S2C2 cryoEM Image Processing Workshop June 12, 2020 (Webinar)
Q&A
Webinar ID: 948 6863 3490

Q		i have a question from previous day's talks . if you are not sure about how many conformational states your molecule have , is there any minimum number of random initialisation that you should start withbin Ab intio reconstruction. if this has been answered earlier, you can write to me offline . thanks !
	A	If you are not sure how many states there are (which will generally always be the case at the start of a project) you should always try ab-initio reconstruction multiple times, with 1, 2, 3, 5, etc classes, and inspect the results to try and see how many different conformations you find. At some point, you won't find any more conformations that are distinct and well-resolved. At some point the reconstructions will end up becoming degenerate with differnet views being separated into different classes rather than different conformations. Even after using ab-initio reconstruction, you should still use heterogeneous refinement and 3D Variability to separate your classes into further subclasses or to discover continuous heterogeneity.
Q		In case of large protein-RNA complex, can the grouping help to separate the RNA-protein interface from the rest of the protein? Is there any problem if the RNA is smaller in size compared to the protein?
	A	usually for this you want to do only a small amount of grouping, or better yet, use grouping by connectivity; you can follow the regions corresponding to the RNA and group them manually pretty easily if they are nicely resolved
Q		Regarding segmentation with Seger (via Chimera): I understand that parameters to be used for segmentation are very dependent on what we want to segment in the map, but is there any advised (/default) values to be used if one wants to segment the map in obvious chains, or is it always a play&check approach?
	A	yes it is play&check but the default parameters usually are ok; you can adjust parameters to remove dust or to create larger/smaller regions, otherwise the results tend to be similar; for grouping by connectivity, play with the threshold of the map, as that can affect what parts are connected
Q		Does this different fitting gives some kind of score? How to check which method work best for a particular map?
	A	usually they give a cross-correlation score for the best fit; it can be hard to compare scores from different methods because they may use different paramters; usually it is best to compare results visually, e.g. by opening the fitted models in Chimrea to see if they fit in the same place
Q		For ligand-protein interaction studies what will be the critical limit of resolution?
	A	i would say at least ~4Å so that you should see side chains; the higehr the resolution the better though
Q		do we need to do energy minimisation?or is it all intergrated in the fitting software?
	A	the fitting software usually includes some sort of energy minimization
Q		In phenix what would be the inputs, does it take the .mrc files and .pdb as the input model for the refinement or we need to change file types
	A	yes, it can take .mrc and .pdb files, the only other parameter is the resolution, where you want to use the estimated resolution of your map
Q		How do you deal with the flexible domains in protein, when you apply sharpening?
	A	if the domains are flexible you may not see them very well, so sharpening may add more noise; you may want to do the oposite of sharpening (Gaussian smoothing or filtering) to see them better; e.g. use Chimera -> Volume Data -> Volume Filtering
Q		How to generate a model when there are no prior structures of the protein available?
	A	you can try homology modeling tools like Phyre2, MODELLER, I-TASSER, and Swissprot. You just upload a sequence and it will give you a structure if there are similar sequences that have structures in the PDB
Q		Is the resolution is average for a structure in cryoEM as resolution is not uniform throughout the structure?
	A	yes usually you get the average resolution through the map or the masked part depending on how the mask is generated
Q		What solutions are there to situations, in which a disordered core of a protein shows as a hole in a low resolution cryo-EM map and causes problems to crystal model fitting?
	A	you can try flexible fitting to find different conformations of that variable regions (e.g. with MDFF)
Q		when calculating map-model FSC, is simulated map calculated from the model using global resolution or local resolution?
	A	great question! we use global resolution, but this is an interesting idea to use local resolution…
Q		Theorectically, can cryo-EM tell the chemical property , e.g. the kind of metal. of the each atom from the density?

	A	i'm not sure about the theory but in practice I think this would be quite hard because they all tend to look the same at the resolutions we get; so you may have to rely on distances to nearby atoms and coordination geometry (see our Q-score paper in bioRxiv or Nat. Meth. for a bit of intro to this) https://www.biorxiv.org/content/10.1101/722991v2
Q		might be a naive question so in case of heterogeneous sample or having multiple conformations the fitting is done for each state separately?
	A	yes, you usually start with rigid fitting and then do flexible fitting if needed; you can compare the flexible fit to see how much the model has changed
Q		Can you use the linear regression of Q-score to resolution to normalize the Q-score and compare important residues (let's say catalytic ones) for proteins at different resolutions?
	A	interesting question; you could normalize by dividing by the the expected Q-score for the resolution, values of 1 or larger mean less mobility (higher resolvability), whereas values lower than 1 mean more mobility (less resolved)
Q		Do you use 1.43 cut-off or 0.5 for your resolution vs Q-score plot? Do you get better correlation if you use the 0.5 cut-off?
	A	interesting question! i used .143, maybe i should try .5 too….
Q		How is the RNA identified in the EM map of TMV?
	A	basically by segmentation or building the model; because the model was available I simply used the model in this case
Q		How the partial occupancy is addressed in calculating the Q scores? or its not taken in account?
	A	it's not taken into account; typically lower occupancies may not lead to lower q-scores, it's more the rate of decrease in density that is being considered
Q		In continuation to my earlier question, how do we identify an oligonucleotide in the EM map, if we are not sure about the binding site on the protein?
	A	usually you have to try to build the model that best matches the map; you can try to segment it first by connectivity and pull out just the part that looks like RNA rather than protein (this is what we did for the CRISPR-Csm map)
Q		On a segmented map: Is it possible to change the contour level of only one segment of the map?
	A	good question, i've been tyring to put this in for some time, but unfortunately the only way right now is to resegment the entire map at a different threshold; to increas the threshold is possible, first you can use the 'Extract' tool to extract the segment, then it becomes a separate map where you can increase the threshold
Q		Why waters in EM not as clear as in X-ray maps?
	A	possibly a few reasons: 1) waters may have more positional variations in cryoEM, so they blur out; 2) water molecules may arrange differently in the ice condition vs crystal condition
Q		can we calculate Qscore during refinement?
	A	typically we do it before and after refinement, i'm not too sure it's too useful during
Q		Is there a way to do occupancy refienement,?
	A	not that I know of, again B-factor/ADP refinement is usually done by phenix.refine; it may do occupancy refinement if you start off with occupancines that are other than 1, but check with the phenix team for more details; also you may want to check with the CCP4/MRC team that develops Refmac
Q		If there is a partial occupancy of the domain in an enzyme, is there a way to apply occuapncy refienemnt in cryo-EM
	A	no that i know off… occupancy usually has to do with multiple conformations, so you could build these multiple conformations by hand and then assign them occupancies perhaps
Q		Some comments: 1. Mask application is definitely improve the refinement. It is the same as "solent flattening" in x-ray structures. There are some parameters, like Ks and Bs need to be optimized. 2. Fitting model against map, The "resolution" of the CryoEM map is not the same as X-ray data resolution. The "local resolution" may be correct concept for cyoEM map. So, when refining the model, the model need apply restraints for the low resolution region. 3. Q-score is a good criteria to fitting individual atom to the electron density map (in atomic resolution). But it cannot prevent from "overfiting", For low resolution region, we have to apply restraints, Q-score is invalid. Higher Q-score does not means a good model. 4. The problem is that what is the data error" in CryoEM map? What statistics tool do we use to describe? Push to higher and higher "high resolution" for CryoEM maps, seemed a just improving the sign to noise ratio. 5. We need develop something like the free R factor for model assessment

	A	Thanks for the comments; yes I think we have to evaluate the map statistics and using the model or multiple models may be such a way to evaluate the quality and limitations of the map
Q		When using Segger to generate regions for extraction to use for Masks, do you hare any suggestions on parameters to use during extraction (fall off, gaussian filter, etc)?
	A	i guess it depends on how you are using the mask; for calculating FSCs, you definitely want to use the fall-off or gaussian filter, with a widht of at least 10 voxels or so; for use in other software such as cryoSparc or Relion a simple 0/1 hard mask may be ok because they may smooth it automatically perhaps?
Q		how to search structure below 5 angstrom by subtomogram averagign?
	A	rigid fitting may be best, possibly also good to use refinement or flexible fitting if the model fits ok but looks a bit different
Q		Greg, Have you tryied"omit map" to re-generate the map region, and then compare the difference of Q-score?
	A	Neat idea but I haven't tried that...
Q		in case of proteins having a flexible region or a disordered region what kind of masking would you suggest?
	A	if you want to consider the entire portein area then masking after applying smoothing may be best; if you want to focus on just the well-defined part then a tighter mask may be ok too
Q		What is the recommended countour level for Atoms?
	A	i think you have to try different levels by eye; if you mean the 1.2Å map, at high threshold you can see the atoms separate; at lower threshold they merge with nearby atoms
Q		A general question…What is involved in being able to run cryosparc on the cloud if we don't have access to a cluster or a workstation? How well does it work? Is there a preferred cloud computing provider?
	A	CryoSPARC works great on the cloud- in fact we have many users that use Amazon AWS EC2 instances (workstations that you can reserve on-the-fly) with NVIDIA GPUs for cryo-EM data processing. The installation process is exactly the same as if you were sitting right next to the workstation. A few users also chose to use Amazon's Parallel Cluster to use cryoSPARC in a cluster setting, where the cluster itself automatically scales up and down with the load required. For your reference, here is cryoSPARC's hardware and system requirements, which you can use as a reference when considering cloud resources (all the recommendations still apply): https://guide.cryosparc.com/setup-configuration-and-management/hardware-and-system-requirements Also, heres a great article by Dr. Cianfrocco on the topic: https://aws.amazon.com/blogs/publicsector/structural-biologists-learning-cryo-electron-microscopy-have-new-educational-resources-powered-by-aws/ Please feel free to reach out to us on our discussion forum if you have any questions about setting this up! https://discuss.cryosparc.com/
Q		Actually, I mean search structure below 5 angstrom by subtomogram averagig in the EMDB
	A	see this advanced search feature: https://www.ebi.ac.uk/pdbe/emdb/searchForm.html/ You could input for example, lowest A = 5, and EM method = Subtomogram averaging. Greg and others may want to chime in.
Q		Regarding segmentation, Greg, will you please comment on Haruspex from the Thorn lab? It is implemented in CCPEM, I believe. I haven't used it yet, but it seems to be a tool that complements segmentation.
	A	Thanks for the suggestion; Yes it looks like it could be a nice tool to identify secondary structures; I've been meaning to try to do more of this within Segger, maybe in the near future...
Q		Are there tutorials about setting up a workstation for cryoEM, including all the programs to download, what comuting powers are needed, setting up data storage and back up, accessing servers for data processing?
	A	generally the different software packages will list all of these kinds of requirements/information on their respective websites and they may have links to tutorials there. For cryosparc you can find that at guide.cryosparc.com. For cryoSPARC: - hardware and system requirements, see this page: https://guide.cryosparc.com/setup-configuration-and-management/hardware-and-system-requirements - downloading and installing: https://guide.cryosparc.com/setup-configuration-and-management/how-to-download-install-and-configure - accessing the user interface: https://guide.cryosparc.com/setup-configuration-and-management/how-to-download-install-and-configure/accessing-cryosparc
Q		where can we post questions regarding cryo SPARC and image processing in future?

	A	you can post on the discussion forum at discuss.cryosparc.com or contact feedback@structura.bio. Documentation is available at https://cryosparc.com/docs and guide.cryosparc.com
Q		Thank you Saara and Stephan for the answer about the computing power. So for example and to get an idea of the computing power needed, can I process ribosome cryoEM data on my macbook pro computer, or whould I need to have external storgae and access to servers?
	A	The main requirement for "computing" data in cryoSPARC is having an NVIDIA GPU on your system- newer Macbook Pro's don't have NVIDIA GPUs anymore, but if you do actually have one, it's possible. Although for a more long-term cryo-EM data processing solution, it's best to acquire a proper workstation and external storage.
Q		In case of Negative staining data processing, which motion correction job to use? Patch-motion job gives me an error, after I import data with Import Micrograpgs job.
	A	Please take a look at our tutorial that covers parameters to select when working with negative stain data: https://cryosparc.com/docs/tutorials/negative-stain-phase-plate. Also you don't need to run motion correction on micrographs- motion correction requires "movies", or .mrc files with multiple frames. In this case, you can go straight to Patch CTF Estimation in cryoSPARC.
Q		What symmetry was used with the ATPase or was it C1?
	A	ATPase is C1. All data processing information will be on the Workshop's homepage: https://cryoem.slac.stanford.edu/s2c2/s2c2-cryoem-image-processing-workshop-agenda-june-10-12-2020-zoom-only-participation
Q		A technical question about the small solule protein (50 kDa) data processing. I'm trying different methods to push the streptavidin reconstruction better, but I fail to get a correct 3D-reconstruction in cryoSPRAC with default parameters in refinement/NU-refinement even using the same particles imported from relion (reach to 2.6A). When process from scratch in cryoSPARC, the 2D in cryoSPARC look as good as in relion, the ab intio shows a correct but relatively low resolution structure. However the refinement get worse than the ab intio map. Are there any specific parameters need to be adjusted for small proteins? Thanks!
	A	In this case, you should set the abinitio reconstruction to have it's max resolutino to be higher - maybe 4 or 6 A instead of 12. Then in the refienement, set the "initial lowpass resolution" to a higher resolution also - maybe to 15 A or 10 A. This will allow refinement to lock on to higher resolution details and resolve the correct structure instead of getting stuck at low resolution.
Q		Can cryoSPARC delete multiple jobs at a time?
	A	It's not possible to delete multiple jobs unless you'd like to delete an entire workspace or project

QUESTIONS LIVE ANSWERED

At the most basic level, what is sharpening? does it enhance the resolution or just make the map better for visulaization
grouping by connectivity vs grouping by smoothing are chosen based on your molecule and how it changes the final model ? - Live answered and following answer: grouping by connectivity works better at high resolutions, but grouping by smoothing can be used at all resolutions with good results (as long as there aren't too many protruding narrow segments)
It seems like there are a few ways to fit models into EM densities. Is there a resolution limit for your structure that needs to be considered when using these different approaches?

Are there any big no-no's in fitting models into cryo-EM data?
can a molecule be subjected to both rigid as well flexible fitting and both the data can then be overlapped? - Live answered and following answer: yes, usually you start by rigid fitting, then do flexbile fitting; you can compere the results to see how much the flexible fitting has changed the molecule
what is your recommendation for contour levels of the map when fitting the model? any threshold needs to be conisdered?
what constrains to use to avoid the overfitting ,what is MDFF
the map values, they can differ by the contour you choose. what value of contour is used
Practically, what we expect about the Q-score for the whole map?
Which program used to calculate atom profile and Q-score?
Is a different set of library parameter is being used for Q score of high resolution map, since these are electron potential but not electron density map?
*so occupancy and the Q score have direct positive coorelation ?

In FSC figure: what is plotted in 3-sigma and 1/2-bit curves? what information contained in these curves?
Are the depostion authors/PI the same as the literature authors/PI?
Is there the option to evaluate the model the same metrics but in different resolution range?
What is this challenge pipeline about? Is this what we need to do for deposition?
Based on how the refinement is behaving, are there rationale thoughts behind choosing mask dilation and padding and the resolution at which to start mask input? I often see erratic or non explainable behavior.
Is the service also available for non-US groups?
My protein 40 kDa forms dimer. 2D classification does not align even after reducing max alig resolution to 8 or 10 A. What strategies can I use to align well? Also, is it work going directly to Ab intitio construction and refinement directly?
If we collect the data at S2C2, can we access the computational system to process the data remotely? Or do we need our own computational systemn for data processing?

In reconstruction of viruses, is there any advantage of using C1 symmetry? does it remove restrictions imposed by I symmetry?
Can cryosparc deal with helical particle?
1. My sample is a small membrane protein in a tight nanodisc. I get a mediocore resolution about 6A. I suspect the density coming from the scaffold protein affect the alignment in a bad way. Using NU refinment I still see significant density from the scaffold protein. Any suggestion for processing strategy?
2. Any metrics/parameters to predict final map quality (resolution) from raw dataset? To know if it is the raw data quality or the processing strategy I use that is limiting my final result.
based on your talk about non-uniform refinement, it seems that flexibility is no longer a limit for maximum resolution that we can achieve with cryoEM - it seems that, if you have a very flexible protein, you just need to collect a lot (a huge lot) of images and you'll be fine. What will be the limits for resolution, now? Can you please comment on this.

can I classify the particles regarding to a small peptide density only? So if one part of the particles include this peptide, the other doesn't?
Do you applay different type of masking strategy if you are proccesing a membrane protein that is in nano-disk?
I have a practical question about cryosparc. If I have a known crystal structure that acounts for half then the protein content of my sample, at what stage should I use this data and how.
How to address data collection for a elongated membrane protein in the nanodisc with a flexible hinge midway having only side view ?
Q about heterogenous samples where my complex oligimerizes to form say e.g. 300 kDa, 600 kDa structures. Can I pick different particles depending upon their dimensions?

Could you please explain when to use Deep Picker (Topaz jobs)?
any recommmendaed resources or tools for integrating or correlating single molecule flourescence imaging data with cryo-EM imaging?
Patch CTF question: Do CTF estimations for micrographs with larger carbon edges behave weird in patch CTF estimation? AND How does patch CTF estimation work with phase plate data?
In final stages of reconstruction we use calibrated pixel size, is the difference between px size 1.18 and 1.11 considered significant? does it affect the final map model and corresponding structure?
Any recommendations on how to handle particle picking/inspection on lacey carbon grids?
Is there a pipline to pick, classify, and reconstruct random conical tilt data?
Could you briefly describe choosing the Fourier crop box size again for particle extraction as we did in the T20S tutorial? How can this be changed later?
Could you please explain what are the laser phase plates? And which cases they may be useful for?
can we use rectangular box for particle picking?

NO RECORDED ANSWERS NOTED

Is there a way to apply sharpening only to rigid part of the protein?

3sigma - noise threshold 1/2 bit - sensibel information above noise
Hi, could you discuss how to set up a working station for cryoEM analysis?
Recently, I've read a manuscript draft about a membrane protein composed of a transmembrane domain and cytosolic domain. The authors were able to determine the TM region and they claim that the cytosolic domain is COMPLETELY invisible in the final model and in 2D classes due to its high flexibility. I was wondering why the flexible regions are not fuzzy/blurry (e.g. like in the nicely demostrated case of unfoldase). How the complete invisibility could be explained? Have you ever encountered such cases?

If i want to see the variability in my dataset by 3D variability, how many particles i should have for running it? and which is the minimum resolution needed?
Question for the CS team - could you briefly describe choosing the Fourier crop box size again as we did in the T20S tutorial? How can this be changed later?
The xx axis of the GSFSC plot of any refinement is limited at the right side (high resolution values) by a certain value. In other words, there is a specific range presented for the xx axis of the plot. What makes the plot to present only up to some high-res value? Is it the nyquist limit? or I have a GSFSC plot where the curves go beyond that max of the represented xx axis and I don't see the curve crossing the 0.143 threshold. What does it mean? It happened when I downsampled particles.

In case of ResLog Analysis job, how can one read and understand the results? What does different curves represent? Which curve and plot to finally interpret?
While inspecting the particle picks, how does the change in box size affect the 2D classification results?
How is soft mask generated?
to Ali - is there any way (like the non-uniform refinement in 3D) to get a better alignment of a small membrane protein completely inside the micelle to make 2D classification work?
do you apply different masking for proteins that can be present in different multimetric states like tetramer or dimer?
How to compare CryoSPARC,EMAN2,Relion, the advantage and disadvantage?
Should it go to better than 7.2A?
I have a complex where a chain keeps their shape but rotates/has an angular motion in respect to the anchor point that connects it to the complex.
I believe the right way is to try to follow the cryoSPARC Local Refinement Tutorial (Case Study: Yeast U4/U6.U5 tri-snRNP). I have tried it but seems to be quite challeging to get a better resolved flexible chain. Do you have any tricks/hints/tips on how to approach this tutorial? Where to be more careful? Mask generated with volume eraser (Chimera) didn't seem to yield a good result...
I have performed heterogeneous refinement to separate confirmational changes, but still in structure I see both the confirmations. Can you suggest how to improve this?

Can we run cryoSPARC on the cloud? If so, what does that involve? Does it work well? Is there a preferred cloud computing provider?
Did you ever try to use cryoSPARC with Zernicke phase plate data?
Do you have plans to build a pipeline for tomography data in cryosparc?
A question for Ali: In 3D variability analysis, we obtain variability components that, as you mentioned are uncorrelated but not independent, so all of these motions are taking place at the same time and might be integrated. My question is what is the correct way to go about interpreting the output from 3DVA? What information can be obtained concretely?

I have an elongated protein where the N' and C' halves are thought to move as semi-rigid bodies relative to each other around a hinge region, with a large range of motion (100Å or so). Assuming the EM images will show a range of conformations between the two halves, is it possible to align the EM images in two sets, each focusing on the individual halves, where you would obtain separate maps of the two stable halves independent of each other?
How to address data collection for a relatively elongated asymetric membrane protein (380A length and 25A width) constituted in the nanodisc with a flexible hinge midway of long axis having side view prefered orientation?
Will CryoSPARC add the Ewald Sphere correctiton soon? Or do you have suggestions where to do so using cryosparc results?
For the elongated-shape protein, there are still issues on particle picking. It's hard to pick it well in center by relion or CisTEM in my case. I need to manually recenter the particles. Is any way to pick it well by cryoSPARC?
to Ali: can you think of an example in which enforcing symmetry of heterologous refinement is a good idea?
Will CryoSPARC add the Ewald Sphere correctiton soon? Or do you have suggestions where to do so using cryosparc results?

As Ali pointed out there are various side view for this protein
can I use tilted series data collection?