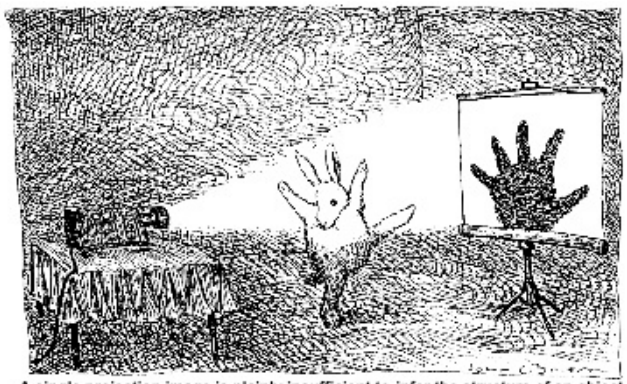
II: Single particle cryoEM - an averaging technique

Radiation damage limits the total electron dose that can be used to image biological sample. Thus, images of frozen hydrated macromolecules are very noisy, with extremely low signal-to-noise ratio (SNR).

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.

An image is the projection of a 3D object



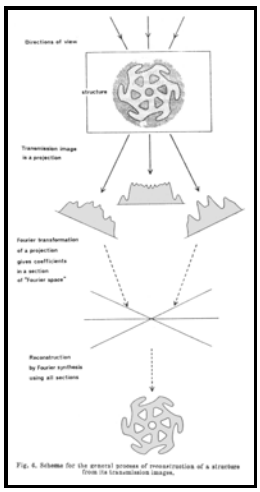
A single projection image is plainly insufficient to infer the structure of an object.
John O'Brien; © 1991 The New Yorker Magazine

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)

Cryo-EM techniques

• Electron crystallography: for membrane protein resolution achieved: 2.5Å bacteriorhodopsin, Aquaporin-0 (recently: 1.9Å aquaporin-0) averaging of molecules which form 2D crystal;

• Single particle cryo-electron microscopy: protein in soluble form; resolution achieved: intermediate 20 Å ~ 3 Å averaging of many single molecules with the same structure;

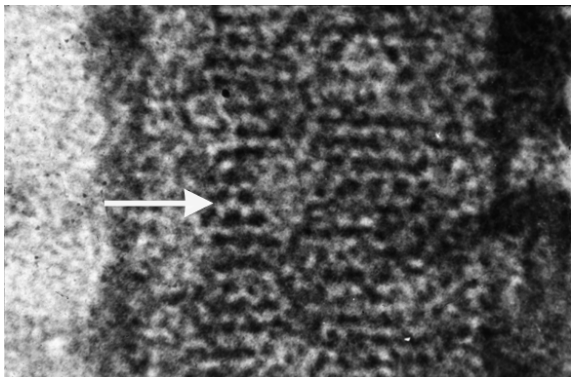
• cryo-electron tomography: large cellular organelle, whole cell etc. Resolution achieved: > 3nm no averaging, 3D reconstruction of single biological object;

Reading list:

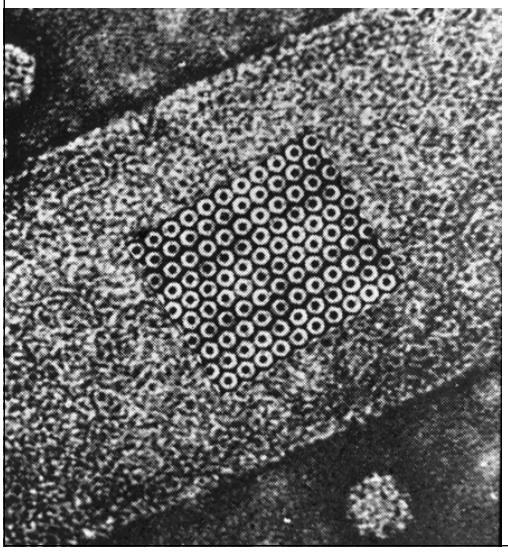
R. Henderson: “The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules”. (1995) Quarterly Reviews of Biophysics, **28**, 171-193

Image averaging

Cryo-EM images are very noisy; have extremely low signal-to-noise ratio. Averaging of a large number of images are necessary to improve the SNR.



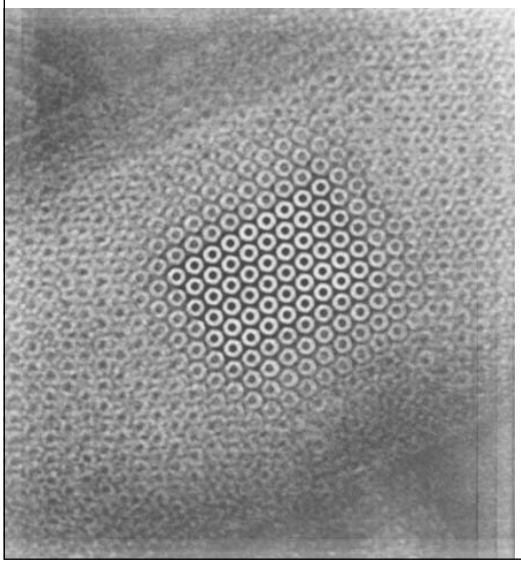
Averaging in darkroom



Photographic image superposition (averaging) by Roy Markham, who shifted image and added to the original in darkroom.

The trick is to know decide much and which direction to shift the image for superposition.

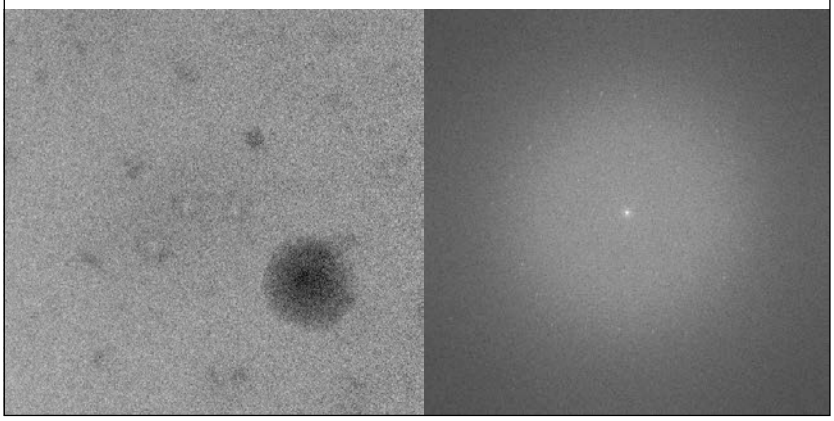
Averaging in computer.



David DeRosier used Markham's lattice to determine how much to shift, and performed averaging by using Adobe Photoshop.

Averaging in 2D crystals

How much and which direction to ship the image can be determined easily from FT of the image of a 2D crystal.



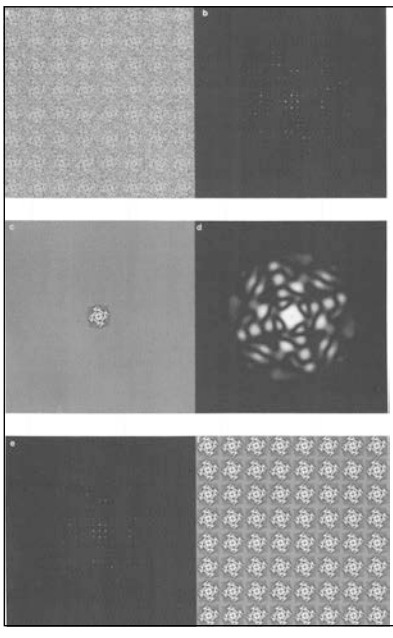


Image averaging in 2D crystal

In 2D crystal, one can extract amplitudes and phases from peaks of FT (contributed by the identical repeats of structural motif) and ignore everything in between peaks (contributed by the random noise). A reverse Fourier Transform using extracted amplitude and phases will give us an averaged features. This is equivalent to the averaging.

It is easy to perform averaging in 2D crystal. The molecules in the 2D crystal are identical in composition and orientation.

Single particle 3D reconstruction

Single particle 3D reconstruction is a technique based on averaging. Many images of the same molecule at random orientations are needed. Every individual image is very noisy with unknown orientation.

* to provide sufficient views covering entire Fourier space;
* to improve signal-to-noisy ration (SNR);
* all images needed to be aligned with each other;

Therefore: the resolution of a 3D reconstruction is dependent on:

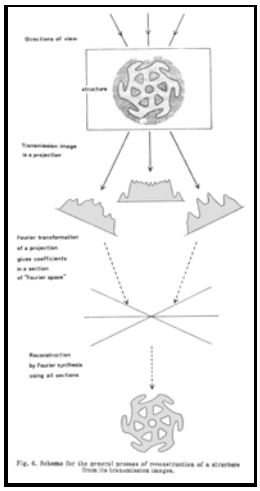
* resolution of individual images;
* total number of images;
* accuracy of alignment;
* complete coverage of Fourier space;

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

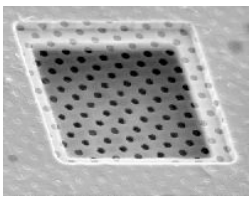
Hart, R.G. (1968) "Electron microscopy of unstained biological material: the polytropic montage" *Science* ***159*** 1464-1467



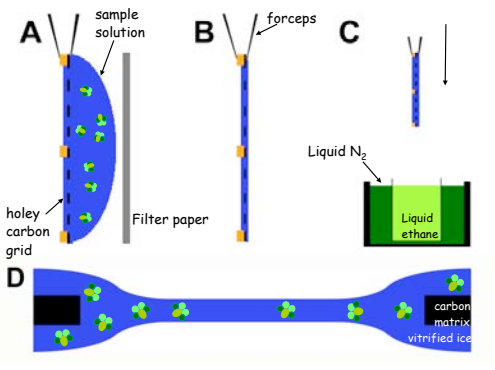
DeRoiser and Klug (1968)

Frozen hydrated specimen preparation

Adrian M, Dubochet J, Lepault J & McDowall AW (1984) Cryo-electron microscopy of viruses. *Nature* **308**, 32-36.

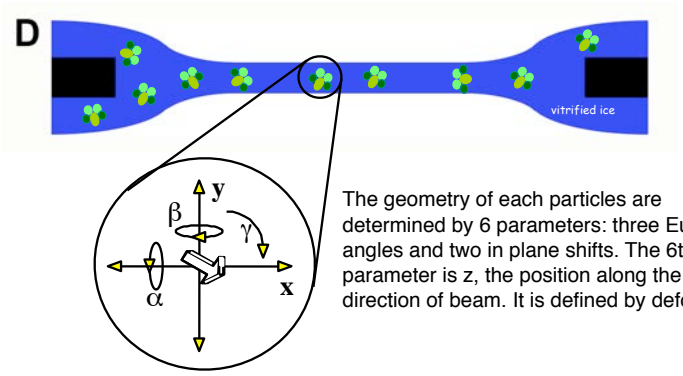


Quantifoil grid



Plunge freezing

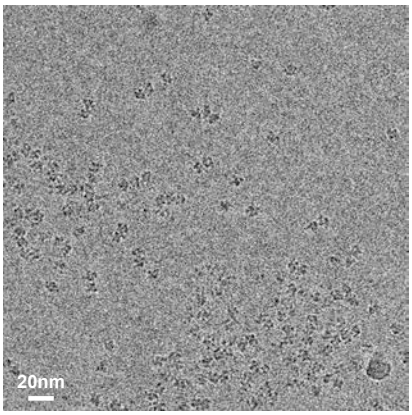
Protein molecules embedded in vitrified ice as single particles



The geometry of each particles are determined by 6 parameters: three Euler angles and two in plane shifts. The 6th parameter is z, the position along the direction of beam. It is defined by defocus.

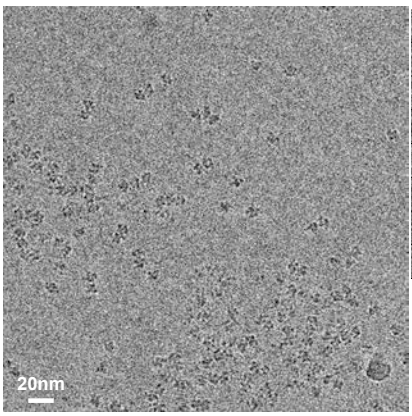
Single particles are randomly oriented in vitreous ice

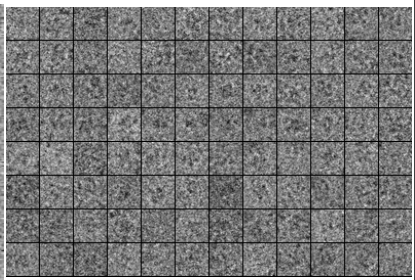
Image of frozen hydrated TfR-Tf complex



* Each micrograph contains images of many particles in random orientations;

Image of frozen hydrated TfR-Tf complex

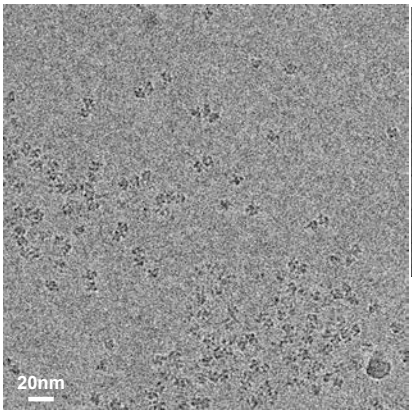


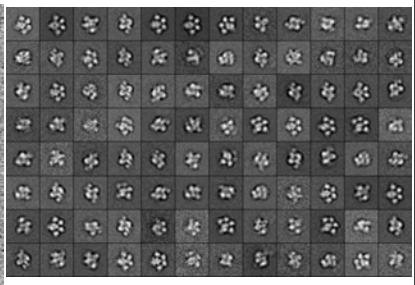


* Each micrograph contains images of many particles in random orientations;

* These particles are computationally selected and boxed out, but individual particle is very noisy, with very low signal-to-noise ratio (SNR);

Image of frozen hydrated TfR-Tf complex



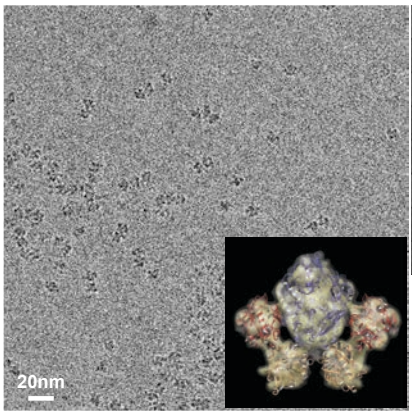


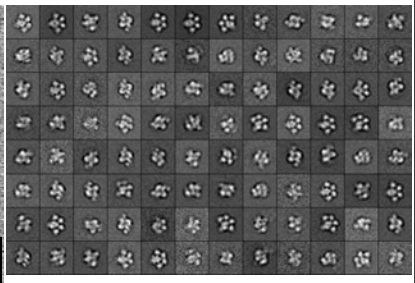
* Each micrograph contains images of many particles in random orientations;

* These particles are computationally selected and boxed out, but individual particle is very noisy, with very low signal-to-noise ratio (SNR);

* Averaging images of identical particles enhances SNR;

3D reconstruction is determined by averaging many individual images at different orientations





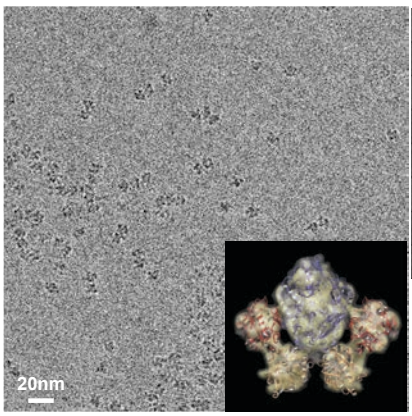
* Each micrograph contains images of many particles in random orientations;

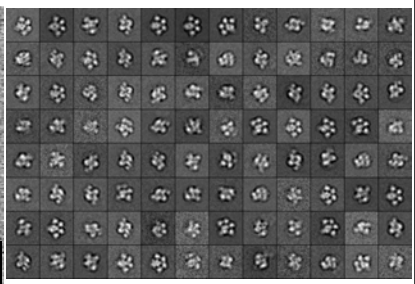
* These particles are computationally selected and boxed out, but individual particle is very noisy, with very low signal-to-noise ratio (SNR);

* Averaging images of identical particles enhances SNR;

Analogous to X-ray crystallography:
* averaging by computation instead of by crystallization;
* accuracy of alignment is similar as order of packing molecules in crystals;
* number of particle images is similar as the size of the crystals;

3D reconstruction is determined by averaging many individual images at different orientations





* Each micrograph contains images of many particles in random orientations;

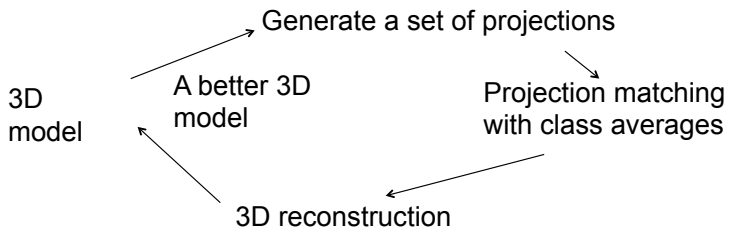
* These particles are computationally selected and boxed out, but individual particle is very noisy, with very low signal-to-noise ratio (SNR);

* Averaging images of identical particles enhances SNR;

Analogous to X-ray crystallography:
* averaging by computation instead of by crystallization;
* accuracy of alignment is similar as order of packing molecules in crystals;
* number of particle images is similar as the size of the crystals;

Iterative refinement procedure

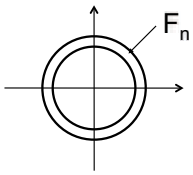
Iterative refinement procedure, using reference model based projection matching:



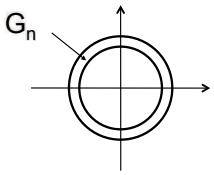
A least square approach to find the best solution that matches all data.

Resolution estimation

In single particle cryoEM the resolution is often estimated by Fourier Shell Correlation.

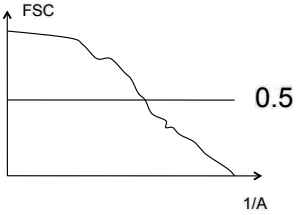


Reconstruction 1



Reconstruction 2

$$FSC(R) = \frac{\sum_{n \in R} F_n G_n}{\left\{ \sum_{n \in R} |F_n|^2 \sum_{n \in R} |G_n|^2 \right\}^{1/2}}$$



Resolution criterion

* What is the criterion to estimate resolution?

"Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy", Peter Rosenthal and Richard Henderson, JMB, 2003, **333**, 721-745.

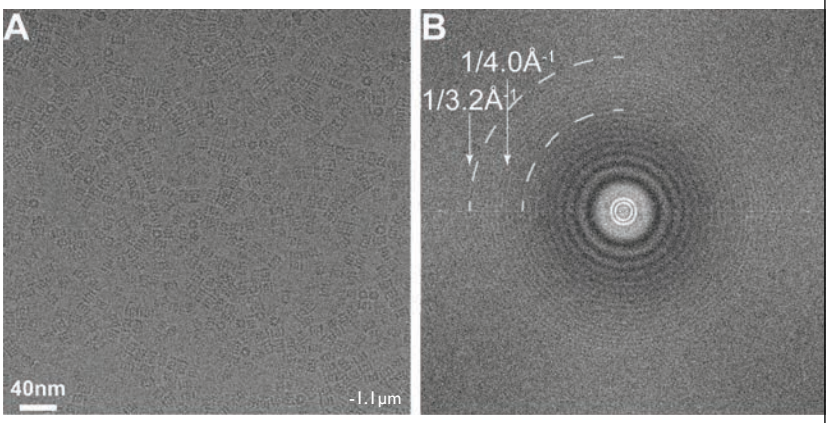
* How to prevent model induce bias and overfitting?

"Prevention of overfitting in cryo-EM structure determination", Sjors Scheres and Shaoxia Chen, Nature Methods, 2012, **9**, 853-854.

Resolution determinants

* Image quality - sufficient image contrast and sufficient high-frequency SNR;
* Alignment accuracy -
* Number of particles -
* Homogeneity - both conformational and compositional

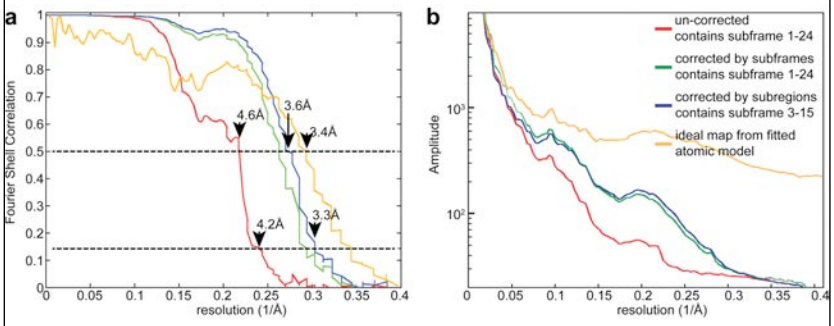
K2 image of frozen hydrated protein samples, T20S



* 300kV, 31kX mag, ~10e-/pixel/sec; ~1.2Å/pixel, 25e-/Å2, 3.5sec exposure;

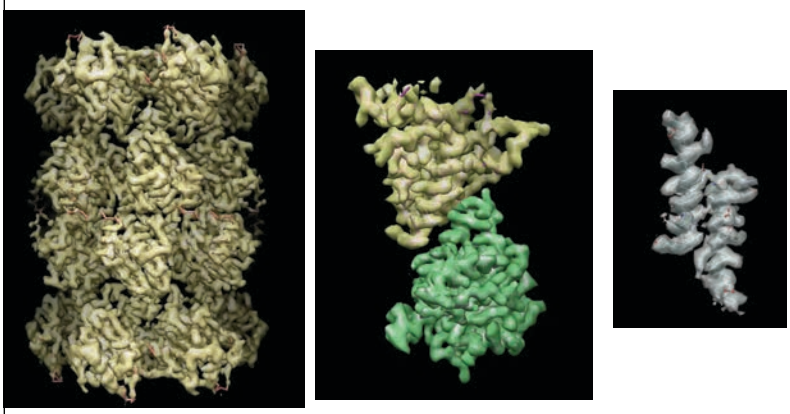
Xueming Li, Kiyoshi Egami

3D reconstruction of T20S proteasome



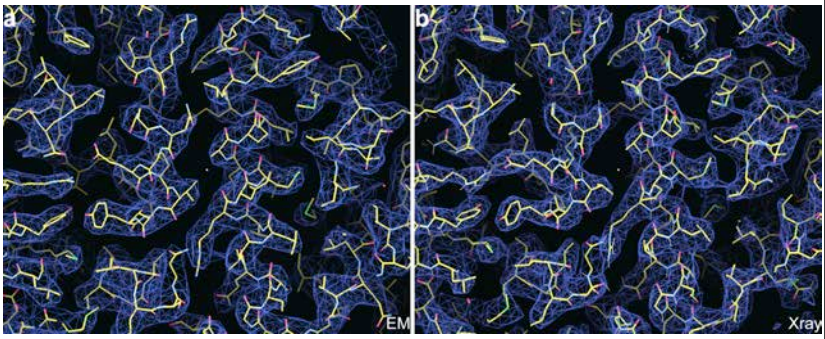
Fourier Shell Correlation curves and amplitude plot of T20S proteasome reconstruction

3D reconstruction of T20S proteasome



* We determined a 3D reconstruction of archaeal 20S proteasome to the resolution of ~3.3 Å, comparable to the resolution of X-ray crystal structure, 3.4Å.

3D reconstruction of T20S proteasome



Coulomb potential density map
(density map)

Electron density map
(density map)

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

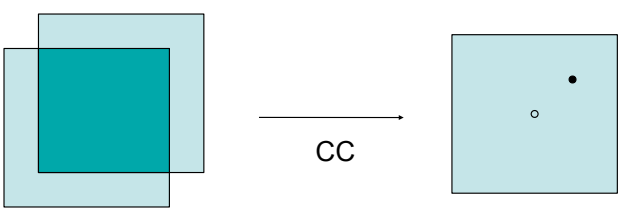
Note that: $-1 < \rho < 1$

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?

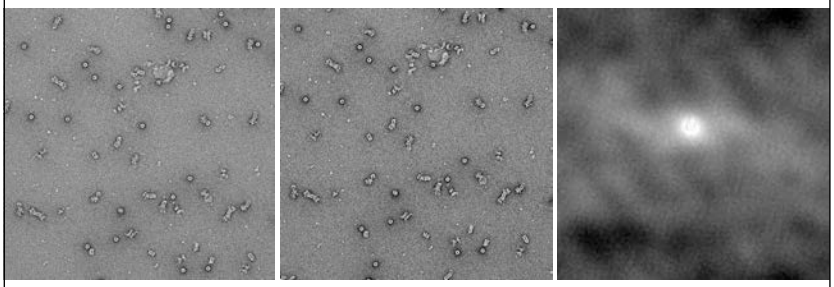
Calculating the cross-correlation

Cross-correlation theorem:

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

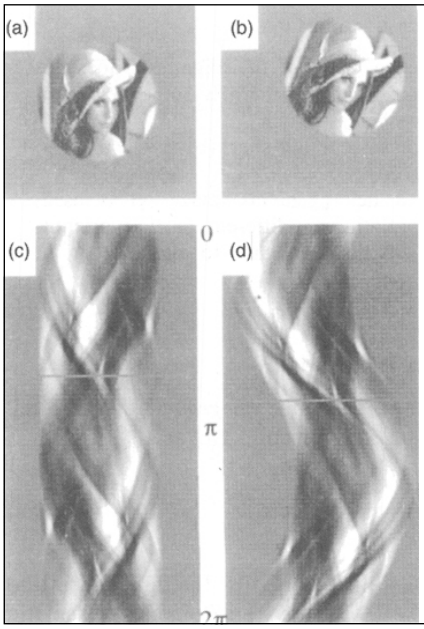
This formula enable us to calculate the cross-correlation between two images easily.

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

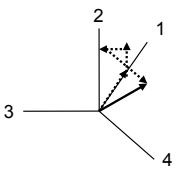
Classification - a process of dividing a set of images into subsets with similar features.

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



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

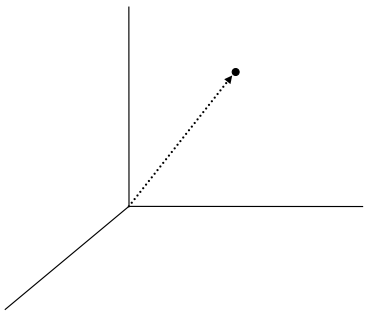
$$\vec{a}_i \perp \vec{a}_j (j \neq i; j = 1,..m);$$

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.

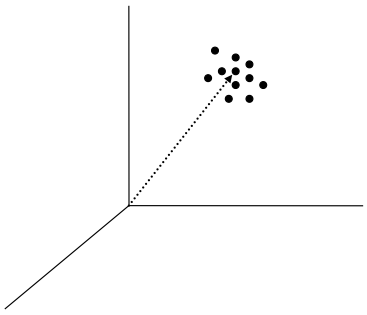
An image without any noise is represented by a point.



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.

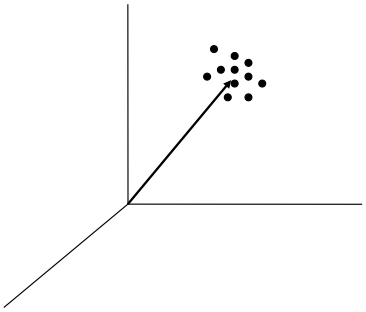
An image without any noise is represented by a point.

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



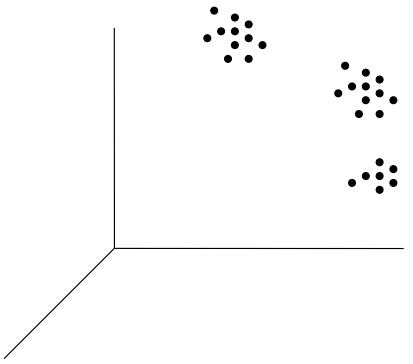
A data set is represented as a cloud in the hyperspace. The center of the cloud is the average of the all images in the data set.

The center of the loud is the average.



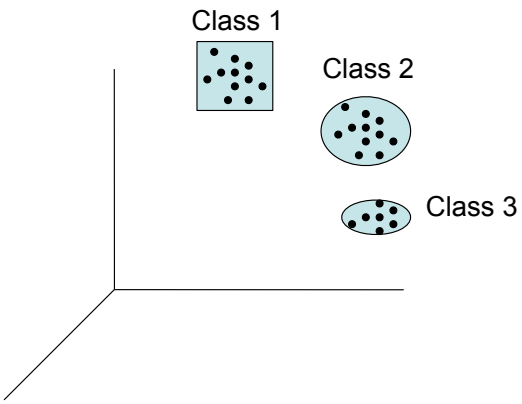
Classification

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



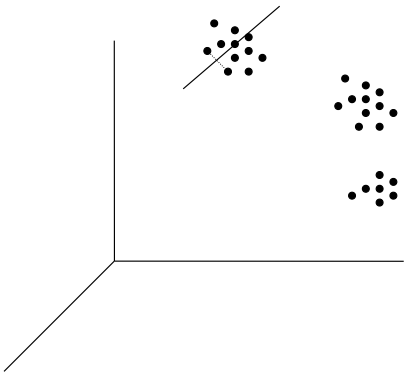
Classification

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



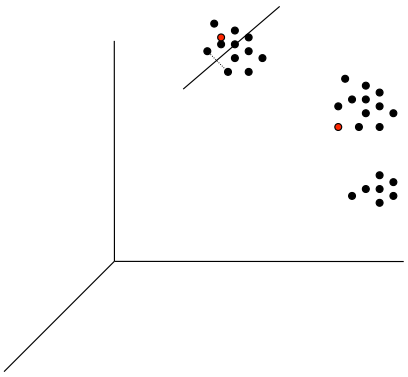
Classification

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



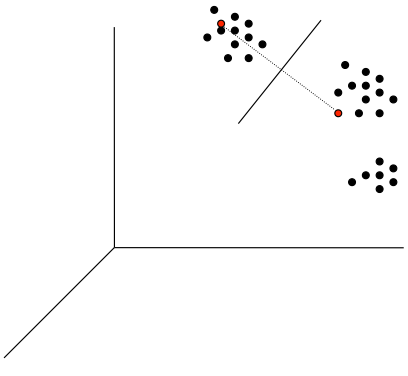
Classification

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



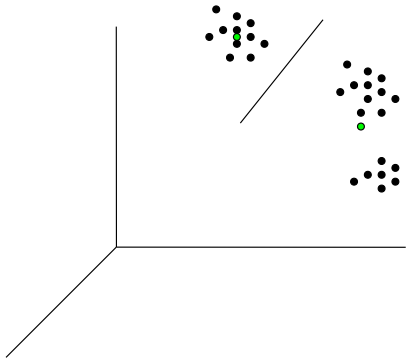
Classification

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



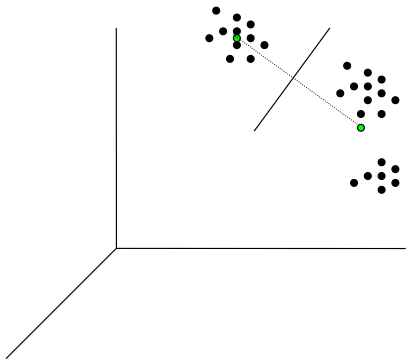
Classification

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



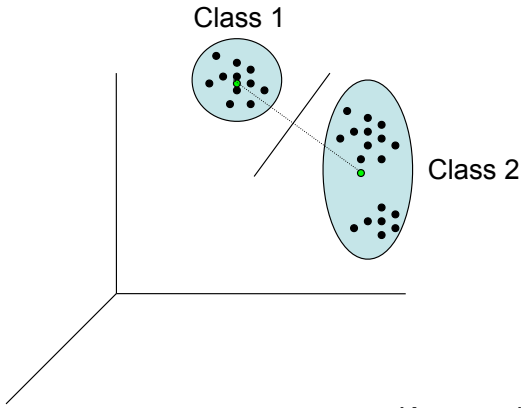
Classification

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

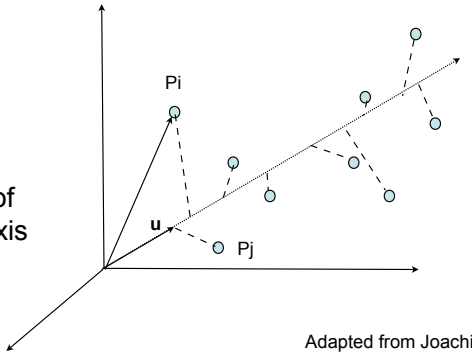
Multivariate statistical analysis

Making patterns emerge from data

Multivariate statistical analysis:

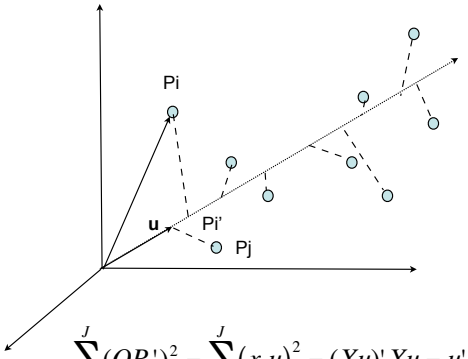
- Principal Component Analysis
- Correspondence Analysis

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

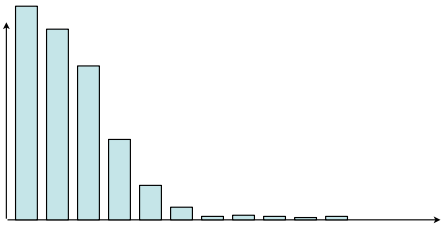
Eigenvector-eigenvalue equation

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

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

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

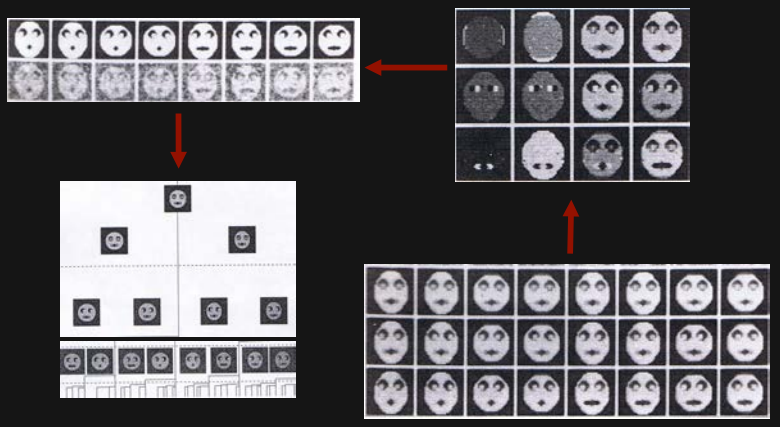
Significant factors:



Classification based on eigenvector/eigenvalue clustering;

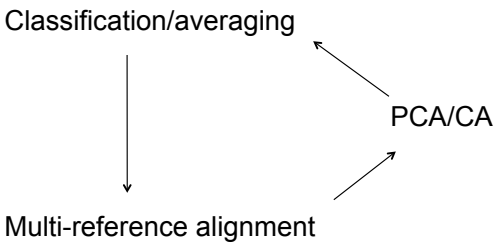
Multivariate statistical analysis & classification of images

Principle Component Analysis



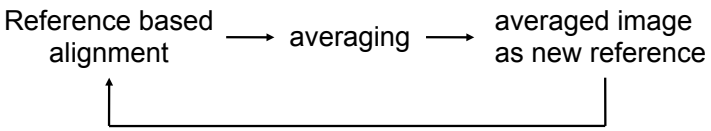
Iterative MRA and classification

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?

Maximum Likelihood approach

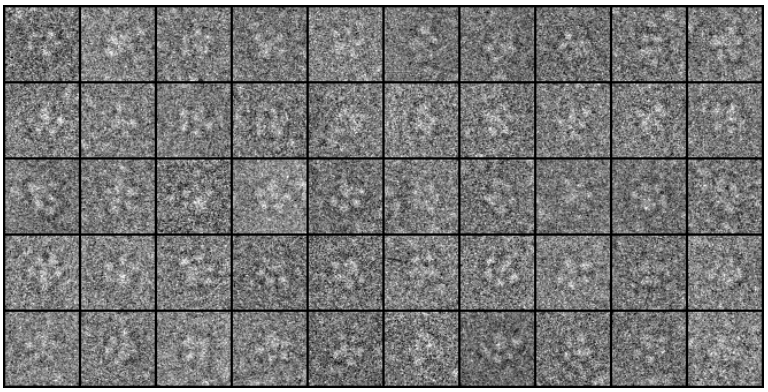
The iterative refinement procedure based on cross correlation is equivalent to a least square optimization procedure.

* Maximum Likelihood approach:
Given a set of images X, we would like to maximize is the probability $P(\Theta|X)$ that this model Θ is the correct one.

Sigworth, et al. “An introduction to maximum-likelihood methods in cryo-EM” Method in Enzymology, Cryo-EM, part C.

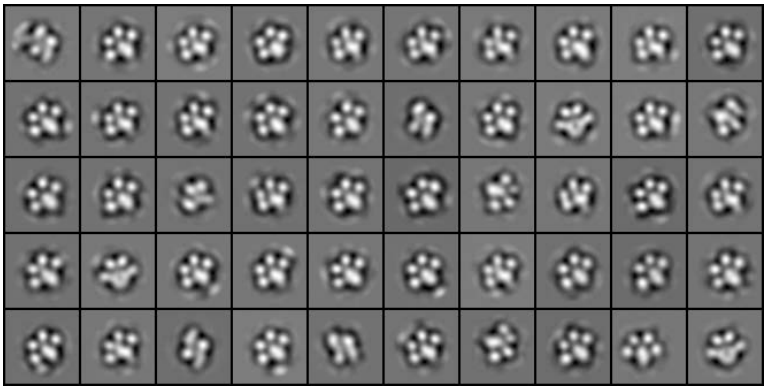
Maximum likelihood algorithm is now implemented in a number of programs, including RELION, XIMP, FREALING, etc.

Individual TfR-Tf Complexes in Vitrified Ice



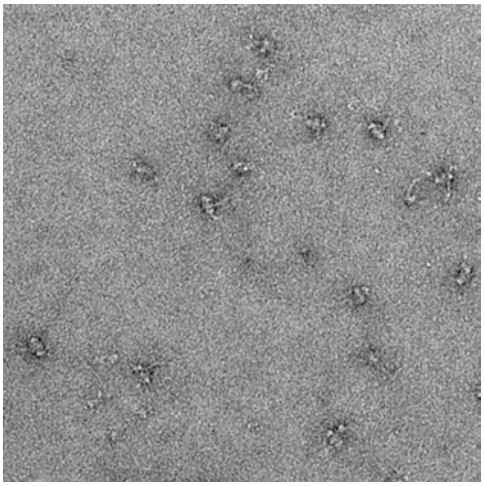
50 out of 36,266 particles

Class Averages of Vitrified TfR-Tf Complexes

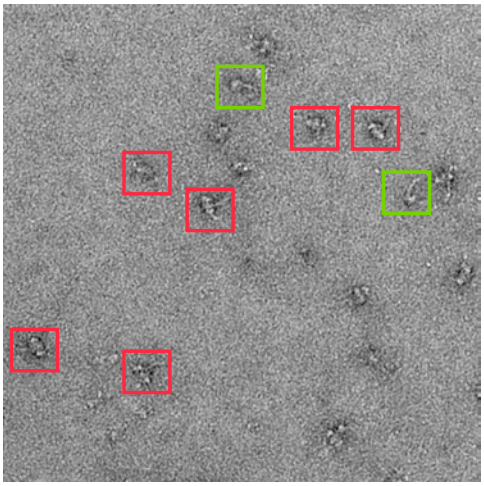


50 out of 200 classes

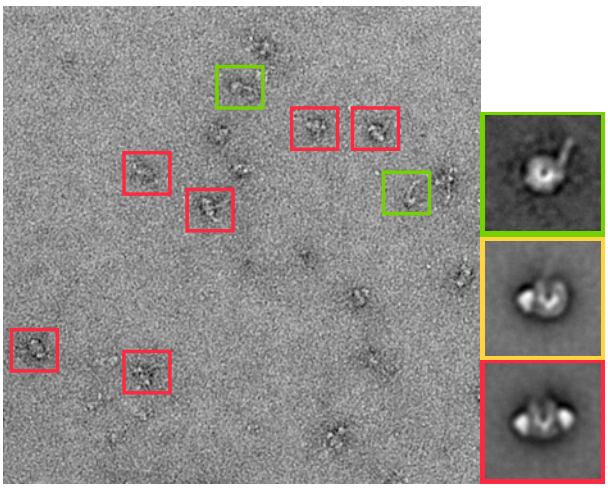
SNF2h-nucleosome complex



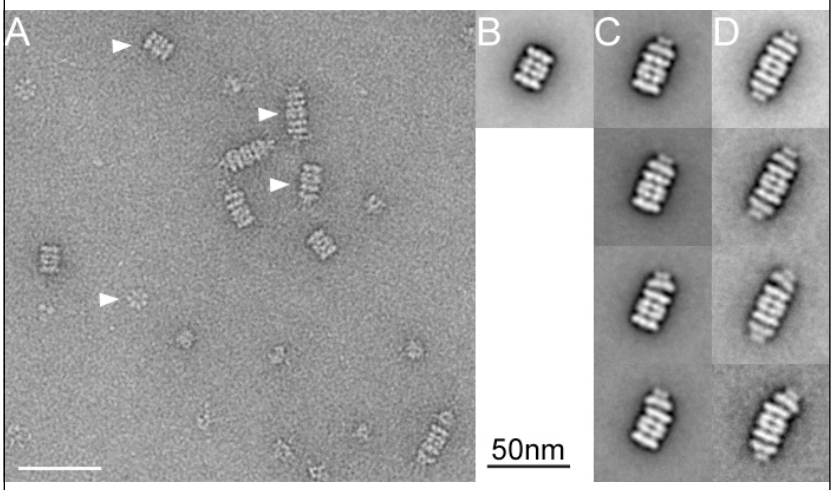
SNF2h-nucleosome complex



SNF2h-nucleosome complex

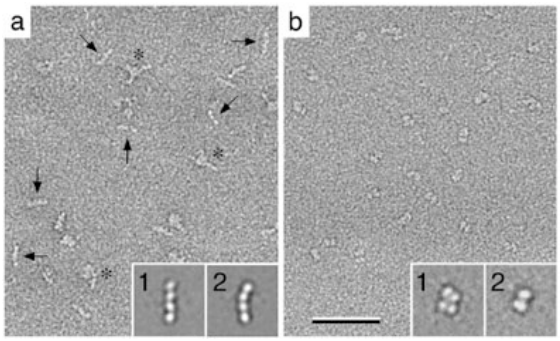


A simple case of using image alignment and classification

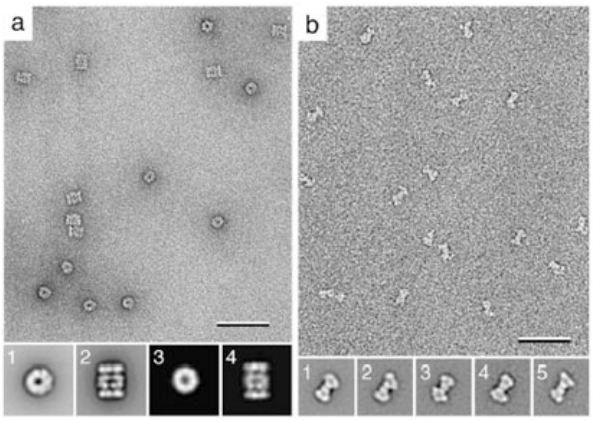


20S-PAN complex

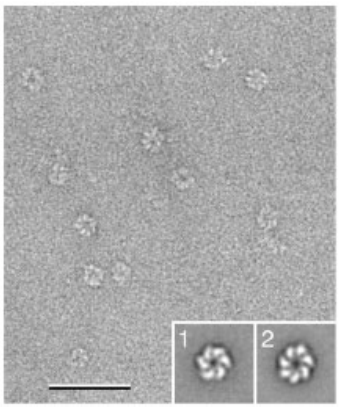
A number of examples



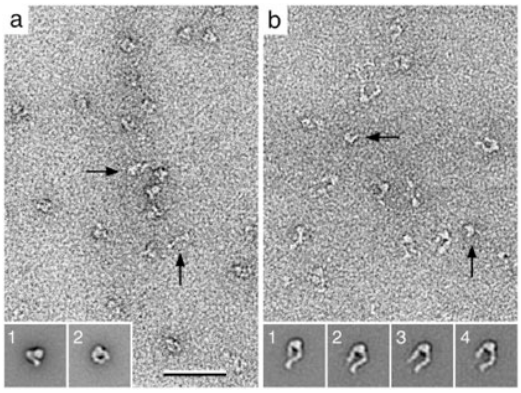
A: integrin $\alpha_5\beta_1$ headpieces and a fibonectin fragment (~40kDa). B: human transferrin (~70kDa, each domain is about ~17kDa).



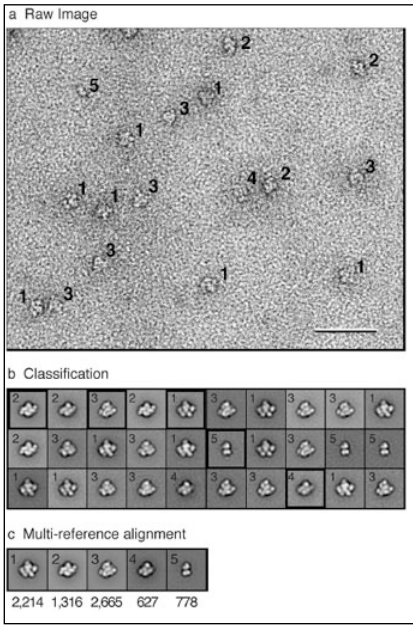
A) Yeast 20S proteasome; B) yeast Sec23p/Sec24p complex;



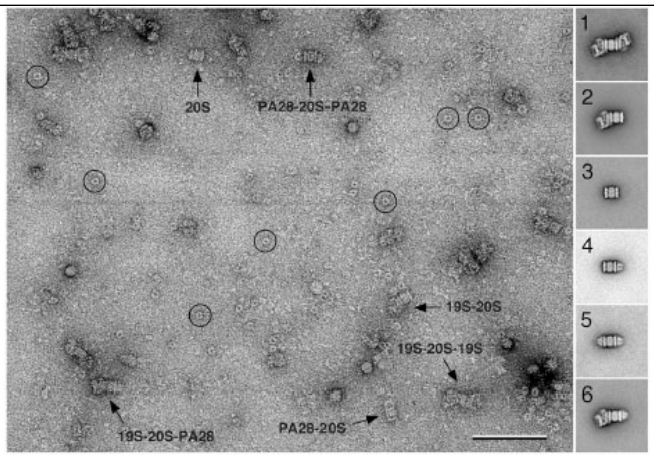
Heterogeneous population of bacteriophage T7 helicase/primase;



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



Human transferrin receptor-transferrin complex



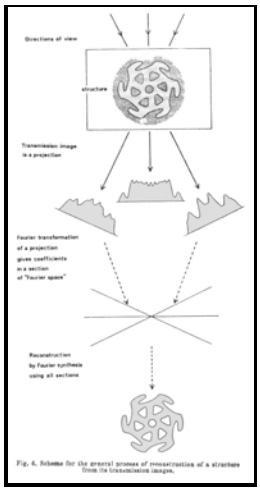
A heterogeneous sample of 20S proteasome, 19S regulator complex and PA26 activator.

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

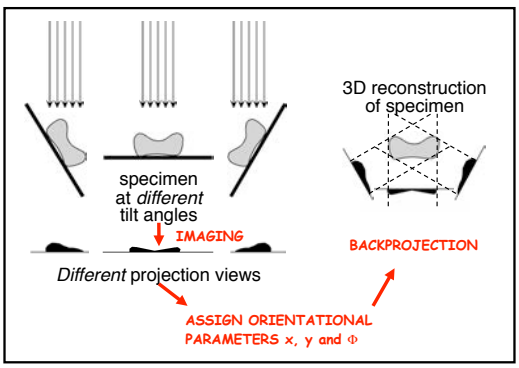


DeRoiser and Klug (1968)

3D reconstruction

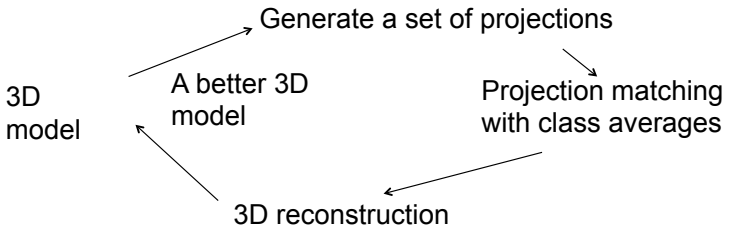
Assume we have already a number of class averages, they represent the projections of a 3D object in different orientations. And we know (can determine) these relative orientations of each class averages. We can reconstruct the 3D object - 3D reconstruction.

Back projection:



How to determine the relative orientations?

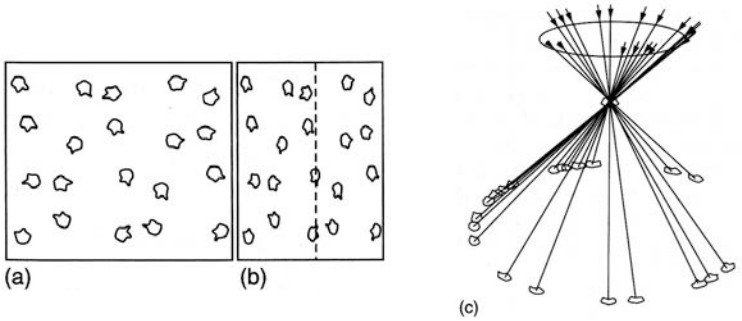
Model based projection matching: by matching (ccc) of class averages with the calculated projections of the 3D object in known directions - a refinement procedure.



Question: Where do you get the first model?

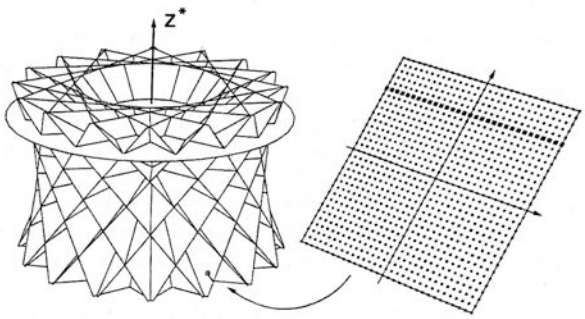
Random conical tilt

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

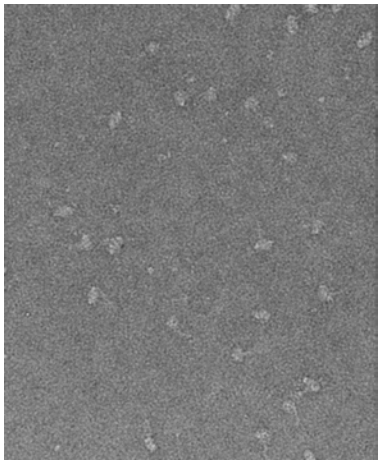


From Joachim Frank

Random conical tilt

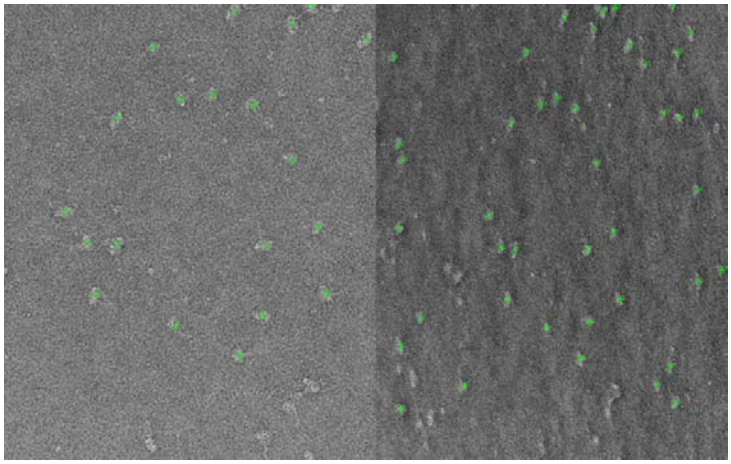


Random conical tilt



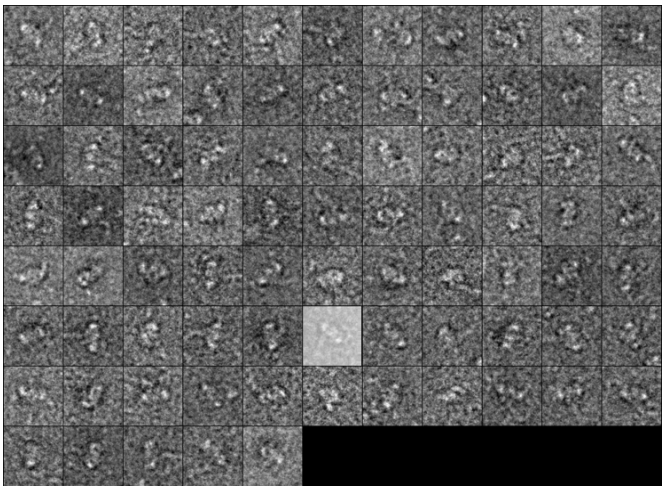
untilted image

Random conical tilt



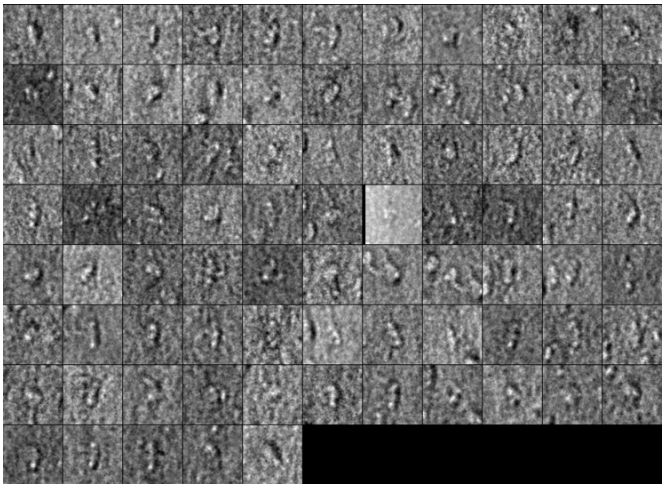
untilted image tilted image

Random conical tilt



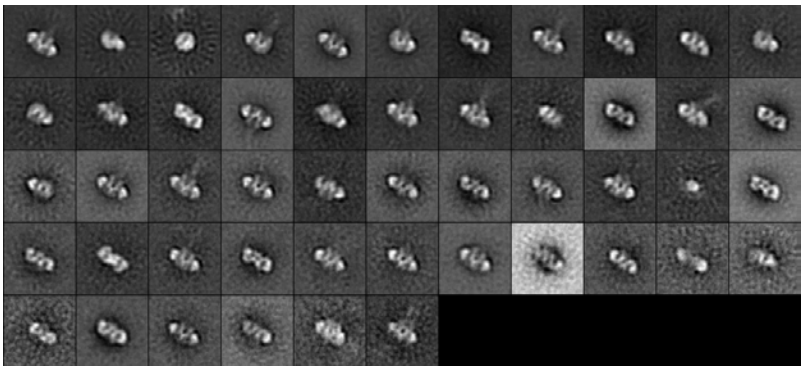
particles from untilted images

Random conical tilt



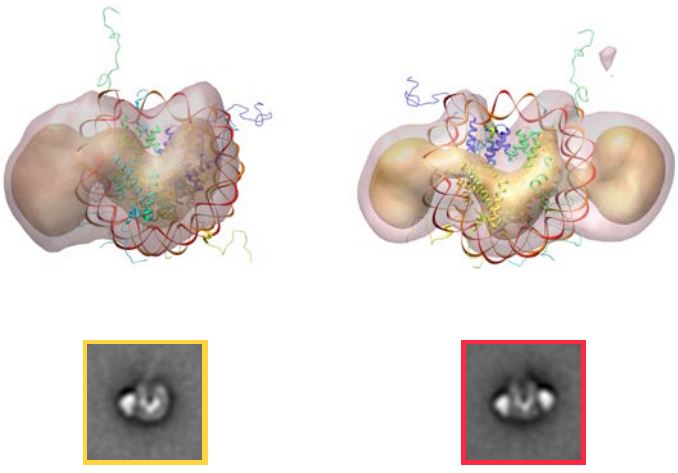
particles from tilted images

Random conical tilt

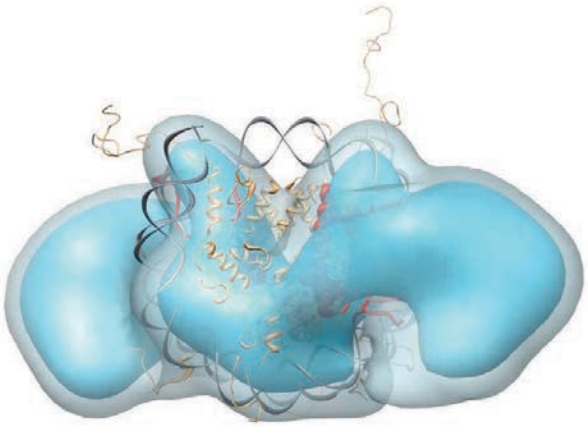


class averages of untilted images

Random conical tilt 3D reconstruction

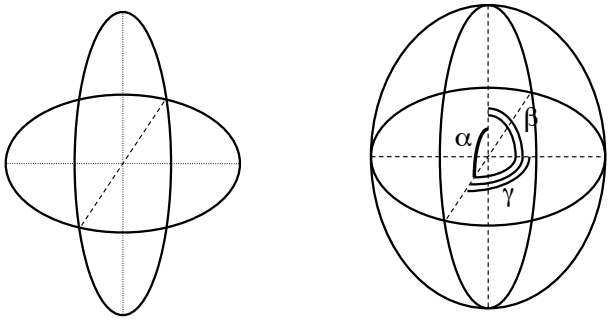


Nucleosome-SNF2h complex

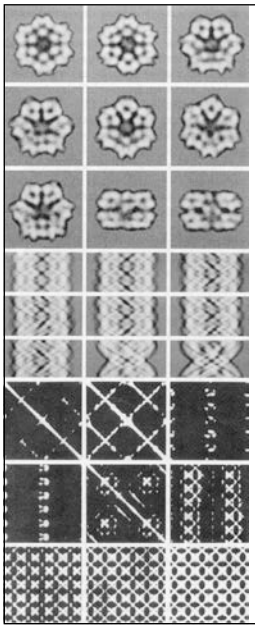


Common line/angular reconstitution

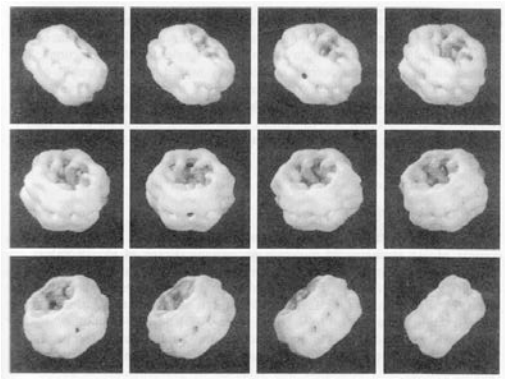
Any two central sections in the 3D Fourier space across each other will have a common line.



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



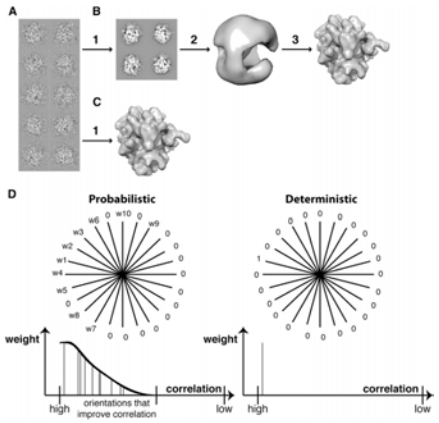
Common line/angular reconstitution



3D reconstruction of frozen hydrated *Lumbricus terrestris* erythrocruorin

From Schatz 1992

Probabilistic initial 3D model generation



Hans Elmlund, et al. (2013) Structure, 21, 1299-1306.