Performance of Different Algorithms on Clustering Molecular Dynamics Trajectories

Chenchen Song

Abstract

Different types of clustering algorithms are applied to clustering molecular dynamics trajectories to get insight about possible conformations for molecules. . The algorithms covered include K-means, EM (multivariate Gaussian), single-linkage, centroid-linkage, average-linkage, complete-linkage. Root-Mean-Square-Deviation(RMSD) is used as metric. Performances of algorithms are analyzed and compared based on Davies-Bouldin index(DBI) and Pseudo-F static (pSF).

1. Introduction

Molecular dynamics simulation methods produce trajectories of molecule configuration snapshots as a function of time. The time scale of chemical process is ns, while the time scale of molecule internal freedom is *fs*, thus a sufficient simulation trajectory will need to contain at least $O(N^6)$ configurations. For such large amounts of data, machine learning algorithm becomes helpful to extract useful information from the datasets.

One type of information we hope to get is the conformation substates of a molecule, which falls into the clustering problem. Using clustering algorithm to help analyze molecular dynamics trajectories is actually not a new idea, and can date back to 1993. Since then, a number of papers about applying different types of clustering algorithms on MD have been published. Thus, in this project, I will focus on comparing the performance of different clustering algorithm.

2. Method

Details of the methods have been carefully discussed in the milestone report. Here we only give a brief review.

(1) Similarity Metric

Instead of normal Euclidean norm, RMSD will be used as metric, which first tries to align two molecule as much as possible before calculating Euclidean distance. By using RMSD, we can eliminate the effect from translational and rotational motion of molecule.

(2) Algorithm

K-means and multivariate Gaussian(will be called EM for short) have been introduced in class.

The linkage methods are different in how the distance between clusters is defined. Single(edge)-linkage uses the shortest inter-cluster point-to-point distance. Centroid-linkage uses the distance between cluster centroids. Average-linkage: uses average distances between individual points of the two clusters. Complete-linkage uses maximal point-to-point distance.

Due to their different definition of critical distance, latter we will see that they sometimes can have very different behavior.

Single-linkage, complete-linkage, and average-linkage only need to calculate a metric matrix of size N^2 at the very beginning. Other methods need to update the positions of centroids and relative distances from points to centroids during each iteration. Because our metric is not a simple Euclidean distance, the latter methods will be more time-consuming.

(3) Performance Metric

DBI is defined as

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left\{D_{ij}\right\}, D_{ij} = \frac{\bar{d}_i + \bar{d}_j}{d_{i,j}} \qquad (1.1)$$

where d_{ij} is distance between centroids and $\bar{d}_i$ is the average distance between points in some cluster with the centroid of that cluster.

pSF is defined as:

$$pSF = \frac{SS_B}{SS_w}\frac{(N-k)}{(k-1)} \qquad (1.2)$$
$$SS_B = \sum_{i=1}^{k} n_i \| m_i - m \|^2, SS_w = \sum_{i=1}^{k}\sum_{x\in c_i} \| x - m_i \|^2$$

where m_i is centroid and m is the overall mean of data.

Usually, lower DBI and higher pSF reflects compact and well-separated DBI. But one should be careful when using these indices. For example, DBI is affected by cluster count, we should only compare DBI values when the number of clusters is similar.

3. Results and Analysis

Molecular dynamics trajectories are generated by Terachem package. RMSD calculation and molecular alignment is performed using VMD package. If the methods only requires N^2 metric matrix as discussed in the previous sections, then the clustering is performed by MATLAB statistic box. Otherwise, the clustering is performed by *Cluster3.0*, where we have added our metric into the clustering library. DBI and pSF are calculated by MATLAB.

3.1 Clustering points in 2D-plane from uniform distribution

First, different clustering methods are applied to a hundred points in 2D plane which are sampled from uniform distribution. The points don't have any internal

structures, thus the clustering results will only reflect the properties of different algorithm.

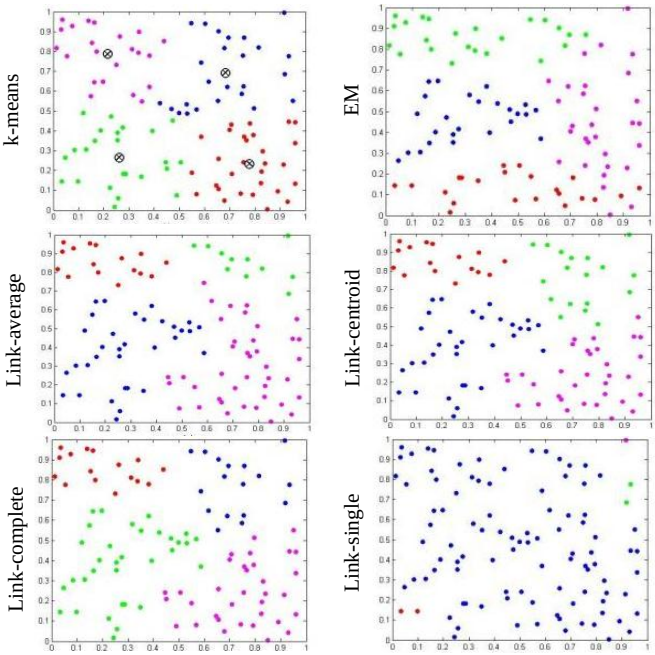
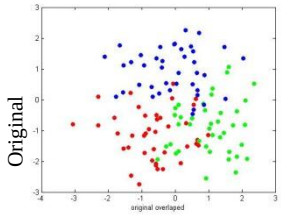


Figure 1. Clustering results for points in 2D-plane drawn from uniform distribution

From Fig.3, the following properties can be observed.
(1) Most of the methods tend to naturally and equally partition the points into four blocks. This is especially true for k-means.
(2) Single-linkage (or edge-linkage) almost classifies all the points into a same cluster. This might because the critical distance of single-linkage is defined as the nearest points between two clusters, thus single-linkage may be very sensitive to cases where points are close to each other and no clear border exists.
(3) EM method is the only one that produces clusters with very different shapes. This might be EM method does clustering based on probabilistic assumptions while all other methods are based on geometric structures.

3.2 Clustering points on 2D-plane from overlapped Gaussian distribution.

In the second step, clustering methods are applied to points sampled from three independent Gaussian distributions. The mean and covariance of 2D Gaussian is tuned so that the three distributions are overlapped.
The reason to test on overlapped points is that in MD simulations, the trajectories are generated consequently, thus adjacent configurations are usually very similar.



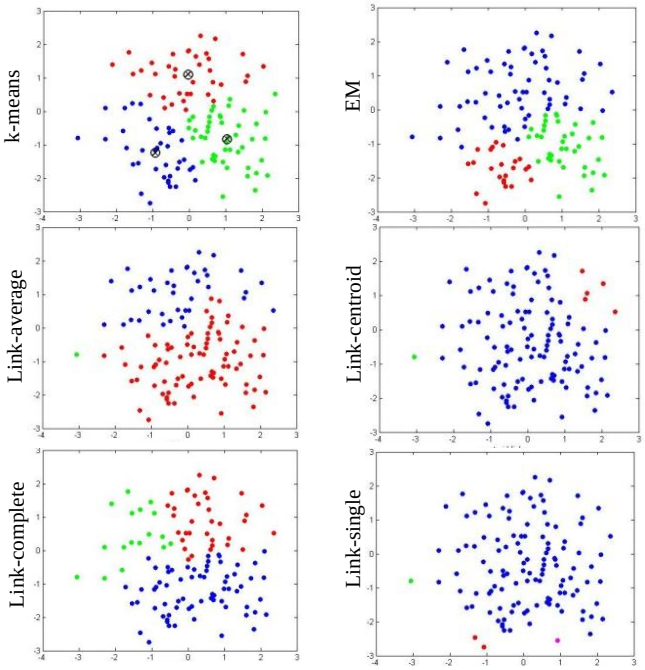


Figure 2. Clustering results for points in 2D-plane drawn from three equally sized overlapped Gaussian distribution.

Compared to the original figure, the following properties can be noticed.
(1) The success of k-means may again due to the internal property of k-means to produce clusters with same size and same shape.
(2) EM can get pretty good result if the underlying probability distribution is very close to Gaussian.
(3) Centroid linkage doesn't work well, perhaps because the centroids are ill-defined.
(4) Single-linkage seems to fail for this circumstance again where points have no clear borders.

3.3. Clustering Artificial MD data: Four equally sized clusters

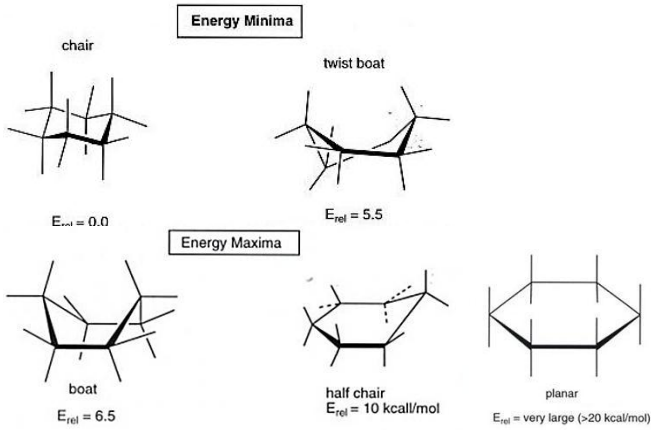


Figure 3. Illustrations of typical cyclo-hexane conformations.

Cyclo-hexane is used as a test model. Four clusters are chair, twisted boat, boat , and half-chair. To generate each cluster, take chair as an example, we start from a

chair configuration, control the timestep to 0.01fs, control the temperature to very low and only proceed 100 steps. Repeat this procedure several times, we can guarantee to generate a cluster with typical chair configurations.

Because the five different clusters are generated independently and artificially, there's actually a very clear border between different clusters.

The figure shows the expected idealized behavior including minima in the DBI, maxima in pSF when clustering number is equal to the optimal value of 4.

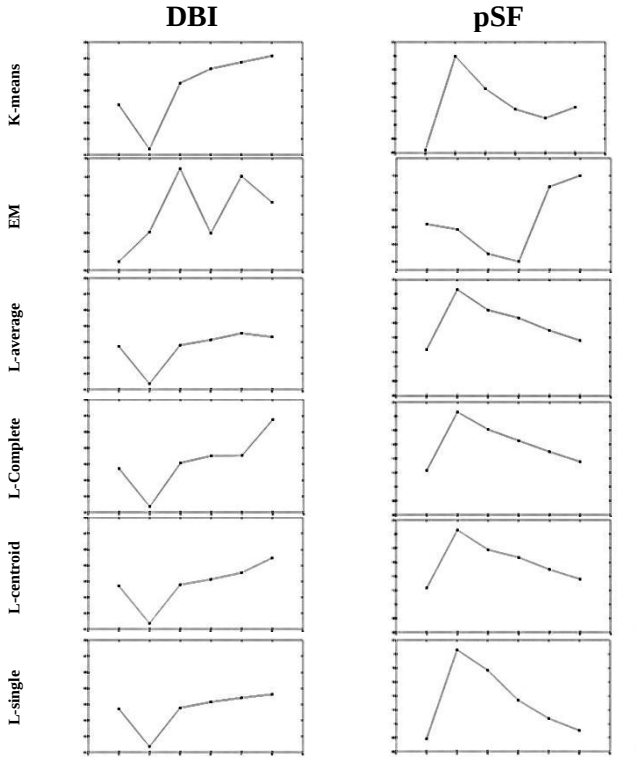


Figure 4.DBI and pSF for clustering artificial MD trajectories of cyclo-hexane. Optimal clusters are four equally sized clusters. X-axis range from 3 to 8.

By checking the assignments, it is found that for this test with equal clustering size and clear border, most of the algorithm perform very well except EM. This might be because Gaussian distribution is not a good assumption for how configurations distribute within each cluster around the centroid.

3.4 Clustering artificial MD data: Five differentially sized clusters.

In real MD simulations, the sizes of clusters can be very different, because the lower the energy is, the higher probability it will appear during simulation. To mimic this property, in this test, we build an artificial MD data by combining: 2 planar structures, 15 half-chair structures, 30 boat structures, 50 twist-boat structures, and 100 chairs. The order is also consistent with the energy order.

This set is much difficult to cluster as it has both very small clusters with small variance and relative large clusters with large variance.

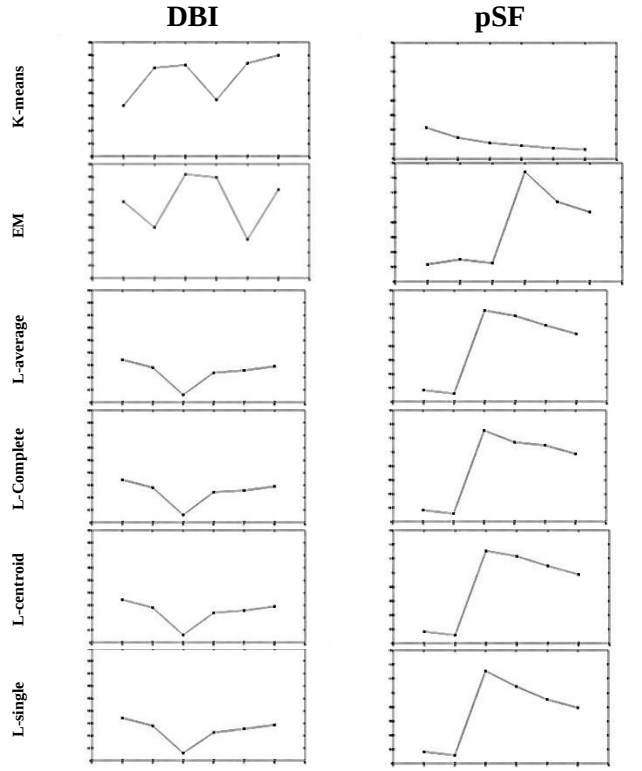


Figure 5.DBI and pSF for clustering artificial MD trajectories of cyclo-hexane. Optimal clusters are five distinct sized clusters. X-axis range from 3 to 8.

Table 1. Size of clusters for different algorithm with different number of clusters.

Method	#cluster	Cluster size
Kmeans	4	46 47 50 54
	5	13 41 46 47 50
	6	10 15 20 25 45 82
L-single	4	2 15 30 150
	5	2 15 30 50 100
	6	2 15 (2, 28) 50 100
L-average	4	2 15 30 150
	5	2 15 30 50 100
	6	2 15 (12, 18) 50 100
L-centroid	4	2 15 30 150
	5	2 15 30 50 100
	6	2 15 (12, 18) 50 100
L-complete	4	2 15 30 150
	5	2 15 30 50 100
	6	2 15 (5, 25) 50 100
EM	4	34 66 108 192
	5	21 32 47 100 200
	6	43 49 51 57 100 100

It can be noticed that when configurations are not uniformly separated, the metrics are less consistent and not necessary gets optimal value at optimal cluster numbers.

From the table and figure the following properties could be observed

(1) For this tests where clear border exists but cluster sizes are very different, all the linkage methods are able to recover the correct assignment at optimal cluster number 5. When cluster number increases, different methods then split clusters in different way due to their different distance definition.

(2) K-mean exhibits a strong tendency to cluster points into equal size, thus doesn't give good performance for this tests.

(3) EM also gives pretty bad result, possibly due to same reason as previous tests.

3.5 Clustering real MD data.

Unlike the artificial trajectories in which configurations are clearly separated, configurations from adjacent steps in real MD simulations are very similar and will make clustering to be more difficult.

Cyclohexane is again used as the test examples. Multiple trajectories are generated starting from half-chair structure. After clustering each trajectories individually, two distinct typical types of behavior is shown below.

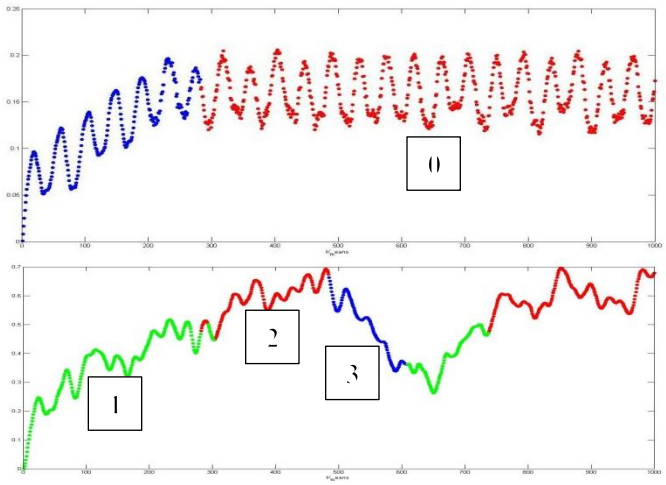
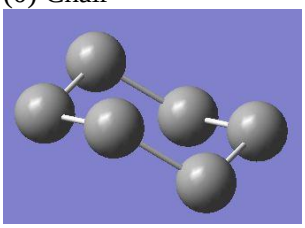
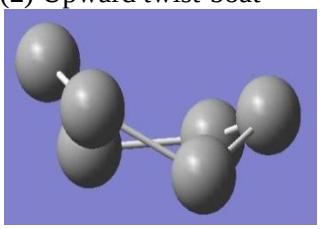


Figure 6. Two distinct types of behavior and clustering results when clustering real MD trajectories of cyclo-hexane starting from half-chair conformation.

(0) Chair

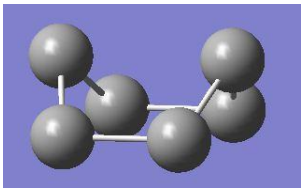


(2) Upward twist-boat



(3) boat

(4) Downward twist-boat



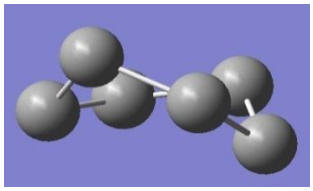


Figure 7. Centroids for each cluster.

To explain this behavior, the energy profile for a continuous changing cyclohexane is checked.

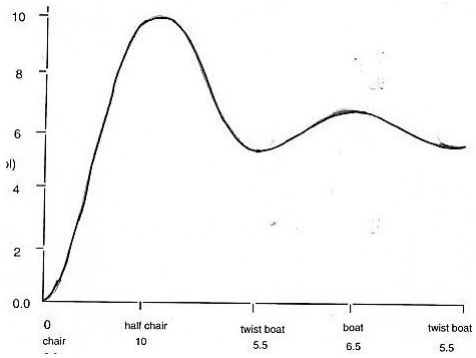
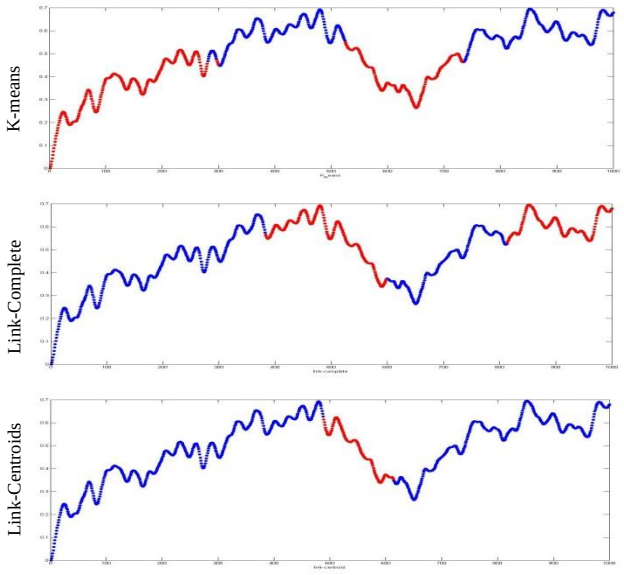


Figure 8. Cyclohexane energy profile.

It can be seen from the figure that, if starting from half-chair, it can either goes to the left, the chair region(if the plat part flips downwards) or goes to the right, the boat region(if the plat fart flips upwards). Once it steps to the left, it will be separated from the other part by a high energy barrier. Similar for stepping into the right region.

All methods give the almost the same clustering when cluster number is three. If we set cluster number to two, methods will have distinct results.



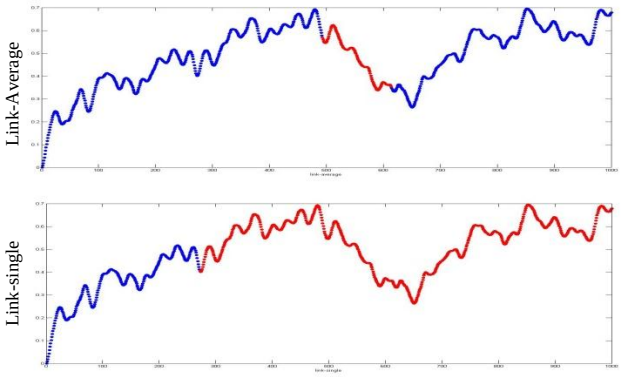


Figure 8. Different clustering results from different algorithm when number of cluster is 2.

It can be seen from the figure that :
(1)K-means and linkage-complete method merges upward and downward twist-boat. These two methods emphasize more on the difference between boat and twist-boat.
(2) Link-centroids and link-average merge boat and upward twist boat. and weigh more on the difference between reversed configurations.
(3) Link-single doesn't preform very well for this tests, perhaps because it cannot handle circumstances where clear border is absent.

3.6 Clustering protein trajectories

Finally, clustering method is applied to protein trajectories. Complete-linkage method is used, because it performs well in the previous tests and only requires a N^2 metric matrix, without necessity to compute updated centroids during each iteration. Optimal clustering number is picked out by pSF indices.
Considering the available computing ability, we apply a pulling force to the protein so that the structure will change more rapidly. Only the coordinates of backbones (carbon, nitrogen, oxygen) are passed into the clustering program. Results are shown below.

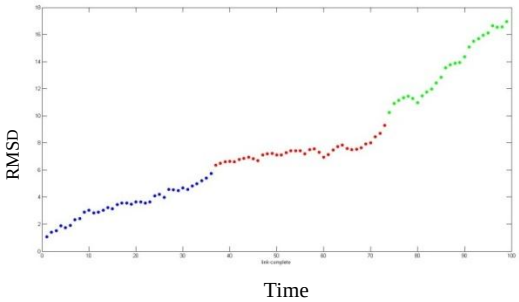


Figure 9. Clustering results for protein.

The three clusters clearly show the transformation from folded, half-unfolded to completely unfolded under the pulling force we apply to the system.

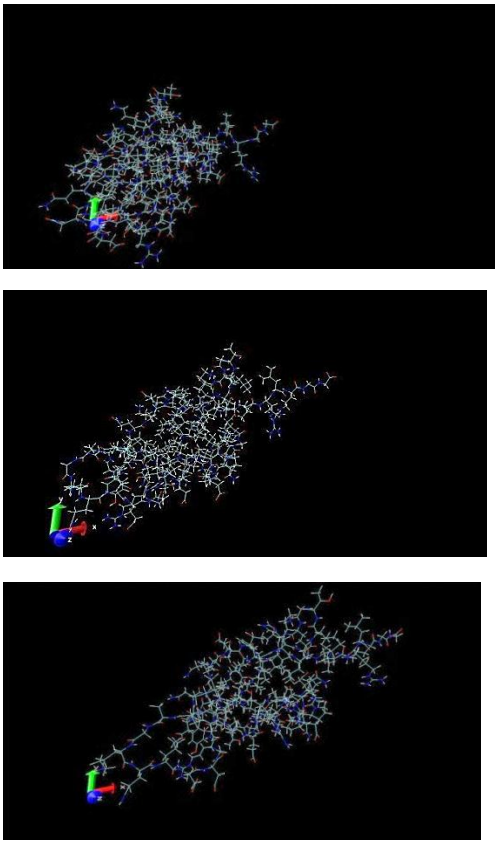


Figure 10. Centroids for three clusters.

Summary

In this project, we use several tests to compare and analyze the performance of different clustering methods under various conditions. The following properties can be summarized from our observations:
(1) K-means tends to produce blocky clusters of similar size. Thus when cluster sizes are similar, K-means gives very good performance. But it usually fails for clusters with distinct sizes.
(2) Single-linkage is very sensitive to closely spaced points. As a result it may be fragile to the presence or absence of single point.
(3) Multivariate Gaussian doesn't work well for most of the tests, perhaps because the assumption of Gaussian distribution is not good in our problem
(4) Centroid-, average-, and complete-average gives quite consistency good results through the results. The latter two doesn't require updating centroids during each iteration, thus may be candidates for clustering molecular dynamics trajectories.

References

1. Karpen, M. E.; Tobias, D. J.; Brooks, C. L. .*Biochemistry*, **1993**, 32, 412-420.

2. Kabsch, W..*Acta Crystallogr*. **1976,** 32:922–923.

3. Ramanathan,A; Yoo, J.O.; Langmead, C.J.; *J. Chem. Theory Comput.***2011**, 7, 778–789

4. Shao, E, et.al. *J. Chem. Theory Comput.***2007**,3,2312-2334