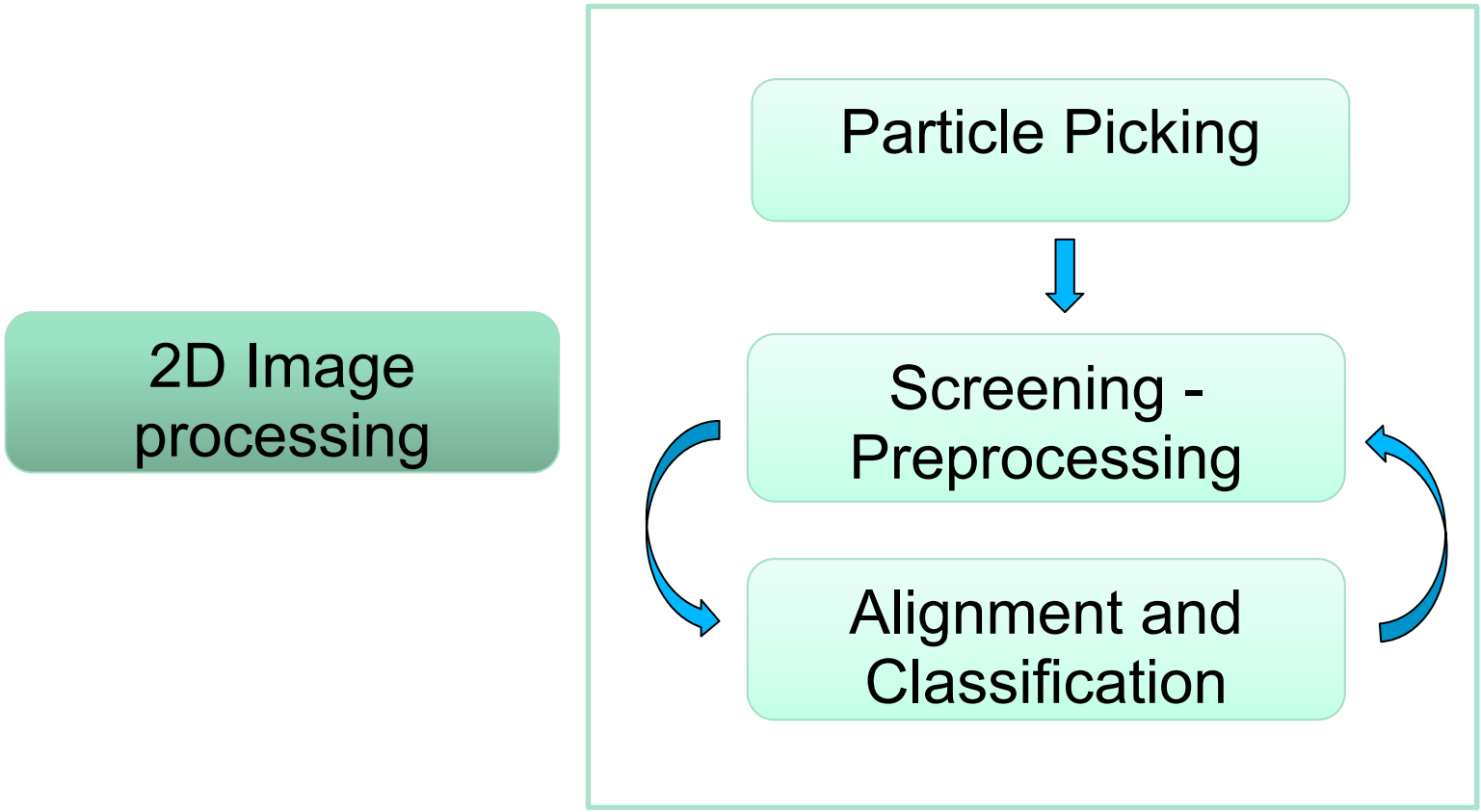
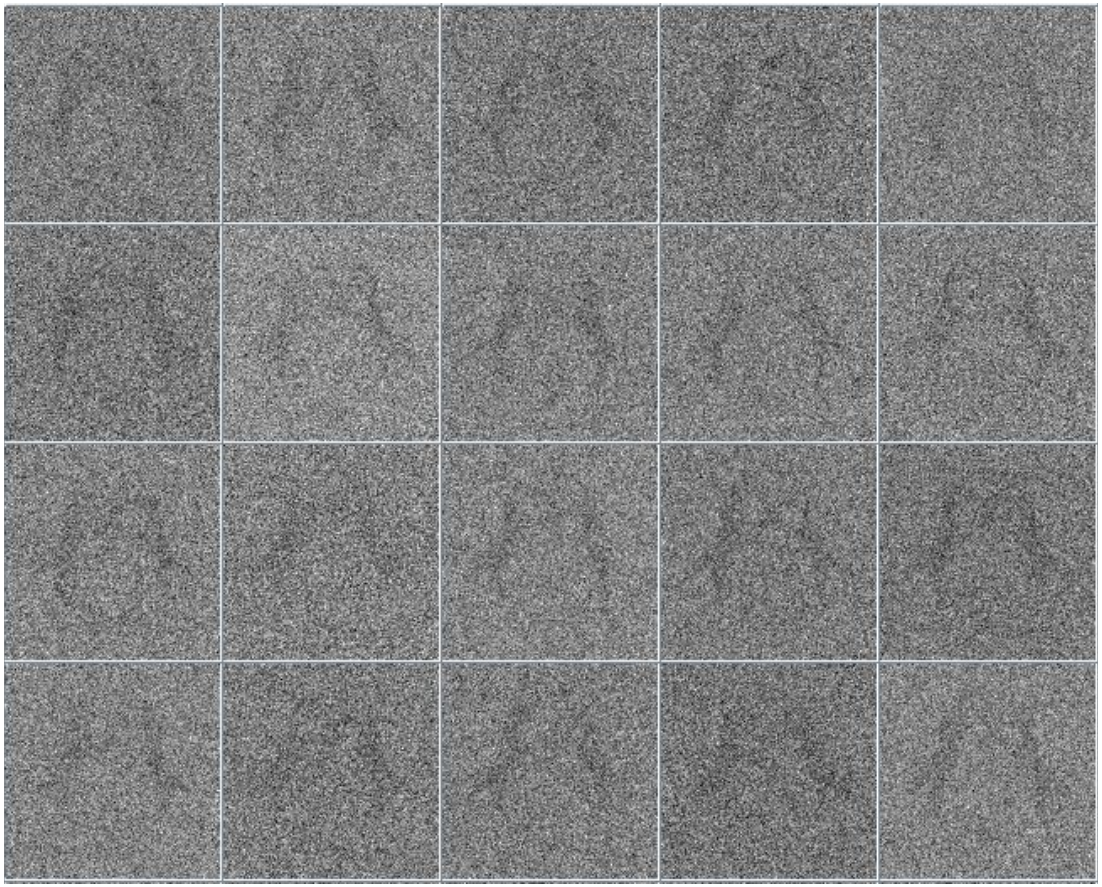
2D alignment and classification

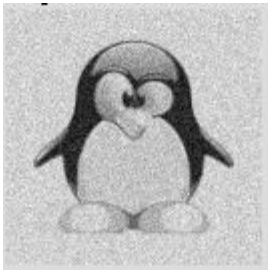
Biocomputing Unit,
Instruct Image Processing Center, CNB-CSIC
J.M. de la Rosa Trevín

2D image processing for Single Particles



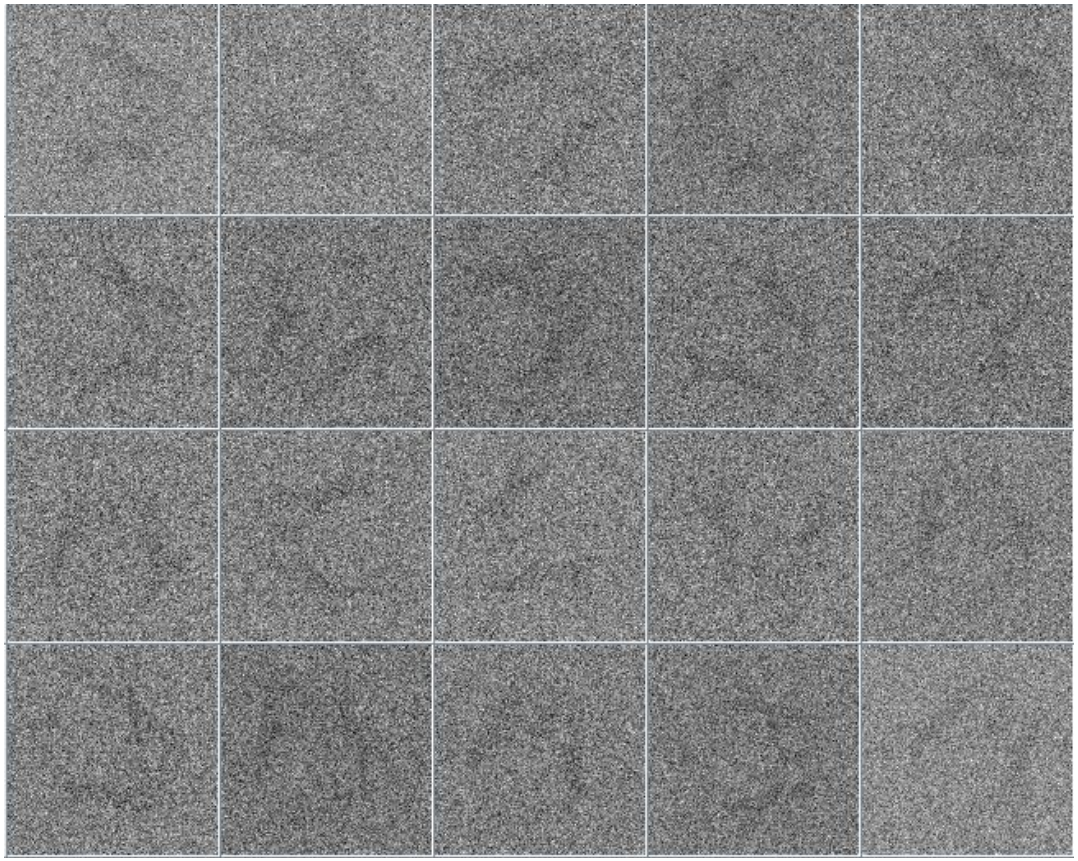
2D Averaging

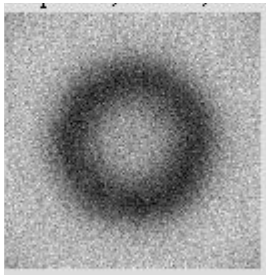




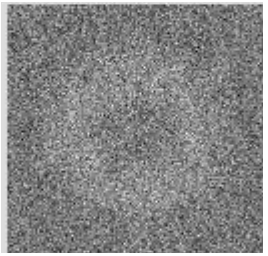
Average image

Alignment





Average image



Standard deviation

CSIC

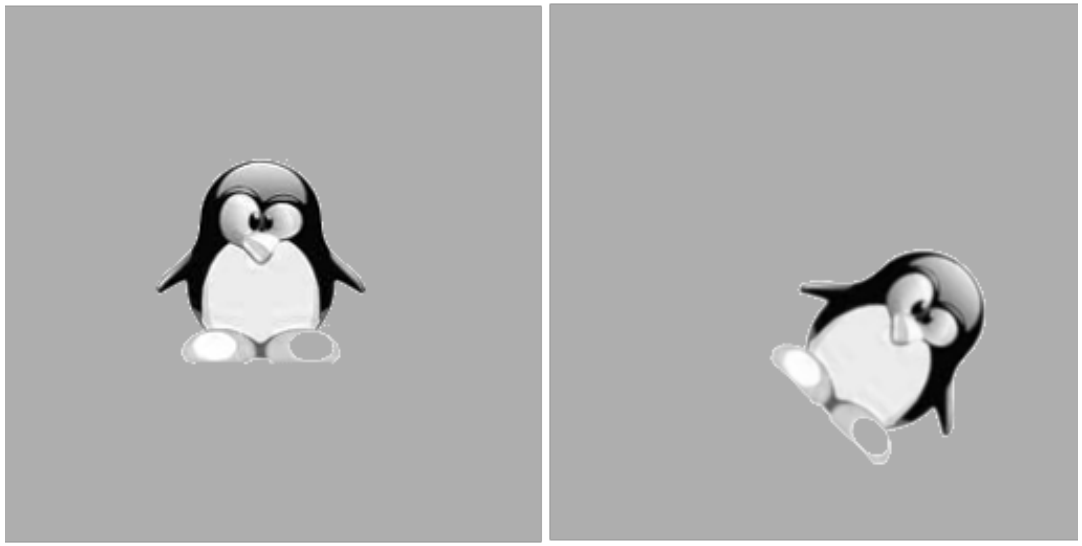
instruct
Integrating
Biology

BO COMPUTING Unit

instruct
image
Processing
Center

Translational and Rotational Cross-Correlation

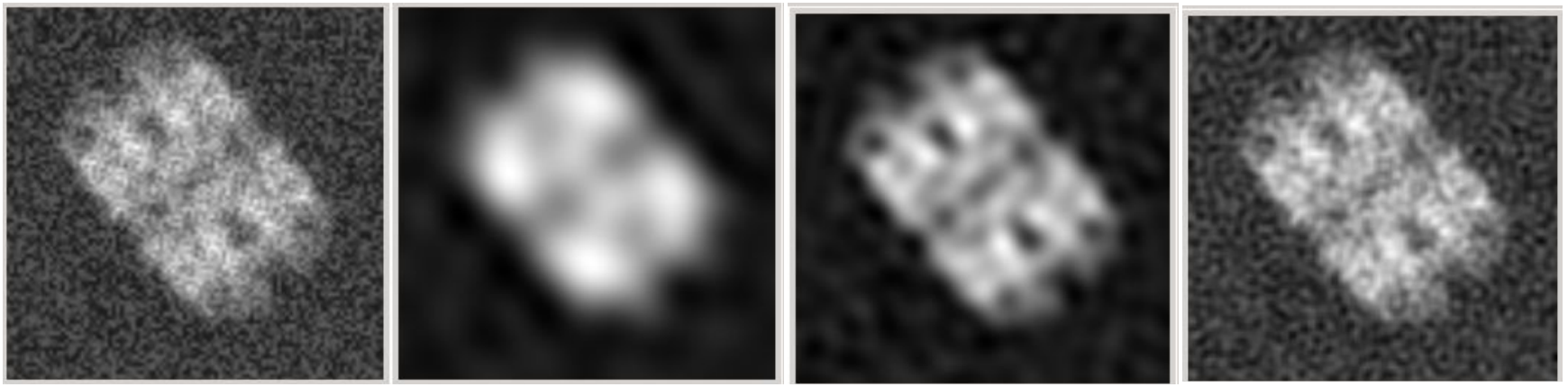
Three free parameters:
in-plane angle, x and y shifts



Preprocessing/Filters

Fourier “Top-hat” low-pass filter

Original image | radius=0.07 | radius=0.13 | radius=0.13



Reference-base alignment produce bias

Reference-free alignment

Classification

- Multivariate Data Analysis (MDA, Spider, Imagic and EMAN)
- Self Organized Maps (SOM, Xmipp)
- Maximum Likelihood (ML2D, Xmipp)
- Robust Clustering in 2D (CL2D, Xmipp)
- Iterative Stable Alignment Clustering (ISAC, SPARX/EMAN)

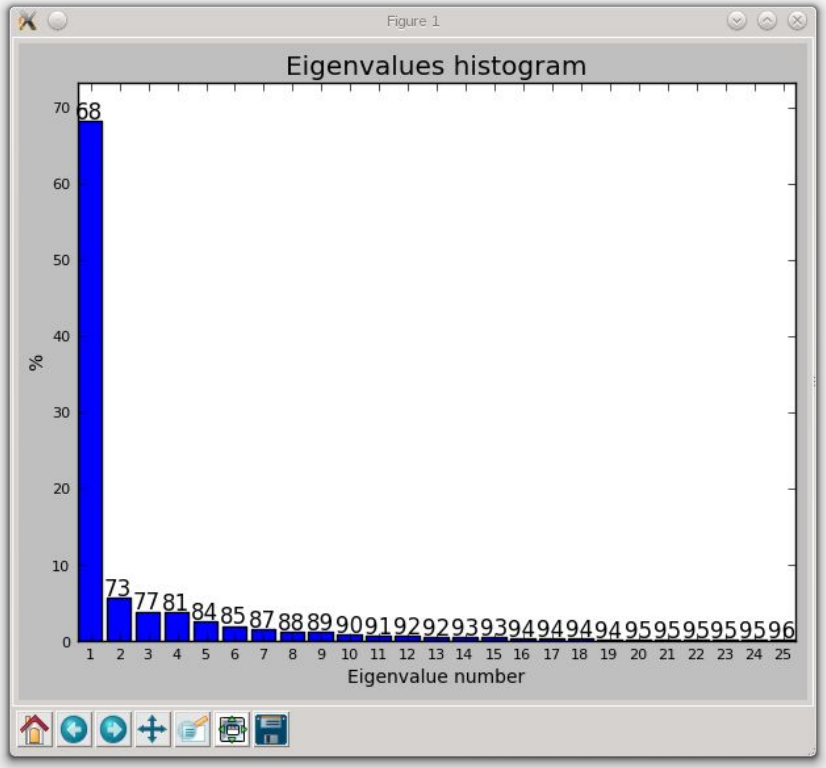
Multivariate Data Analysis (MDA, also known as MSA)

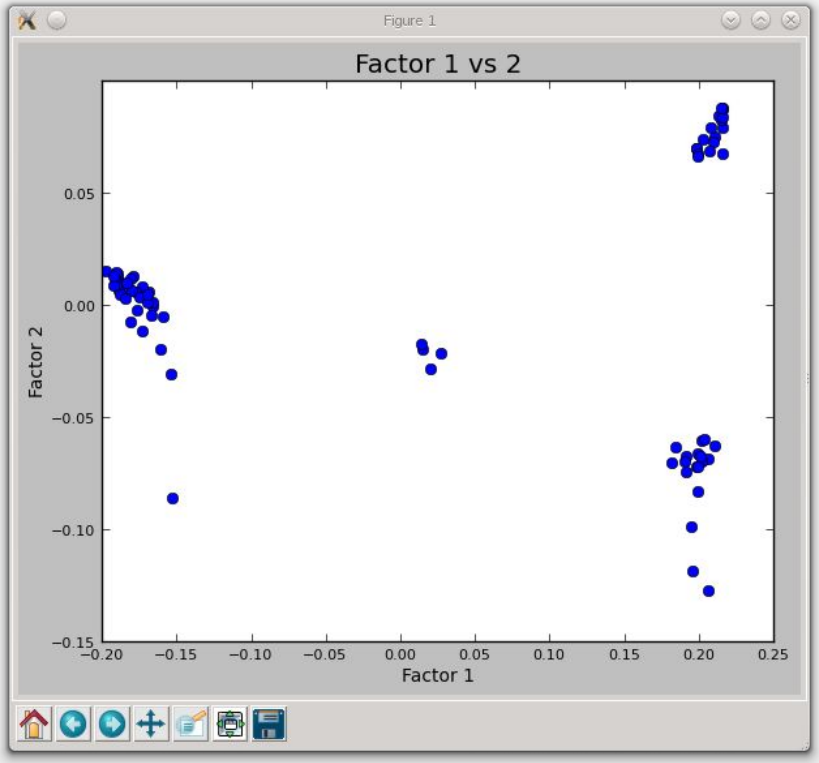
- We have a Set of N images (pi, i=1…N)
- Each image is seen as a point in a multidimensional space (each pixel is a coordinate) Xij {j=1…J}
- Image set is represented as a cloud in the (hyper) space
- All images have been previously aligned

Making patterns emerge from data

- Correspondence Analysis (CA, introduced by Benzecri in 1969)
- Principal Component Analysis (PCA)

Factorial coordinates and factorial maps



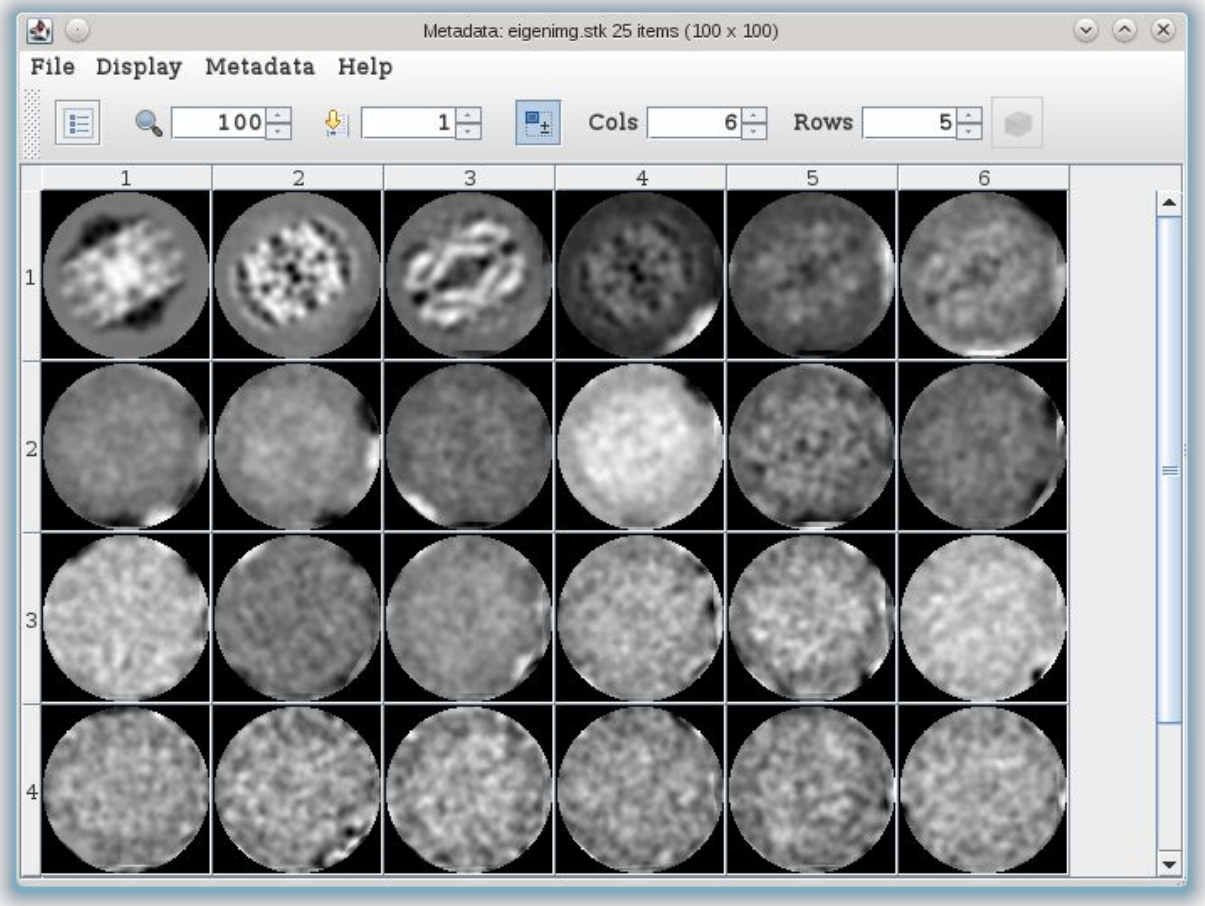


CSIC

instruct Integrating Biology

instruct image Processing Center

Images can be “reconstituted” using the factors

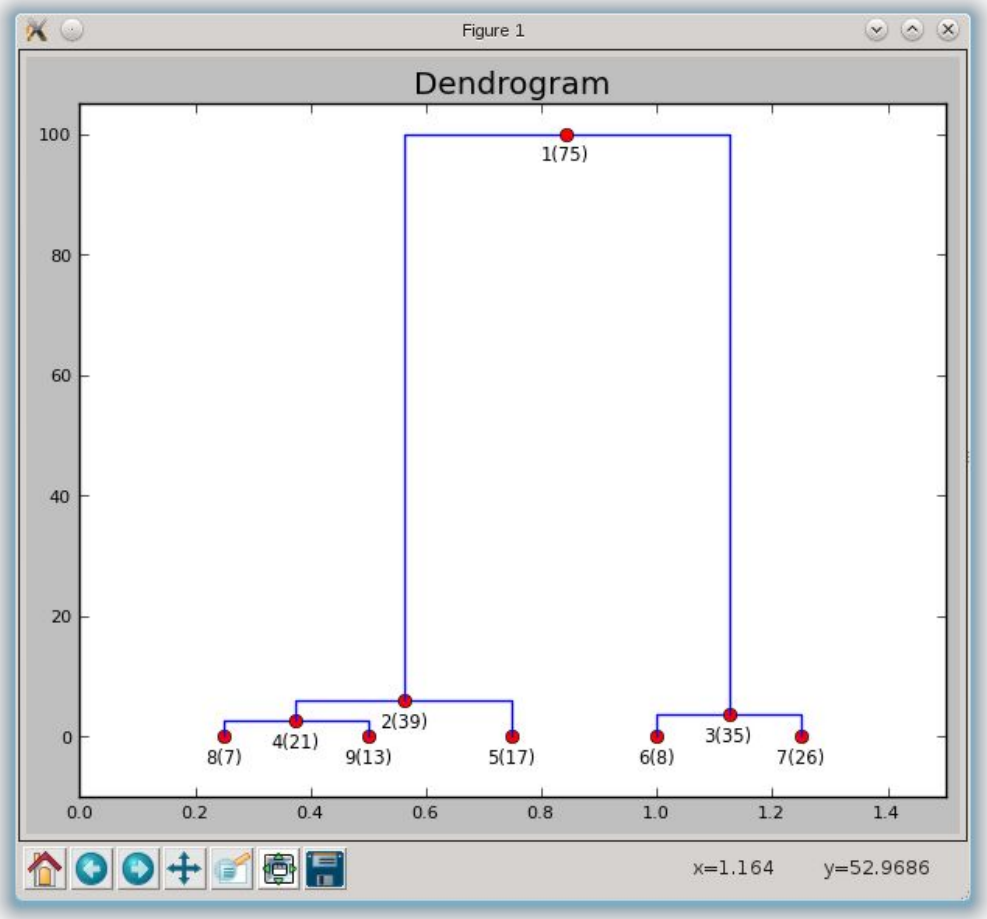


CSIC

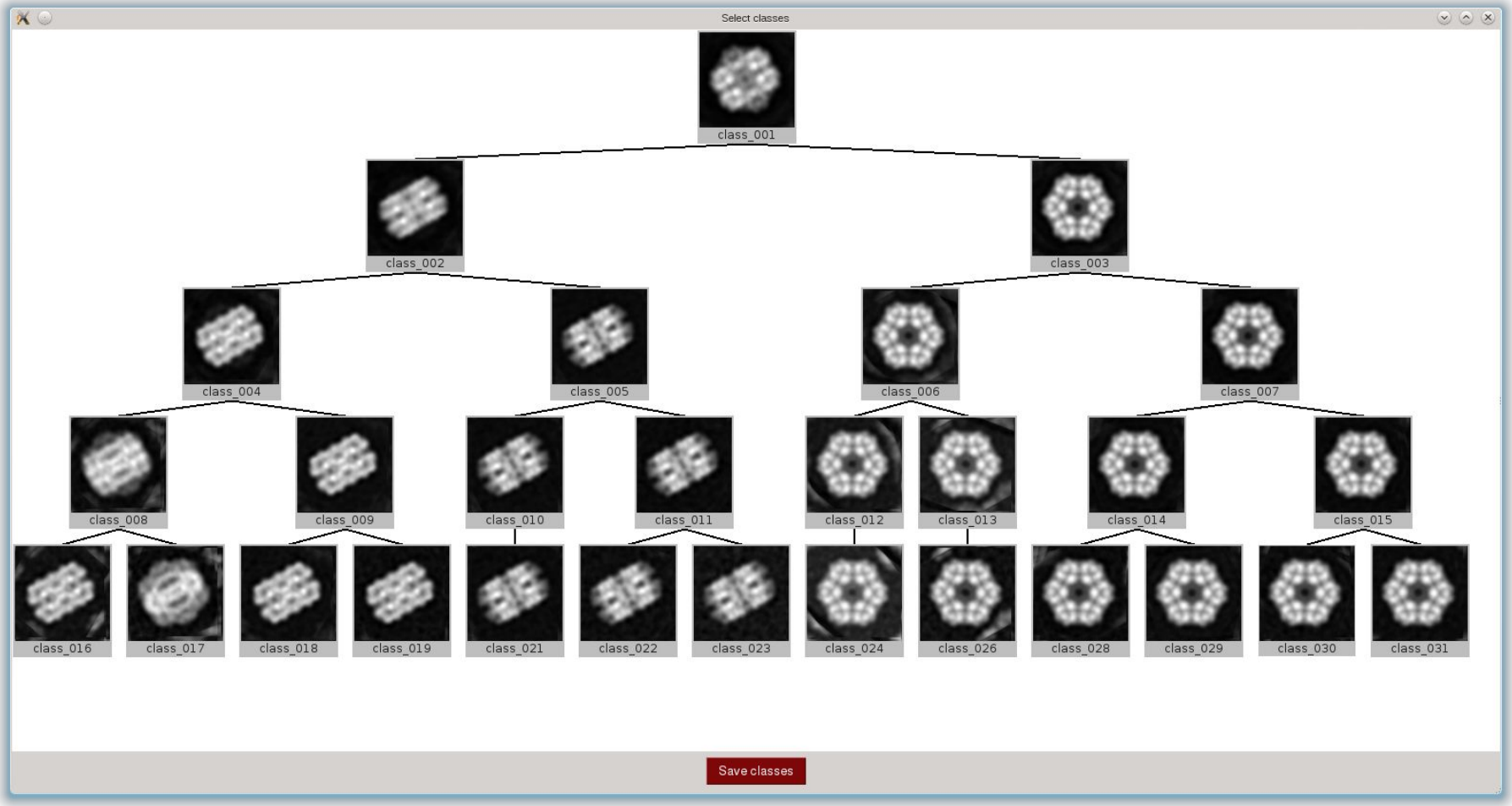
instruct Integrating Biology

instruct image Processing Center

Use of Explanatory Tools



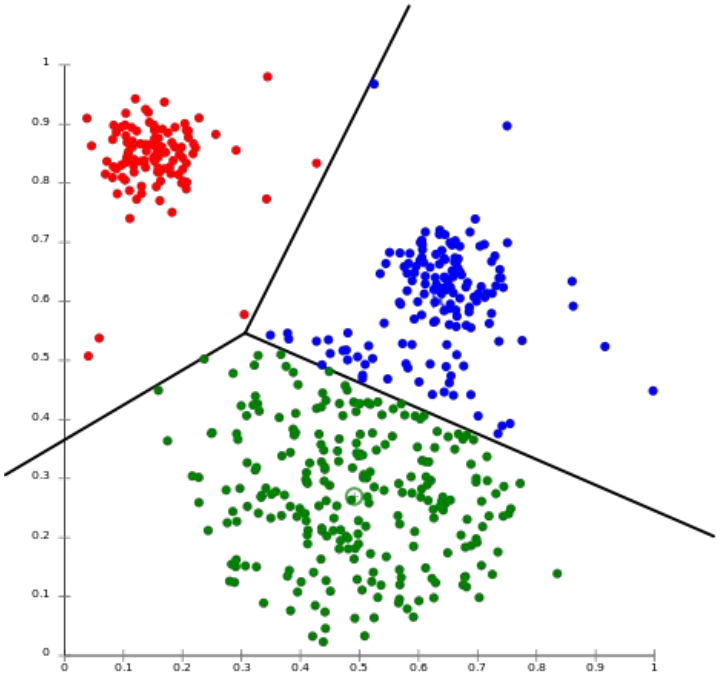
Use of Explanatory Tools

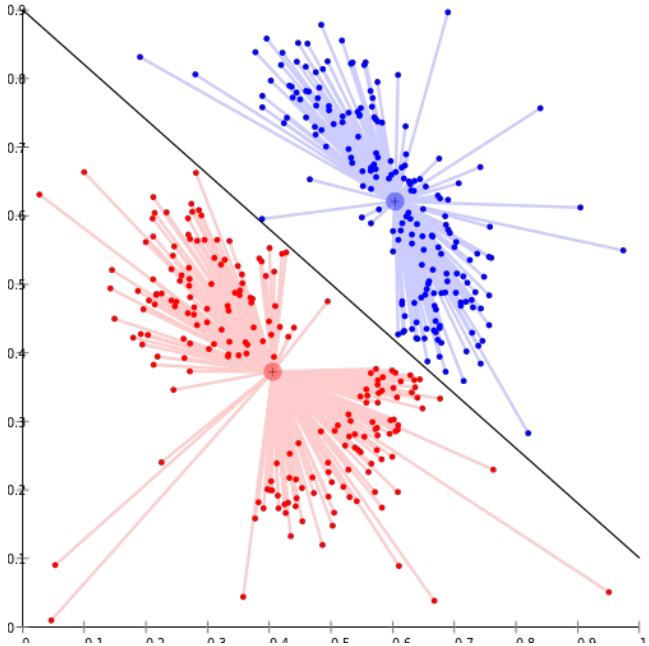


K-means clustering algorithm

1. Have as input the number of clusters K
2. Pick randomly K objects from the population as starting centers
3. Compute the distance of each object to each center and assign to the closest one.
4. Update each center by averaging the objects assigned.
5. Repeat

K-means clustering





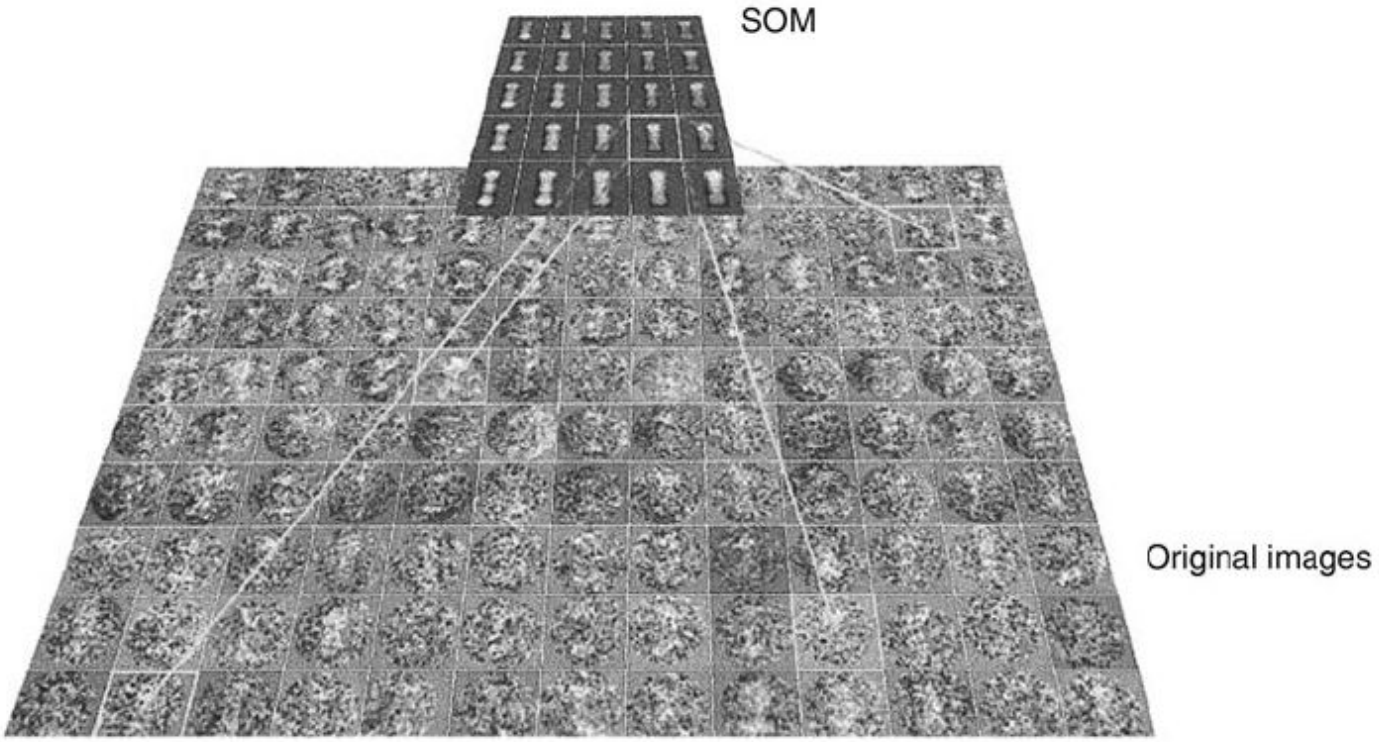
K-means properties

- Simple algorithm and work very well if clusters are separated and guessed K is correct.
- Cluster tends to be (hyper) spherically shaped
- Converge only to a local minimum.
- The results depends on the initial “seeds” (to overcome this, *dynamic clouds*, introduced by Diday in 1971)

Hierarchical Clustering

- Cluster tends to be (hyper) spherically shaped
- The results depends on the initial “seeds” (to overcome this, *dynamic clouds*, introduced by Diday in 1971)
- Hybrid clustering: add a “postprocessor” to break “early marriage” (VanHeel 1984)

Self-Organized Map (SOM)

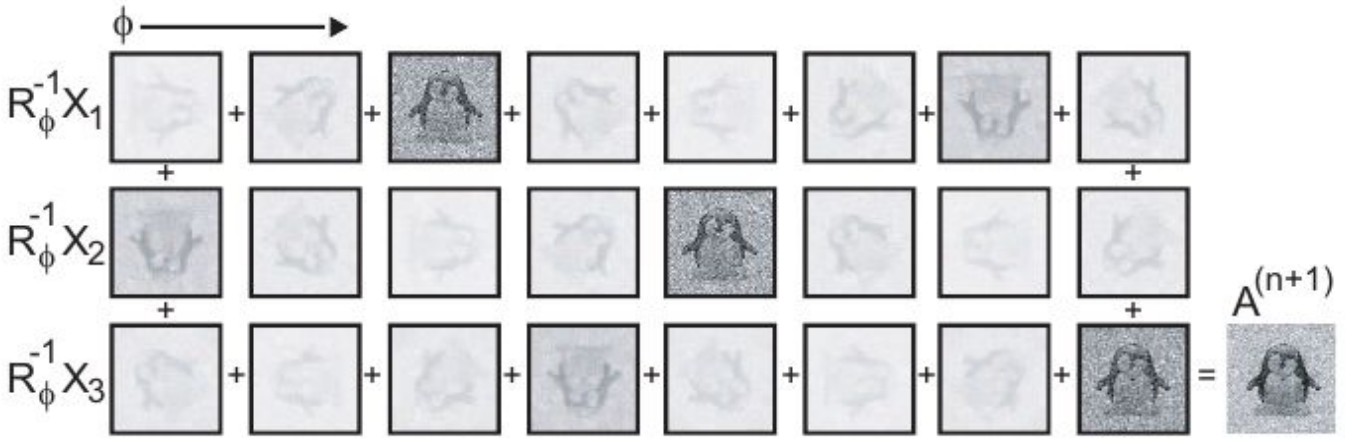


Maximum Likelihood (ML2D)

- Introduced in 1998 by Sigworth for a single reference
- Better model of the noise statistics.
- Extended by Scheres for multireference alignment and 3D.
- Has a reduced sensitivity to initial seeds and extract the underlying signal from much noisier images

Maximum Likelihood (ML2D)

- Each image contributes to each class with certain probability (fuzzy classification)



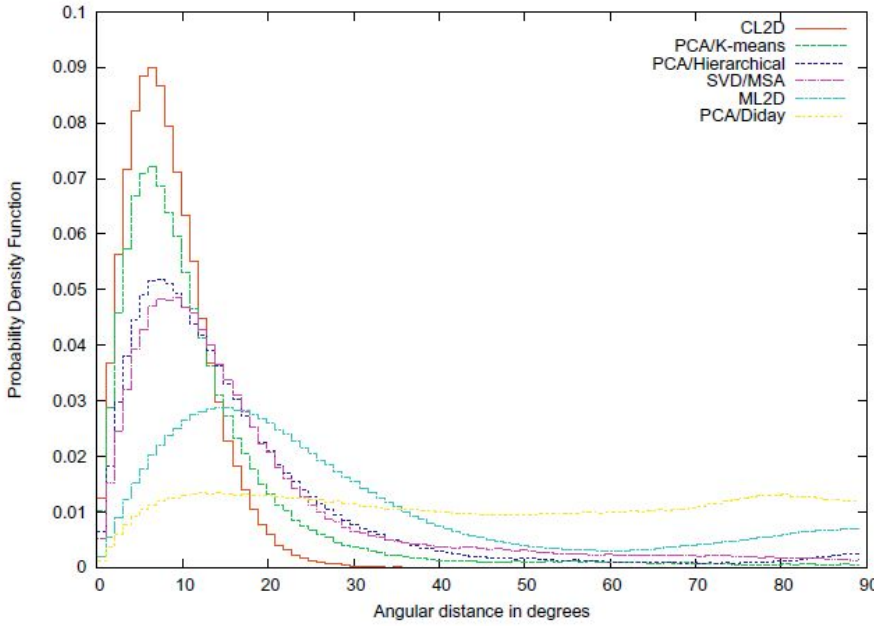
ML2D conclusions

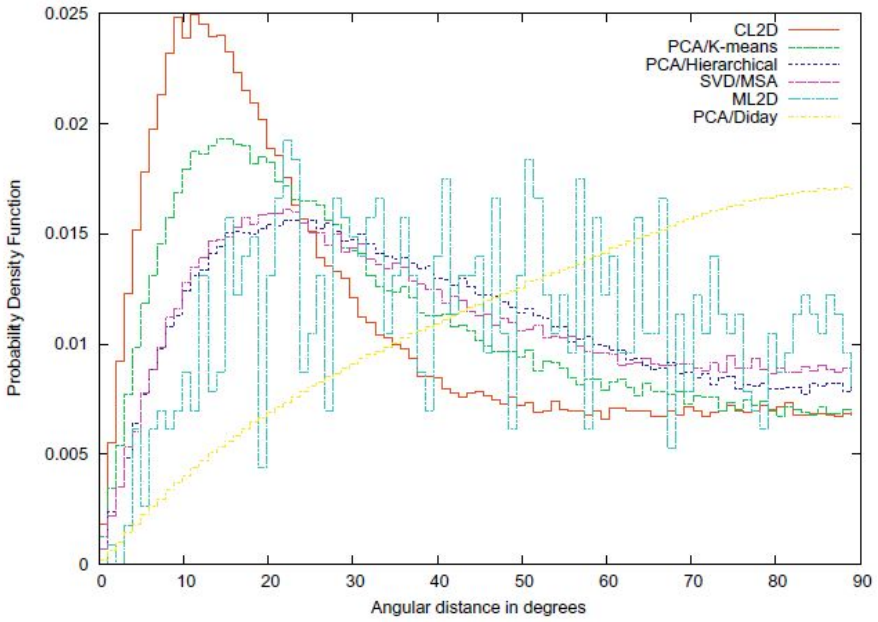
- Has a reduced sensitivity to initial seeds and extract.
- Extract the underlying signal from much noisier images
- Big classes tend to “attract” more particles, leading to empty classes
- Demand a lot of computational resources

Robust clustering (CL2D)

- This clustering approach allows to use correntropy or cross-correlation as similarity measure.
- Avoids the creation of small or empty groups
- Is a divisive clustering to try to avoid get trapped into local minima.
- Assignment of an image to a class is not compared only to image-class measure, but also to other class members

CL2D, results on simulated data





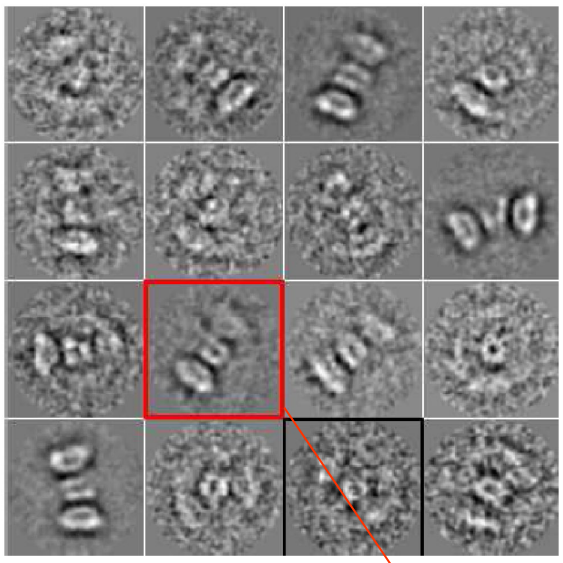
CSIC

instruct
Integrating
Biology

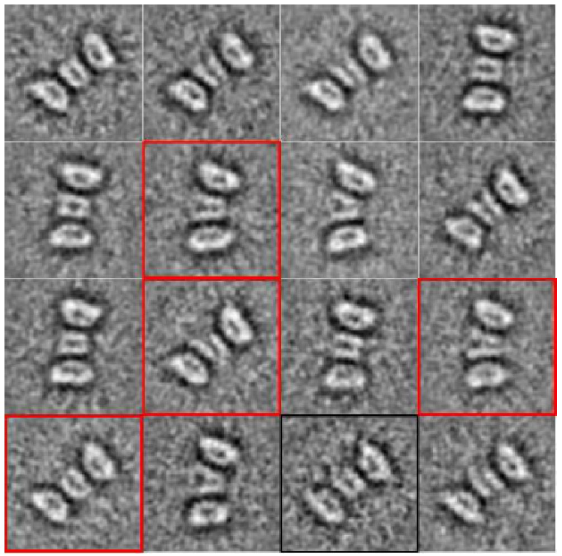
BO COMPUTING Unit

instruct
image
Processing
Center

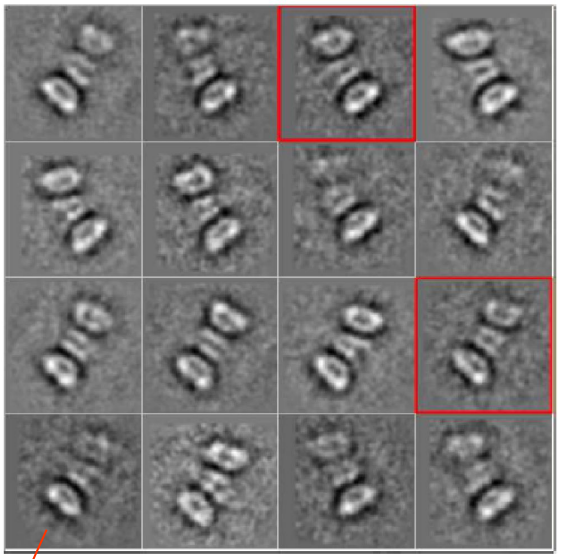
CL2D on SV40 of large-T antigen



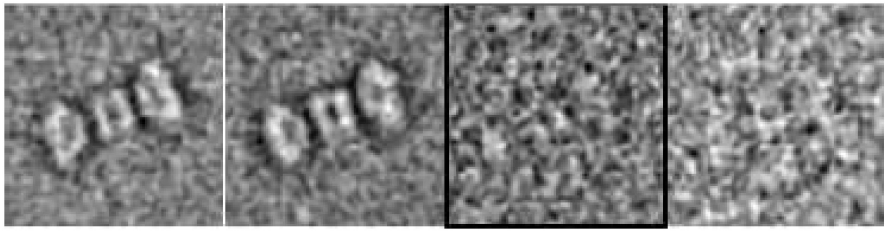
ML2D



CL2D



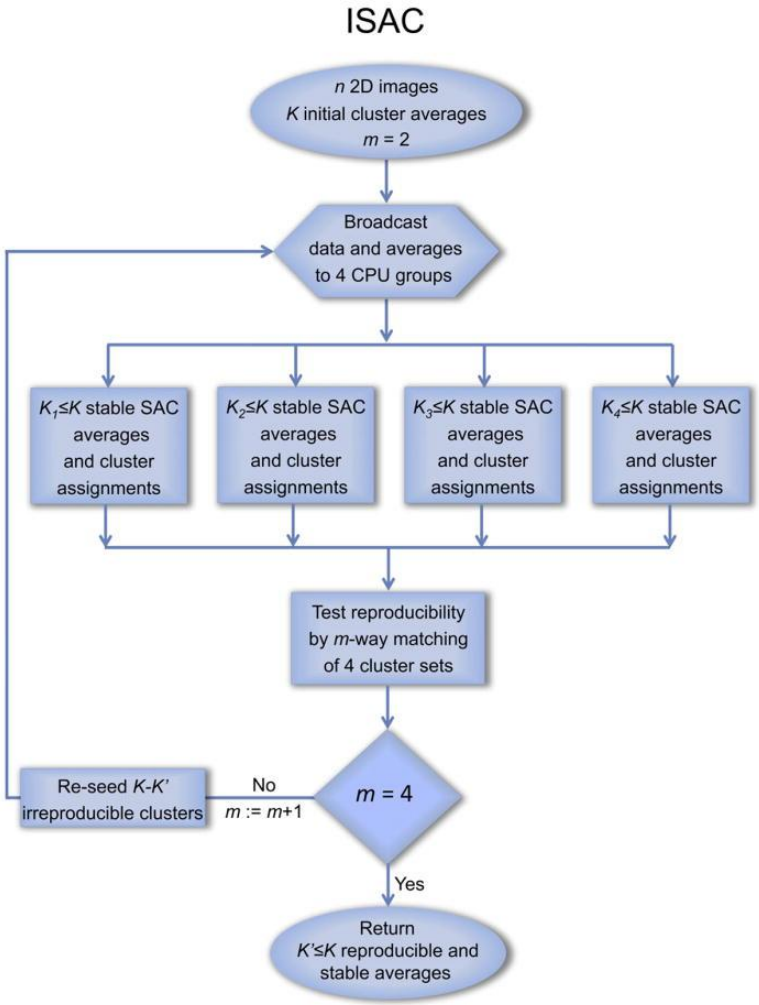
SVD/MSA

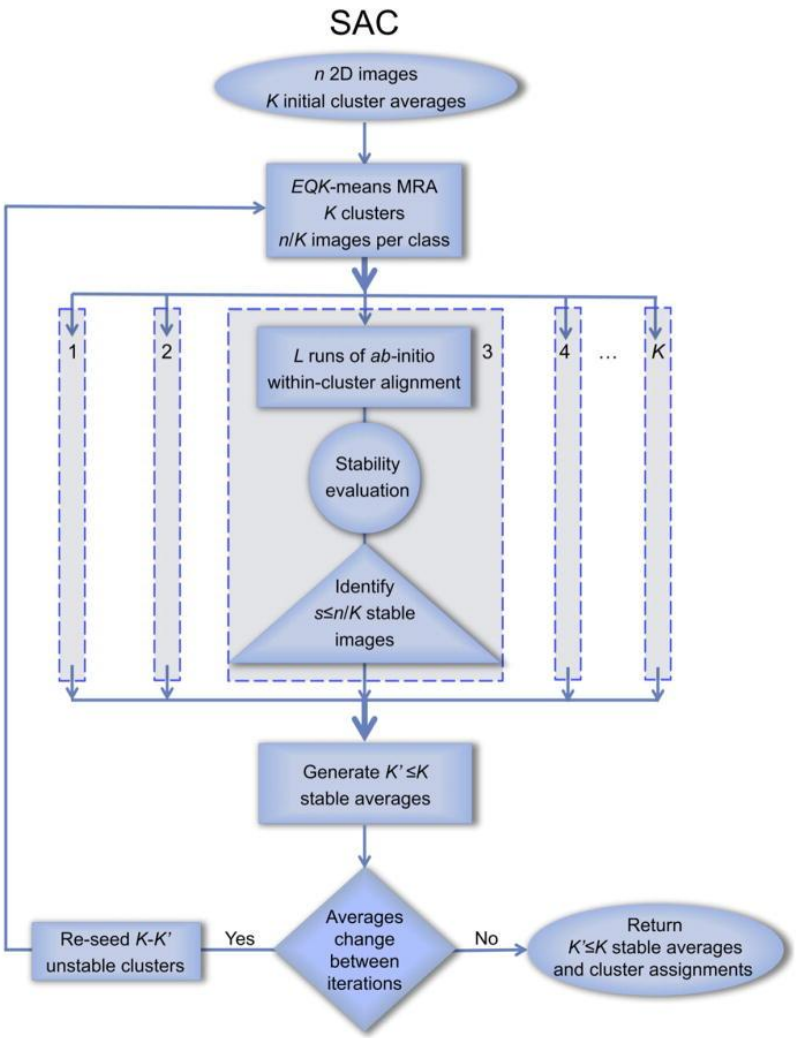


Iterative Stable Alignment and Clustering (ISAC)

- EQK-means forces all cluster to have the same number of members.
- Stability of alignment parameters is validated over L repetitions.
- Reproducibility of several multireference clustering results is checked.

ISAC algorithm





CSIC

instruct
Integrating
Biology

BCU Unit

...ruct
image
Processing
Center

ISAC conclusions

- ISAC is a simple approach based on stability and reproducibility.
- Results are validated, increasing the reliability.
- Requires few user parameters:
    - Number of classes
    - Number of elements per cluster
    - Number of alignment repetitions (L)
- Extra validations requires more computational time.

General Conclusions

- There are not “silver bullets” for all kind of data and problems.
- Is important to know the advantages and disadvantages of each algorithm.
- We should dedicate more effort to validation and reproducibility.
- Software and algorithms still have room for a lot of improvements.

instruct
Integrating
Biology

www.structuralbiology.eu

Follow us on twitter **@instructhub**

CSIC

instruct
Integrating
Biology

BIO COMPUTING Unit

instruct
image
Processing
Center