GCMB Clustering

Note Title 5/30/2007

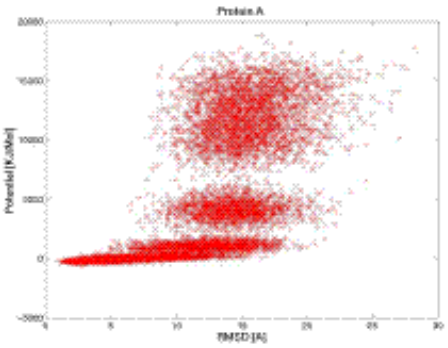
Motivation for clustering in computational biology

Configuration space

Simple clustering methods
- k-means
- k-medoids

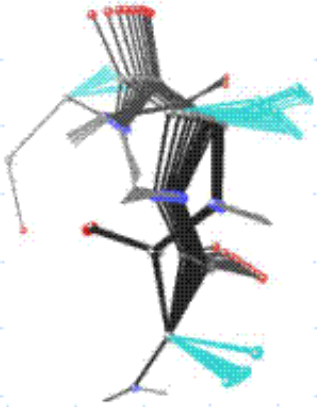
Clustering is hard

Motivation



Computers produce more examples than we can possibly study.

Random search algorithms and experimental measurements both produce many small variations of the "same thing".



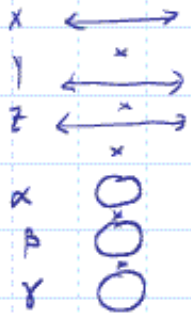
Configuration space

An obvious mathematical idea, named in robotics.

Represent an object having d degrees of freedom by a point in d-dimensional space.

e.g. an object in 3d has 3dof translation & 3dof rotation

so $p \in R^3 \times S^3$



Configuration space

An obvious mathematical idea, named in robotics.

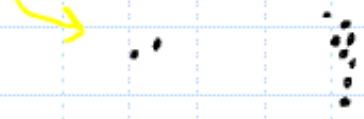
Represent an object having d degrees of freedom by a point in d-dimensional space.

e.g. an object in 3d has 3dof translation & 3dof rotation

so $p \in R^3 \times S^3$

This backbone fragment has six ϕ, ψ and one bond angle dofs

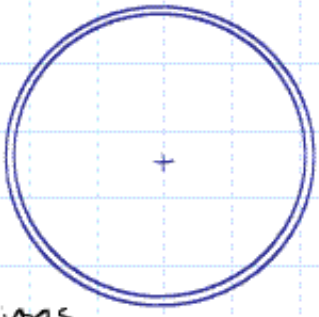
S^7 or $[a,b] \times S^6$



Can cluster points in configuration space

Warning: Curse of dimensionality

For points in a ball of radius 1,
What fraction of the points are
within 0.01 of the boundary?



d	fraction
2	1.990%
4	3.940%
8	7.726%
16	14.854%
32	27.502%
64	47.440%
128	72.375%
256	92.369%
512	99.418%
1024	99.997%
2048	100.000%

} protein backbones easily have this many degrees of freedom

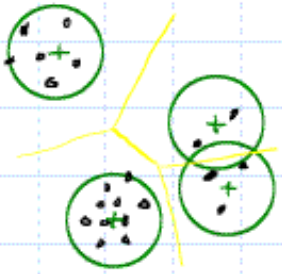
In high dimensions,
everything is far, far away.
The points we care about often
live on lower dimensional manifolds.

point clustering by distance

Definitions:

for points p, q distance $d(p,q) = \|p-q\|$
for point p, set Q distance $d(p,Q) = \min_{q \in Q} d(p,q)$

Given P, we could define clusters
by giving k "centers" $A = \{a_1 \dots a_k\}$
Cluster $C_i = \{p \in P \mid d(p,A) = d(p,a_i)\}$
i.e assign each point to nearest center
Cluster radius = $\max_i \max_{p \in C_i} d(p,a_i)$

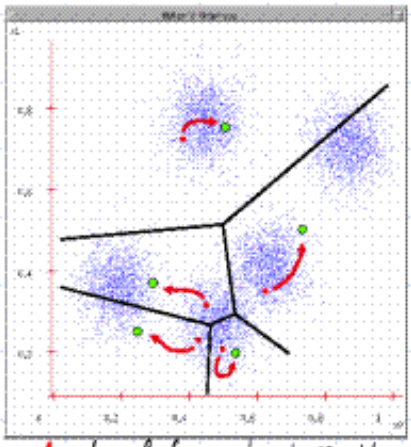


k-means clustering [MacQueen67]

Given $k>0$ and points P, find k cluster centers

k-means heuristic:

1. randomly choose initial $A=\{a_1 \dots a_k\}$
2. repeat
   3. assign clusters $C_i=\{p\in P \mid d(p,A)=d(p,a_i)\}$
   4. assign centers $a_i = \frac{\sum_{p\in C_i} p}{|C_i|}$
5. until convergence.



red pts define clusters, then move to green centroids

www.autonlab.org/tutorials/kmeans11.pdf

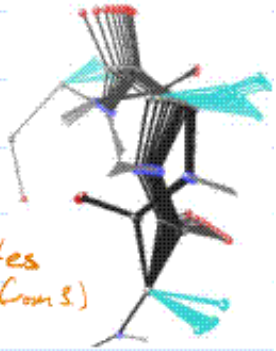
k-means clustering [MacQueen67]

Given $k>0$ and ~~points~~ protein backbones P, find k cluster centers

Need a distance $d(p,q)$ and an average $\frac{p+q}{2}$ to use k-means.

k-means heuristic:

1. randomly choose initial $A=\{a_1 \dots a_k\}$ from P
2. repeat
   3. assign clusters $C_i=\{p\in P \mid d(p,A)=d(p,a_i)\}$ (can use RMSD for coordinates)
   4. assign centers $a_i = \frac{\sum_{p\in C_i} p}{|C_i|}$ ← average coordinates (after rotation Coms)
5. until convergence.



(Justification for 4; [SW06])

Theorem 1. The weighted sum of squared distances for all pairs equals the weighted sum of squared distances to the average structure $\overline{S}$:

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m} w_k \left\| p_{ik} - p_{jk} \right\|^2 = n\sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik} - \overline{p}_k \right\|^2 .$$

k-~~means~~ medoids clustering [MacQueen 67]

To avoid averaging structures...

Given $k>0$ and points P,
find k cluster centers from P

k-means heuristic:

1. randomly choose initial $A=\{a_1 \ldots a_k\} \subset P$
2. repeat
   3. assign clusters $C_i=\{p \in P \mid d(p,A)=d(p,a_i)\}$
   4. assign centers $a_i \in P$ that minimizes $\sum_{p \in C_i} d(a_i,p)$
5. until convergence.

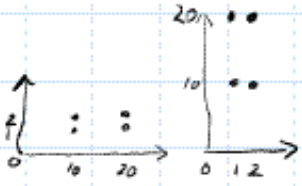
Cluster centers are chosen from P.

Clustering is hard

The curse of dimensionality.

Almost all variations are NP-hard;
heuristics & approximations are needed.

Reparameterizing the space can change clusters.



Our perception of clusters can be
based on higher-level organization
that is hard to capture mathematically.