCK -->

HADDOCK\_DIR=/home/gzundert/haddock2.2-em/<BR>

PROJECT\_DIR=/home/gzundert/Haddock\_cryoem/Virus-antibody/example<BR>

N\_COMP=2<BR>

PDB\_FILE1=/home/gzundert/Haddock\_cryoem/Virus-antibody/example/3c5x\_segidA\_renum.pdb<BR>

PDB\_FILE2=/home/gzundert/Haddock\_cryoem/Virus-antibody/example/antibody\_segidB.pdb<BR>

PROT\_SEGID\_1=A<BR>

PROT\_SEGID\_2=B<BR>

CRYO-EM\_MAPFILE=/home/gzundert/Haddock\_cryoem/Virus-antibody/example/subunit.xplor<BR>

AMBIG\_TBL=/home/gzundert/Haddock\_cryoem/Virus-antibody/example/ambig.tbl<BR>

RUN\_NUMBER=example<BR>

submit\_save=Save updated parameters<BR>

</h4><!-- HADDOCK -->

</body>

</html>

The run-directory was setup by invoking HADDOCK-EM in the directory where *new.html* is present:

>>> haddock-em

The last step was changing the parameters of interest in *run.cns* in the run-directory, which include requesting cryo-EM restraints in general (*cryo\_em\_rest*), the resolution of the data (*map\_resolution*), and the coordinates of the centroids (*xcom\_n, ycom\_n* and *zcom\_n*). Also, the size (*xlength, ylength* and *zlength*) and number of voxels (*nc, nr* and *ns*) in each dimension need to be explicitly added, though this will be automated in future versions. See below for the parameters and their associated values

{===>} cryo\_em\_rest = true;

{===>} map\_resolution = 21.0;

{===>} xcom\_1 = 11.16;

{===>} ycom\_1 = 29.76;

{===>} zcom\_1 = 219.48;

{===>} xcom\_2 = 44.64;

{===>} ycom\_2 = 29.76;

{===>} zcom\_2 = 293.88;

{===>} nc = 54;

{===>} nr = 36;

{===>} ns = 64;

{===>} xlength = 200.88;

{===>} ylength = 133.92;

{===>} zlength = 238.08;

All other parameters were left to their default values. Nonetheless, for the sake of clarity we will go through the relevant ones for HADDOCK-EM. *Cryo\_em\_kscale* and *wa* are the force constants of the centroid based distance restraint and cross-correlation potential, respectively. These values were determined during the optimization of the benchmark. Since there is no ambiguity in the placement of the centroids, i.e. it is known to which centroid each subunit should go, the centroid-based distance restraints were left as unambiguous (*ambi\_1* and *ambi\_2*). And finally, the scoring weight parameters (*w\_ccc\_0, w\_ccc\_1* and *w\_ccc\_2*) used in the it0-, it1 and itw-HADDOCK scores were left to default values, as these were also optimized on the benchmark. The parameters with their default values are shown below

{===>} cryo\_em\_kscale = 50;

{===>} wa = 15000;

{===>} ambi\_1 = false;

{===>} ambi\_2 = false;

{===>} w\_ccc\_0 = -4.0;

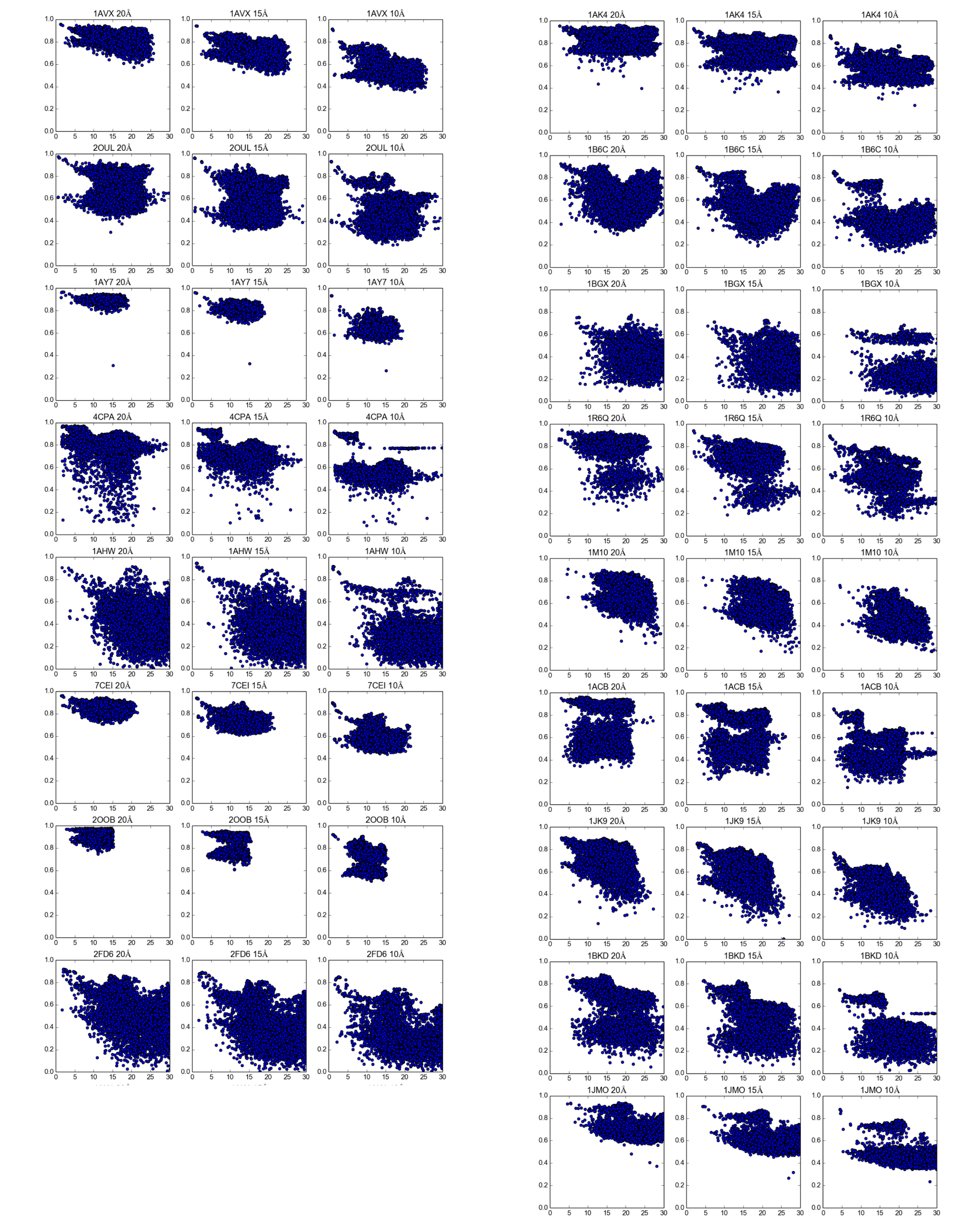
{===>} w\_ccc\_1 = -100;

{===>} w\_ccc\_2 = -100;

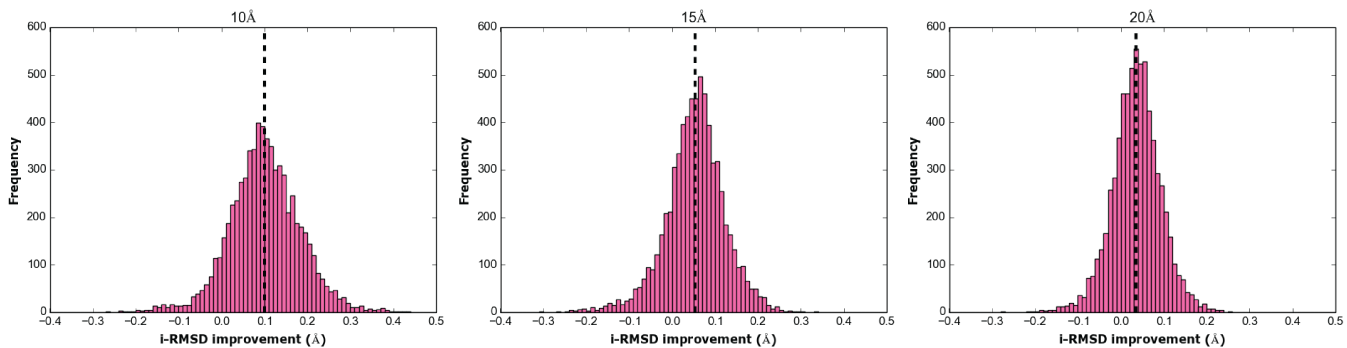
These are all the parameters that are involved in setting up a HADDOCK-EM run. To finish, the docking run was started by invoking HADDOCK again in the directory of *run.cns*.

>>> haddock-em

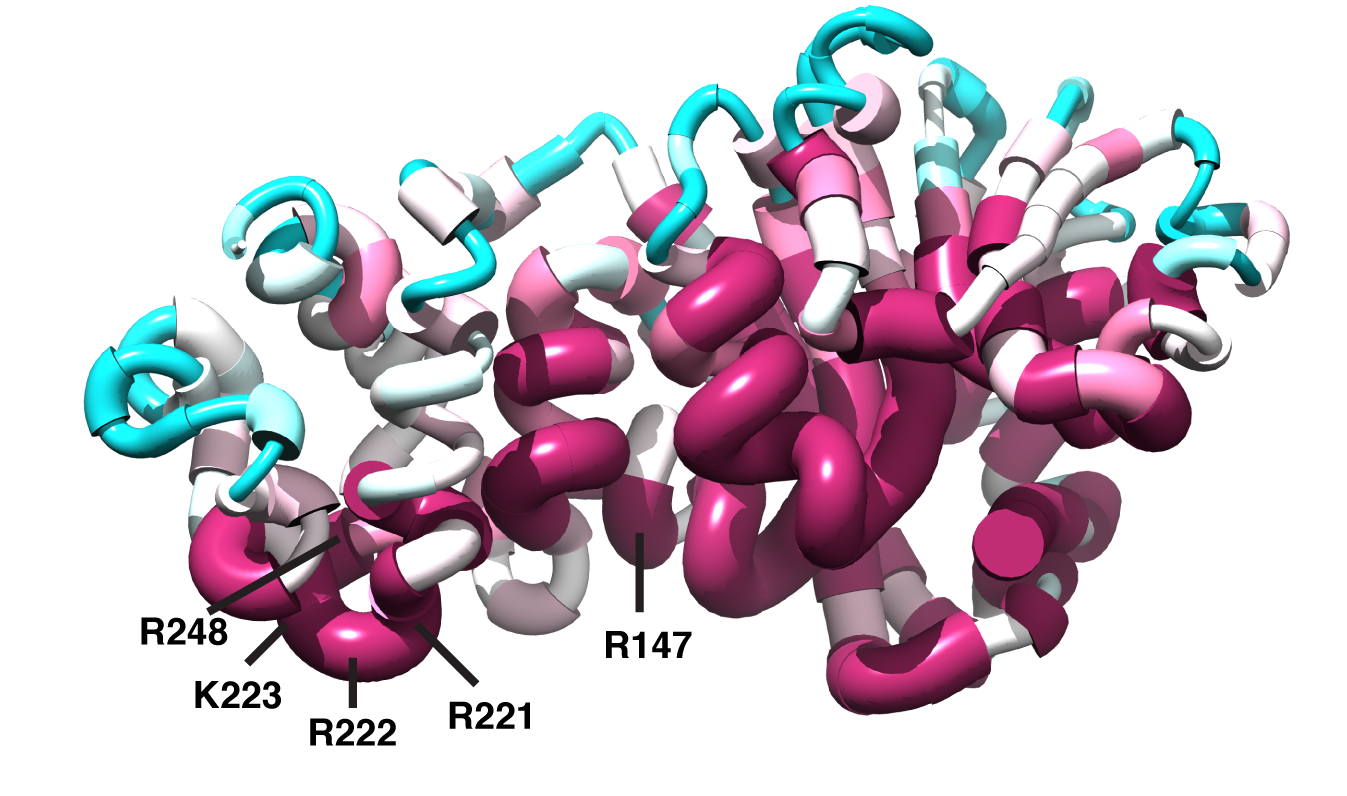
**Supplemental Figures**

****

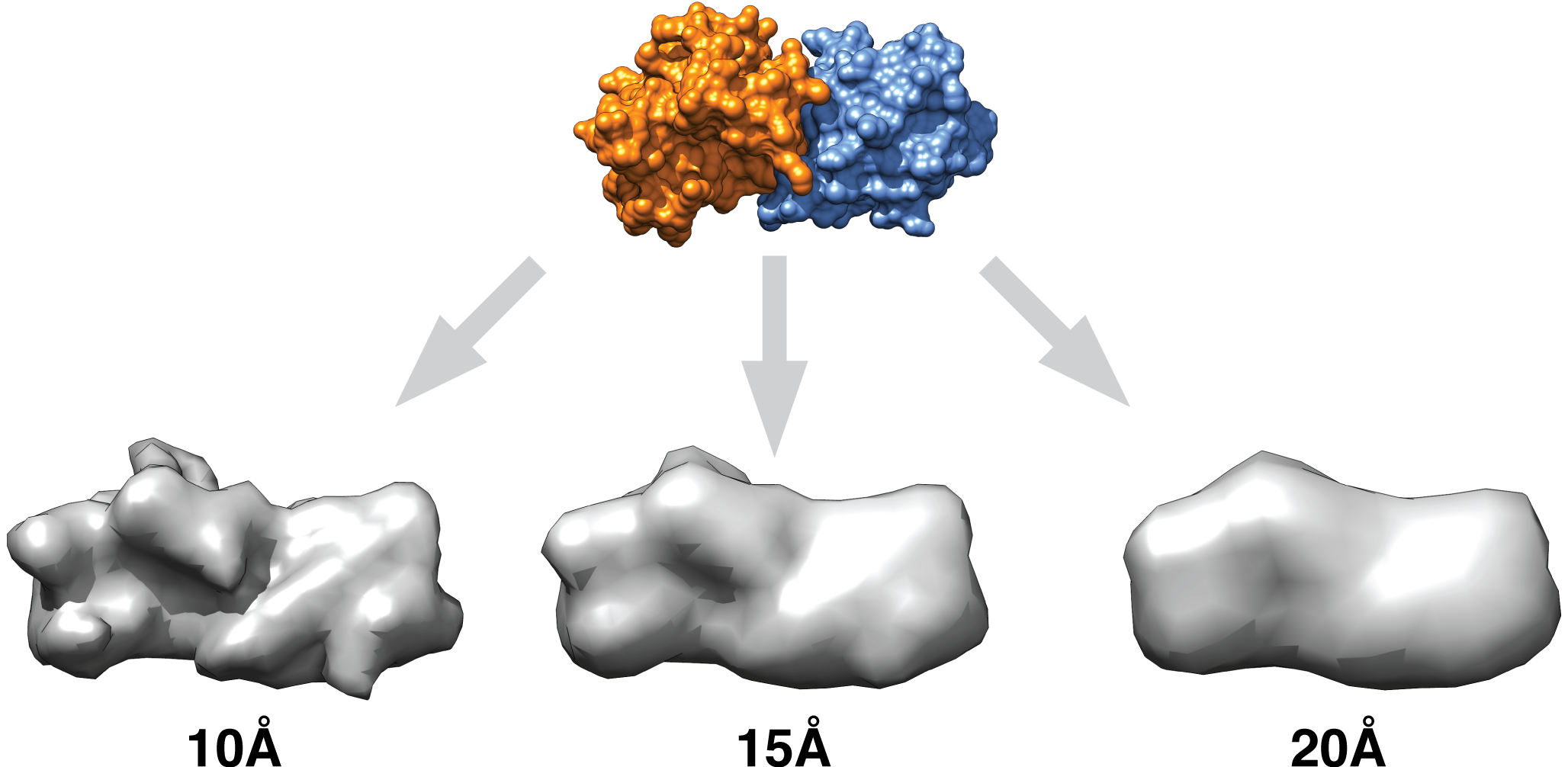
**Figure S1 – Local cross correlation scores for all 17 complexes using simulated cryo-EM data.** The local cross correlation score is plotted against the i-RMSD compared to the native complex for all 17 complexes using simulated 20Å (left), 15Å (middle) and 10Å (right) resolution cryo-EM data.

****

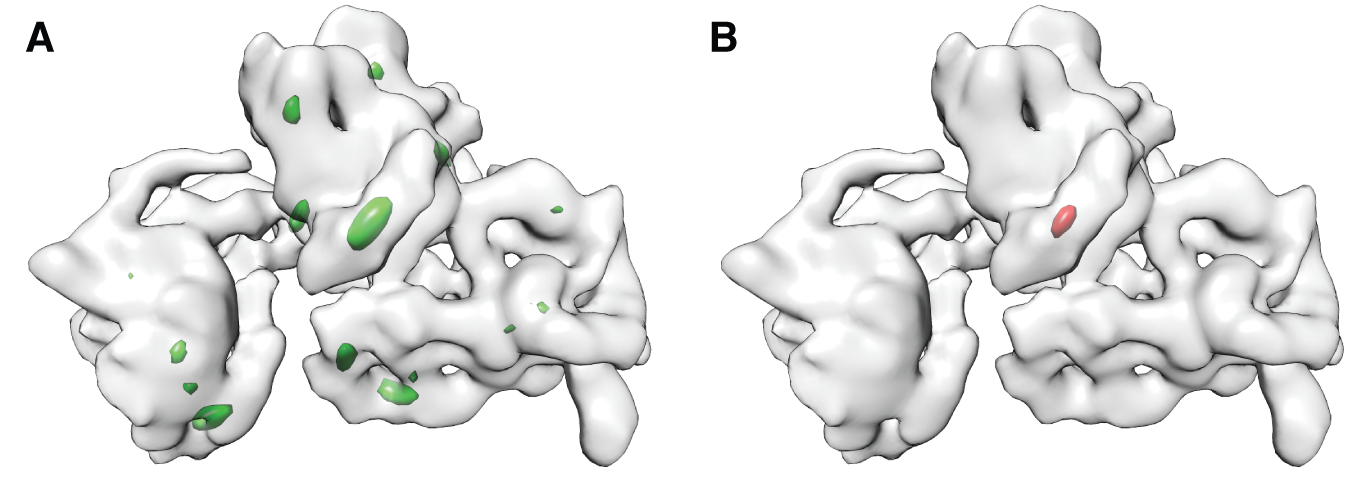
**Figure S2 – Effect of the itw-flexible refinement stage with cryo-EM restraints on i-RMSD.** The i-RMSD improvement (i-RMSD it1 – i-RMSD itw) for all refined complexes after itw when using 10Å **(A)**, 15Å **(B)** and 20Å **(C)** resolution data. The dashed vertical line in each figure represents the average i-RMSD improvement.

****

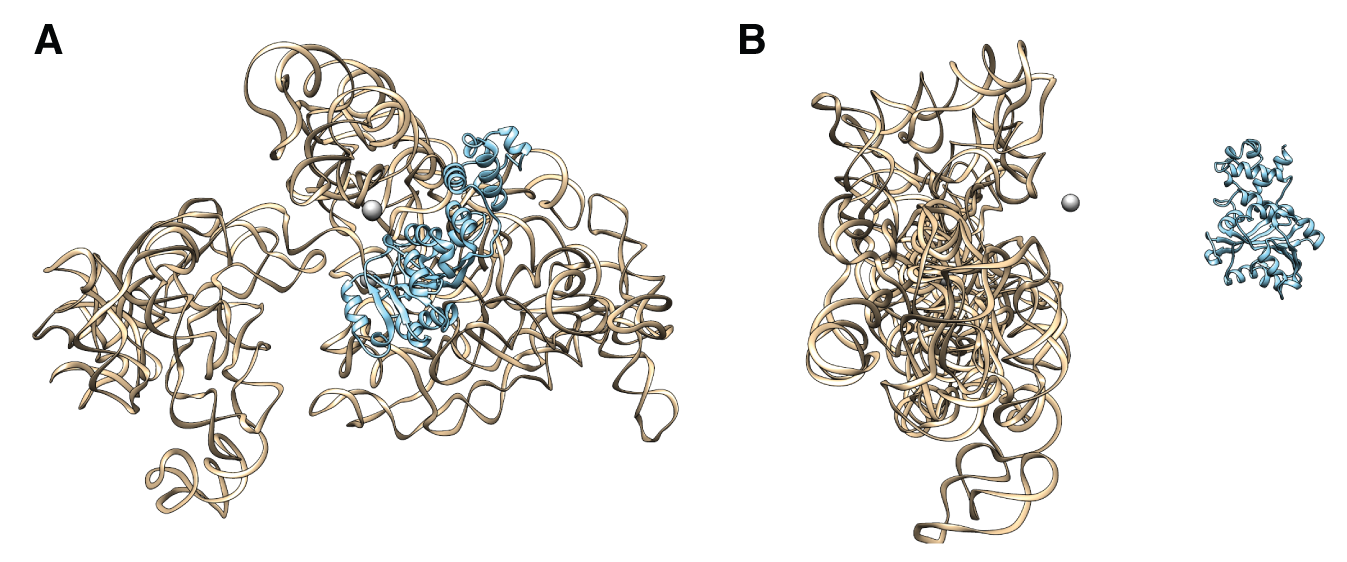
**Figure S3 – Worm representation of KsgA showing the conservation score of each residue.** The conservation score is higher for thicker and purple residues and lower for thinner and blue residues. Conservation scores were determined using the ConSurf webserver (Ashkenazy *et al.*, 2010).

****

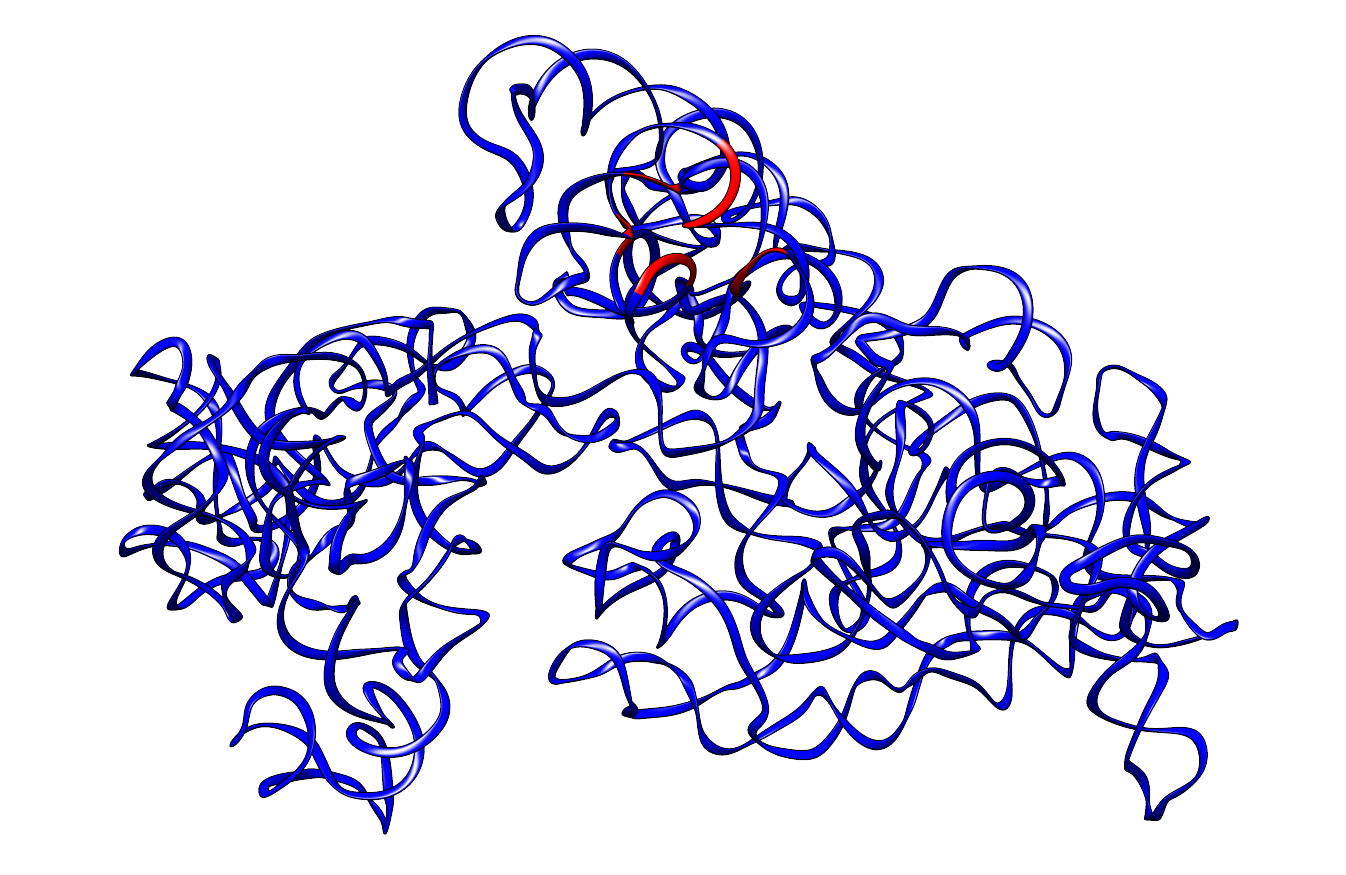
**Figure S4 – Example of simulated cryo-EM data generated for benchmarking HADDOCK-EM.** A surface representation of the 7CEI complex is shown on top. Under it, three iso-surfaces are shown for simulated cryo-EM data at 10Å, 15Å and 20Å resolution.

****

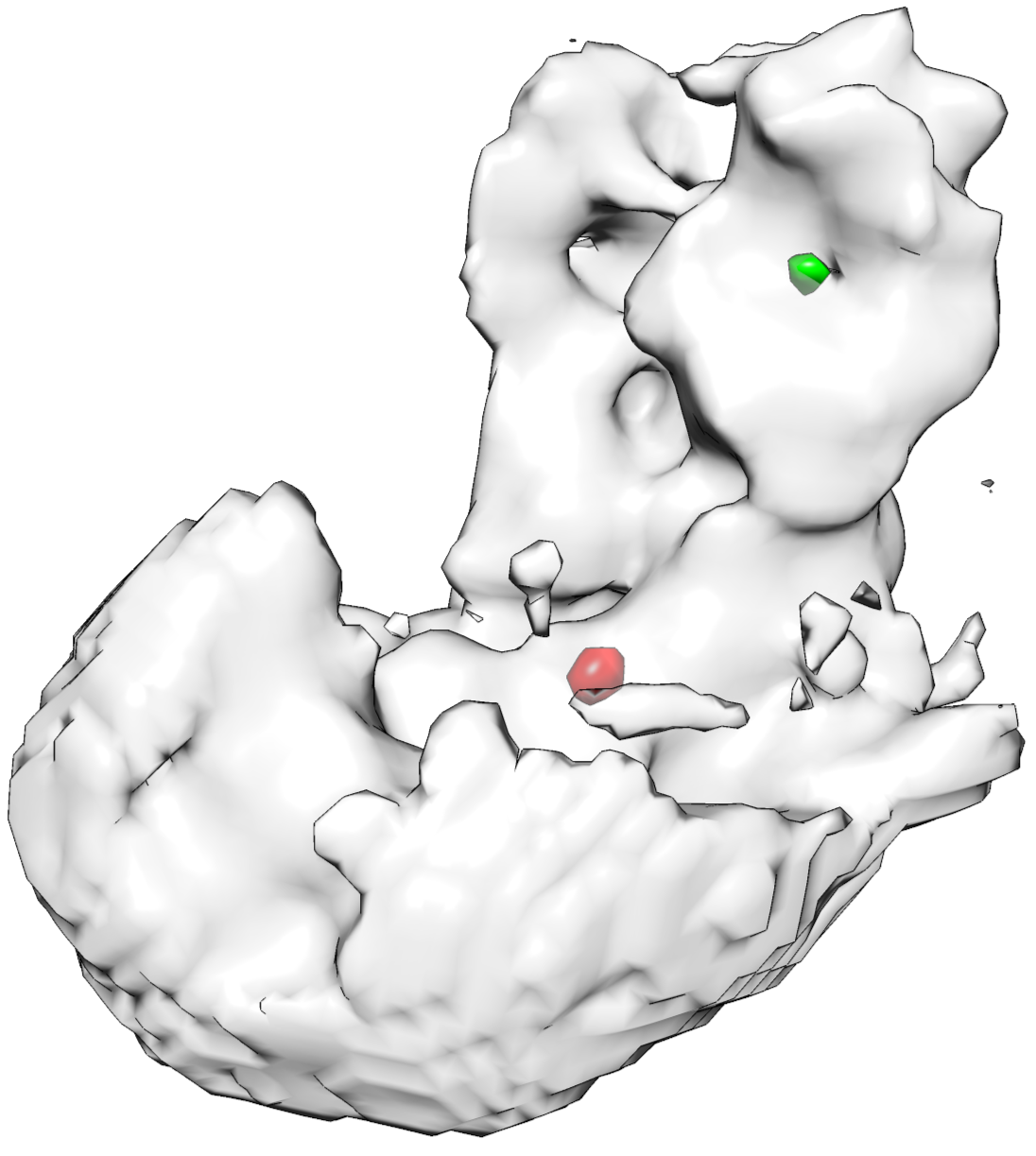
**Figure S5 – Determination of centroid position of KsgA in the cryo-EM density of 30S E.Coli ribosome.** Iso-contour of the 30S ribosome in gray and the iso-contour of local cross correlation values at 0.5 **(A, green)** and 0.6 **(B, red)** as a result of the full-exhaustive search. The centroid was positioned on the maximum correlation value within the iso-contour shown in **B.**

****

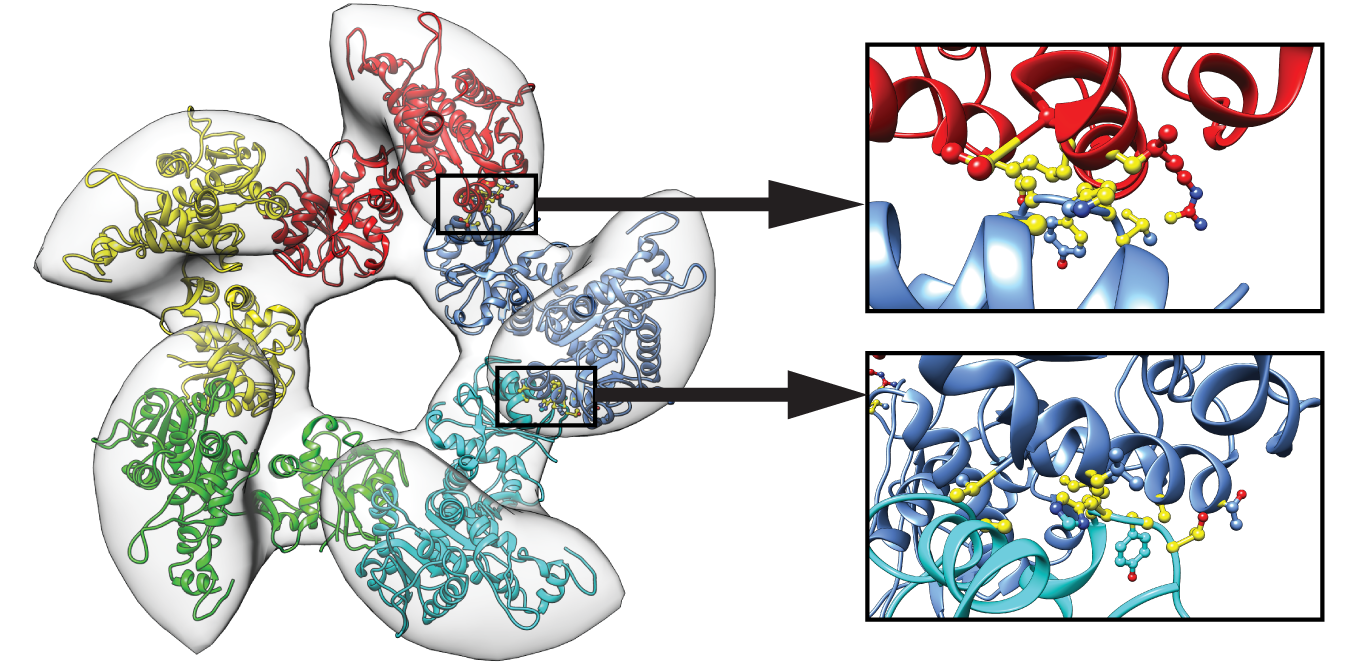
**Figure S6 – Initial placement of the 16S rRNA and KsgA during it0.** A front- **(A)** and side-view **(B)** of the 16S rRNA and KsgA initial setup as was used during the rigid body docking stage.

****

**Figure S7 – Active residues of the 16S rRNA used during the docking.** A ribbon representation of the 16S rRNA with the active residues 768 – 773, 781, 782, 801 – 803, 899 – 902, 1512 – 1516 and 1523 shown in red.



**Figure S8 – Determination of centroid positions for the Dengue-virus envelope protein and antibody.** Iso-contour of a subunit part of the 21Å resolution cryo-EM data of Dengue virus (grey), showing regions of high local cross correlation values (0.35) for the envelope protein (red) and antibody (green).



**Figure S9 – Current deposited model of the large terminase complex.** A ribbon representation of the current deposited terminase complex (4bij). Multiple clashes (yellow ball-and-sticks) are observed when zooming in on the interfaces of the subunits.

**Supplemental Tables**

**Table S1.** **Number of acceptable solutions in the top 400 after each docking stage, using simulated cryo-EM data at 10Å, 15Å and 20Å resolution.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PDB | Number of acceptable solutions in top 400 | | | | | | | | |
| 10Å | | | 15Å | | | 20Å | | |
| It0 | It1 | Itw | It0 | It1 | Itw | It0 | It1 | Itw |
| 1ACB | 13 | 19 | 19 | 7 | 8 | 7 | 4 | 4 | 4 |
| 1AHW | 26 | 42 | 42 | 13 | 17 | 17 | 5 | 8 | 8 |
| 1AK4 | 78 | 81 | 81 | 37 | 40 | 40 | 3 | 8 | 8 |
| 1AVX | 38 | 51 | 53 | 13 | 17 | 18 | 5 | 6 | 6 |
| 1AY7 | 23 | 28 | 28 | 10 | 16 | 16 | 9 | 9 | 9 |
| 1B6C | 40 | 47 | 47 | 21 | 28 | 28 | 7 | 12 | 12 |
| 1BGX | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1BKD | 0 | 5 | 5 | 0 | 1 | 1 | 0 | 0 | 0 |
| 1JMO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1M10 | 0 | 2 | 2 | 0 | 1 | 1 | 0 | 1 | 1 |
| 1R6Q | 54 | 57 | 57 | 12 | 30 | 30 | 3 | 4 | 4 |
| 1JK9 | 118 | 223 | 224 | 64 | 84 | 95 | 21 | 23 | 23 |
| 2FD6 | 140 | 147 | 148 | 82 | 97 | 97 | 37 | 42 | 42 |
| 2OOB | 15 | 15 | 15 | 3 | 3 | 4 | 0 | 0 | 0 |
| 2OUL | 49 | 62 | 64 | 28 | 32 | 32 | 9 | 10 | 10 |
| 4CPA | 90 | 105 | 108 | 85 | 88 | 91 | 65 | 70 | 72 |
| 7CEI | 75 | 90 | 92 | 27 | 48 | 49 | 13 | 19 | 20 |

**Table S2.** **Number of acceptable solutions generated with displaced centroids using simulated cryo-EM data at 10Å, 15Å and 20Å resolution.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PDB | **Acceptable or better after itw** | | | | | | | | | | | |
| 10Åa | | | | 15Åa | | | | 20Åa | | | |
| 0Åb | 3Åb | 5Åb | 7Åb | 0Åb | 3Åb | 5Åb | 7Åb | 0Åb | 3Åb | 5Åb | 7Åb |
| 1ACB | 19 | 12 | 26 | 11 | 7 | 6 | 22 | 4 | 4 | 8 | 1 | 3 |
| 1AHW | 42 | 44 | 30 | 32 | 17 | 21 | 20 | 22 | 8 | 10 | 10 | 9 |
| 1AVX | 53 | 48 | 57 | 36 | 18 | 23 | 28 | 19 | 6 | 11 | 15 | 16 |
| 1JK9 | 224 | 198 | 189 | 191 | 95 | 78 | 77 | 28 | 23 | 26 | 10 | 33 |
| 1M10 | 2 | 0 | 3 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |

a Resolution of simulated cryo-EM data.

b Displacement of each centroid into a random direction.

**Table S3.** **Number of acceptable solutions in top 400, 10 and 1 after water refinement stage using *ab initio* HADDOCK (HADDOCK-CM), and HADDOCK-EM using simulated 20, 15 and 10Å resolution cryo-EM data.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PDB** | **Number of acceptable solutions after itw** | | | | | | | | | | | | | | | |
| **HADDOCK-CM** | | | | **HADDOCK-EM 20Å** | | | | **HADDOCK-EM 15Å** | | | | **HADDOCK-EM 10Å** | | | |
| **Total** | **Top 400** | **Top 10** | **Top 1** | **Total** | **Top 400** | **Top 10** | **Top 1** | **Total** | **Top 400** | **Top 10** | **Top 1** | **Total** | **Top 400** | **Top 10** | **Top 1** |
| **1AVX** | 0 | 0 | 0 | 0 | 6 | 6 | 5 | 1 | 18 | 18 | 10 | 1 | 53 | 53 | 10 | 1 |
| **2OUL** | 2 | 2 | 1 | 1 | 10 | 10 | 9 | 1 | 32 | 32 | 10 | 1 | 64 | 64 | 10 | 1 |
| **1AY7** | 1 | 1 | 0 | 0 | 9 | 9 | 8 | 1 | 16 | 16 | 10 | 1 | 28 | 28 | 10 | 1 |
| **4CPA** | 1 | 1 | 0 | 0 | 72 | 72 | 8 | 1 | 91 | 91 | 10 | 1 | 108 | 108 | 10 | 1 |
| **1AHW** | 0 | 0 | 0 | 0 | 8 | 8 | 5 | 1 | 17 | 17 | 10 | 1 | 42 | 42 | 10 | 1 |
| **7CEI** | 4 | 4 | 1 | 0 | 20 | 20 | 5 | 1 | 49 | 49 | 10 | 1 | 92 | 92 | 10 | 1 |
| **2OOB** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 15 | 15 | 0 | 0 |
| **2FD6** | 2 | 2 | 0 | 0 | 42 | 42 | 10 | 1 | 97 | 97 | 10 | 1 | 148 | 148 | 10 | 1 |
| **1AK4** | 1 | 1 | 0 | 0 | 8 | 8 | 3 | 0 | 40 | 40 | 10 | 1 | 81 | 81 | 10 | 1 |
| **1B6C** | 0 | 0 | 0 | 0 | 12 | 12 | 6 | 1 | 28 | 28 | 10 | 1 | 47 | 47 | 10 | 1 |
| **1BGX** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1R6Q** | 0 | 0 | 0 | 0 | 4 | 4 | 2 | 1 | 30 | 30 | 10 | 1 | 67 | 67 | 10 | 1 |
| **1M10** | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 |
| **1ACB** | 0 | 0 | 0 | 0 | 4 | 4 | 1 | 0 | 7 | 7 | 0 | 0 | 19 | 19 | 8 | 1 |
| **1JK9** | 0 | 0 | 0 | 0 | 23 | 23 | 7 | 1 | 95 | 95 | 10 | 1 | 224 | 224 | 10 | 1 |
| **1BKD** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 5 | 5 | 5 | 1 |
| **1JMO** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Supplemental References**

Ashkenazy H., Erez E., Martz E., Pupko T. and Ben-Tal N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucl. Acids Res. *38*, W529-W533.

Van Zundert, G.C.P. and Bonvin, A.M.J.J. (2015) Fast and sensitive rigid body fitting into cryo-EM density maps with PowerFit. AIMS Biophysics (to be published).