III: Single particle cryoEM - practical approaches

Single particle EM analysis can be performed at both 2D and 3D.

Single particle EM (both negative stain and cryo) is to extract structural information (both 2D and 3D) of macromolecules by averaging a large number of molecules without crystals.

A single particle image data set is a collection of images, each contains projection images of one molecules. The orientations and position of particles in all images are different. Before averaging, one needs to:

- judge how similar is the two particles: *cross-correlation coefficient*;
- shifts/rotates one particle to match another by maximizing ccc: *alignment*;
- separate different particles for averaging: *classification*;

Alignment ⟷ Classification

Alignment between two images

Alignment is a process to search the grids to maximize the cross-correlation coefficient between two images. Three parameters are used to define alignment of 2D images: in-plane shift (x,y) and in-plane rotation angle.

Cross-correlation function based alignment:

- In-plane shift can be determined by determine the peak position in the translational cross-correlation function between two images.
- Rotation can be determined by different ways: rotational cross-correlation function, Radon transform.

A digital image is collection of numbers in a grid

3	20	5	-3	4
3	5	34	45	4
0	-2	34	45	6
-1	34	2	3	1
4	5	2	2	0

$$f = \sum_{j=1}^{J} f(\vec{r}_j) = \sum_{j=1}^{J} f(m_j, n_j)$$

3	2	5	-3	4
25	2	4	2	4
0	34	45	5	6
-1	32	40	2	1
35	3	2	2	0

$$g = \sum_{j=1}^{J} g(\vec{r}_j) = \sum_{j=1}^{J} g(m_j, n_j)$$

Cross-correlation coefficient

Cross-correlation coefficient is a measure of similarity and statistical interdependence between two data sets. The mathematic definition of cross-correlation coefficient is:

$$\rho = \frac{\sum_{j=1}^{J} \left[f_1(\vec{r}_j) - \langle f_1 \rangle \right] \left[f_2(\vec{r}_j) - \langle f_2 \rangle \right]}{\left\{ \sum_{j=1}^{J} \left[f_1(\vec{r}_j) - \langle f_1 \rangle \right]^2 \sum_{j=1}^{J} \left[f_2(\vec{r}_j) - \langle f_2 \rangle \right]^2 \right\}^{1/2}}$$

Where: $\langle f_i \rangle = \frac{1}{J} \sum_{j=1}^{J} f_i(\vec{r}_j)$

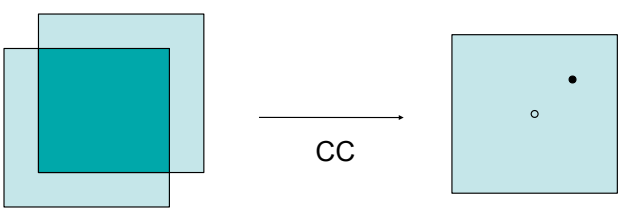
Note that: $-1 < \rho < 1$

Cross-correlation function

The cross-correlation function is the most important tool for alignment of two images.

The mathematic definition of cross-correlation is:

$$f * g = \int_{-\infty}^{\infty} f(t) g(t - \tau) d\tau$$



Q: what happens if shift is more than half of the image size?

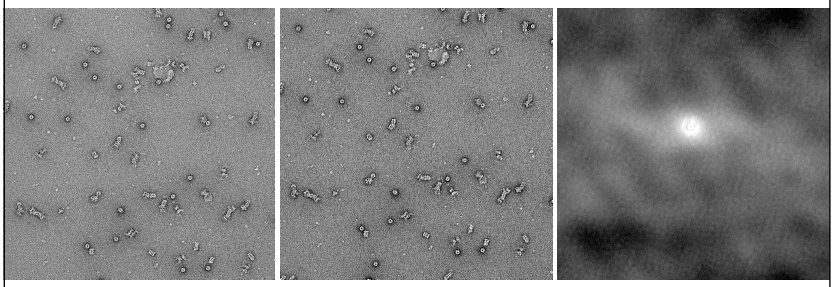
Calculating the cross-correlation

Cross-correlation theorem:

$$f * g = \int_{-\infty}^{\infty} f(t)g(t-\tau)d\tau = F\left\{F(f)\cdot F^{-1}(g)\right\}$$

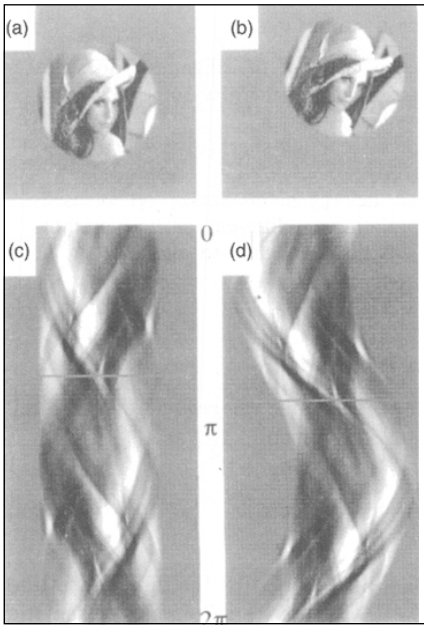
This formula enable us to calculate the cross-correlation between two images easily.

How cross-correlation looks like



-1μm -1.5μm CCF

The image size is 1024X1024. The peak in the CCF is at (445,500). How much is the shift?



Radon transform

Radon transform is an efficient way for determining angular relationship between two images, but it only works well in images with high SNR.

More about the cross-correlation function

- Peak searching in the cross-correlation function;
  search for a peak is not just finding the point of highest value in the CCF.

- Keep in mind that one can calculate cross correlation between any two images, and will always find a point with highest value.

- Cross-correlation based alignment and averaging always enhance the features of the reference image.

Classification

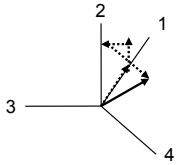
Classification - a process of dividing a set of images into subsets with similar features.

One can perform classification based on CCC to determine if the images are similar with each other;
But for a very large data set of very noisy images (> 50,000 images)?

Hyperspace

An image of m×m pixels can be represented by a vector (or end point of a vector) in the hyperspace of m×m dimensions.

3	2
2	3



$$f = (f_1, f_2, ..., f_m) = \sum_{i=1}^{m} f_i \vec{a}_i$$ Where: $|\vec{a}_i| = 1;$

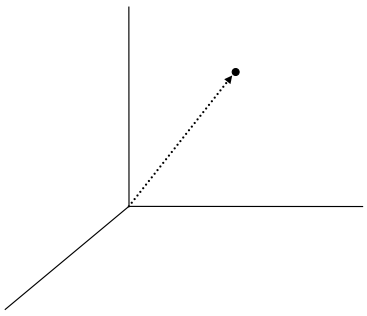
$$\vec{a}_i \perp \vec{a}_j (j \neq i; j = 1,..m);$$

Similar to the cross-correlation coefficient, the distance between two spots in the hyperspace represents the difference between two images.

A data set is represented as a cloud in the hyperspace. The center of the cloud is the average of the all images in the data set.

A data set is represented as a cloud in the hyperspace. The center of the cloud is the average of the all images in the data set.

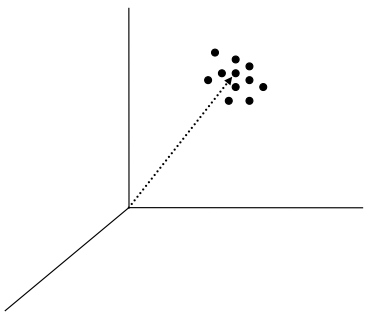
An image without any noise is represented by a point.



A data set is represented as a cloud in the hyperspace. The center of the cloud is the average of the all images in the data set.

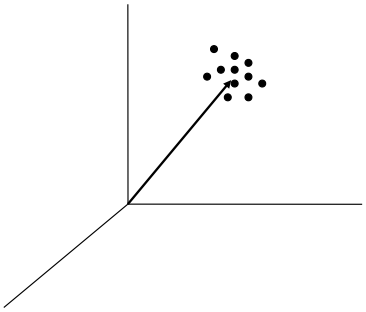
An image without any noise is represented by a point.

Adding random noise to the image expand the point into a cloud.



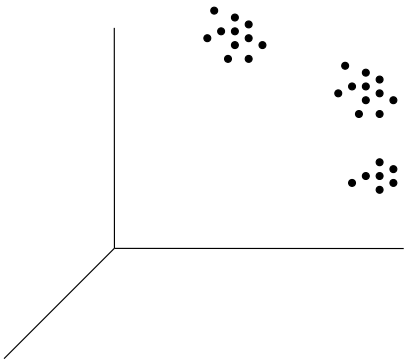
A data set is represented as a cloud in the hyperspace. The center of the cloud is the average of the all images in the data set.

The center of the loud is the average.



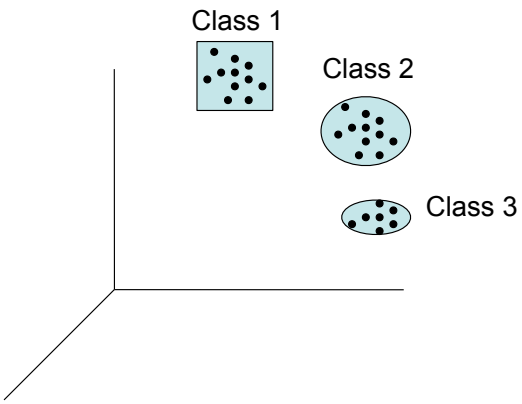
Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



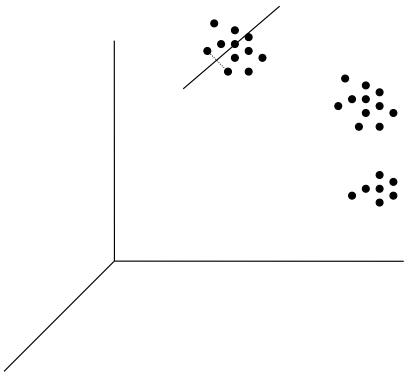
Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



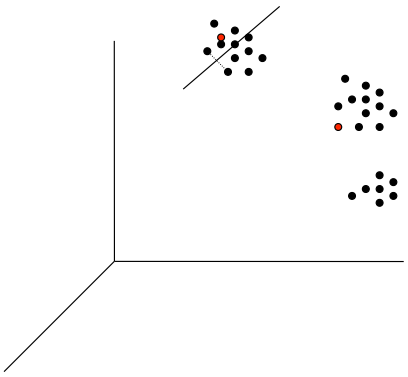
Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



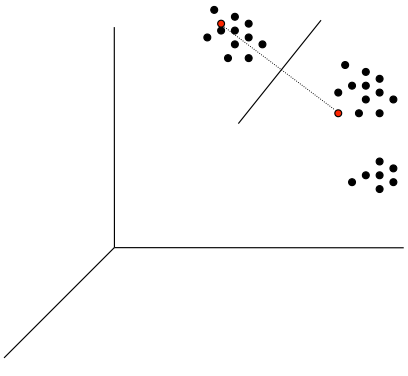
Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



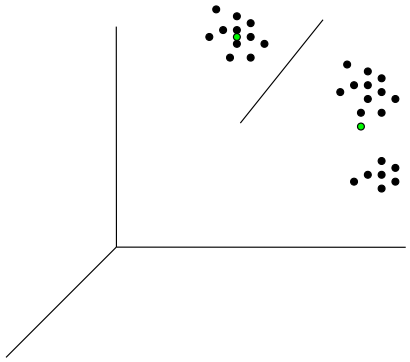
Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



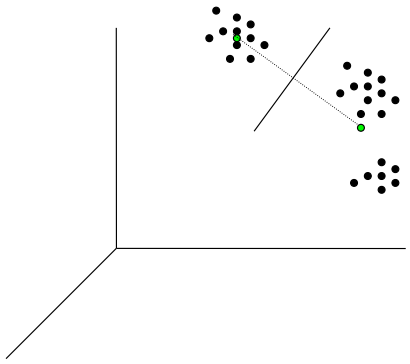
Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



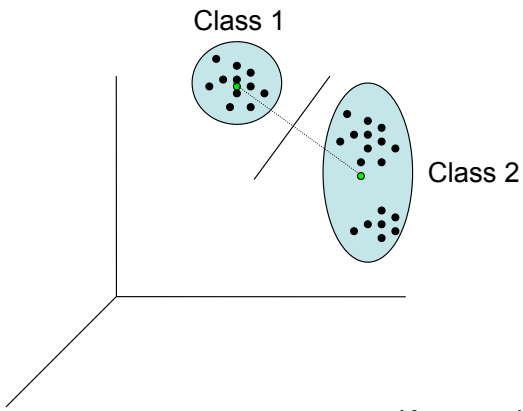
Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



Classification

Assume images are aligned with each other. The clouds of particles can be grouped into different groups - classification.



K-mean classification

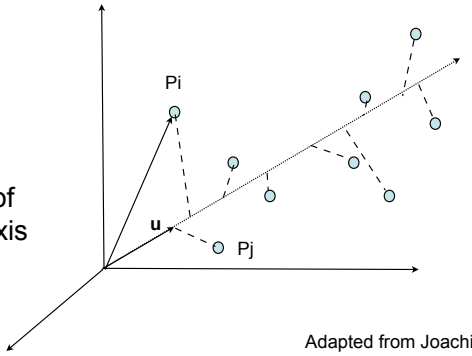
Multivariate statistical analysis

Making patterns emerge from data

Multivariate statistical analysis:

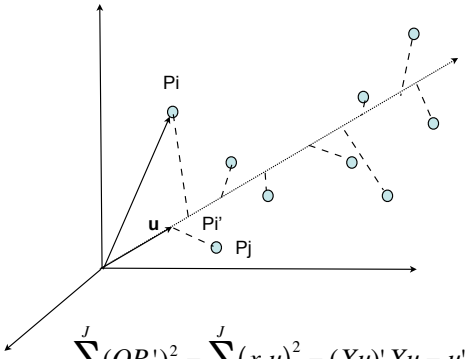
- Principal Component Analysis
- Correspondence Analysis

Definition of principal axis



Adapted from Joachim Frank

Principal component analysis (PCA)



$$\sum_{j=1}^{J}(OP_i')^2 = \sum_{j=1}^{J}\left(x_i u\right)^2 = (Xu)'Xu = u'X'Xu \longrightarrow \max$$

with constraint: u'u = 1 X: coordinate matrix

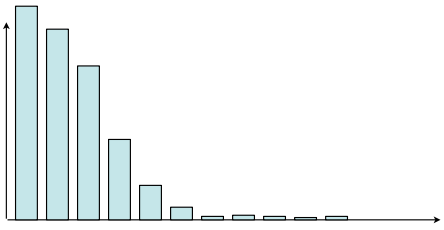
Eigenvector-eigenvalue equation

$$\mathbf{Du} = \lambda\mathbf{u}$$

where $D = (X - \overline{X})'(X - \overline{X})$

Solution of this equation generate a set of eigenvectors and eigenvalues.

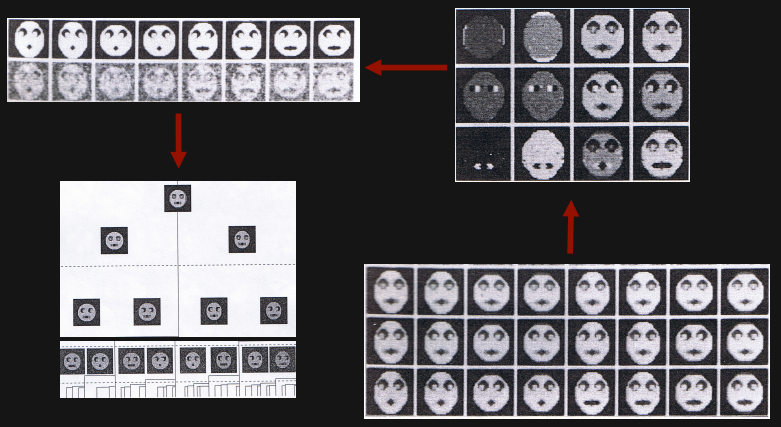
Significant factors:



Classification based on eigenvector/eigenvalue clustering;

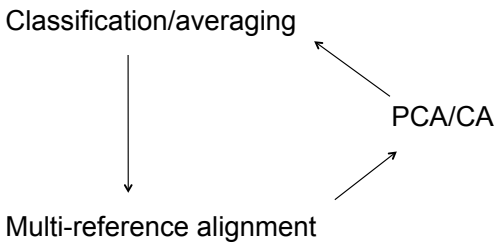
Multivariate statistical analysis & classification of images

Principle Component Analysis



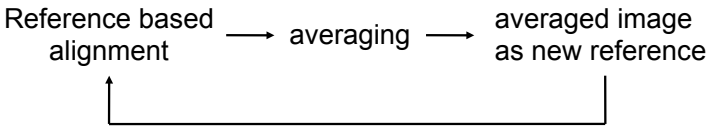
Iterative MRA and classification

For a heterogeneous data set (multiple structural conformation and/or protein compositions presented in one data set), iterative cycles of classification/multi-reference alignment is performed.



Iterative alignment

Assume a data set of identical particles of different in-plane rotation:



Q: during the iterative alignment the new reference is the averaged image of previous alignment cycle, but what are the images used for the alignment in the next cycle? The original images or the images after the alignment?

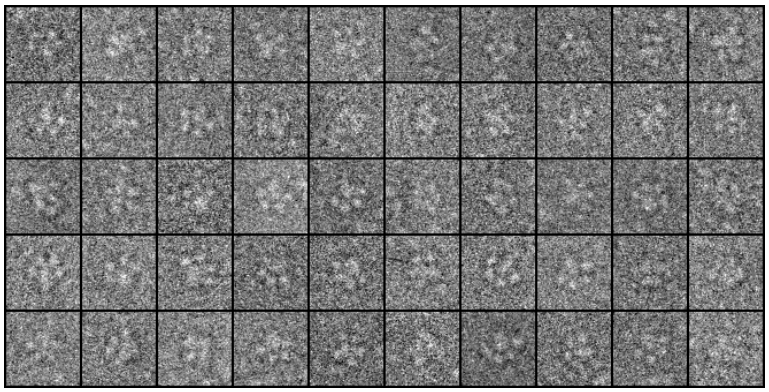
Demonstration of reference induced bias

Note: The averaged image after reference based alignment is strongly biased towards the reference.

100 images 1000 images reference

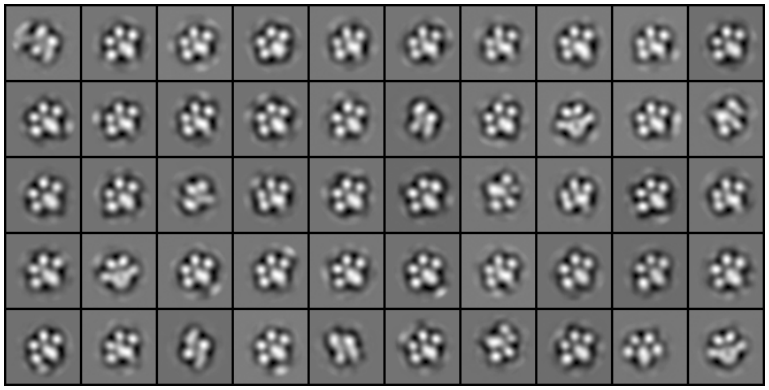
From Niko Gorigorieff

Individual TfR-Tf Complexes in Vitrified Ice



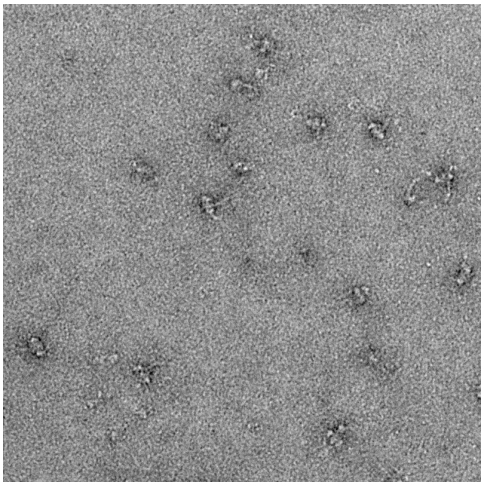
50 out of 36,266 particles

Class Averages of Vitrified TfR-Tf Complexes

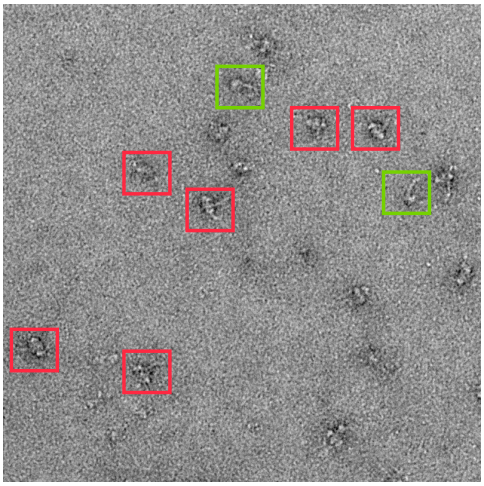


50 out of 200 classes

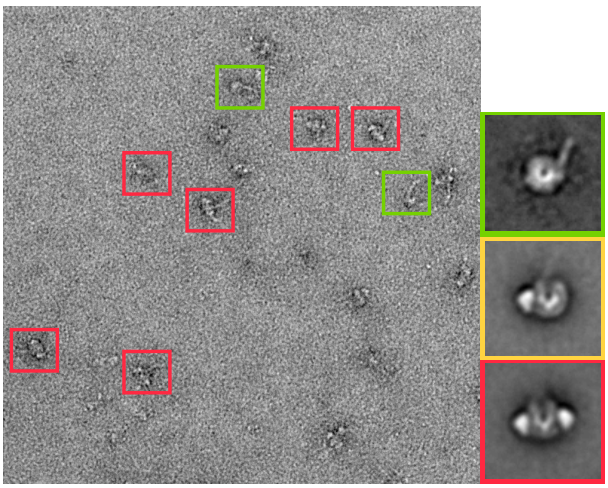
SNF2h-nucleosome complex



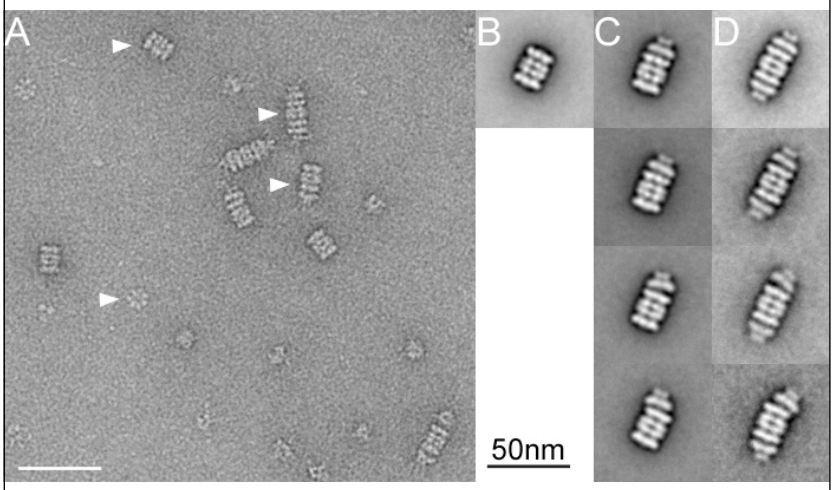
SNF2h-nucleosome complex



SNF2h-nucleosome complex

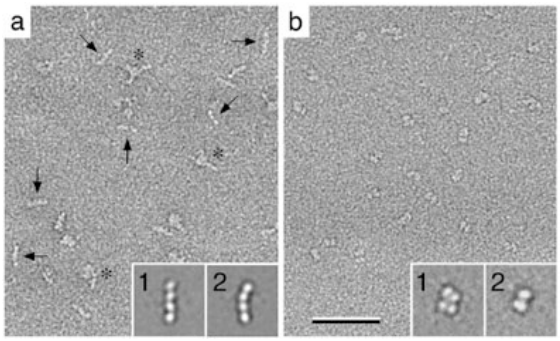


A simple case of using image alignment and classification

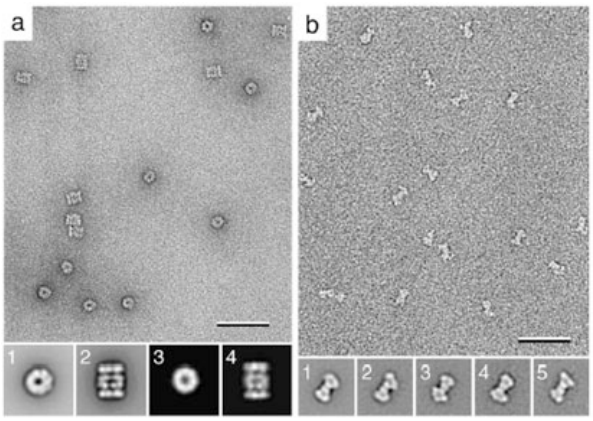


20S-PAN complex

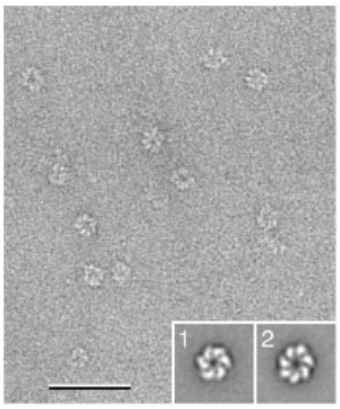
A number of examples



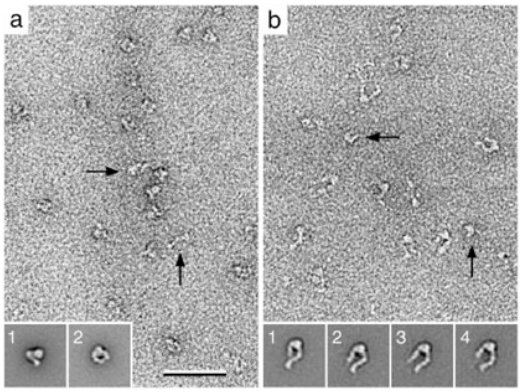
A: integrin $\alpha_5\beta_1$ headpieces and a fibonectin fragment (~40kDa). B: human transferrin (~70kDa, each domain is about ~17kDa).



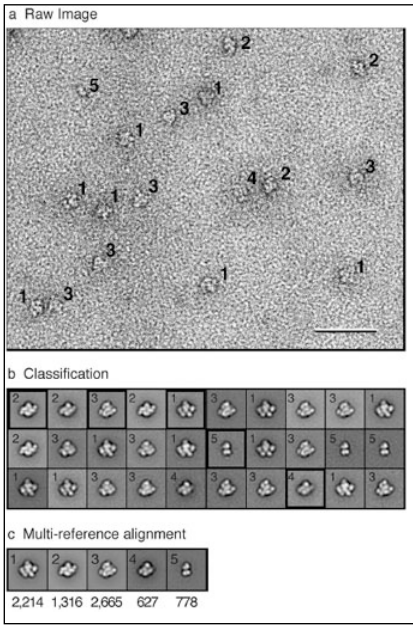
A) Yeast 20S proteasome; B) yeast Sec23p/Sec24p complex;



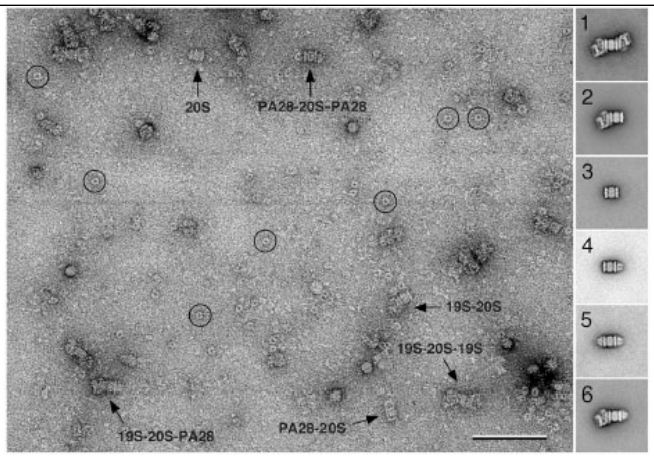
Heterogeneous population of bacteriophage T7 helicase/primase;



Integrin $\alpha_v\beta_3$ in the preseence of inactivating Ca^{2+} (a) and activating Mn^{2+} (b).



Human transferrin receptor-transferrin complex



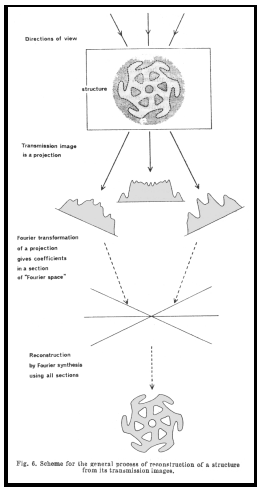
A heterogeneous sample of 20S proteasome, 19S regulator complex and PA26 activator.

Fourier Central section theorem

Central Section Theorem :
Fourier transform of a 2D projection equals the central section through its 3D Fourier transform perpendicular to the direction of projection.

DeRosier, D. and Klug, A. (1968) "Reconstruction of three dimensional structures from electron micrographs" *Nature* ***217*** 130-134

Hart, R.G. (1968) "Electron microscopy of unstained biological material: the polytropic montage" *Science* ***159*** 1464-1467

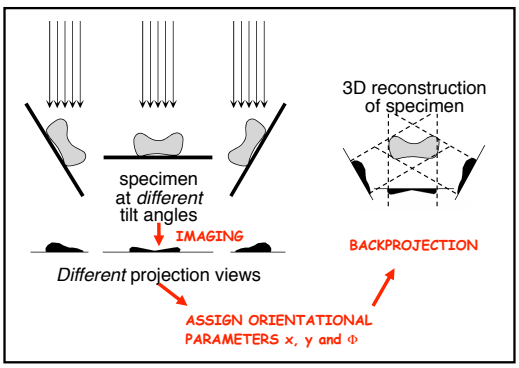


DeRoiser and Klug (1968)

3D reconstruction

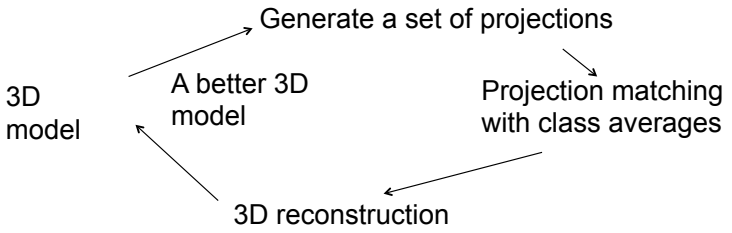
Assume we have already a number of class averages, they represent the projections of a 3D object in different orientations. And we know (can determine) these relative orientations of each class averages. We can reconstruct the 3D object - 3D reconstruction.

Back projection:



How to determine the relative orientations?

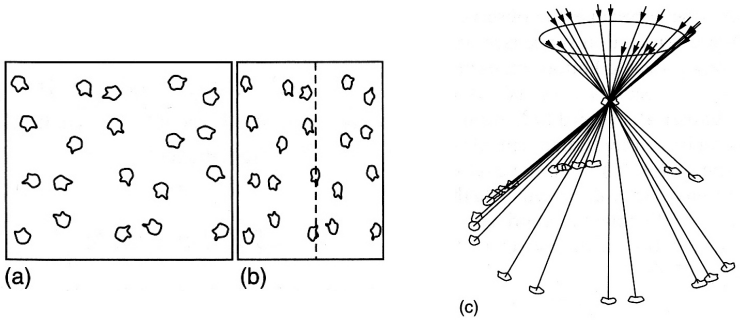
Model based projection matching: by matching (ccc) of class averages with the calculated projections of the 3D object in known directions - a refinement procedure.



Question: Where do you get the first model?

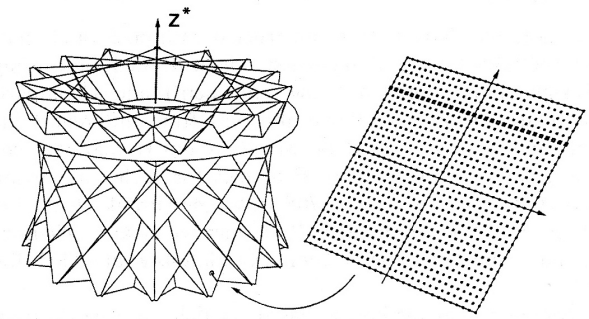
Random conical tilt

A pair of images are taken from the same specimen area for the random conical tilt 3D reconstruction.

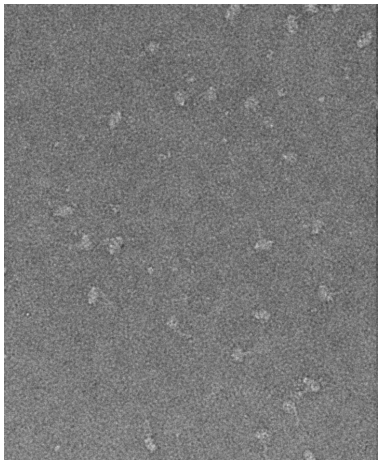


From Joachim Frank

Random conical tilt

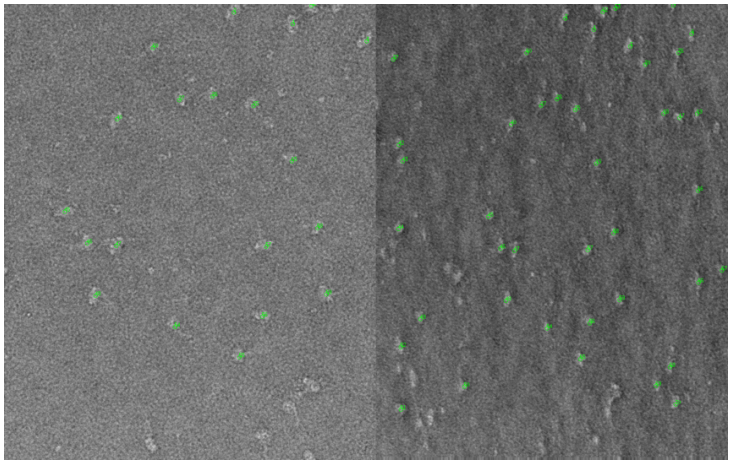


Random conical tilt



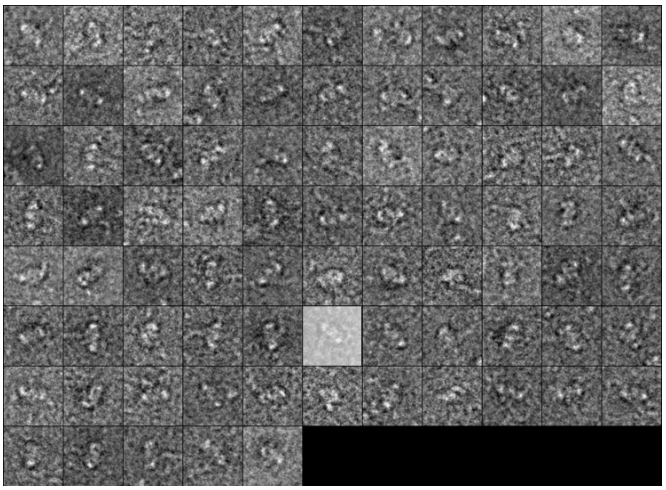
untilted image

Random conical tilt



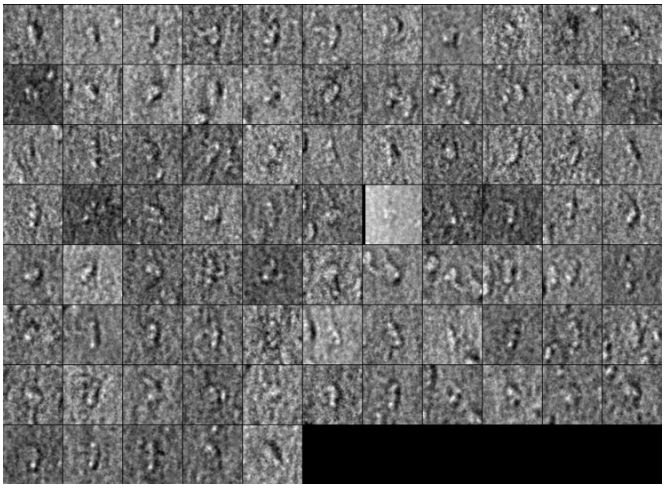
untilted image tilted image

Random conical tilt



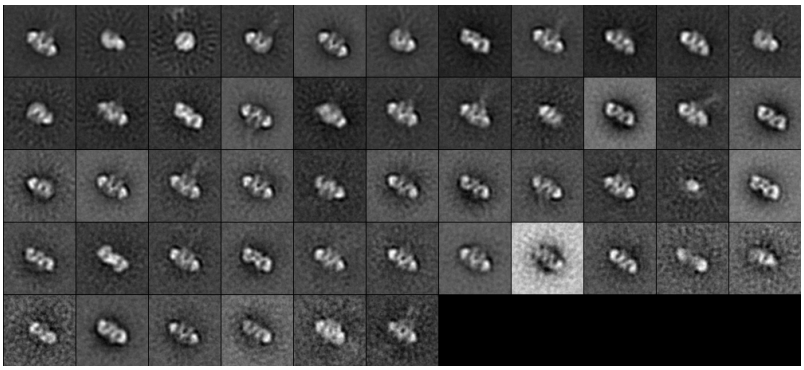
particles from untilted images

Random conical tilt



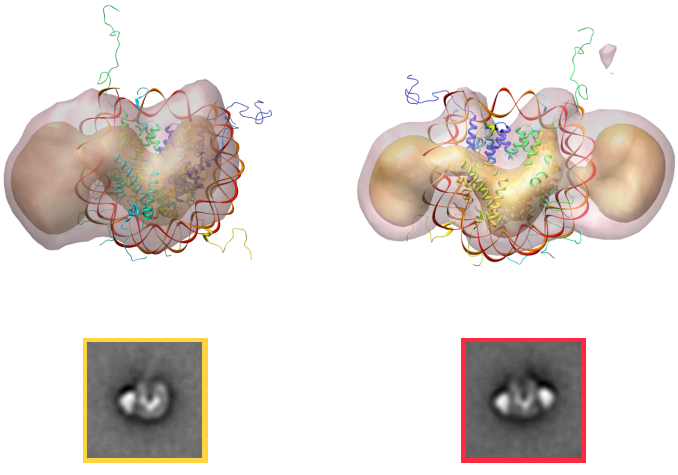
particles from tilted images

Random conical tilt

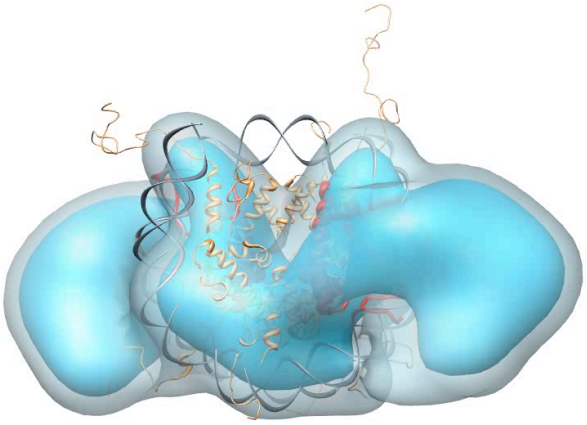


class averages of untilted images

Random conical tilt 3D reconstruction

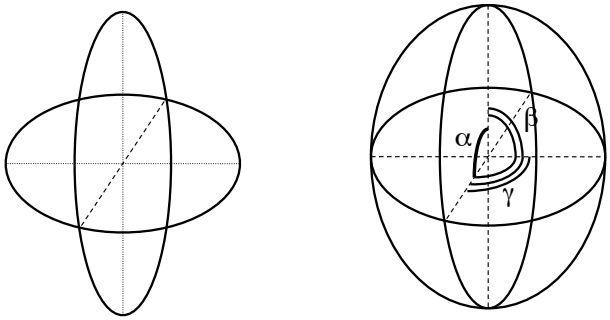


Nucleosome-SNF2h complex

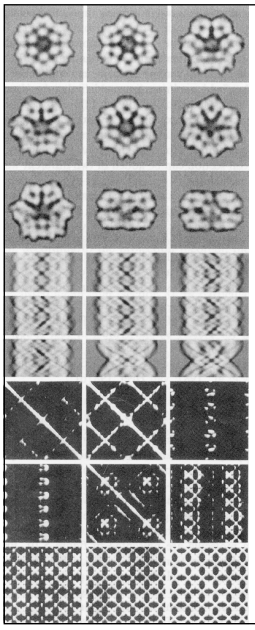


Common line/angular reconstitution

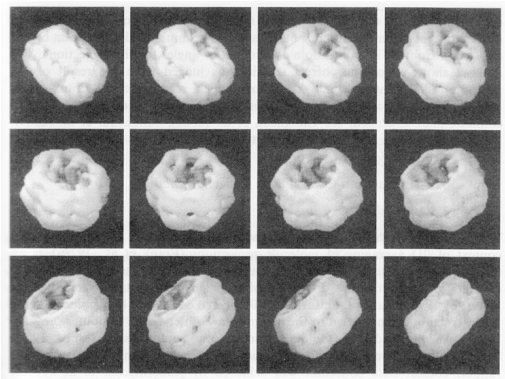
Any two central sections in the 3D Fourier space across each other will have a common line.



With angles between three central section determined, eular angles of any images can be determined by common line.



Common line/angular reconstitution



3D reconstruction of frozen hydrated *Lumbricus terrestris* erythrocruorin

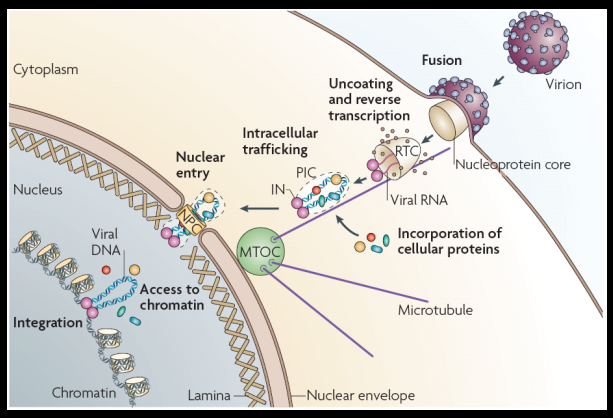
From Schatz 1992

A problem that we have to deal with all the time: images of two particles are different, are they from different type of particles or from different views of the same particle?

In principle, common line approach should give us an answer. But in real world, it may not be.

IN project

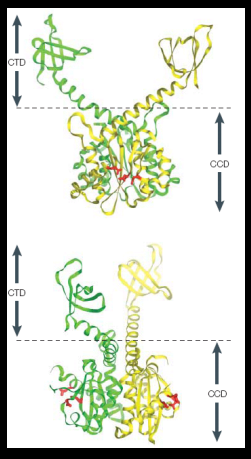
Early stage of retroviral infection: integration into host chromatin

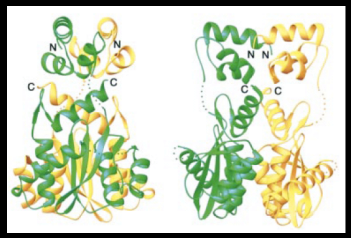


- Shenping Wu, (Angela Lai), Akram Alian, Sarah Griner and Bob Stroud

Suzuki, Y. and Craigie, R. 2007. *Nature Reviews Microbiology.*

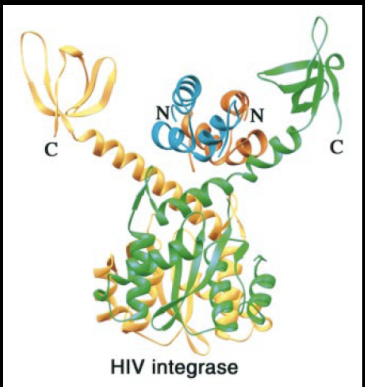
Crystal structure of integrase dimer





Cheng, J.C.H., et al. 2000. *PNAS.*
Wang, J.Y., et al. 2001. *EMBO J.*

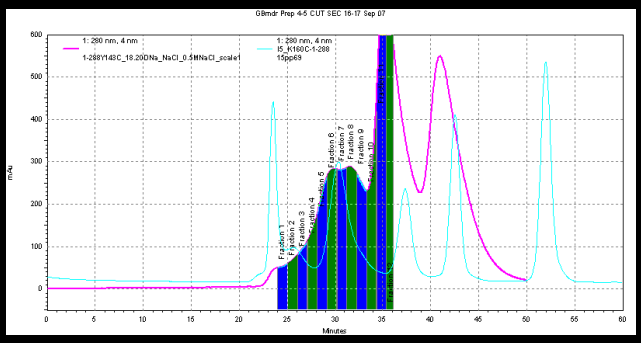
Superposition of the two structures generates the model of the full length integrase dimer



Wang, J.Y., et al. 2001. *EMBO J.*

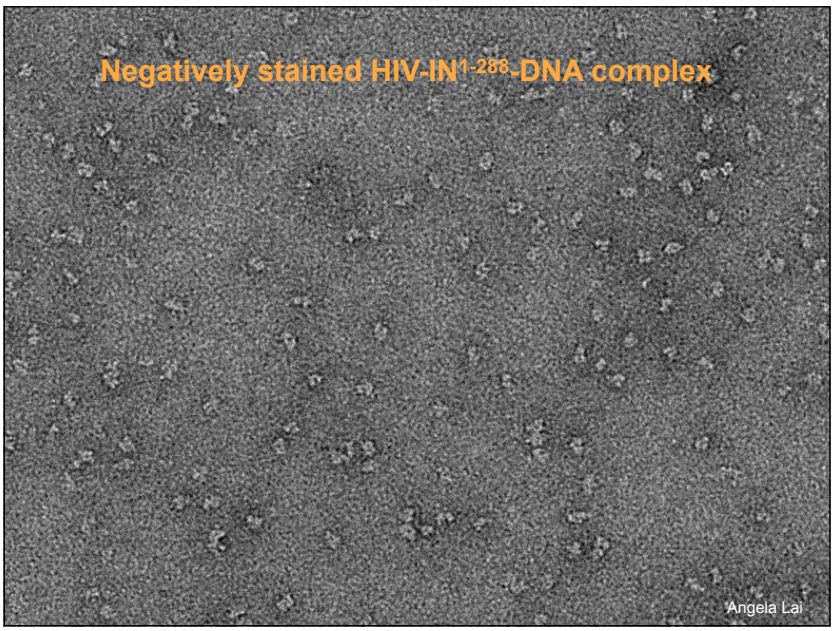
Our goal

To obtain a structure of tetramer full length integrase, in complex with DNA, by single particle EM.



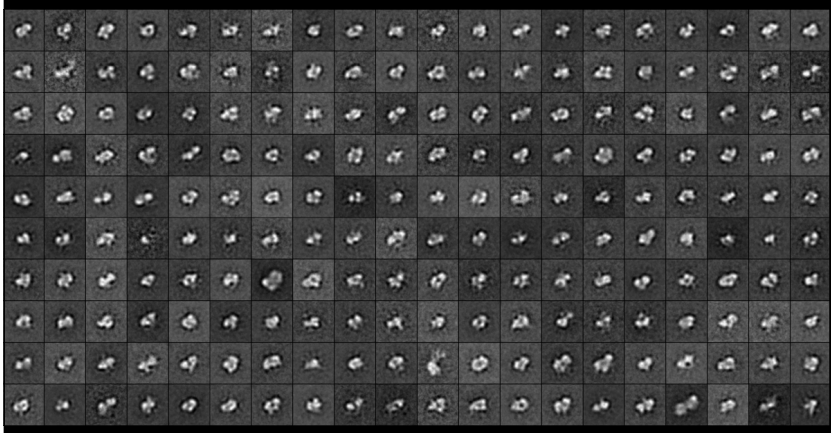
HIV-IN$^{1\text{-}288}$-DNA Purification Chromatograph, Sarah

Negatively stained HIV-IN$^{1\text{-}288}$-DNA complex



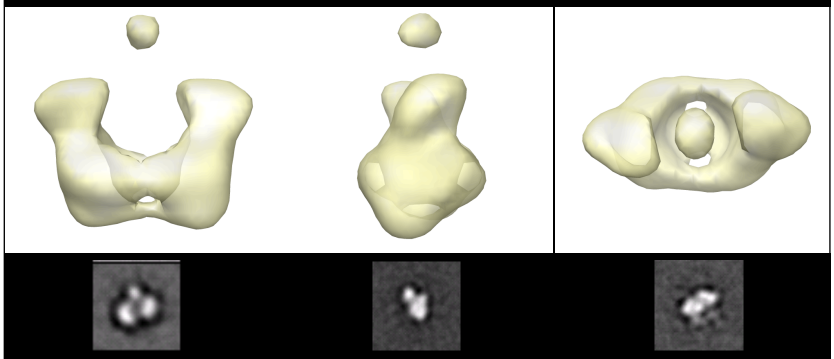
Angela Lai

Class averages of integrase particle



Angela Lai

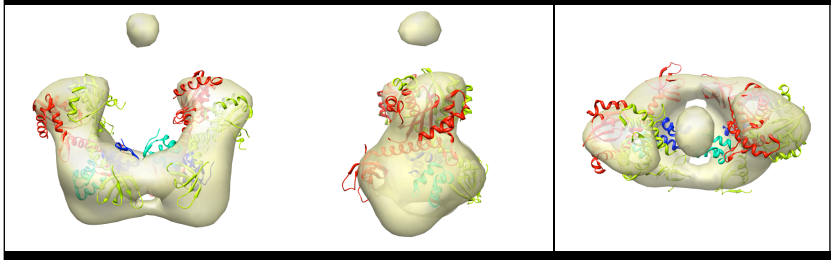
3D reconstruction of Integrase



To obtain this structure, we have to make a assumption that all particles we selected are integrase tetramers.

Angela Lai

Placing atomic structure into 3D reconstruction



By placing atomic models of integrase into 3D reconstruction, we generated a model.

* Problem: the size exclusion column cannot really separate tetramer and dimer, and we are not sure if the particles contains DNA or not.

Angela Lai

Technical questions to think about:

* procedure of calculating the 3D reconstruction
* how to determine which protein subunit goes to which domain?
* how to dock the atom structure into a low resolution structure?