



Technical Note

Building the atomic model of a boreal lake virus of unknown fold in a 3.9 Å cryo-EM map

Luigi De Colibus^{a,*}, David I. Stuart^{a,b}^a Division of Structural Biology, University of Oxford, The Henry Wellcome Building for Genomic Medicine, Headington, Oxford OX3 7BN, UK^b Diamond Light Source Ltd, Harwell Science & Innovation Campus, Didcot OX11 0DE, UK

ARTICLE INFO

Keywords:

Cryo electron microscopy
Model building
Density modification
Virus structure

ABSTRACT

We report here the protocol adopted to build the atomic model of the newly discovered virus FLiP (*Flavobacterium* infecting, lipid-containing phage) into 3.9 Å cryo-electron microscopy (cryo-EM) maps. In particular, this report discusses the combination of density modification procedures, automatic model building and bioinformatics tools applied to guide the tracing of the major capsid protein (MCP) of this virus. The protocol outlined here may serve as a reference for future structural determination by cryo-EM of viruses lacking detectable structural homologues.

Thanks to recent advances in detector efficiency, microscope stability and image processing, the 3D structure of biomolecular complexes spanning a wide range of sizes, from haemoglobin to microtubules and intact viruses, can be reconstructed at near-atomic resolution (Bai et al., 2013; Brilot et al., 2012; Campbell et al., 2012, 2015; Khoshouei et al., 2016; Liu et al., 2010; Nogales and Scheres, 2015; Scheres, 2012; Walls et al., 2016). Notably, determination of structures of icosahedral viruses in the 5–3 Å resolution range by cryo-EM is becoming increasingly straightforward. In this resolution range, it is possible to identify the fold of the capsid subunits and the secondary structure elements of protein subunits can be clearly discerned (Jiang and Chiu, 2007); however, building an atomic model can be very challenging when homologous structures determined at higher resolution are not available.

The virus used as an example in this report, FLiP (*Flavobacterium* infecting, lipid-containing phage), has a circular ssDNA genome and possesses an internal lipid membrane enclosed in the icosahedral capsid. Its genome showed limited sequence similarity to other known viral sequences. However, the structure of the viral major capsid protein, elucidated at near-atomic resolution using cryo-EM, turned out to be strikingly similar to those observed in dsDNA viruses of the PRD1–adenovirus lineage (Abrescia et al., 2004), characterised by a

major capsid protein bearing two beta-barrels. The strong similarity between FLiP and another member of the lineage, bacteriophage PM2 (Abrescia et al., 2008), extends to the capsid organisation (Fig. 1a) (pseudo $T = 21$ dextro) in spite of the differences in the genetic material packaged and the lack of significant sequence similarity (Laanto et al., 2017).

Virus purification and cryo-EM data collection were performed as described in Laanto et al. (2017). Here we illustrate all the steps involved in model building of FLiP major capsid protein (FLiP MCP). We focus on the MCP because the other portions of the capsid, including spikes and internal capsid proteins, were at lower resolution and there were no homologous structures available which would have made the model building even more challenging and error prone than the building of FLiP MCP.

Searching with default parameters for structures homologous to FLiP MCP with the HH-pred server (Soding, 2005), no reasonable hits were found, whereas using HHsenser (Soding, 2005) one hit was found, comprising 38% sequence identity with a structural protein of the *Cellulophaga* phage phi48:2, but no structure was available, making it impossible to start model building based on a pre-existing protein model. We therefore had to devise a novel strategy to build the atomic model of FLiP MCP in the cryo-EM map determined at an overall re-

* Corresponding author.

E-mail address: luigi@strubi.ox.ac.uk (L. De Colibus).

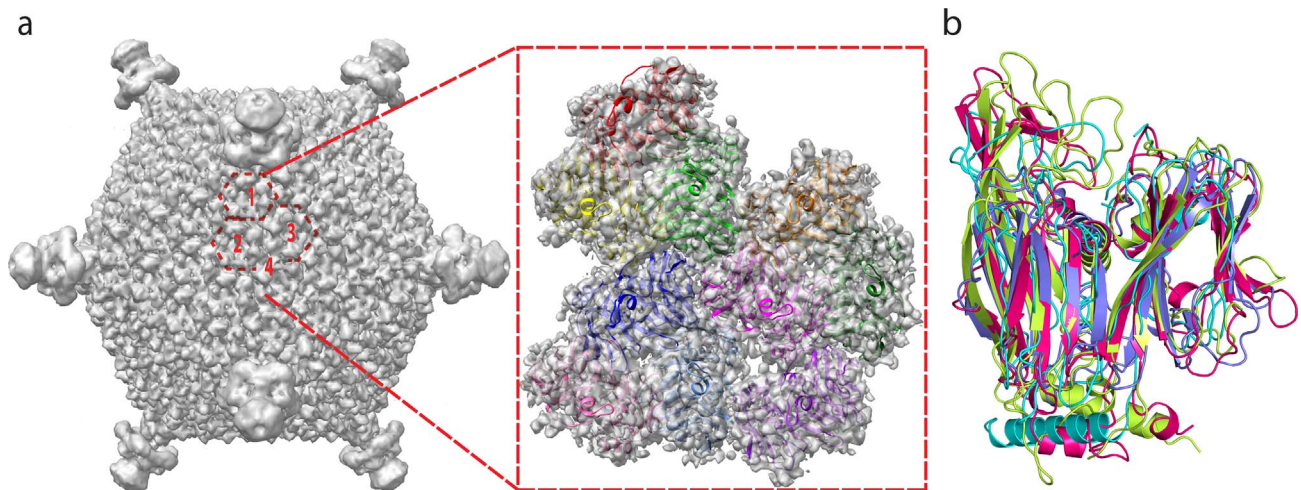


Fig. 1. (a) Cryo-EM map of entire particle (EMDB: EMD-3771) and atomic structure of the Asymmetric Unit (AU) fitted into the density. 10 chains of PM2 Major Capsid Protein (MCP) compose the AU. (b) Structures superposition of MCP proteins identified by DALI on the incomplete model of FLiP MCP (FLiP MCP in cyan, STIV MCP in pink PDBid: 2BBD PRD-1 MCP in lime, PDBid: 1W8X, PM2 MCP in purple, PDBid: 2VVF).

solution of 4 Å (Laanto et al., 2017). We note that the FLiP cryo-EM map was less well resolved on the surface of the virus, while the local resolution in the capsid interior was approaching 3.9 Å (a map of the resolution is provided in Laanto et al., 2017), so the map used to perform the model building and refinement was low-pass filtered at 3.9 Å (Laanto et al., 2017) with the image_handler tool in Relion (Scheres, 2012).

It was apparent from the cryo-EM map that the capsid of FLiP resembled that of bacteriophage PM2 (Abrescia et al., 2008). The PM2 major capsid protein structure was instrumental in defining roughly the boundaries of the asymmetric unit (AU) and the location of jelly roll domains whose characteristic shape allowed us to establish the hand of the map but it could not be used for further steps in the process of model building because FLiP MCP has a different structure. Manual placement of 10 copies of the PM2 model in the cryoEM map enabled identification of the volume containing an AU of the FLiP virus and also the correct hand of the EM map. These 10 copies of the MCP in the AU were arranged in three trimers plus an extra copy close to the icosahedral threefold axis. The AU volume was excised from the particle density by fitting 10 individual chains of the PM2 capsid shell protein (Abrescia et al., 2008) (PDBid:2VVF) and cutting the corresponding region of the map with the phenix_box_map program (Adams et al., 2010). Extra density, corresponding to the lipidic envelope and the nucleic acid, was removed in two steps. First, a mask with a radius of 4 Å around the protein atoms was generated in CCP4 ncsmask (Winn et al., 2011), using 10 individual chains of the PM2 capsid shell protein as fitted in the map. Then, this mask was used in CCP4 mapcut (Winn et al., 2011) to carve out the density corresponding to the FLiP MCP in the AU (Fig. 1a). In absence of an atomic model to guide carving out the density corresponding to the FLiP MCP in the AU, an alternative strategy could be to run Segger (Pintilie et al., 2010) or phenix.segment_and_split_map (Adams et al., 2010) on the cryo-EM map to segment it in order to identify the AU.

Initial attempts to generate an *ab initio* model with PHYRE2 server (Kelley et al., 2015) and I-Tasser server (Wu et al., 2007; Zhang, 2009; Roy et al., 2010) were made but in both cases the output models had very low scores. This result was most likely due to the absence of homologous

protein sequences and structures of FLiP MCP, and these servers heavily rely on this source of information to generate a reliable 3D *ab initio* model. Secondary structure elements were placed in the icosahedrally averaged map in the AU by Resolve (Terwilliger, 2004) within Phenix (Adams et al., 2010). All the secondary structure elements placed in the AU were gathered and folded back to a single copy by acting on them with the Resolve (Terwilliger, 2004) non crystallographic symmetry (NCS) operators relating the 10 copies in the AU. The NCS operators were found searching in the cut AU map, using phenix.find_ncs_from_density (Adams et al., 2010; Terwilliger, 2013). These operators were subsequently employed to perform NCS averaging.

The operator to map molecule *j* onto molecule 1 is defined by matrix \mathbf{M}_{1j} and vector \mathbf{t}_{1j} :

$$(x_i, y_i, z_i) = \mathbf{M}_{1j} \cdot (x_j, y_j, z_j) + \mathbf{t}_{1j} \quad (1)$$

Multiplying both sides of the equation by the inverse matrix (\mathbf{M}_{1j}^{-1}), Eq. (2) is obtained:

$$\mathbf{M}_{1j}^{-1} (x_i, y_i, z_i) = \mathbf{1} \cdot (x_j, y_j, z_j) + \mathbf{M}_{1j}^{-1} \mathbf{t}_{1j} \quad (2)$$

Rearranging:

$$(x_j, y_j, z_j) = \mathbf{M}_{1j}^{-1} \cdot [(x_i, y_i, z_i) - \mathbf{t}_{1j}] \quad (3)$$

These transformations have been performed using the following server: <http://www.calcul.com/show/calculator/matrix-multiplication;3;3;3;1>.

The NCS operators were applied to the different NCS copies using CCP4 PDBSET (Winn et al., 2011) and whenever a secondary structure element was recognized as overlapping one already belonging to reference copy 1 then the longer element was retained as belonging to copy 1, and the shorter element discarded. Non-overlapping elements were simply added to the list belonging to copy 1. In this way the 10-fold redundant set of secondary structure elements was reduced to a single set belonging to reference copy 1, which contained the most complete set of secondary structural elements.

Model building was performed using both the original icosahedrally averaged cryo-EM map and the NCS-averaged map at 3.9 Å (Fig. 2a,c). The tracing of the main chain of each copy of FLiP MCP was aided by

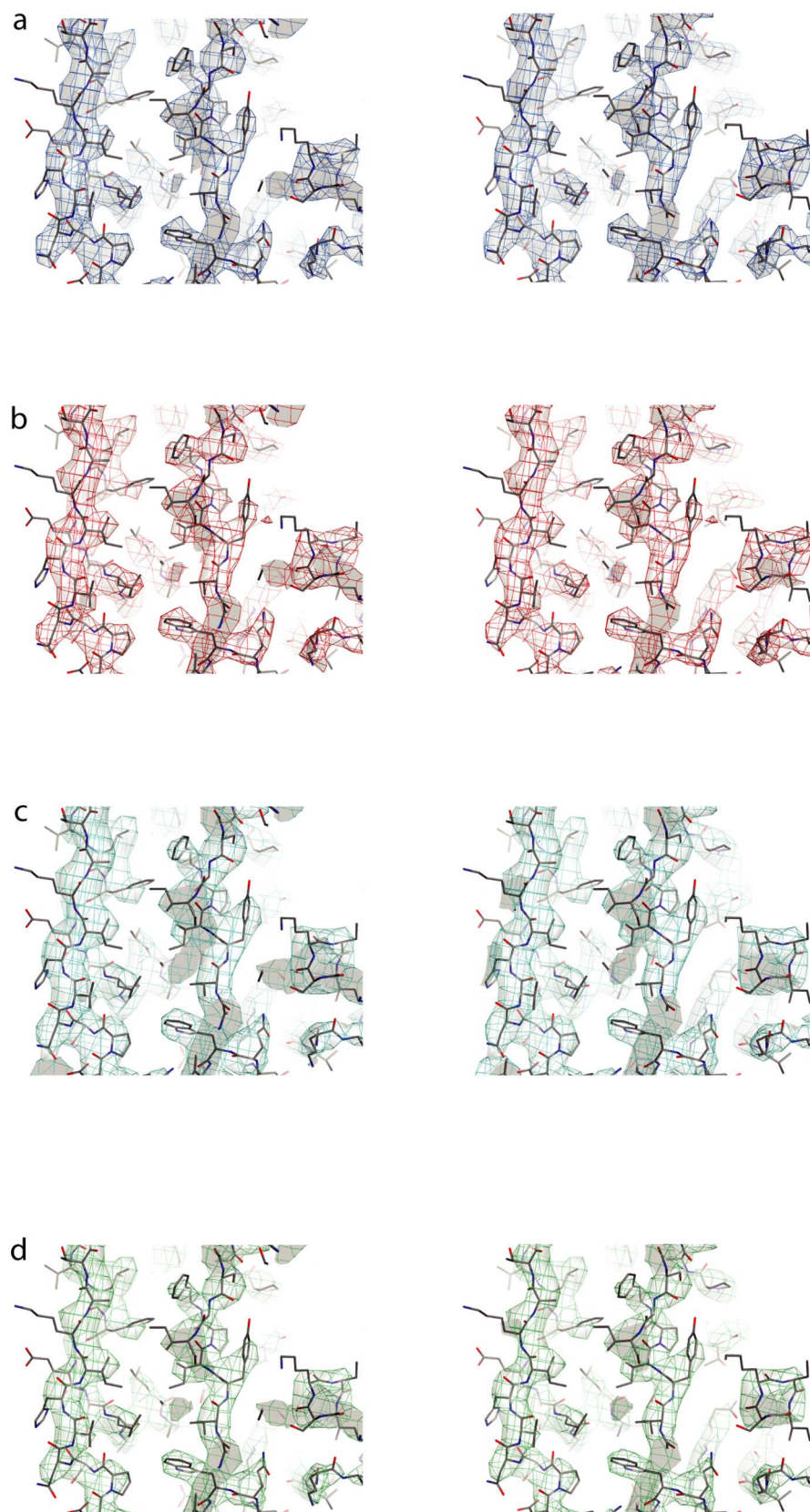


Fig. 2. (a) Stereo representation of the cryo-EM density with the atomic model fitted in. (b) Stereo representation of B-factor sharpened map with the atomic model fitted in. (c) Stereo representation of NCS-averaged map with the atomic model fitted in. (d) Stereo representation of B-factor sharpened and NCS-averaged map with the atomic model fitted in. All the maps are contoured at 2σ .

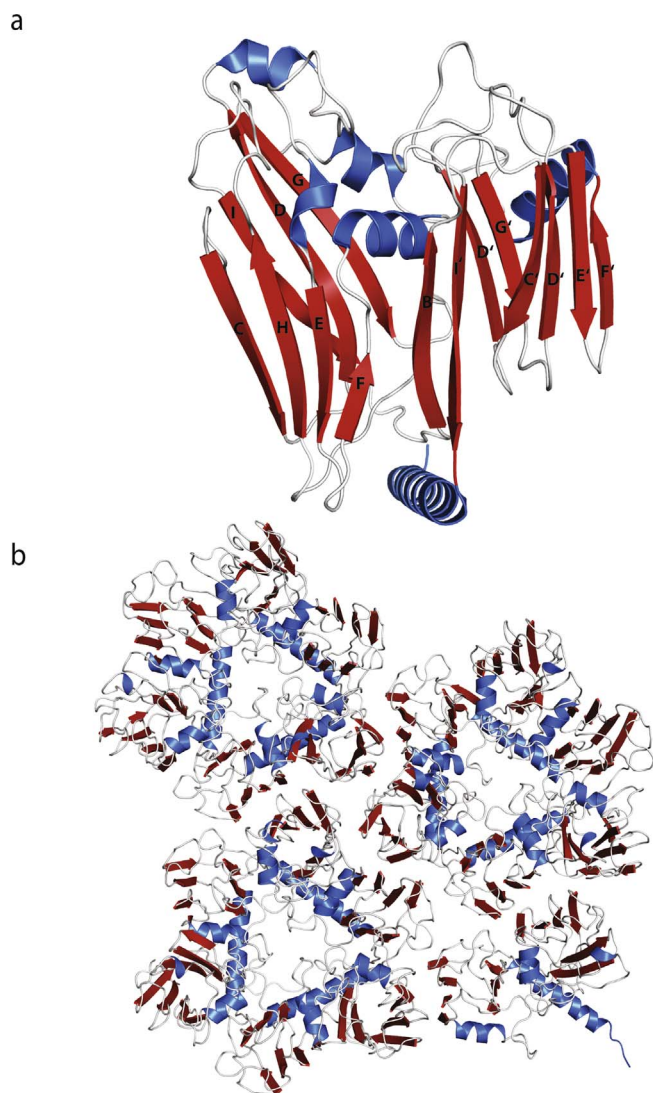


Fig. 3. (a) FlIP MCP atomic model (PDBid: 5OAC) in cartoon representation with secondary structure elements in different colors (strands in blue, helices in red and loops in white). (b) FlIP asymmetric unit. The strands in the two beta barrels are labeled B-I and C'-I'.

the map skeleton command in Coot (Emsley et al., 2010). All the collected secondary structure elements making up NCS reference copy 1 were joined and extended into an initial polyalanine model accounting initially for 267 residues out of the protein's 311 residues using Coot (Emsley et al., 2010) with a model map correlation coefficient (CC) around the atoms of 0.4530, this value was calculated in Phenix (Adams et al., 2010). This initial incomplete model showed two jelly-roll domains and it was further extended to 282 residues (CC = 0.5030). This model was used as a probe in DALI (Holm and Rosenstrom, 2010), searching for other viral proteins with similar folds, in order to facilitate the domain assignments and aid main chain tracing. The viral proteins with similar fold with top scores were: *Sulfolobus* turreted icosahedral virus 1 (STIV) MCP (Khayat et al., 2005) (PDBid: 2BBD, PDBid: 3J31; $\text{rmsd}_{\text{Ca}} = 3.3 \text{ \AA}$ over 236 residues and a Z-score value of 13.7) and PRD1 MCP (Abrescia et al., 2008) (PDBid: 1W8X; $\text{rmsd}_{\text{Ca}} = 3.9 \text{ \AA}$ over 244 residues and a Z-score value of 13.0). Both structures together with PM2 MCP (Abrescia et al., 2008) (PDBid: 2VVF; $\text{rmsd}_{\text{Ca}} = 3.0 \text{ \AA}$ over

203 residues and a Z-score value of 11.1) were superposed on the FlIP incomplete model, allowing it to be extended to 298 residues out of 311 and enabling clear identification of the jelly-roll domains (Fig. 1b).

Sequence assignment and rebuilding of the reference chain was performed using Coot (Emsley et al., 2010) by employing the following maps: 1) NCS-averaged map (averaged using Phenix, Adams et al., 2010), 2) sharpened and blurred maps (calculations performed in both REFMAC (Brown et al., 2015; Murshudov et al., 2011) and Phenix (Adams et al., 2010)), 3) B-factor sharpened and further NCS-averaged map, calculated using Phenix (Adams et al., 2010). Portions of these maps are shown in Fig. 2b,c,d. This model was then NCS-expanded, by applying the NCS operators found in the map with `phenix.find_ncs_from_density` (Adams et al., 2010; Terwilliger, 2013), to reconstitute the AU. The individual copies were then rigid body refined into the density and the NCS updated (the movements were very small). The AU was re-built and refined in Rosetta release version 2016.32.58837 using protocols optimized for cryo-EM maps (DiMaio et al., 2015). The refinement was performed with optimization of the density fit using `elec_dens_fast` function (with `-denswt = 40`, empirically chosen after several trials on the basis of the Molprobit (Chen et al., 2010) scores of the refined models). Selection of the best fitting structure and structure relaxation were carried out using the `-FastRelax` flag. The best-scoring model as estimated by density fit and geometry was selected and used in Coot (Emsley et al., 2010) to guide further model building and optimization.

Regions of the single protomer where the density was difficult to interpret were removed from the model and rebuilt with Rosetta-CM (Song et al., 2013) using `-denswt = 40`. The model for the reference copy was then inspected in Coot (Emsley et al., 2010) to perform further model building (Fig. 3a) before replicating onto the other NCS copies to reconstitute the AU (Fig. 3b). The AU was manually fitted in the full particle density map in Chimera (Pettersen et al., 2004) followed by entire icosahedral particle generation. The viral capsid shell volume was excised from the particle density with the same protocol described above for the AU. The model for the full particle was refined against the same map with `phenix.real_space_refine` (Adams et al., 2010; Afonine et al., 2013) using icosahedral constraints, secondary structure restraints and reciprocal space B-factor refinement. The refined model was validated by Molprobit (Chen et al., 2010), EMRinger (Barad et al., 2015) and CaBLAM (Williams et al., 2013) through Phenix (Adams et al., 2010). Molprobit (Chen et al., 2010) compares local geometric features describing an all-atom model to those from high-resolution crystal structures. EMRinger (Brown et al., 2015) validates both model geometry and density-fit at sidechain level. CaBLAM (Williams et al., 2013) identifies errors in peptide plane orientation and uses the relatively reliable Ca trace information to identify probable secondary structure disguised by these errors. All the metrics values are good for such resolution, in particular the EMRinger (Brown et al., 2015) score of 2.32 for the entire particle is superior to the typical score of 1.0 for a 4 \AA map (Fig. S1). The refinement statistics are reported in Table 1. The workflow developed during this project is summarized in the Fig. 4. Variations on this workflow might be useful for other complexes showing local symmetry. We hope that our work will help other scientists to tackle the difficult task of analyzing large macromolecular complexes solved by cryo-EM at medium resolution.

Author contributions

L.D.C. performed research; D.I.S. supervised the project; L.D.C. in discussion with D.I.S. wrote the manuscript. All authors read and approved the manuscript.

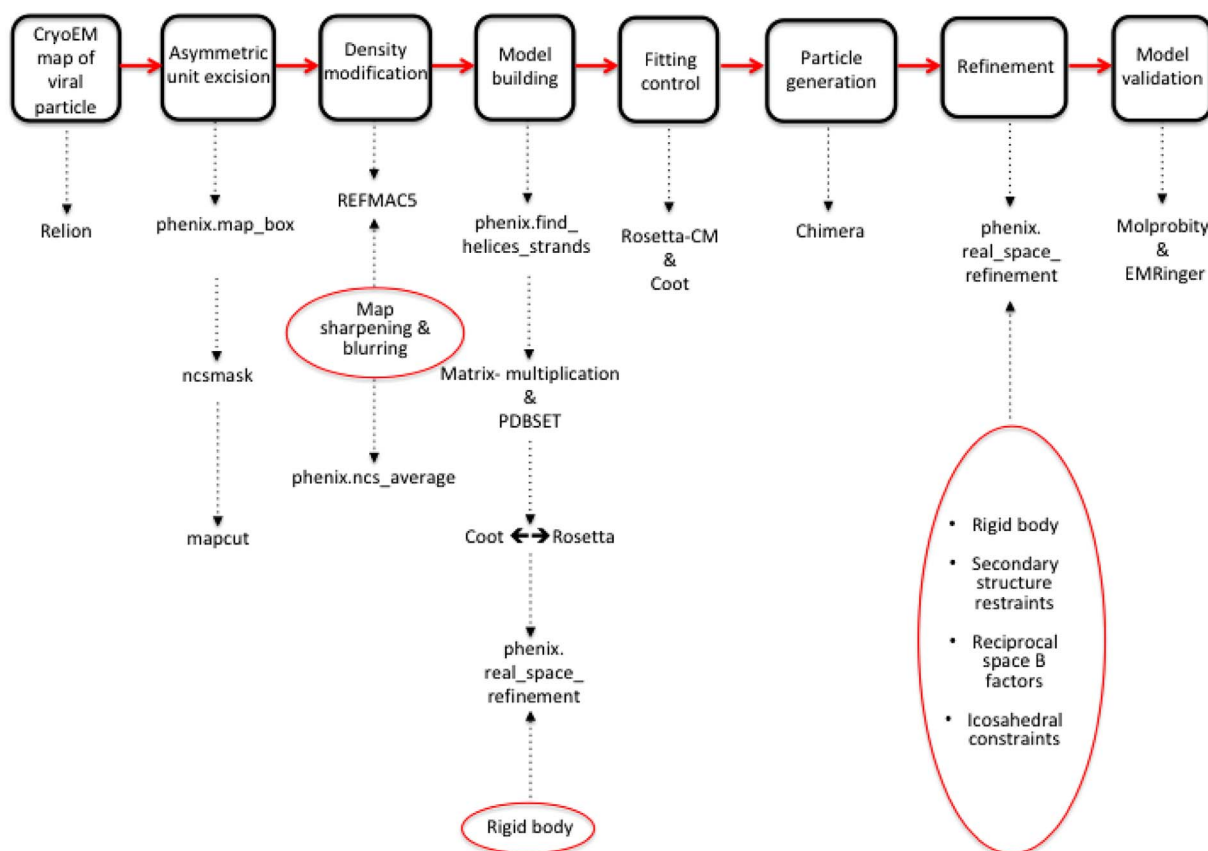


Fig. 4. Workflow showing all the steps performed to build the FLIP MCP atomic model.

Acknowledgments

We would like to thank Professor Juha T. Huiskonen for providing cryo-EM map and FSC curve, Juha and Dr. Lotta-Riina Sundberg for the structure project, Professor Frank DiMaio and Dr. Pietro Roversi for very helpful discussion on building and refinement with cryo-EM maps.

Grant acknowledgements

This work was supported by Medical Research Council Grant MR/N00065X/1 (to D.I.S.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jsb.2017.10.010>.

References

- Abrescia, N.G.A., Cockburn, J.J.B., Grimes, J.M., Sutton, G.C., Diprose, J.M., et al., 2004. Insights into assembly from structural analysis of bacteriophage PRD1. *Nature* 432, 68–74.
- Abrescia, N.G.A., Grimes, J.M., Kivela, H.M., Assenberg, R., Sutton, G.C., et al., 2008. Insights into virus evolution and membrane biogenesis from the structure of the marine lipid-containing bacteriophage PM2. *Mol. cell* 31, 749–761.
- Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., et al., 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* 66, 213–221.
- Afonine, P.V., Headd, J.J., Terwilliger, T.C., Adams, P.D., 2013. New tool: phenix.real-space_refine. *Comput. Crystallogr. Newslett.* 4, 43–44.
- Bai, X.C., Fernandez, I.S., McMullan, G., Scheres, S.H., 2013. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *eLife* 2, e00461.
- Barad, B.A., Echols, N., Wang, R.Y., Cheng, Y., DiMaio, F., et al., 2015. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* 12, 943–946.
- Brilot, A.F., Chen, J.Z., Cheng, A., Pan, J., Harrison, S.C., et al., 2012. Beam-induced motion of vitrified specimen on holey carbon film. *J. Struct. Biol.* 177, 630–637.
- Brown, A., Long, F., Nicholls, R.A., Toots, J., Emsley, P., et al., 2015. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. Sect. D* 71, 136–153.
- Campbell, M.G., Cheng, A.C., Brilot, A.F., Moeller, A., Lyumkis, D., et al., 2012. Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *Structure* 20, 1823–1828.
- Campbell, M.G., Veesler, D., Cheng, A., Potter, C.S., Carragher, B., 2015. 2.8 Å resolution reconstruction of the *Thermoplasma acidophilum* 20S proteasome using cryo-electron microscopy. *eLife* 4.
- Chen, V.B., Arendall 3rd, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., et al., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* 66, 12–21.
- DiMaio, F., Song, Y., Li, X., Brunner, M.J., Xu, C., et al., 2015. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* 12, 361–365.
- Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K., 2010. Features and development of Coot. *Acta Crystallogr. Sect. D* 66, 486–501.
- Holm, L., Rosenstrom, P., 2010. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* 38, W545–549.
- Jiang, W., Chiu, W., 2007. Cryoelectron microscopy of icosahedral virus particles. *Methods Mol. Biol.* 369, 345–363.
- Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858.
- Khayat, R., Tang, L., Larson, E.T., Lawrence, C.M., Young, M., et al., 2005. Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18944–18949.
- Khoshouei, M., Radjainia, M., Phillips, A.J., Gerrard, J.A., Mitra, A.K., et al., 2016. Volta phase plate cryo-EM of the small protein complex Prx3. *Nat. Commun.* 7, 10534.
- Laanto, E.M.S., De Colibus, L., Marjakangas, J., Gillum, A., Stuart, D.I., Ravanti, J.J., Huiskonee, J.T., Sundberg, L.R., 2017. Virus found in a boreal lake links ssDNA and dsDNA viruses. *Proc. Natl. Acad. Sci. U.S.A.* 114 (31), 8378–8383.
- Liu, H.R., Jin, L., Koh, S.B.S., Atanasov, I., Schein, S., et al., 2010. Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. *Science* 329, 1038–1043.
- Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., et al., 2011. REFMACS for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* 67, 355–367.
- Nogales, E., Scheres, S.H.W., 2015. Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol. cell* 58, 677–689.
- Petersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., et al., 2004. UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612.

- Pintilie, G.D., Zhang, J., Goddard, T.D., Chiu, W., Gossard, D.C., 2010. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* 170, 427–438.
- Roy, A., Kucukural, A., Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738.
- Scheres, S.H., 2012. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* 180, 519–530.
- Soding, J., 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960.
- Song, Y., DiMaio, F., Wang, R.Y.-R., Kim, D., Miles, C., et al., 2013. High-resolution comparative modeling with Rosetta CM. *Structure (London, England: 1993)* 21, 1735–1742.
- Terwilliger, T., 2004. SOLVE and RESOLVE: automated structure solution, density modification and model building. *J. Synchrotron. Radiat.* 11, 49–52.
- Terwilliger, T.C., 2013. Finding non-crystallographic symmetry in density maps of macromolecular structures. *J. Struct. Funct. Genomics* 14, 91–95.
- Walls, A.C., Tortorici, M.A., Bosch, B.J., Frenz, B., Rottier, P.J., et al., 2016. Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* 531, 114–117.
- Williams, C.J.H.B., Richardson, D.C., Richardson, J.S., 2013. CaBLAM: Identification and scoring of disguised secondary structure at low resolution. *Comput. Crystallogr. Newslett.* 4, 33–35.
- Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., et al., 2011. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* 67, 235–242.
- Wu, S., Skolnick, J., Zhang, Y., 2007. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.* 5, 17.
- Zhang, Y., 2009. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins* 77 (Suppl. 9), 100–113.