Lecture 7 **MSA and ML classification**

Dan Clare

Multidimensional space
Statistical analysis
Eigen vectors
Cluster analysis: Classification
MSA and eigen images
ML alignment and classification

Image processing for cryo microscopy

1 - 11 September 2015

FEI COMPANY EMBO Leica MICROSYSTEMS JEOL Direct Electron

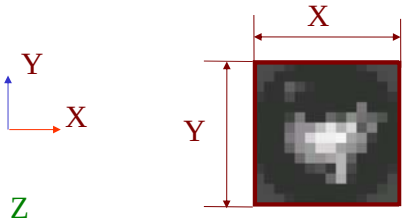
Practical Course
Birkbeck College London

Multidimensional space

One dimensional space

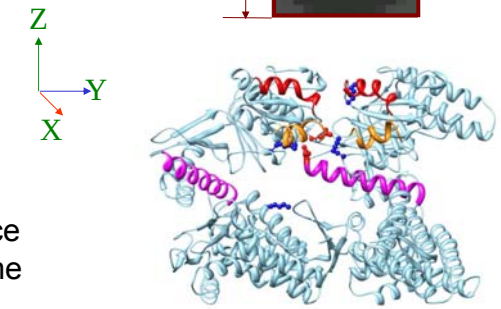
Two dimensional space

X,Y coordinates



Three dimensional space

X,Y,Z coordinates

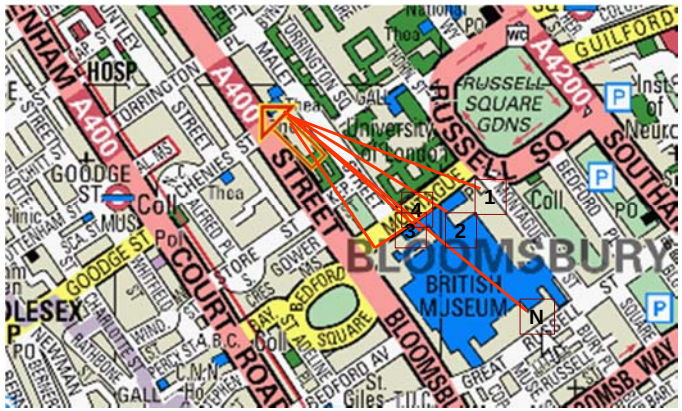


Four dimensional space
X,Y,Z coordinates + time

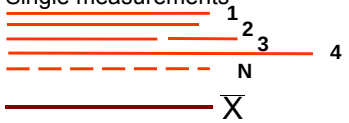
Statistical analysis

Statistical analysis

One dimensional space - ONE MEASUREMENT



Single measurements



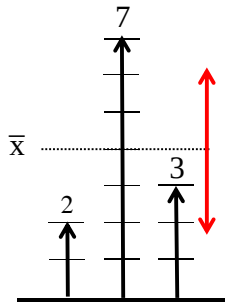
The average distance

$\overline{X} = (X_1+X_2+X_3+\ldots+X_N)/N$

Elena Orlova

Some "basic statistics"

example for 3 measurements



n = 3

$\Sigma x_i = 2+7+3 = 12$

$\overline{x} = 1/n \Sigma x_i = 1/3 (2+7+3) = 4$

$\sigma = \sqrt{1/n \Sigma (x_i - \overline{x})^2} = \sqrt{14/3} = 2.16$

standard deviation

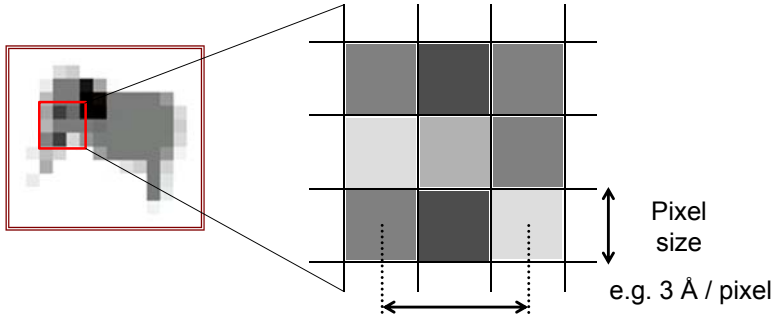
Statistical analysis

Two dimensional space - TWO MEASUREMENTS

1– the length of the fish ; 2- the weight of the fish

Statistical analysis

EM images are 2D projections of our 3D object which are composed of pixels, each of the pixels correspond to a grey level and therefore each pixel represents an individual measurement.



Minimum resolvable distance = 2 x pixel size (6 Å)

So for each EM image there are box size^2 number of measurements to compare ! That is a lot of measurements

How can a large number of variables (different types measurements) be reduced to a small number of **'important'** parameters?

Principal component analysis was designed to reduce the number of variables, to find the most significant (principal) variations in the measurements.

Factor analysis is aimed to find variations in a number of original variables, using a small number of factors. Principal components can be used as the factors. In the analysis it is assumed that EACH original variable can be expressed as a SUM of the factors, taken with different coefficients. In mathematical language it is a linear combination:

$$M_n = a_{1n}F_{1n} + a_{2n}F_{2n} + a_{3n}F_{3n} + \ldots + b_n$$

After finding the principal components of the data, classification must be performed.

Cluster analysis is the identification of groups of similar objects. This type of analysis is used for the classification of images.

The most common implementations of cluster analysis in EM are:

K-means (Sparx, Spider, EMAN, Xmipp)

Hierarchical ascendant classification - HAC (Imagic, Spider)

- The aim of statistical analysis and classifications is to group images that are similar so that, when they are averaged, the signal to noise ratio is improved.

Eigen vectors

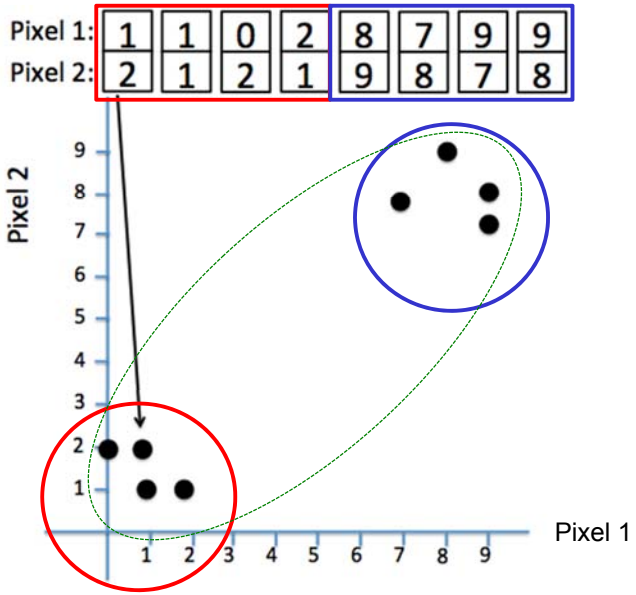
Statistical analysis of 2 pixel images

We have images that consist of
2 pixels with different grey values

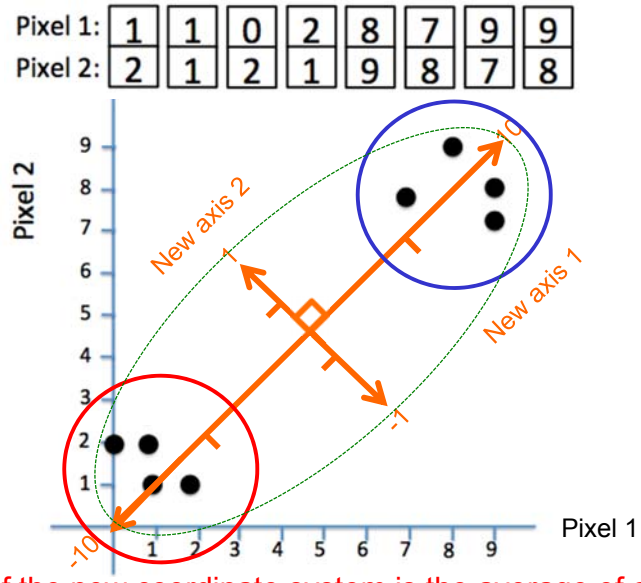
Pixel 1:	1	1	0	2	8	7	9	9
Pixel 2:	2	1	2	1	9	8	7	8

How to make a (stupid) computer do this?

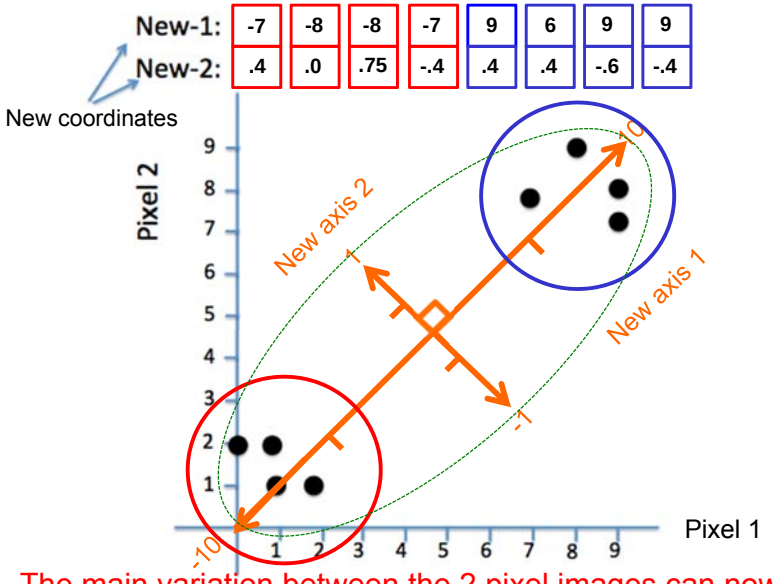
Sjors Scheres



Distance between points in hyperspace (2D) defines similarity, close points represent similar images



Origin of the new coordinate system is the average of all the images. X-axis is rotated so that it describes the most variability. Y-axis is orthogonal and describes the next most significant variation



The main variation between the 2 pixel images can now be described by the new X-axis (1st eigen vector) reducing the number of measurements required to differentiate the images.

Eigen vectors correspond to the axes of a new system of coordinates. The eigen vectors are orthogonal as X, Y, and Z are perpendicular to each other in the usual 3D system of coordinates.

In multivariate statistical analysis (MSA) eigen vectors correspond to the principal components of the measurements.

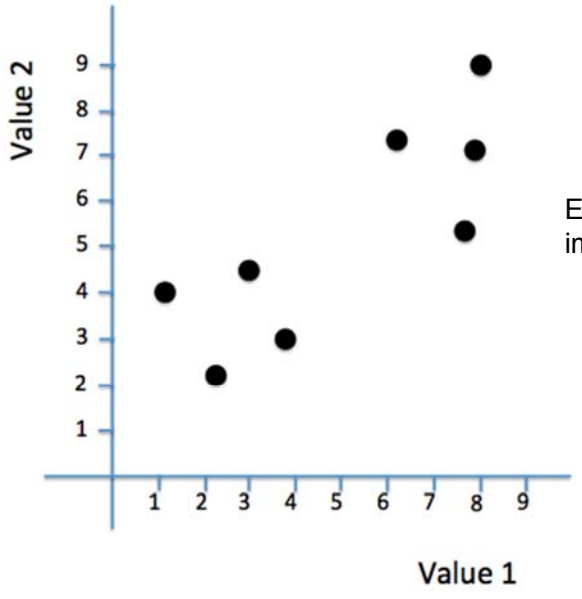
Eigen values are coefficients that describe the length of eigen vectors.

The eigen value is proportional to the variance in that direction

The aim of MSA is to describe the shape of the data cloud in hyperspace using eigen vectors.

Cluster Analysis: Classification

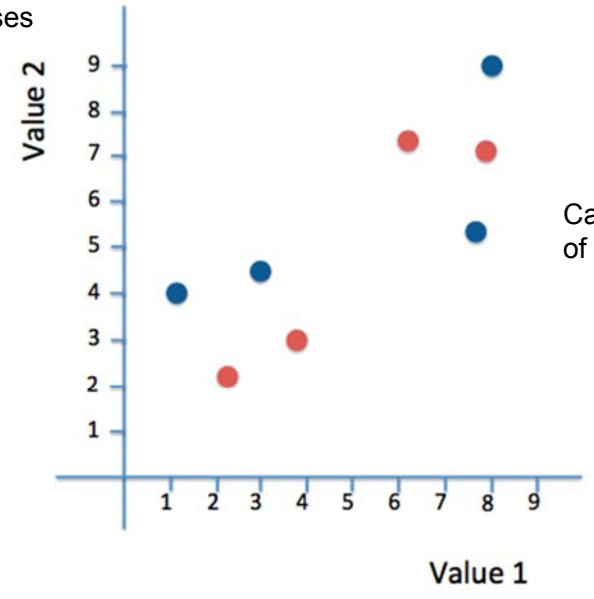
Classification: K-means



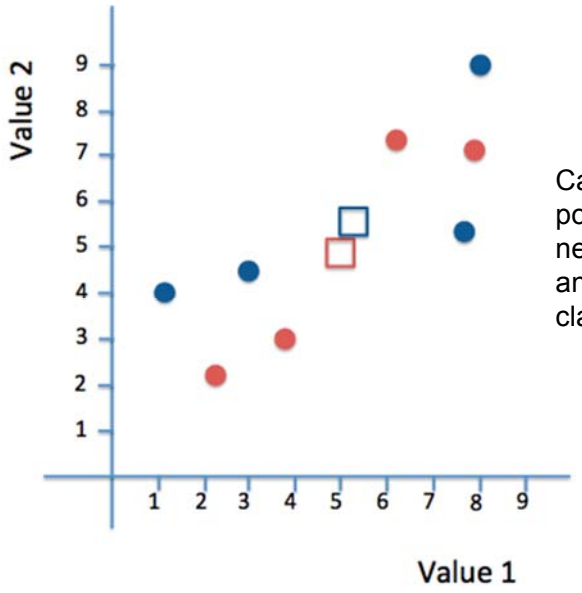
Each point is an image

Sjors Scheres

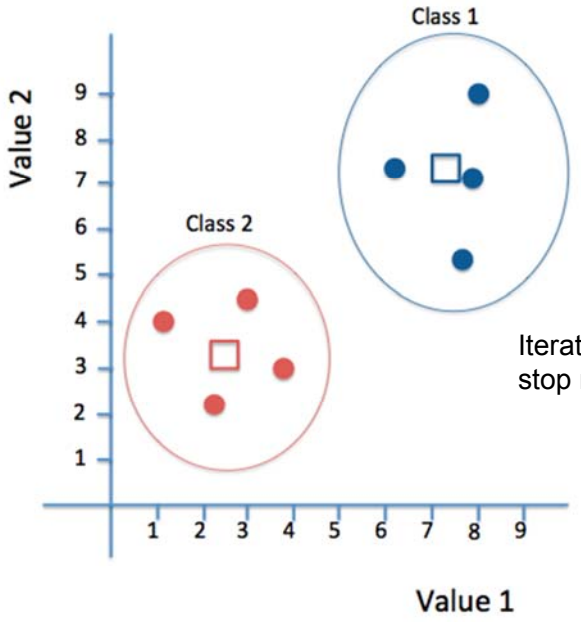
The user chooses a number of classes (K), in this example K=2, then the algorithm randomly assigns each point to one of the classes



Calculate averages of the two classes

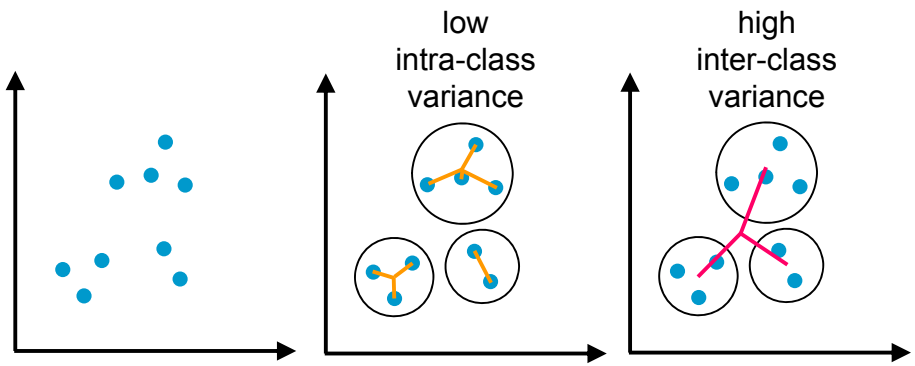


Calculate for each point which is the nearest average and recalculate class average



Iterate until points stop moving classes

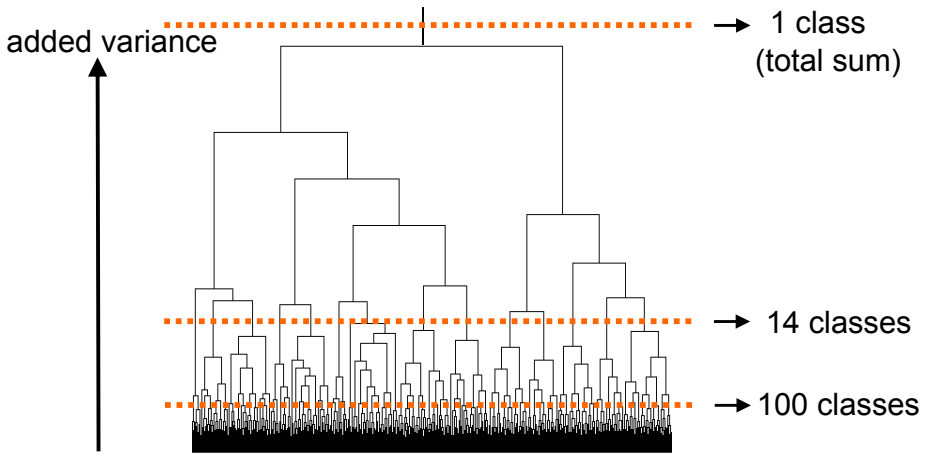
Classification: HAC



The Ward criterion (Ward 1982) is used to minimise intra-class variance while maximising inter-class variance and is the criterion used when pair-wise merging classes to form a HAC tree

Bruno Klaholz

Classification: HAC



HAC + moving elements refinement

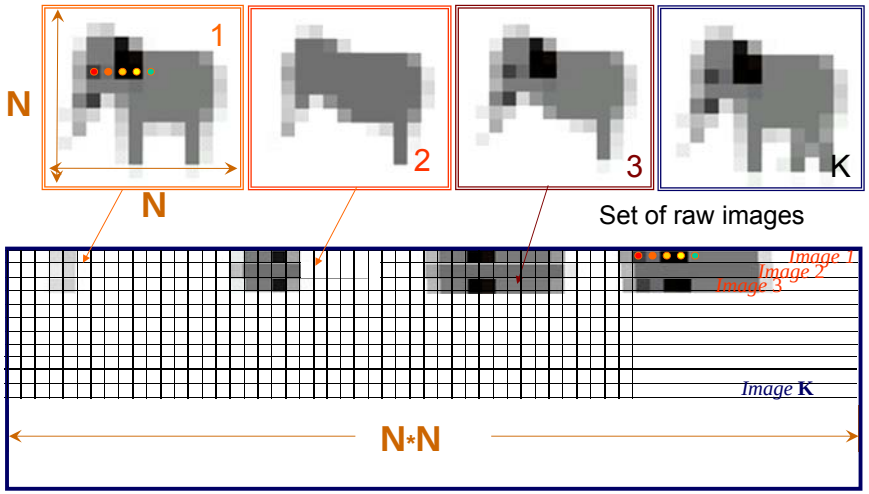
The user selects how many classes are required and the HAC tree is cut at this point

Classification

- K-means works better in low dimensional space but can also work in high dimensional space
  - Few eigen values per particle
  - Entire particle images
- K-means is reasonably fast
- Features of the K-means approach
  - Optimal solution is not guaranteed: local minima
  - Local minima more of a problem as dimensionality increases
  - Try multiple times with different random starts
- HAC is slower than K-means for large data sets
- Once images are merged they are stuck in assigned classes
  - This may not give the lowest intra-class variance
- Moving elements refinement after HAC allows images to move
  - This should give the classes the lowest intra-class variance

MSA and eigen images

Multivariate statistical analysis (MSA)



Matrix of all raw images

Elena Orlova

Set of raw (*aligned*) images

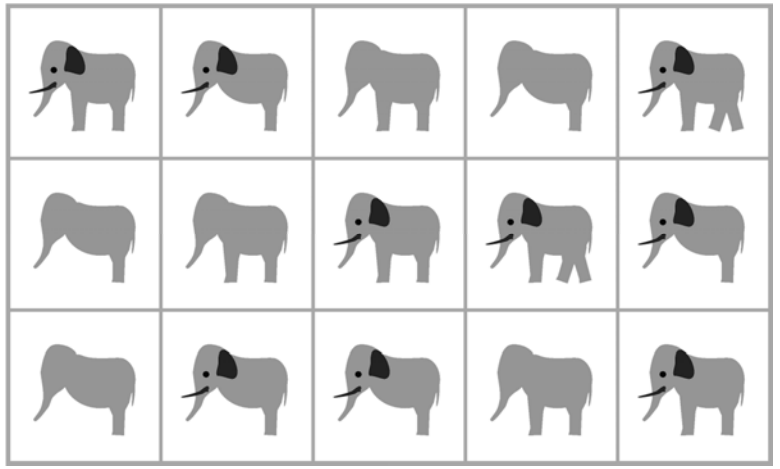
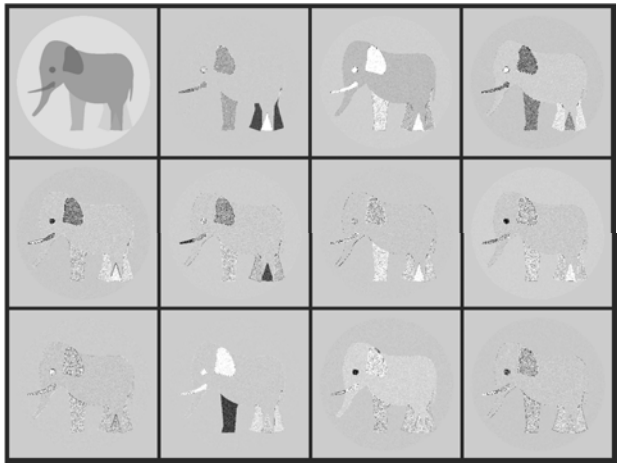
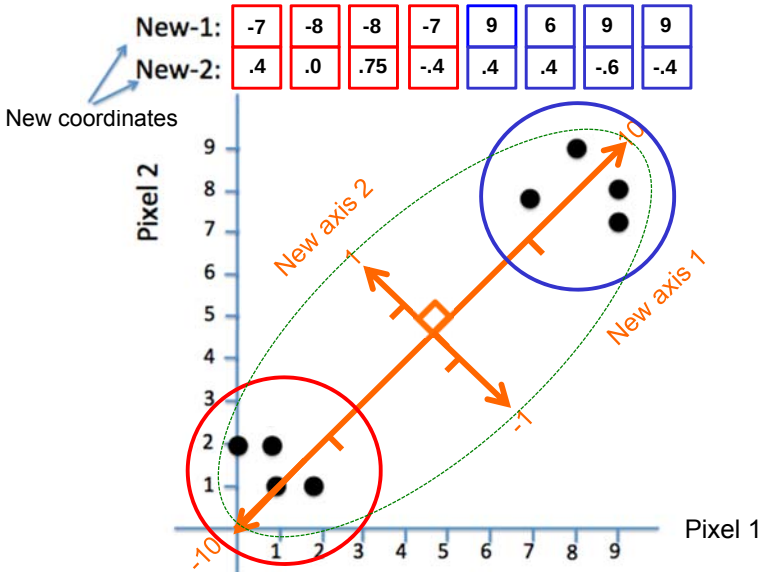


Image alignment is crucial for MSA

Eigen images (first 12)

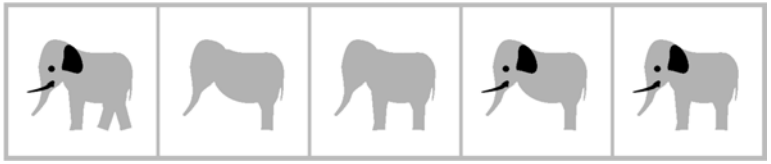


Eigen images are equivalent to the eigen vectors displayed in the original coordinate system. The eigen images are ranked according to their eigen values with the biggest first



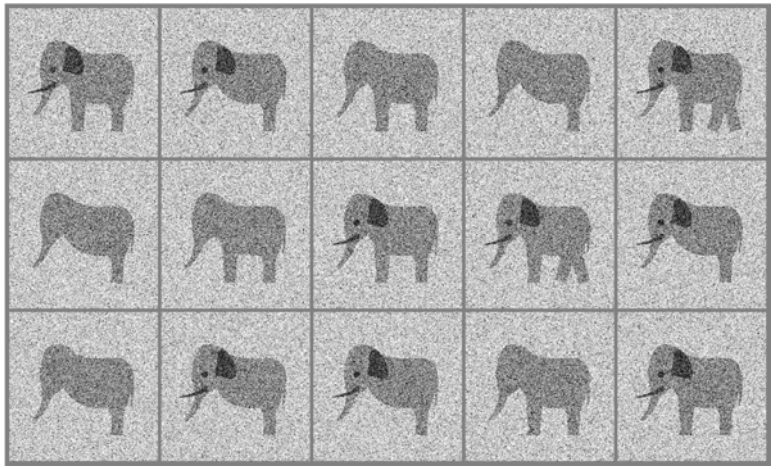
The main variation between the 2 pixel images can now be described by the new X-axis (1st eigen vector) reducing the number of measurements required to differentiate the images.

Class averages
24 eigenimages are used



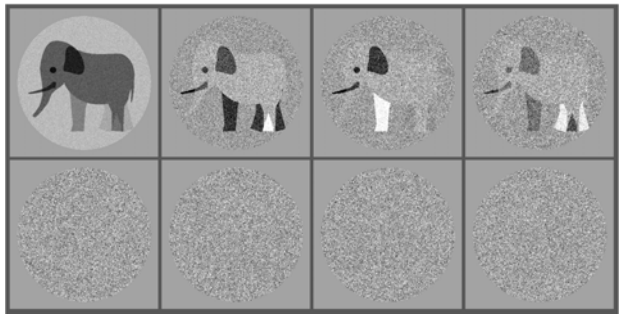
Classes look very nice and reflect the different elephants in the aligned images, but what about noise?

Set of raw (aligned) images with noise



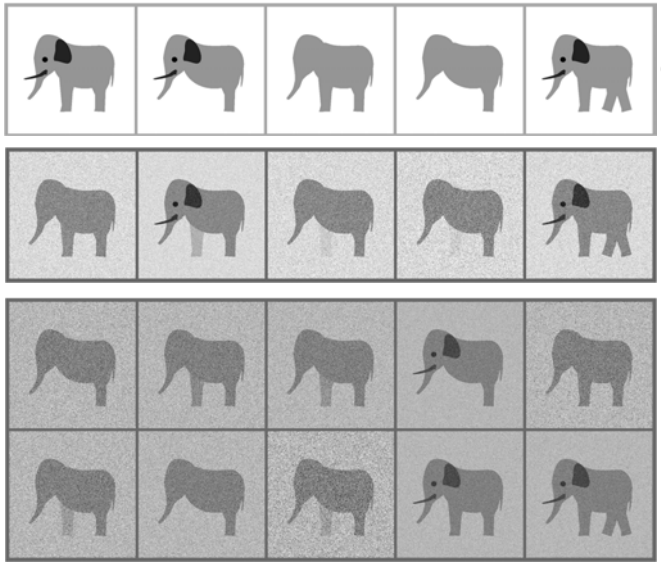
The noise in our EM images comes from the imaging system and particularly the low dose we have to use to image biological material.

Eigen images (first 8)



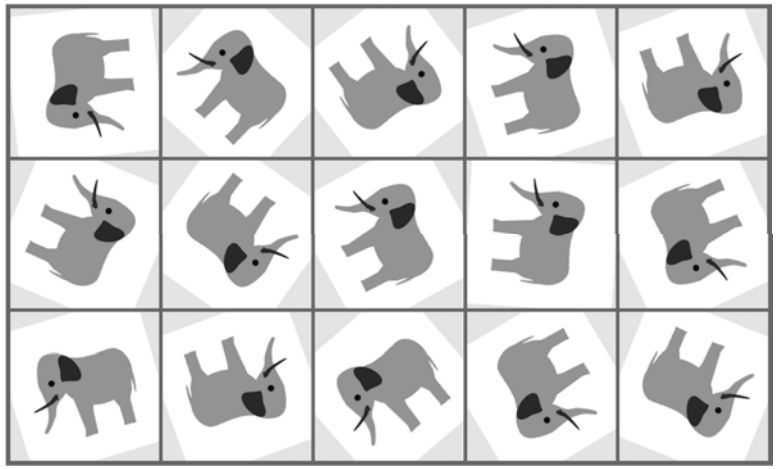
The noise makes it more difficult for us to distinguish fine details in the eigen images, such that the number significant eigen images are greatly reduced, which makes classification less accurate

Classification of images



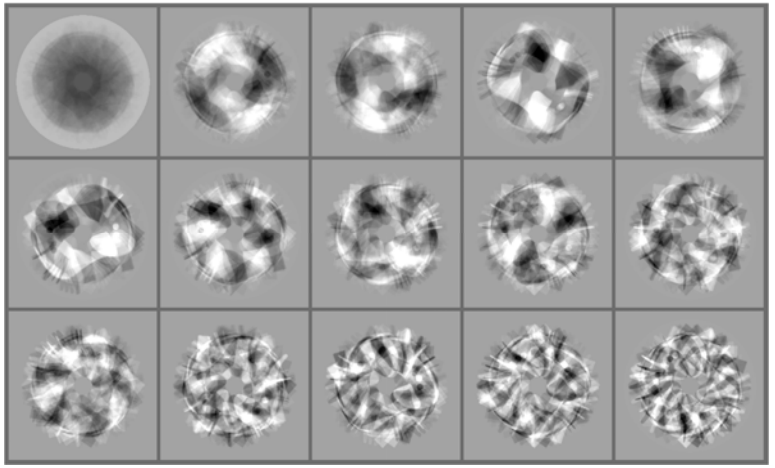
In the 5 classes calculated with the noisy images you can see some artifacts due to averaging of different elephants in the same class

Set of raw not aligned images randomly rotated



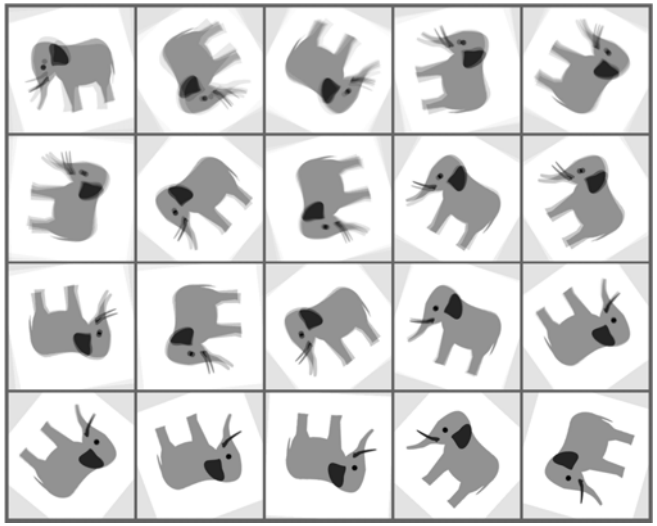
The next problem is that MSA requires the images to be well aligned (One exception to this). So what happens if the images are rotationally misaligned?

Eigen images (first 15)



The eigen images a pretty difficult to interpret!

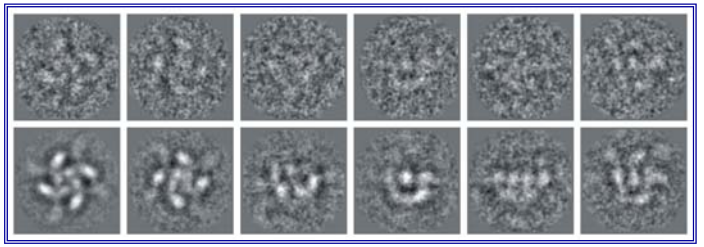
Classification of images



Some of the classes are good and can be used to realign the data such that a second round of MSA would give better class averages (Iterative procedure)

Symmetry determination using eigen images

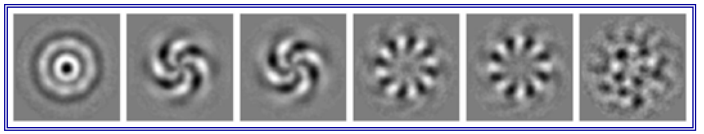
α-Latrotoxin *Tetramer - 520 kDa*



Raw images

Class averages

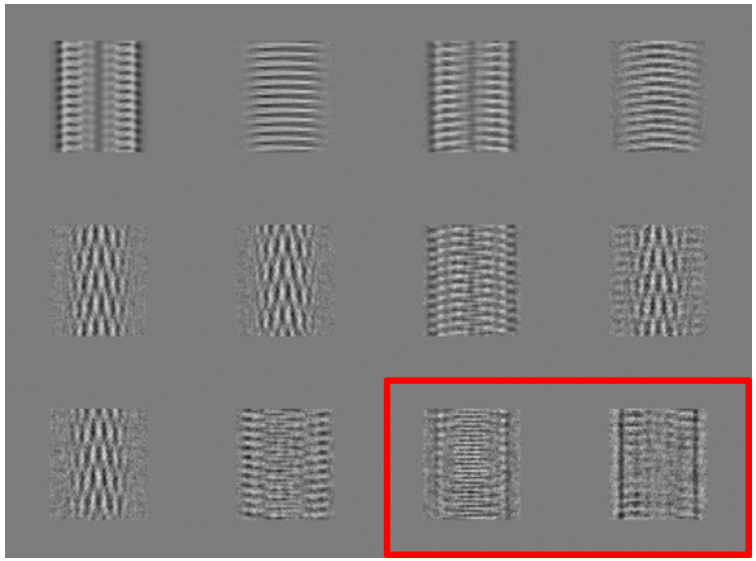
4 4 8 8



Eigen images of end-views

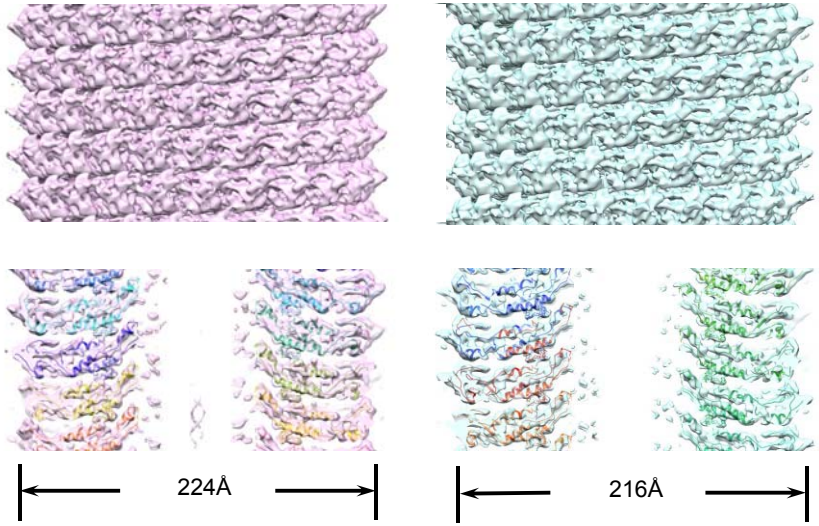
Orlova et al, 2000

Size separation of BSMV capsids



Eigen images 11 and 12 suggested a width difference between capsids

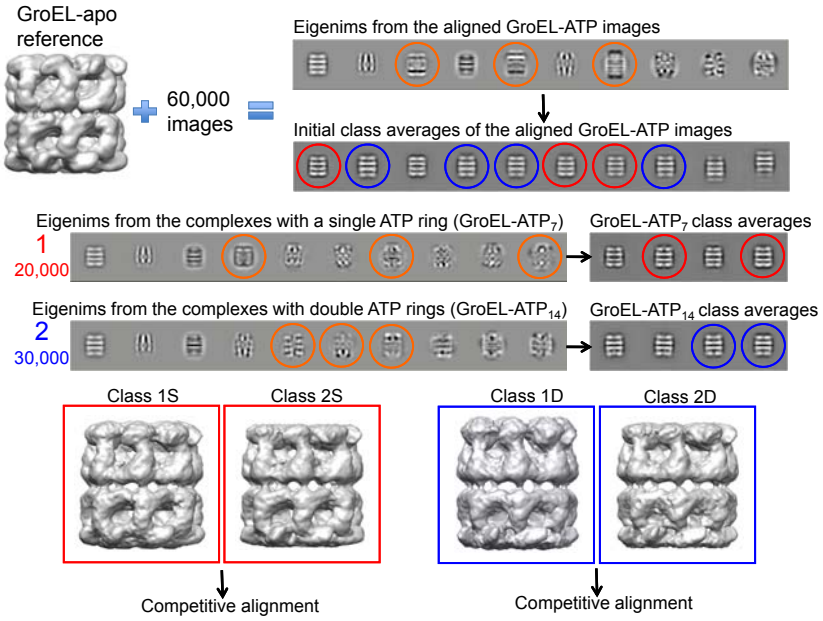
Reconstruction of wide and narrow BSMV



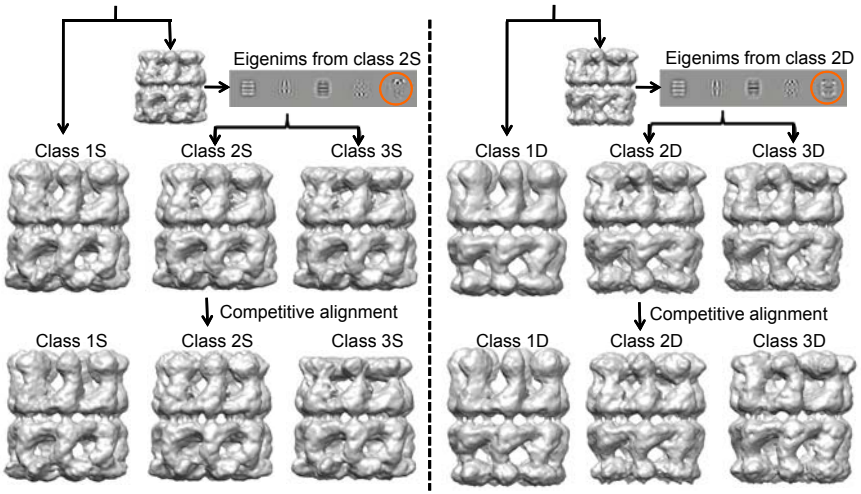
After initial separation using MSA and classification the structures were refined by competitive alignment

Clare et al, 2015

MSA and competitive alignment GroEL-ATP



MSA and competitive alignment GroEL-ATP



After multiple rounds of competitive alignment and MSA analysis there were 3 stable structures for each of the ATP_7 and ATP_{14} data sets.

(Clare et al., 2012)

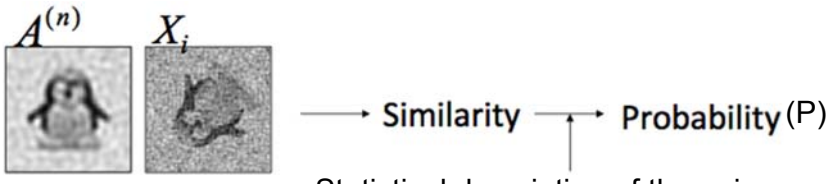
Maximum Likelihood

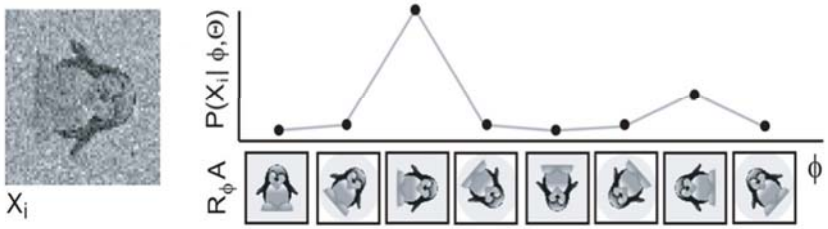
Maximum-likelihood approaches

- Probability-weighted assignments
- First described for use in EM by Fred Sigworth (JSB 1998)
  - For 2D-alignment to a single reference
- Extended for 2D and 3D classification (2005-2010)
  - XMIPP
- New software package for 2D and 3D ML alignment and classification (2012)
  - Relion
  - Fourier-space data model (coloured noise)
  - CTF-correction
- 3D ML-based classification (2013)
  - FREALIGN

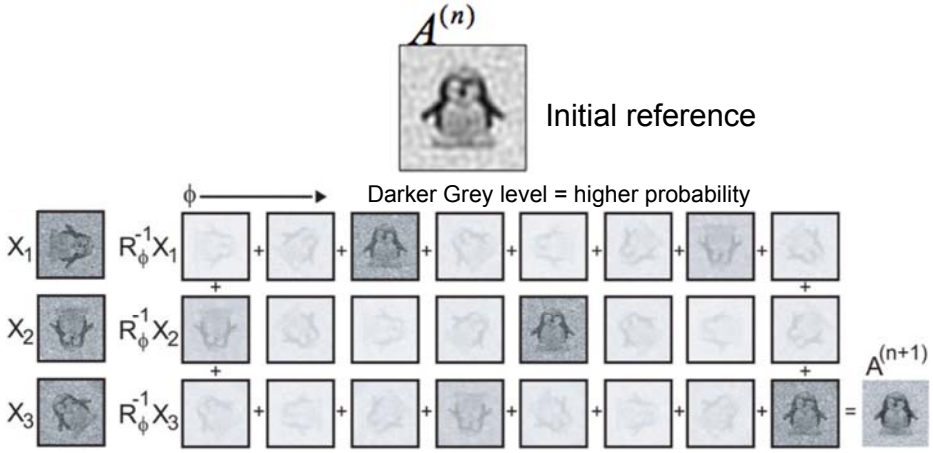
Sjors Scheres

2D Maximum-likelihood



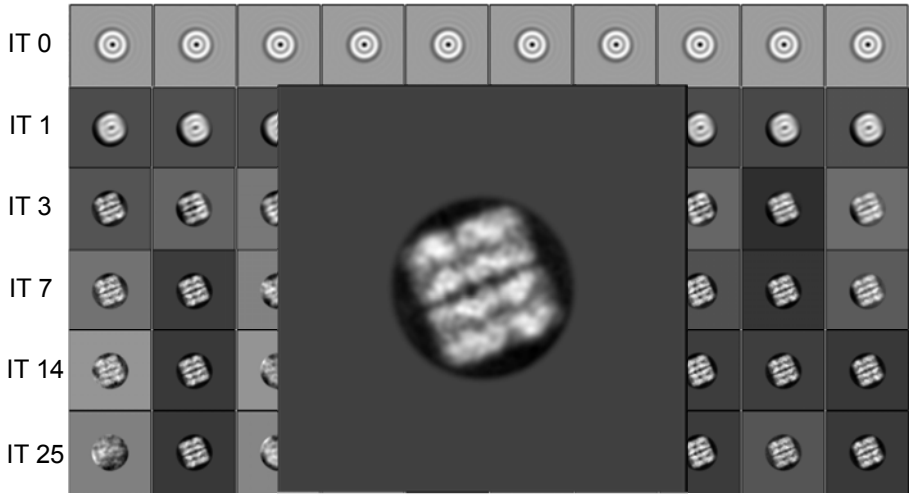


In this example the reference is rotated*



The average generated after ML is weighted according to the probability that the data matches the reference with the alignment parameters determined (in this example just different in plane rotations). This procedure is iterated until convergence at which time the probability is maximised for the alignment parameter that gives the best match between data and reference.

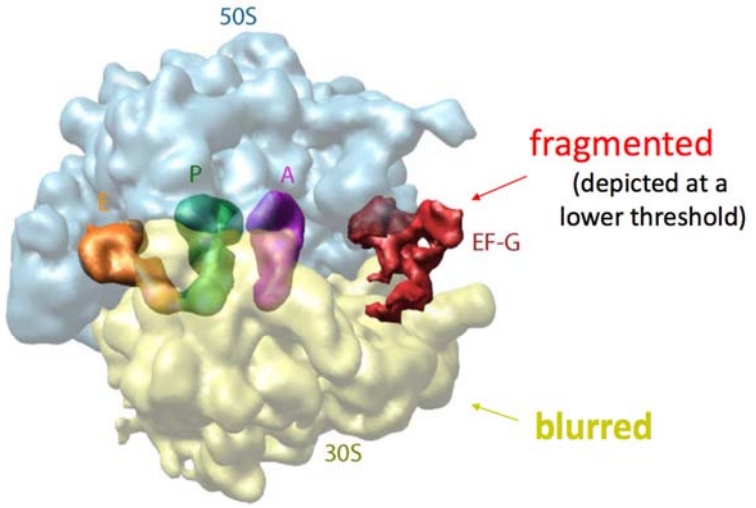
Reference-Free 2D ML alignment and classification



Alignment and classification done at the same time! Used to sort out bad particles in the data set and can be used to separate different sub-populations of images.

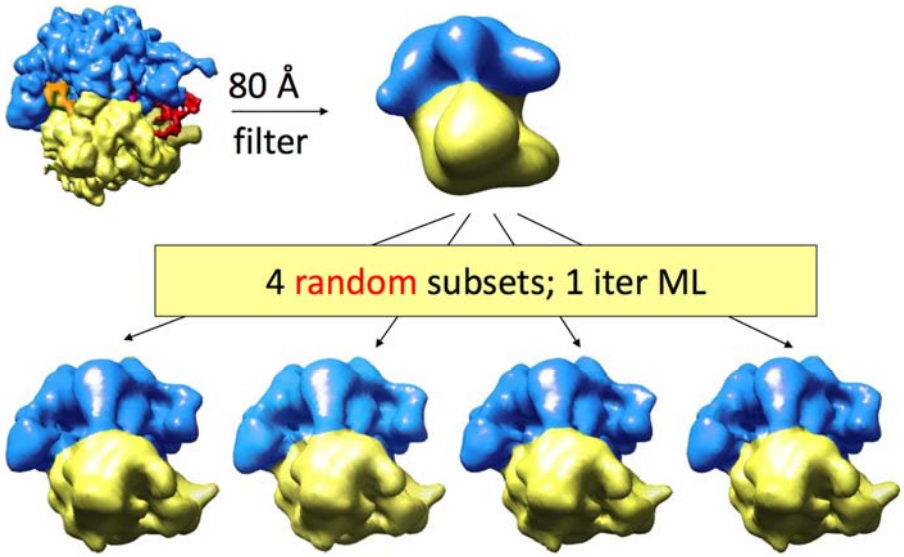
3D ML classification

Preliminary ribosome reconstruction: 91,000 particles and 9.9Å



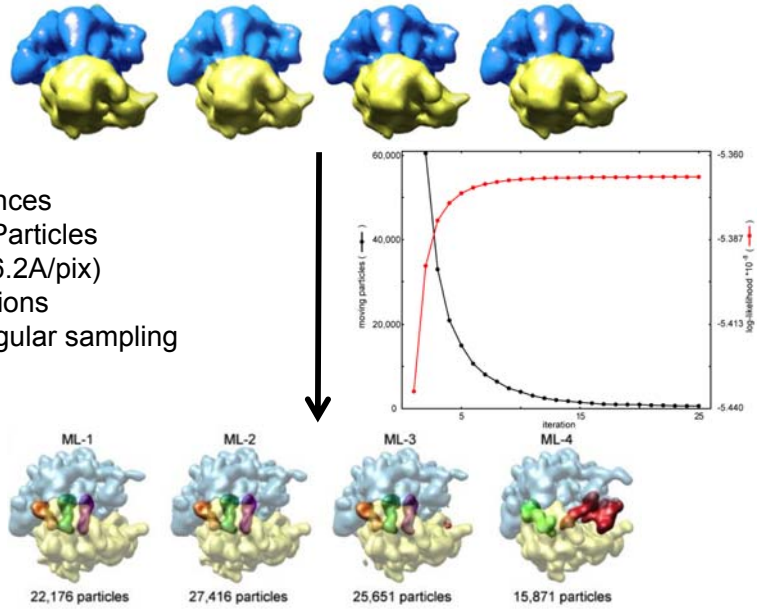
Sjors Scheres in collaboration with Haixiao Gao and Joachim Frank

3D ML seed generation



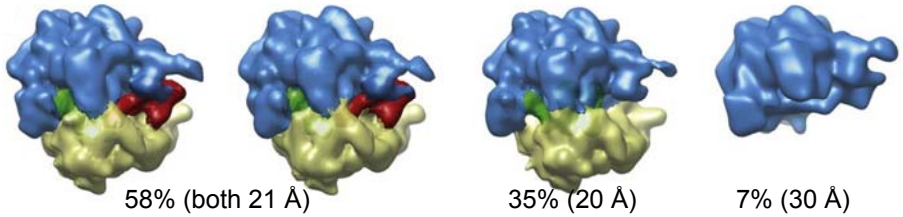
3D ML classification

- 4 references
- 91,000 Particles
- 64x64 (6.2A/pix)
- 25 iterations
- 10° angular sampling



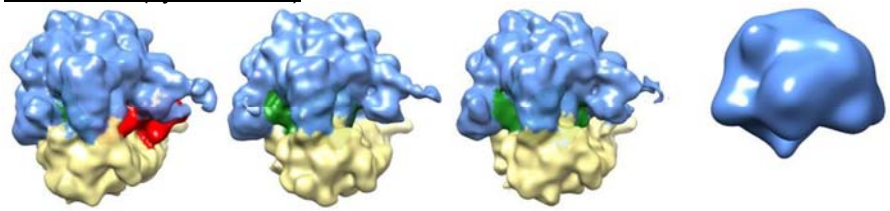
3D ML classification comparison

Relion (Scheres 2012)



58% (both 21 Å) 70S-EFG 1T-RNA | 35% (20 Å) 70S 3T-RNA | 7% (30 Å) 50s

FREALIGN (Lymkis 2013)



43% (17 Å) 70S-EFG 1T-RNA | 23% (18 Å) 70S 3T-RNA | 24% (18 Å) 70S 2T-RNA | 10% (40 Å) 50s

Similar results for both ML methods!

Maximum-likelihood characteristics

- ML can still get trapped in local minima and can also be dependent on starting model
- Compared to CC-based alignment:
  - Better convergence behavior
    (In practice you can start from random classes/orientations)
  - Much slower and needs more computing
  - ML = CC-based refinement for noiseless data
- Limited user interaction
  - only choose the number of classes in 2D and 3D

References (MSA cluster analysis)

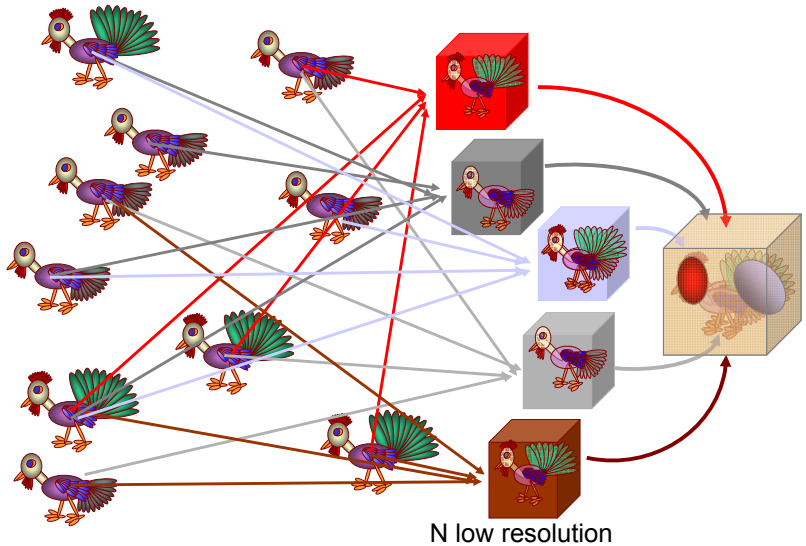
- Manly, B.F. (1998) Multivariate statistical methods. A primer. *Charmen & Hall,* London
- Everitt, B.S. (1995) Cluster analysis. *Arnold*, London
- van Heel, M. (1989). Classification of very large electron microscopical image data sets. *Optik* **82**, 114-126
- Frank J. (1996) Three-dimensional microscopy of macromolecular assemblies. *Academic Press*, San Diego.
- Frank J. (1990) Classification of macromolecular assemblies studied as 'single particles'. *Quart. Rev. Biophys.* **23**, 281-329.
- Ward, J.H. (1982) Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc*. **58** 236-244.
- van Heel, M. (1981) Use of multivariate statistics in analysing the images of biological macromolecules, *Ultramicroscopy* **6**,187-194.
- van Heel M and Frank J (1984) Multivariate Statistical Classification of Noisy Images (Randomly Oriented Biological Macromolecules), *Ultramicroscopy* **13**, 165-183.
- van Heel M et al (2009) Multivariate Statistical Analysis in Single Particle (Cryo) Electron Microscopy. An electronic text book: "Electron microscopy in Life Science", 3D-EM Network of Excellence, Editors: A. Verkley and E. Orlova (2009).

References (ML)

- Sigworth, F.J. (1998) A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biol.* **122**, 328–339.
- Scheres, S.H.W. et al (2005) Maximum- likelihood multi-reference refinement for electron microscopy images. *J. Mol. Biol.* **348**, 139–149.
- Sigworth, F.J. et al (2010) Maximum-likelihood methods in cryo-EM. Part I: theoretical basis and overview of existing approaches. *Methods in Enzymology* **482**, 263-294.
- Scheres, S.H.W (2010) Maximum-likelihood methods in cryo-EM. Part II: application to experimental data. *Methods in Enzymology* **482**, 295-320.
- Scheres, S.H.W (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519-530.
- Scheres, S.H.W (2012) A Bayesian view on Cryo-EM structure determination. *J. Mol. Biol.* **415**, 406-418.
- Lymkis, D et al., (2013) Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, 377-388.

Bootstrap method 1-

Start from a large data set that already has angles assigned. From this set create many maps from random populations of particles. Then determine the variance of each voxel.



N low resolution models; N > 100

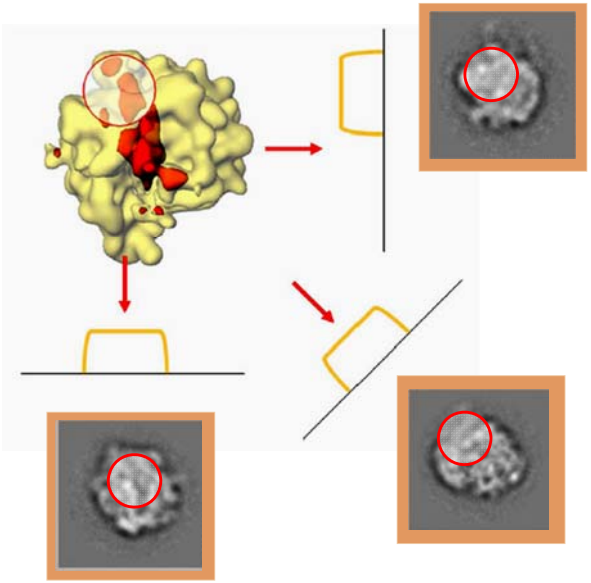
Penczek et al., 2006

Bootstrap method 2-

Select a region of variance in 3D and map its position on the 2D projections: focused classification

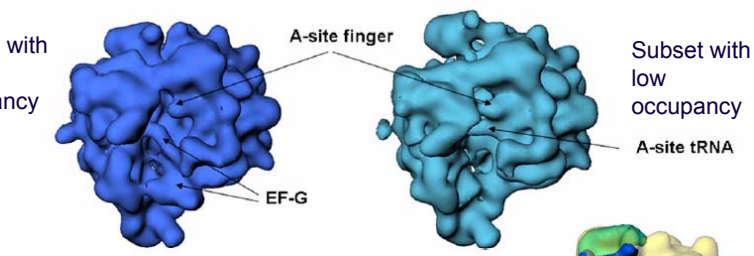
Focused classification:

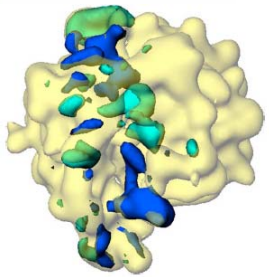
- Create a spherical mask around the variable region
- Project spherical mask in all orientations
- Classify each orientation group using only pixels within the spherical mask
- Use the classes to separate the structural variants in each orientation
- Calculate new 3D templates for multi-reference alignment



Bootstrap method 3-

The alignment and therefore angles of the original projection images are refined using the multiple 3D templates, until stable structures are reached.





Map of whole data set (yellow transparent surface) with positive difference densities superposed. Light blue/green: high occupancy map subtracted from low occupancy map.
Dark blue: low occupancy map subtracted from high occupancy map.