Clustering algorithms

Kass, Itamar
30 October 2008

What is the problem?

Typical MD produced thousand to millions of configurations. e.g. Saving every 5 ps will end up with 200 configuration in 1 ns.

If one is interested in the structural properties which does not depends on time, this dependent add a level of complexity which should be removed.

Clustering

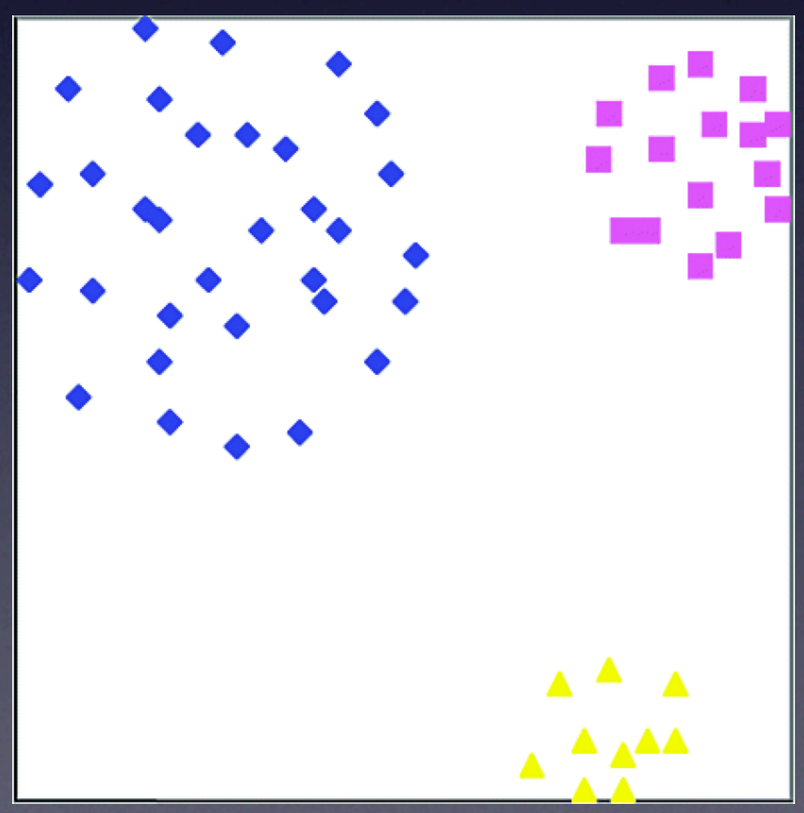
Clustering is a data sieving technique that can be applied to any data set if one has a function which measures the distances between any two points.

Clustering can lower the complexity of of the structural information, revealing patterns which are hidden, at the expanse of dynamic properties.

e.g. low energy states, regardless of when they occur.

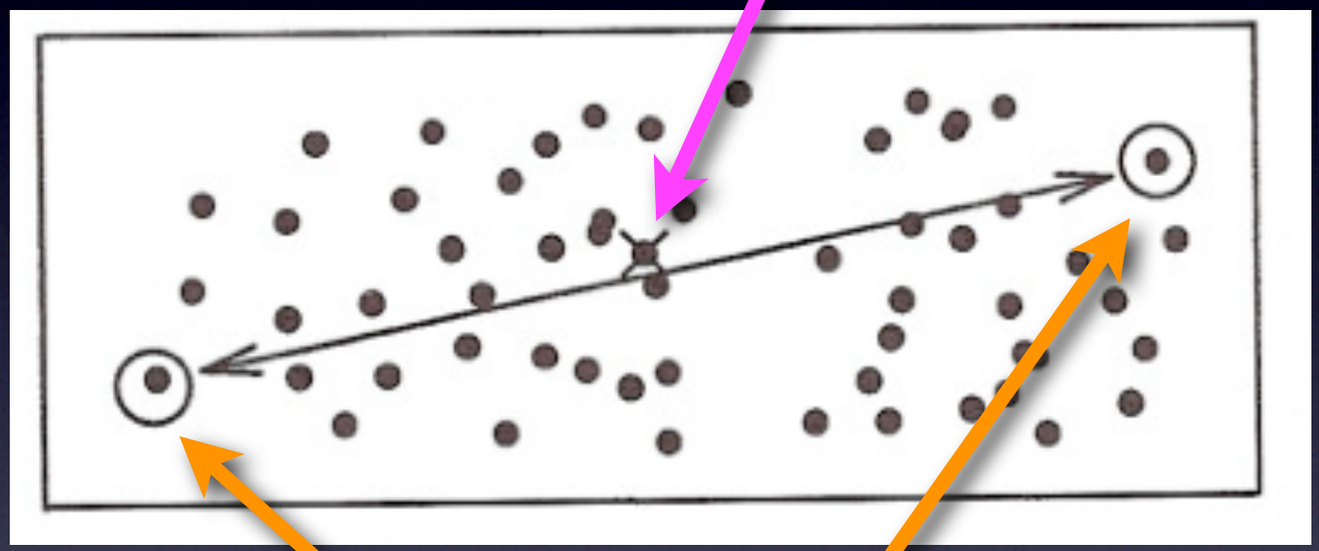
Clustering

Clustering will gather all similar points into one group (cluster). All the points in one cluster are similar to each other then to points in other clusters.



Few concepts

centroid



eccentrics

Different algorithms

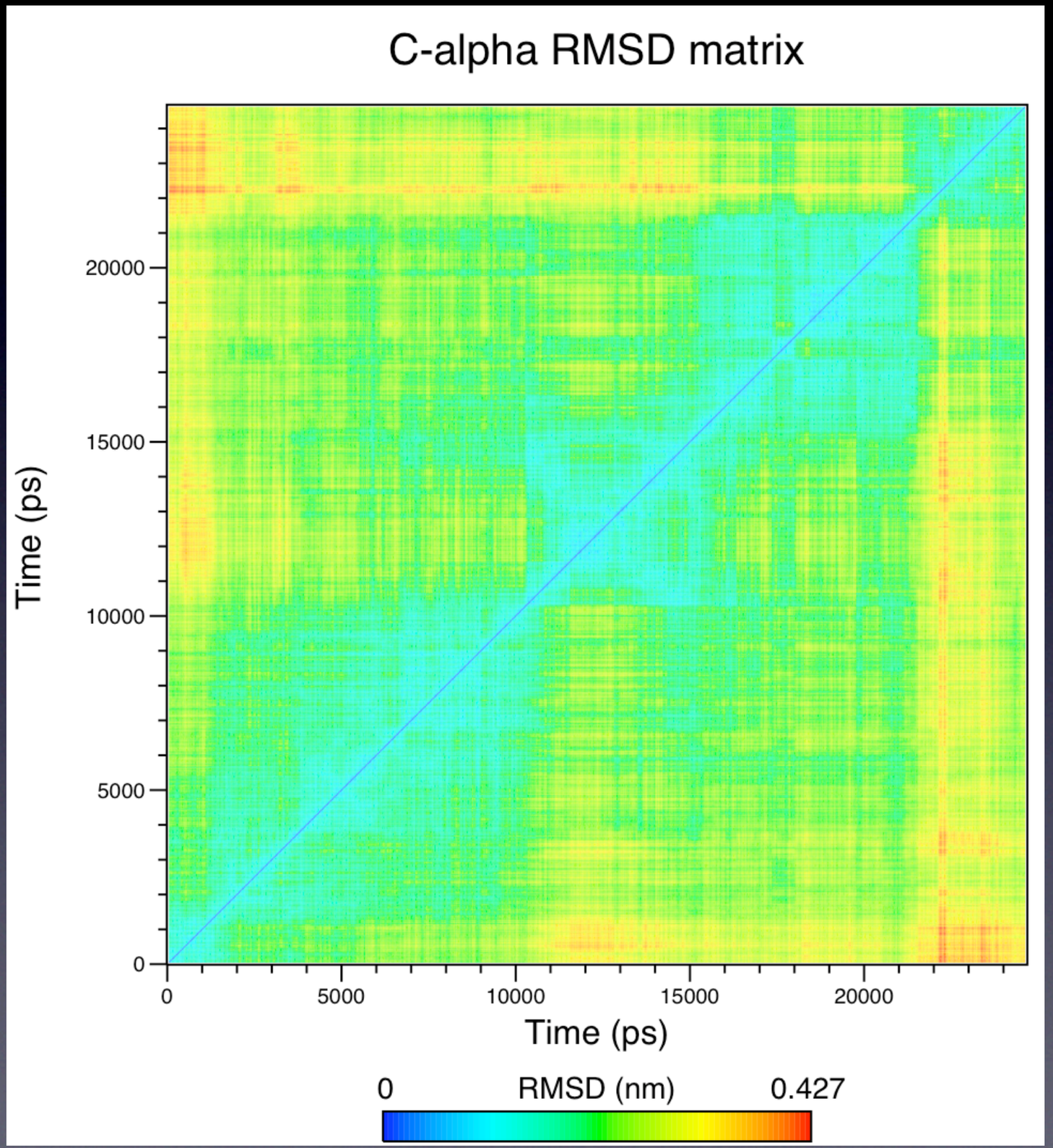
- Top→Down (Hierarchical)
- Bottom→Up (Single-linkage, Centroid-linkage, Average linkage ...)
- Refinement (Means)
- Jarvis-Patrick Clustering

Step 0 - Similarity matrix

Any clustering process starts with the building of a similarity matrix. In which one uses a distance function to calculate the distance between any two points.

- RMSD between structures
- RMSD between intramolecular distances
- Energy differences
-
-
-

Step 0 - Similarity matrix



Top → Down

Hierarchical

1. Assign all points to a single cluster.
2. Calculate the diameter of size of the clusters.
3. Split the biggest cluster around the two eccentric points A and B. All points closest to A are assigned to one child cluster and all points closer to B fall into the other.
4. If the desired cluster count has been reached, stop; otherwise, go to Step 2.

Top→Down

GROMOS (Daura *et al.* Angew. Chem. Int. Ed. 1999, 38, pp 236-240).

1. Calculate for each point the number of other points for which the RMSD is less then a given cutoff (neighbors).
2. The point with the highest number of neighbors as together with all its neighbors forms a cluster.
3. The points of this cluster are thereafter eliminated from the pool of not assigned points.
4. If the pool is empty, stop; otherwise, go to Step 1.

Bottom → Up

Single-linkage

1. Assign each point to its own cluster.
2. Calculate all the inter-cluster point-to-point distances between each cluster.
3. Choose the two clusters that have the shortest inter-cluster distance and merge.
4. If the desired cluster count has been reached, stop; otherwise go to Step 2.

Refinement

Means

1. Assign a given number of points as seeds such that the distance between the all points will be the largest available.
2. Assign the remaining points to the cluster which has the shortest distance to the cluster centroid.
3. Refine all points. For each point, remove it from its cluster unless it is the single member; update the centroid; reassign that point according to the new cluster centroids.
4. Repeat Step 2 until no new change occurs or for n times.

Jarvis Patrick

1. Determine the M nearest neighbors for each point.
2. Assign two points to the same cluster if:
    I. they are contained in each other's neighbor list.
    II. they have at least P nearest neighbors in common.

What you can find at GROMACS

- Single-linkage (Bottom→Up)
- Jarvis-Patrick
- Monte Carlo
- Diagonalization
- GROMOS (Top→Down ; Daura et al.)