Abstract

Motivation: Discovering sequence motifs involves grouping relatively similar protein segments from a huge collection of protein sequences. A granular computing strategy combined with K-means clustering algorithm was previously proposed for the task, but, the strategy requires a manual selection of biologically meaningful clusters, which are undesirable clustering methods.

Results: We utilize sparse non-negative matrix factorization (SNMF) to cluster large protein datasets. We show how to combine the method with Fuzzy C-means algorithm and incorporate bio-statistics to increase the percentage of protein segments clusters with high structural similarity. Our experimental results show that SNMF provides better protein groupings in protein structures maintaining the similarities in protein sequences.

Sparse Non-Negative Matrix Factorization (SNMF) and clustering

Discovering motifs has two steps; clustering, and motif finding. This project focuses on clustering a huge protein profile segments so that a center of each cluster can be a meaningful sequence motif.

Non-negative matrix factorization (NMF) is the process to factor a data matrix into two non-negative matrices so that each data can be represented as addition of basis vectors.

Each sequence profile (Fig. 1) is rearranged into a vector and the whole data forms a data matrix $A$. In SNMF (Eq.1), $A$ is factored by a basis matrix $W$ and a coefficient matrix $H$.

$$\min_{W,H} \frac{1}{2} \left\{ \| A - WH \|_F^2 + \eta \| W \|_F^2 + \beta \sum_{j=1}^{m} \| H(:,j) \|_2^2 \right\}$$

subject to $W \geq 0$, $H \geq 0$.

Eq.1 SNMF with sparseness on $H$

Fuzzy C-Means

Fuzzy C-Means (Eq.2) is soft clustering algorithm so that it divides the whole data into a set of small number of subsets allowing some overlapping, to make parallel computation possible. SNMF is more expensive than K-means, and it applies better for the problem with smaller number of clusters.

Therefore, we applied FCM hierarchically twice (Fig.2) before applying SNMF so that each smaller data set can be clustered into a small number of subsets.

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{C} H_{ij}^m \| X_i - C_j \|_2^2$$

Eq.2 Fuzzy C-Means (FCM) Clustering Algorithm

Chou-Fasman parameters

Improved K-means (Zhong et al.) and Greedy K-means (Chen et al.) used supervised methods by managing the initial points to improve the secondary structure similarity in each cluster. This is undesirable as clustering should be unsupervised. Since, initial points affect little the SNMF results, we do not need initial manipulation. However, we need a way to give the data structural information related to its primary sequence.

Therefore, we incorporate the Chou-Fasman parameter (Tb.1) which is the statistics of amino acid per a secondary structure, introduced by Chou and Fasman.

Fuzzy C-Means (Eq.2) is soft clustering algorithm so that it divides the whole data into a set of small number of subsets allowing some overlapping, to make parallel computation possible. SNMF is more expensive than K-means, and it applies better for the problem with smaller number of clusters.

Therefore, we applied FCM hierarchically twice (Fig.2) before applying SNMF so that each smaller data set can be clustered into a small number of subsets.

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{C} H_{ij}^m \| X_i - C_j \|_2^2$$

Eq.2 Fuzzy C-Means (FCM) Clustering Algorithm

Chou-Fasman parameters

Improved K-means (Zhong et al.) and Greedy K-means (Chen et al.) used supervised methods by managing the initial points to improve the secondary structure similarity in each cluster. This is undesirable as clustering should be unsupervised. Since, initial points affect little the SNMF results, we do not need initial manipulation. However, we need a way to give the data structural information related to its primary sequence.

Therefore, we incorporate the Chou-Fasman parameter (Tb.1) which is the statistics of amino acid per a secondary structure, introduced by Chou and Fasman.