What Does It Give Us?

- Generates a TSET & PSET.
- It is difficult to exactly get a TSET of specified size.
- As a result we need to vary c by trial and error.
- Once you have the TSET, randomly select a CVSET (if required) from it.

Dividing a Data Set

- Both the training & prediction set should be representative of the whole data set.
- Ideally, prediction set should also mirror the training set.

Methods for Set Selection

- Random.
- Activity Sampling.
- Clustering Methods.
  - KSOM.
  - K - means algorithm.
  - Kennard Stone.
  - Maximum Dissimilarity.

Problems with Clustering

- Different clusters have different density of points.
- Closeness of TSET & PSET is not gauranteed.

Sphere Exclusion

- Use *probe spheres* to set a similarity limit.
- Radii of the spheres is given by,

$$R = c\left(\frac{V}{N}\right)^{1/K}$$

- Depends on a user defined constant, c, called the *Disimilarity Level*.

Algorithm

1. Select compound with highest activity and add to TSET.
2. Construct sphere centered at this point, radius R.
3. All compunds within the sphere go into the TSET.
4. Exclude the points selected in 3 from the dataset.
5. If there are no more compounds, exit.

Algorithm

6. Calculate distance between all remaining compunds and all constructed sphere centers.
7. Select compunds with smallest (or largest) distance and go to step 2.

a

[a] A. Golbraikh et al, *J. Comp. Aid. Mol. Des.*