Expert Systems with Applications 38 (2011) 13198–13207

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Sparse nonnegative matrix factorization for protein sequence motif discovery ☆

Wooyoung Kim [a,*], Bernard Chen [b], Jingu Kim [c], Yi Pan [a,*], Haesun Park [c]

[a] Department of Computer Science, Georgia State University, Atlanta, GA 30302, USA
[b] Department of Computer Science, University of Central Arkansas, Conway, AR 72035, USA
[c] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Keywords:
Sparse non-negative matrix factorization
Clustering
Protein sequence motif
Fuzzy *C*-means
Chou–Fasman parameters

ABSTRACT

The problem of discovering motifs from protein sequences is a critical and challenging task in the field of bioinformatics. The task involves clustering relatively similar protein segments from a huge collection of protein sequences and culling high quality motifs from a set of clusters. A granular computing strategy combined with *K*-means clustering algorithm was previously proposed for the task, but this strategy requires a manual selection of biologically meaningful clusters which are to be used as an initial condition. This manipulated clustering method is undisciplined as well as computationally expensive. In this paper, we utilize sparse non-negative matrix factorization (SNMF) to cluster a large protein data set. We show how to combine this method with Fuzzy *C*-means algorithm and incorporate bio-statistics information to increase the number of clusters whose structural similarity is high. Our experimental results show that an SNMF approach provides better protein groupings in terms of similarities in secondary structures while maintaining similarities in protein primary sequences.

1. Introduction

Proteins are vital parts of organisms, providing structural or mechanical functions and participating in every process within cells such as cell signaling, immune responses, and the cell cycle. Proteins are complete biological molecules in a stable conformation and are made of twenty possible amino acids arranged in a linear chain. The chemical interactions of amino acid residues determine the conformation of proteins and form a relationship between protein sequences and structures. Therefore, understanding the close relationship between protein sequences and structures by discovering its hidden knowledge has been one of the primary interests in bioinformatics research.

A protein sequence motif is a recurring pattern in sequences that is prevalent in a number of proteins. Protein motifs are known to have biological significance such as binding sites and conserved domains. If a sequence motif is in the exon of a gene, it can encode a *structural motif* which is a three dimensional motif determining a unique element of the overall structure of a protein. With this property, sequence motifs can predict other proteins' structural or functional behaviors. Therefore, discovering sequence motifs is a key task to comprehend the connection of sequences with their structures.

PROSITE (Hulo et al., 2004), PRINTS (Attwood et al., 2002) and BLOCKS (Henikoff, Henikoff, & Pietrokovski, 1999; Henikoff, Henikoff, & Pietrokovski, 1999) are currently the most popular motif databases. However, since the sequence motifs from these servers search through the same protein family members, they might carry little information about the consensus region beyond protein families (Zhong, Altun, Harrison, Tai, & Pan, 2005). On the other hand, many software programs for discovering one or more candidate motifs from a number of nucleotide or protein sequences have been developed. These include PhyloGibbs (Siddharthan, Siggia, & van Nimwegen, 2005), CisModule (Zhou & Wong, 2004), WeederH (Pavesi, Zambelli, & Pesole, 2007), and MEME (Bailey & Elkan, 1994). For example, MEME utilizes hidden Markov models (HMM) to generate statistical information for each candidate motif. However, such tools can handle only small to medium scale data sets and inappropriate for huge data sets.

In order to obtain universally preserved sequence patterns across protein family boundaries, we use an extremely large data set collected from various protein families. After collecting a number of protein sequences, their protein family information is ignored in further processing. Therefore, the task of discovering protein motifs is mainly divided into three steps: collecting all the possible protein segments with a fixed window size, clustering the segments, and evaluating the quality of discovered motifs with respect to its structural closeness. Collecting all the possible protein segments are completed in previous studies by Chen, Tai,

☆ This document is a collaborative effort.

* Corresponding authors. Tel.: +1 404 413 5703.

E-mail addresses: wkim@cs.gsu.edu (W. Kim), bchen@uca.edu (B. Chen), jingu@cc.gatech.edu (J. Kim), pan@cs.gsu.edu (Y. Pan), hpark@cc.gatech.edu (H. Park).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.133

Harrison, and Pan (2006), Chen, Tai, Harrison, and Pan (2006), Chen, Pellicer, Tai, Harrison, and Pan (2008), and Chen and Johnson (2009) using a sliding window technique from a protein profile data set. After clustering, evaluating the quality of discovered motifs is conducted by comparing the secondary structures in each cluster.

Therefore, clustering protein segments is the most challenging and crucial task. Previously, *K*-means clustering algorithms with supervised initial points were proposed by Zhong et al. (2005) and Chen et al. (2006), Chen et al. (2006), Chen et al. (2008), and Chen and Johnson (2009). These methods improve on an earlier approach where naive *K*-means algorithm was used by Han and Baker (1983). The improved *K*-means approach proposed in Zhong et al. (2005) increased the number of clusters having high structural homology, by selecting 'good' initial points from a number of preliminary results obtained by using a *K*-means algorithm with random initial seeds. Utilizing a granular computing strategy to divide the original data set into smaller subsets and introducing a greedier *K*-means algorithm (Chen et al., 2006, Chen et al., 2006), or subsequent filtering process with support vector machine (Chen et al., 2008, Chen & Johnson, 2009), Chen et al. further improved the overall quality of the clusters in terms of biological, chemical, and computational meanings. Those high quality of motifs are used to predict local tertiary structure of proteins in Chen and Johnson (2009) as well. However, these clustering techniques are undisciplined, insecure, and computationally expensive. They are actually supervised methods since they plug good initial cluster centers, which are evaluated and selected after several runs, into a final *K*-means algorithm. Also, the selection process requires repeated runs of *K*-means and additional user setups, which increase the computational costs.

In this paper, we propose to use sparse nonnegative matrix factorization (SNMF) (Kim & Park, 2008, Kim & Park, 2007) to cluster the protein segments data set. Originally proposed as a dimension reduction method for nonnegative data, NMF has been successfully applied to several tasks in computational biology described by Devarajan (2008). Areas of application include molecular pattern discovery, class prediction, functional analysis of genes, and biomedical informatics. As an extension of NMF, SNMF which imposes sparsity constraints on the low dimensional factors showed superior results for microarry data analysis with computational efficiency as demonstrated in Kim and Park (2007). Recently, Kim and Park demonstrated that SNMF was able to produce more consistent results than *K*-means with random initial seeds (Kim & Park, 2008), because SNMF tends to converge with any initial setups, while *K*-means algorithm is very sensitive to its initial setups. Additionally, we show how to incorporate a bio-statistics to improve the results with its high structural similarity. Unlike the previous methods, we avoid using the secondary structure of the data being studied in the process of clustering as the structure should be used only for evaluation. Instead, we use Chou–Fasman parameters, statistical information on existing protein data which do not require knowing of the secondary structure of the proteins being studied.

The work in this paper makes four contributions in the study of molecular biology. First, we explore the use of SNMF to a new problem domain, protein profiles. NMF has been used for various data including image, text, microarry gene or protein expression data. As far as we know, this is the first time that NMF has ever been applied to a protein profile data set. Even with the same SNMF algorithm, adjustment of parameters to different data format was a challenge. Second, we adopt Chou–Fasman parameters (Chou & Fasman, 1974; Chou & Fasman, 1978) which give us the statistical relationship between sequence and secondary structure and improve the quality of resulting motifs. It is also shown that the inclusion of Chou–Fasman parameters itself is a powerful tool to improve the quality of clusters even with a *K*-means algorithm with random initial seeds. Third, by applying granular computing strategy, we were able to overcome the issues with obscure assignments with SNMF method for large data sets. The final contribution is designing a new measurement which evaluates the quality of motifs based on a 'statistical' structural information inferred from its primary sequence. With this measurement, we can evaluate its structural significance without loss of sequential similarity.

Combining all the techniques aforementioned, the work conducts the following tasks. First, we use granular computing to split the extremely large collection of protein segments into smaller subsets and then use the Chou–Fasman parameters to add an analyzed structural information. SNMF is applied to each small subset in parallel. Our experimental results demonstrate that SNMF produces better results in terms of their structural agreements than other previous methods conducted in Chen et al. (2006), Chen et al. (2006), and Zhong et al. (2005). The remaining paper is organized as follows. We review some of the previous methods closely related to our method in Section 2. Then in Section 3, we introduce NMF and SNMF, and explain the use of Fuzzy *C*-means clustering and Chou–Fasman parameters. The experiments and the final results are described in Section 4, followed by the conclusion and future research work in Section 5.

2. Related works

DNA or protein sequence motifs have been discovered through the studies of evolutionary conservation by de novo computation with various tools such as MEME (Bailey & Elkan, 1994), CisModule (Zhou & Wong, 2004), PhyloGibbs (Siddharthan et al., 2005) and WeederH (Pavesi et al., 2007). However, these programs take proteins from the same protein families. Therefore, they are unequipped to discover patterns appearing across protein family boundaries. Expecting that protein motifs carrying biological significance can be found from different protein families as well, *K*-means clustering algorithms have been utilized for a large data set of proteins from diverse protein families in Chen et al. (2006), Chen et al. (2006), Chen et al. (2008), Chen and Johnson (2009), and Zhong et al. (2005).

K-means clustering algorithms are efficient for large data sets, but performance is sensitive to initial points and the order of instances. Peña et al. compared four different initialization methods for *K*-means algorithm: *Random, Forgy, MacQueen* and *Kaufman* in Peña, Lozano, and Larrañaga (1999). The random method initializes a partition of *K* clusters at random, while the Forgy method (Forgy, 1965) randomly chooses *K* seeds as initial cluster centers and assigns each data to a cluster of the nearest seed. Macqueen (1967) initialization strategy selects *K* random seeds but assignments follow an order of the seeds. The Kaufman method (Kaufman & Rousseeuw, 2005) successively picks *K* representative instances by choosing the center as the first one. According to the study in Peña et al. (1999), Random and Kaufman methods outperform the other two methods, and the Kaufman method is faster than Random method. However, due to the stochastic nature of the large data used in the work of discovering motifs, Zhong et al. used the Forgy method as a traditional *K*-means in Zhong et al. (2005). Throughout this paper, the *K*-means with random initial seeds refers to the Forgy initialization strategy. Previously, Han and Baker (1983) utilized a *K*-means clustering with a random initial seeds to find protein motifs. Subsequently, Zhong et al. (2005) introduced an improved *K*-means algorithm that greedily chooses suitable initial centers so that final partition can reveal more protein motifs with structural similarity. However, the good initial centers are selected from the resulting clusters obtained through previous *K*-means. Also this method requires two additional user inputs, a threshold for structural similarity *hs*, and a minimum distance between cluster centers *md*. That is, after a number of *K*-means, they select initial points having both produced the clusters whose structural similarity are higher than *hs* and whose distance between already selected initial points in the previous run is farther than *md*. All the selected initial points are applied to the final *K*-means clustering algorithm. Although the *Improved-K-means* method was able to obtain more valuable clusters with higher structural homology (over 60%) than a traditional *K*-means algorithm, this method actually supervised and led the results with manual selection of the cluster centers.

For further improvements, Chen et al. utilized granular computing introduced in Chen and Johnson (2009), Chen et al. (2008), Lin (2000), Yao (2001) and combined improved *K*-means or greedy *K*-means to develop the FIK model (Chen et al., 2006) and the FGK model (Chen et al., 2006; Chen & Johnson, 2009; Chen et al., 2008), respectively. Fuzzy-Improved-*K*-means (FIK) model (Chen et al., 2006) and Fuzzy-Greedy-*K*-means (FGK) model (Chen & Johnson, 2009; Chen et al., 2008, 2006) are granular based learning models used for the same task but for a larger data set than that of the improved *K*-means (Zhong et al., 2005). FIK and FGK both used Fuzzy *C*-means (FCM) algorithm for granular computing. FCM is a soft clustering algorithm which allows a data point to belong to one or more clusters (Dunn, 1973; Bezdek, 1981). FIK and FGK model divided the original data set into smaller subsets with FCM and slightly modified the improved *K*-means (Zhong et al., 2005) to each subset. While improved *K*-means selected initial seeds sequentially, FIK collects all 'candidate' initial seeds from all of the preliminary *K*-means, then selects the ones which frequently appear and are reasonably distant from the other seeds. With FGK, the selection is more greedy by selecting the high quality of clusters first. However, although FIK and FGK models produced better results than the improved *K*-means, they are still manipulating the results by plugging good initial points into a final *K*-means.

3. New approaches

The previous models discussed in Section 2 used *K*-means clustering algorithms with various initialization strategies. Instead of the *K*-means methods, we propose a different clustering algorithm called *sparse nonnegative matrix factorization (SNMF)*. We will first review an original NMF algorithm and introduce an SNMF algorithm which is used for clustering by enforcing sparseness on one of the factor matrices. Then our method will be described in the subsequent sections.

3.1. Nonnegative matrix factorization

Nonnegative matrix factorization (NMF), which was first introduced as a positive matrix factorization (PMF) by Paatero and Tapper (1994), is a matrix analysis that has attracted much attention during the past decade. Besides NMF, there are several matrix factorization methods used in many applications, including principal component analysis (PCA) and vector quantization (VQ). All of those matrix factorization methods represent the data by using vectors and form a data matrix to decompose it into two factor matrices. Ross and Zemel (2006) noted that when data are represented as vectors, parts of the data are interpreted with subsets of the bases that take on values in a coordinated fashion. Although other factorization methods are related to this interpretation in general, only NMF has a sparse and part-based localization property (Lee & Seung, 1997; Lee & Seung, 1999), but under special conditions (Donoho & Stodden, 2004). NMF is considered for high dimensional data where each entry has a nonnegative value, and it provides a lower rank approximation formed by factors whose entities are also nonnegative. NMF was successfully applied to analyzing face images (Lee & Seung, 1999; Li, Hou, Zhang, & Cheng, 2001), text corpus (Pauca, Piper, & Plemmons, 2006), and many other tasks in computational biology (Devarajan, 2008).

Given an $m \times n$ data matrix A, nonnegative factors such as W, H are commonly computed by solving

$$\min_{W,H} \frac{1}{2}\|A - WH\|_F^2 \text{ s.t. } W \geqslant 0, \quad H \geqslant 0, \tag{1}$$

where W is the $m \times k$ bases matrix, H is the $k \times n$ coefficient matrix, and k is usually much smaller than min (m,n). The interpretation of factored matrices, W and H, depend on the domain of application. For instance, if the data matrix A denotes microarry data, the rows correspond to expression levels of genes and the columns correspond to samples representing distinct tissues, experiments, or time points. Thus, $A(i,j)$ describes the ith gene expression level for the jth sample. If the microarry data matrix A is factored into W and H using NMF, each column of W defines a metagene and each column of H represents the metagene expression pattern of the corresponding sample. In this case, the metagenes of W summarize gene behaviors across samples, while the patterns of H summarize the behavior of samples across genes. On the other hand, when clustering data samples, each basis of W can represent a prototype of each cluster and each column of H is the relevance of the data sample corresponding to each prototype.

The nonnegativity of W and H provides a pleasing interpretation of the factorization. Each object is explained by an additive linear combination of intrinsic 'parts' of the data (Lee & Seung, 1999). This property of NMF gives an intuitive meaning and physical interpretation, especially for large-scale data, while the orthogonal components with arbitrary signs in PCA lack their conceptual interpretation. In face image applications with SNMF (Lee & Seung, 1999), the column vectors of W represent each component of the face, that is, nose, eyes, cheeks, etc. In addition to the natural interpretability as a dimension reduction method, NMF has shown favorable performance for clustering tasks. For text clustering, Xu, Liu, and Gong (2003) reported competitive performance of NMF compared to other methods in spectral clustering. Brunet, Tamayo, Golub, and Mesirov (2004) used NMF on cancer microarry data and demonstrated its ability to detect cancer classes.

3.2. Sparse nonnegative matrix factorization

Generally, NMF provides sparse and part-based representations, but this may not always be the case. Li et al. (2001) and Hoyer (2002) presented part-based but holistic (non-local) representations produced by NMF. These results exemplify that nonnegativeness is an insufficient condition to produce sparse representations. Therefore, many studies (Li et al., 2001; Hoyer, 2002; Hoyer, 2004; Gao & Church, 2005) focused on enforcing sparseness explicitly on W, H or both. Kim and Park (2007), Kim and Park (2008) proposed a sparse NMF (SNMF) using a refined formulation with an additional penalty term and proposed an efficient algorithm. SNMF was further studied by Kim and Park (2008), where they demonstrated that SNMF gives more consistent clustering results than *K*-means algorithm.

In this paper, we use SNMF with sparseness enforced on the right factor H used by H. Kim and Park in Kim and Park (2007) to cluster protein profile segments. We note that they (Kim & Park, 2007, 2008) also provided the SNMF with sparseness enforced on the left factor W (SNMF/L), but SNMF/L is useful for representing part-based bases.

Sparse Nonnegative Matrix Factorization
Given a nonnegative matrix A, find nonnegative matrix factors W and H such that;

$$\min_{W,H} \quad \frac{1}{2}\left\{\|A - WH\|_F^2 + \eta\|W\|_F^2 + \beta\sum_{j=1}^{m}\|H(:,j)\|_1^2\right\} \qquad (2)$$

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

where $\|.\|_F^2$ is the square of the Frobenius norm, $\|.\|_1^2$ of the L_1 norm, and $H(:,j)$ is the jth column of matrix H. Regularization using L_1-norm promotes sparse solutions on the factor H. Two parameters, η and β, are involved, where η suppress the Frobenius norm of W, and β regulates balances between the sparseness of matrix H and the accuracy of the factors. In practice, the parameters are adjusted empirically as they are affected by the data size and the number of clusters. Alternating nonnegativity constrained least squares algorithm using the active set method (Kim & Park, 2007) was used to obtain W and H.

In the SNMF application for clustering protein profile data, we represent each protein segment as an m-dimension vector. By arranging n number of data to column-wise, we form an $m \times n$ data matrix A. After deciding the number of clustering k, we use alternating nonnegativity constrained least square algorithm, and factor A into W and H factor matrices. Then each column of H is k-dimensional vector, ith entry representing the relevance of ith cluster of the corresponding sample. Each data is then assigned to the cluster of maximum relevance. Detail description of data representation and experiment steps with SNMF is illustrated in Section 4.

3.3. Granular computing and Fuzzy C-means clustering

Chen et al. (2006), Chen et al. (2006) proposed a granular computing model (Lin, 2000; Yao, 2001) by utilizing Fuzzy C-means (FCM) clustering algorithm as described in Section 2. Granular computing involves the processing of complex information granules arising in the process of data abstraction and the derivation of knowledge from the data. FCM (Dunn, 1973; Bezdek, 1981), known as a common method for granular computing, is a clustering algorithm that allows a data point to belong to more than one cluster. Therefore, FCM is used as a preprocessing step as it splits the data with softer constraints. FCM clusters N data points, x_i's, into C clusters by minimizing the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \|x_i - c_j\|^2, \qquad (3)$$

where m is the *fuzzification factor* and u_{ij} is the degree of participation of x_i into the cluster j with a center c_j. Then the number of clusters for each information granule divided by FCM is computed as

$$C_k = \frac{n_k}{\sum_{i=1}^{m} n_i} \times \text{total number of clusters}, \qquad (4)$$

where C_k is the number of clusters and n_k is the number of members for the kth information granule.

FIK (Chen et al., 2006) and FGK (Chen et al., 2006) models applied FCM with empirically chosen fuzzification factor and the number of clusters, then applied K-means to each information granule with manually chosen initial points. They highlighted that not only the manual selection of initial points, but also the FCM process itself improved the final results due to its pre-filtering work as shown in Table 2.

In the present work, we apply the SNMF method instead of variant K-means algorithms, because a sparse coefficient matrix can assign each data to one of the clusters. Therefore less number of clusters can produce more desirable results with SNMF. If the number of clusters is numerous, then many of the relevant factors hold similar weights, thereby obstructing proper assignments. Fig. 1 shows one example of obscure assignment by comparing with the case of $k = 3$ and that of $k = 45$. The bottom image of Fig. 1 visualizes the H factor with 45 rows. One data shown within a red box has 45 weights, but not one value is promising enough to assign the data to a particular set. The top image of Fig. 1 only has 3 rows and the second value is prominent enough to cluster the data of red box into the second set. Hence, the granular computing with

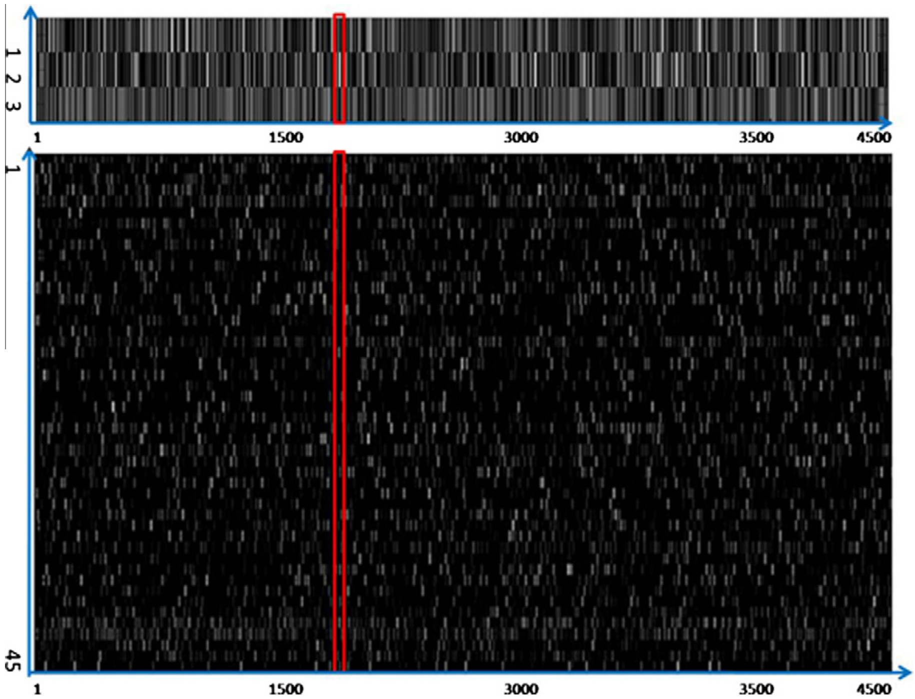

Fig. 1. The top image is the coefficient matrix when $k = 3$ and bottom image is the coefficient matrix when $k = 45$. The y-axis represents the number of clusters and the x-axis is the data point. For a specific data shown as a red vertical box, the assignment of the top matrix is clearer than the bottom matrix, as the second row clearly beats the others. The bottom coefficient matrix has more than 7 non-zero values holding around 10% of the weight each, making a proper assignment difficult.

Table 1
Chou–Fasman parameter.

Symbol and name of Amino Acid	$P(a)$	$P(b)$	$P(t)$	$f(i)$	$f(i+1)$	$f(i+2)$	$f(i+3)$
A: Alanine	142	83	66	0.66	0.076	0.035	0.058
R: Arginine	8	93	95	0.07	0.106	0.099	0.085
D: Aspartic Acid	101	54	146	0.147	0.110	0.179	0.081
N: Asparagine	67	89	156	0.161	0.083	0.191	0.091
C: Cysteine	70	119	119	0.149	0.050	0.117	0.128
E: Glutamic Acid	151	37	74	0.056	0.06	0.077	0.064
Q: Glutamine	111	110	98	0.074	0.098	0.037	0.098
G: Glycine	57	75	156	0.102	0.085	0.19	0.152
H: Histidine	100	87	95	0.14	0.047	0.093	0.054
I: Isoleucine	108	160	47	0.043	0.034	0.013	0.056
L: Leucine	121	130	59	0.061	0.025	0.036	0.07
K: Lysine	114	74	101	0.055	0.115	0.072	0.095
M: Methionine	145	105	60	0.068	0.082	0.014	0.055
F: Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
P: Proline	57	55	152	0.102	0.301	0.034	0.068
S: Serine	77	75	143	0.12	0.139	0.125	0.106
T: Threonine	83	119	96	0.086	0.108	0.065	0.079
W: Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Y: Tyrosine	69	147	114	0.082	0.065	0.114	0.125
V: Valine	106	170	50	0.062	0.048	0.028	0.053

The first column is the name of twenty amino acids with its corresponding one-letter code in parentheses. The next three columns represent the propensities of each amino acid for α-helices ($P(a)$), β-sheets ($P(b)$) or turns ($P(t)$). The rest of the parameters $f(i+j)$'s are the tendencies of the j + 1th position of a hairpin turn, which are generally used to predict a bend.

FCM is a crucial step for clustering with SNMF. Instead of one FCM to divide the data set, we applied FCM hierarchically to avoid data overfitting. We carefully picked the proper fuzzification factor through experiments and strictly enforced the amount of data overlapping for this double FCM process. As a result, we improved the final results in terms of the structural homology and reduced the overall spatial and temporal complexities as well.

3.4. Chou–Fasman parameters

K-means algorithm is known to considerably depend on initial centroids, which can lead to a local optimal solution rather than the global optimal one. Therefore, Zhong et al. (2005), and Chen et al. (2006), Chen et al. (2006), Chen et al. (2008), Chen and Johnson (2009) chose 'favorable' initial points to plug into a *K*-means and increased the number of clusters with high structural similarity. However, the selection of good initial points involve knowing the results in advance. That is, a number of executions of *K*-means algorithm preceded and the resulting clusters are evaluated with its secondary structure similarity. The selection of favorable initial points from the 'good' clusters is then followed. This process is actually a supervised learning method which is undesirable for clustering.

Therefore, we use SNMF to cluster the data set without supervising the procedure. SNMF is proved to be more consistent than *K*-means in the study (Kim & Park, 2008), meaning that with any initial points the results tend to converge closely to a global optimal point. In the experiment, we actually observed that the primary sequence groupings is much better with SNMF than *K*-means with initial random seeds. Computationally, however, the resemblance of primary sequence does not guarantee the similarity of secondary structure in a cluster. To infer the clusters of high structural homology from its primary sequence, we used Chou–Fasman parameters to add a statistical relationship between primary sequence and secondary structure into the data set. Chou–Fasman parameters shown in Table 1 were first introduced by Chou and Fasman (1974), Chou and Fasman (1978). Each amino acid is assigned ***conformational parameters***, $P(a)$,$P(b)$ and $P(t)$, which represent the tendency of each amino acid to alpha helices, beta sheets and beta turns, respectively. The parameters were determined by observing a set of sample protein sequences of known secondary structure. The additional parameters of $f(i)$, $f(i+1)$, $f(i+2)$ and $f(i+3)$ correspond to the frequency with which each amino acid was examined in the first, second, third or fourth position of a hairpin turn. For additional structural information, we compute the tendency of secondary structures based on the frequencies of amino-acid residues at each location, with the three conformational parameters of $P(a)$,$P(b)$ and $P(t)$. For example, if a location at a protein segment consists of 20% of Alanine and 80% of Cysteine, then the relevant secondary structure for the location is 28% of alpha helices, 37% of beta sheets and 35% of beta turns. The statistical structural information is included into the data set to perform SNMF. Details are described in Section 4.1.

4. Experiments

This paper uses the same data in Chen et al. (2006), Chen et al. (2006), Chen et al. (2008), Chen and Johnson (2009), which extended the data used in Zhong et al. (2005). We utilize their measurement and design a new measurement to evaluate clustering results. By reviewing detailed description of data representation and their measurement of previous studies of Zhong et al. (2005), Chen et al. (2006), Chen et al. (2006), Chen et al. (2008), Chen and Johnson (2009), we design a new measurement which can evaluate the quality of overall clusters.

4.1. Data set and data representation

A total of 2,710 protein sequences, none of which shares more than a 25% sequence identity, from Protein Sequence Culling Server (Wang & Dunbrack, 2003) are used in this work. By sliding a window of size 9 through each sequence, we collect more than 560,000 sequence segments. Each segment is represented as the frequency profile constructed from HSSP (Sander & Schneider, 1991), based on the aligned sequences from protein data bank (PDB). The secondary structure of each segment, which will be used to evaluate the results, is also obtained by DSSP (Kabsh & Sander, 1979). Hence, as shown in Fig. 2, each primary sequence segment of length 9 forms a 20×9 matrix, where each location has the frequencies of 20 amino acid residues in the vertical direction.

In this study, we apply FCM algorithms to primary sequences, then we add secondary structure statistics inferred by

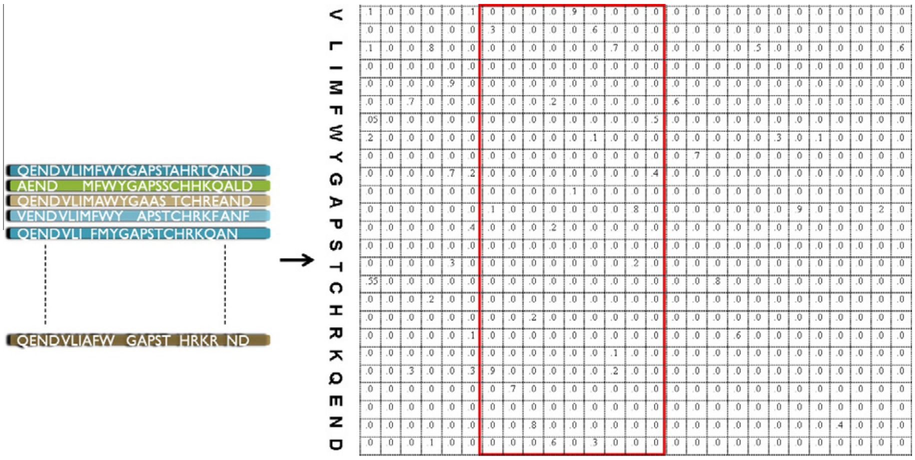

Fig. 2. A number of protein sequences in a protein family obtained from PDB server are aligned on the left. According to the frequencies of twenty amino acids represented as one-letter codes, the proteins are expressed as a profile data on the right figure. Sliding a window of length 9, the 20 × 9 matrix shown inside the red box represents one protein segment data format.

Chou–Fasman parameters to the original data format before applying SNMF. The additional data structure is computed as follows. Let S to be the statistically inferred secondary structure of a 3×9 matrix format with a length of 9 and three types of secondary structure for the helices, beta sheets and beta turns. Let O be the original sequential data shown in Fig. 2, with a 20×9 matrix format. Since $O(i,j)$ represents the frequency of the ith amino acid at the jth location, we want $S(i,j)$ to stand for the probability of the ith second structure at the jth location. We obtain a 3×20 matrix for the Chou–Fasman parameter C, where $C(i,j)$ is the percentage of ith structure for jth amino-acid. Then S is computed as

$$S = C \times O. \tag{5}$$

The final data format for SNMF is the combination of the primary sequence, O, and the computed secondary information, S, forming a 23×9 matrix. That is, each position includes the frequencies of H, E, and C in addition to the frequencies of 20 amino-acid residues.

Finally, each data is unfolded into $23 \times 9 = 207$ dimensional vector and n number of data are formed into $207 \times n$ data matrix A for SNMF. The sparse factor H now directs an assignment of each data sample to one of k clusters.

4.2. Evaluating clustering algorithms

We emphasize that this is an unsupervised learning task, meaning that no prior information about data grouping is given. Hence, after we cluster the data set into similar protein groups, we need biological measurements to evaluate the clusters to discover qualifying motifs. Zhong et al. (2005) suggested a measure of secondary structure similarities in order to capture close relationships between protein sequences and their structure, Chen et al. (2006), Chen et al. (2006) additionally proposed a biochemical measurement, HSSP-BLOSUM62, as well as a computational measurement of the David–Bouldin Index (DBI) measurement. In this paper, we use the secondary structure similarity evaluation which is used in common in the previous studies (Zhong et al., 2005; Chen et al., 2006; Chen et al., 2006), and additionally introduce a new evaluation called sDBI which is the DBI measurement for the computed secondary structure.

Secondary Structure Similarity measure

The structural similarity of each cluster is computed as the following:

$$\frac{\sum_{i=1}^{ws} max(P_{i,H}, P_{i,E}, P_{i,C})}{ws}, \tag{6}$$

where ws is a window size and $P_{i,H}$ is the frequency of the helix at the ith position of the segments in the cluster. $P_{i,E}$, $P_{i,C}$ are defined similarly for beta sheets and turns. After a clustering, each cluster is evaluated with its secondary structure similarity, and clusters with more than 60% similarity are counted, since proteins exceeding 60% structural homology are considered structurally similar (Sander & Schneider, 1991; Zhong et al., 2005). A method producing more clusters with over 60% structural homology will be considered better method with this measurement.

Structural David–Bouldin Index (sDBI) measure

Besides the biological measurement of secondary structure similarity, Chen et al. (2006) used a computational evaluation called David–Bouldin Index (DBI) measure (Davies & Bouldin, 1979), to evaluate the groupings only in terms of their primary sequence. The DBI measure is a function of intra-cluster (within-cluster) distance and inter-cluster (between-cluster) distance. Because a cluster with a relatively larger inter-cluster distance and a relatively smaller intra-cluster distance is more favorable, a lower DBI indicates a better data groupings. Eq. (7) computes the DBI value of a clustering task.

$$DBI = \frac{1}{k}\sum_{p=1}^{k} max_{p \neq q}\left\{\frac{d_{intra}(C_p) + d_{intra}(C_q)}{d_{inter}(C_p, C_q)}\right\}, \tag{7}$$

where $d_{intra}(C_p)$ is the average of all pairwise distances between each member in the cluster C_p and its center, and $d_{inter}(C_p, C_q)$ is the distance of the centers of two cluster C_p and C_q, and k is the number of clusters. All the distance is computed in Hamming distance metric.

However, the DBI of a primary sequence evaluates grouping behavior only in terms of primary sequence. Before we add statistical structure into original data set, when we evaluated its primary sequential grouping with DBI, SNMF showed better results than those of a K-means with random initialization. But, DBI measurement is improper for finding qualifying motifs since good clusters in terms of primary sequences have little biological significance. Therefore, we introduce a new measurement which evaluates its computational clustering results based on the inferred structural information. We call this new measurement *Structural David–Bouldin Index measure*(sDBI). The sDBI follows the same equation as DBI in Eq. (7). The difference is that each cluster consists of the inferred secondary structure S instead of the primary sequence O in Eq. (5). By using sDBI, we can evaluate the overall grouping qualities not restricted to finding a subset of good clusters.

Table 2
Comparison of various clustering methods.

Methods	>60%	>70%	sDBI
Traditional	25.82	10.44	N/A
Improved-*K*	31.62	11.50	N/A
FCM	37.14	12.99	N/A
FIK	39.42	13.27	N/A
FGK	42.93	14.39	7.21
FCM + CF + *K*-means	42.94	13.23	9.07
FCM + SNMF	44.07	12.73	9.85
FCM + CF + SNMF	48.44	16.23	7.05

The first five rows summarize the results of previous methods introduced in Zhong et al. (2005), Chen et al. (2006), Chen et al. (2006). The rest of methods list the experiments conducted in this study. The last result is the best result obtained for both measurements. This result was obtained by using an SNMF which was combined with FCM and Chou–Fasman parameters.

4.3. Experiment steps

Protein motifs discovery using SNMF method follows the subsequent steps. We divide the data set into a number of small size of subsets using FCM hierarchically, for proper clustering task with SNMF method. We first split the data set into ten smaller subsets using FCM, then divide each subset further into much smaller subsets with another FCM. Although the ‘FCM + SNMF’ model, shown in Table 2, increases the percentage of structurally significant clusters, we utilize the conformational parameters of Chou–Fasman table to compute the structural relationship with primary sequence, to improve the results further. Fig. 3 summarizes the experiment steps conducted in this study. To see the impact of Chou–Fasman parameters, we applied a *K*-means with initial random seeds to the combined data set, and provided the result as well.

4.4. Clustering results

We summarize the clustering results in Table 2. Each method is compared with two measurements: the secondary structure similarity and the sDBI. The first column indicates the methods we used in this study as well as the ones from Zhong et al. (2005), Chen et al. (2006), Chen et al. (2006). The second column is the percentage of clusters which have a secondary structure similarity exceeding 60%. The next column is the percentage of clusters having a structural similarity greater than 70%. For structural similarity, a higher percentage is more favorable. The last column indicates sDBI value of each method. With sDBI values, lower values are preferred. The first five methods listed in Table 2 are from Zhong et al. (2005) and Chen et al. (2006), Chen et al. (2006), and they are used to compare with our models. We excluded the results in Chen et al. (2008), Chen and Johnson (2009) since the studies added further filtering procedure after clustering.

‘Traditional’ is a *K*-means with random initial seeds and ‘Improved *K*-means’ is the method studied in Zhong et al. (2005). ‘FCM’ is granular computing combined with a *K*-means with random initial seeds. With FCM, the increased percentage of good clusters having high secondary structure similarity shows a significant improvement over a traditional *K*-means. FIK in the fourth row in Table 2 illustrates further improvement, and FGK model produced the best result among all of the previous models. The sDBI is a new measurement introduced in this paper, and we were unable to provide sDBI values for previous models except FGK, since the resulting clusters of other models were unavailable. Reproduction of the results were also impossible as these results are obtained through lots of experimental trials with different settings. As the result with FGK was obtainable from the authors, we could compute sDBI of FGK result only.

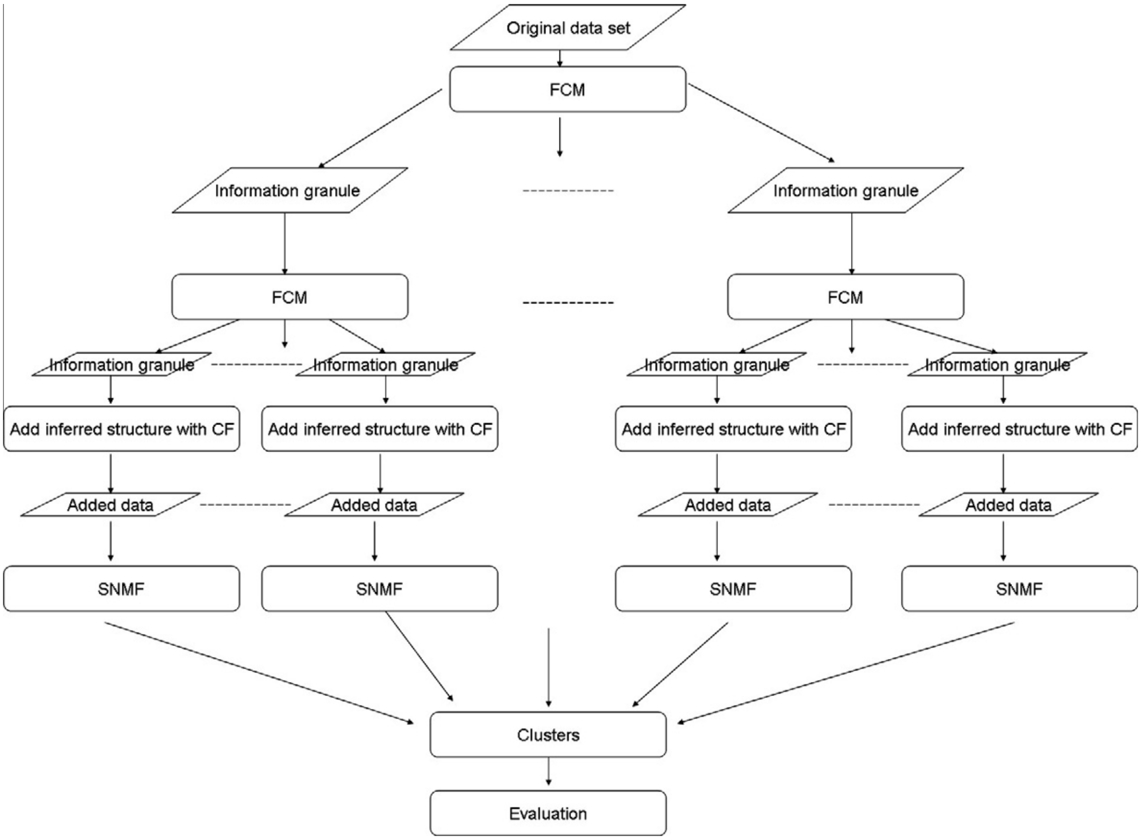

Fig. 3. The above figure summarizes the experiment steps in this study. The original data set of a primary sequence is divided into smaller subsets (information granules) with double applications of FCM. For each subset, secondary structure statistics is inferred with Chou–Fasman parameters and added to each data set. SNMF is finally applied to each subset and the result is evaluated using two evaluation criteria, secondary structure similarity and sDBI.

The rest of Table 2 lists some of the experiments we conducted in this study. ‘FCM + SNMF’ shows the result of applying FCM followed by an SNMF, without the Chou–Fasman parameters. The structural homology indicates that SNMF provides more qualifying motifs than other results with structural similarity measurement. However, it did not beat sDBI value of ‘FGK’ model, requiring another way to improve the clustering result further. Therefore we incorporated secondary structure information computed with Chou–Fasman parameters into the data set and were able to see an improvement on both measurements. To see the influence of Chou–Fasman parameters on *K*-means, we applied this incorporated data to *K*-means with random initial seeds too, and the ‘FCM + CF + *K*-means’ model shows further improvement than FCM, in terms of the structural homology. Since these models are using random initial seeds, we can expect to have further improvement when using greedy *K*-means algorithm on this combined data.

Finally, we further improved both the structural and the computational qualities using the ‘FCM + CF + SNMF’ method (shown in the last row of Table 2). As summarized in Fig. 3, we divided the original data set into much smaller information granules by applying FCM hierarchically, then added secondary structure statistics inferred by Chou–Fasman parameters and primary sequences. Then, we processed SNMF to obtain a sparse coefficient factor for clustering. As a result, the last model bettered the performance of the previous best model, ‘FGK’, for both the structural similarity and sDBI measurements. In conclusion, the ‘FCM + CF + SNMF’ demonstrates that the combination of extended data with structure statistics along with SNMF can discover more structurally meaningful motifs. The result is actually proving that using SNMF, we can obtain more qualifying motifs with proper unsupervised clustering method, without manual setting of cluster centers.

4.5. Sequence motifs

Figs. 4–8 are five different sequence motif examples discovered in this study. They were created using the Weblogo tool (Crooks, Hon, Chandonia, & Brenner, 2004). Weblogo is a web-based tool that generates sequence logos which are graphical depictions of the sequence patterns within a multiple sequence alignment. We illustrate some of the motifs found by our method with sequence logos as they provide a richer and more precise description of sequence similarities than consensus sequences or the previous formats used in Zhong et al. (2005), Chen et al. (2006), Chen et al. (2006). The sequence logos are obtained from the clusters which have over 60% secondary structural similarity, and more than 1,000 protein segments. The exact number of segments and the structural homology are given at the top of each motif image. The motif pattern is represented starting from the N-terminal and the letters stacked at each position demonstrate the type of amino acid which appears with over 8% frequencies in that position. The height of symbols indicates the relative frequency. The letter shown below the x-axis is the representative secondary structure of that position, where *H* is for helix, *E* for sheets and *C* for turns. For example, Fig. 4 is a motif of helix-structure with conserved Alanine (A), and Fig. 6 is a turn-sheet motif and its second position consists of four amino acid (D,G,E,S) with roughly equal frequencies.

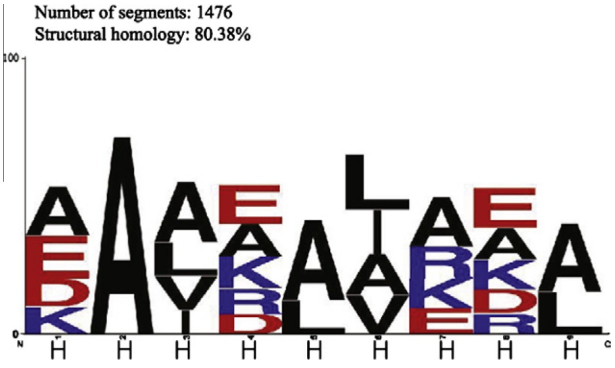

Fig. 4. Helices motif with conserved A.

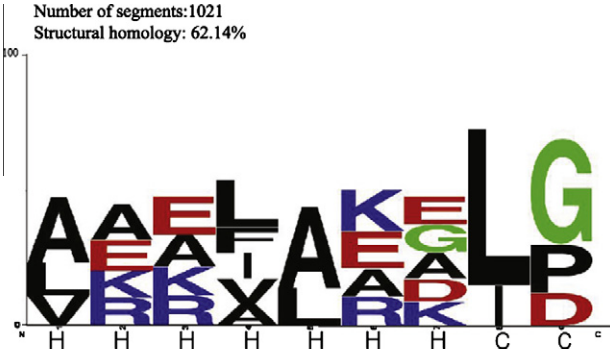

Fig. 5. Helix-turn motif.

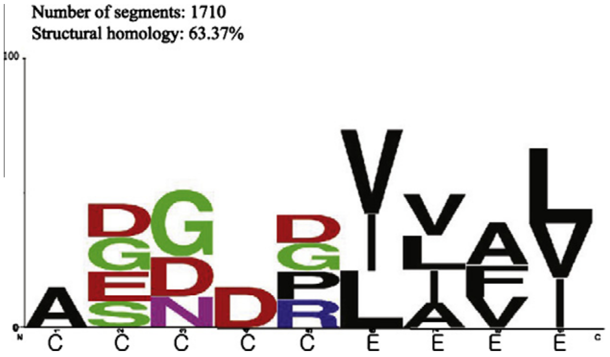

Fig. 6. Turn-sheet motif.

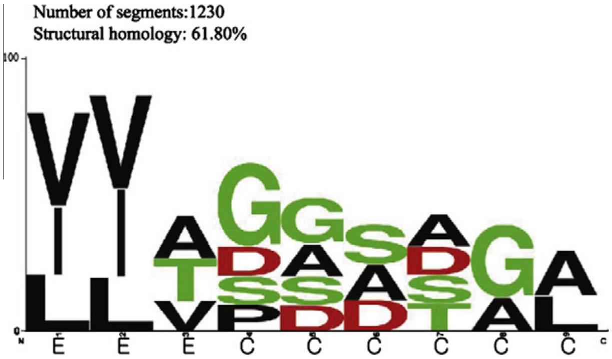

Fig. 7. Sheet-turn motif.

5. Conclusions and future work

In this paper, sparse nonnegative matrix factorization (SNMF) combined with granular computing and inclusion of statistical structure is proposed to discover protein motifs which are universally conserved across protein family boundaries. Discovering high

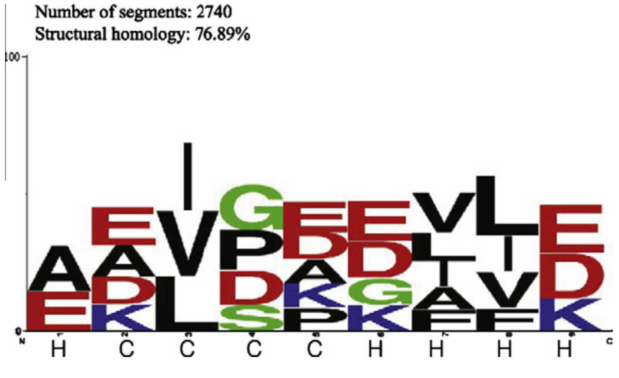

Fig. 8. Helix-turn-helix motif.

quality of protein motifs is very useful in the study of bioinformatics, as the sequence motifs can reveal structural or functional patterns. For example, Chen and Johnson (2009) showed that the sequence motifs can be used to predict protein local tertiary structure. Previous models proposed in Zhong et al. (2005), Chen et al. (2006), Chen et al. (2006), Chen et al. (2008), Chen and Johnson (2009) involve *K*-means clustering algorithms with various initialization strategies. However, in the process of initialization, they used the secondary structure of the data being examined which should be used only for evaluating the results. Therefore, the previous models are undesirable as they are actually supervised clustering methods. Instead, we use an SNMF clustering method since it is more consistent and efficient than *K*-means algorithms with manually selected initial points. In addition, we found that the incorporation of Chou–Fasman parameters plays an important role for this task. Besides the secondary structure similarity measurement, which is limited to selecting a subset of good clusters, we designed a new measurement, sDBI, which evaluates the overall grouping qualities based on the inferred secondary structures and the primary sequences. We also observed that the process with SNMF is less expensive and more meaningful if the size of each subset is reduced with Fuzzy *C*-means preprocessing.

Sparse nonnegative matrix factorization method, however, does have its limitations. For better clustering results, the number of clusters should be small. Otherwise, the presence of many nonzero coefficients holding similar weights make the assignment task obscure. Therefore, an additional dividing process is required, which in turns increases computational complexities and risks of data overfitting. As with *K*-means clustering, the number of clusters need to be determined as a prior parameter for the SNMF method, hindering an automated optimization.

Therefore, our future works include the followings. We need to find a way to decide an optimal number of clusters automatically. Resolving the problem of assigning data to a cluster when there are one or more candidates is another area of future interest. It is also necessary to reduce the computational costs and risks caused by additional dividing steps. We also want to add other evaluation methods, such as functional homology, to qualify the discovered motifs. Finding more biological applications with the protein motifs discovered through this study would be very important future study as well.

Acknowledgement

The work of Yi Pan was supported in part by the National Science Foundation Grants CCF-0514750 and CCF-0646102. Wooyoung Kim was supported by an MBD Fellowship from Georgia State University. The work of Jingu Kim and Haesun Park was supported in part by the National Science Foundation Grants CCF-0621889 and CCF-0732318. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Finally, the authors thank Daniel Lee at Georgia Institute of Technology for his careful reading and commenting of the paper.

References

Attwood, T., Blythe, M., Flower, D., Gaulton, A., Mabey, J., Naudling, N., et al. (2002). PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acid Research, 30*(1), 239–241.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (Vol. 2, pp. 28–36). AAAI Press.

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.

Brunet, J., Tamayo, P., Golub, T., & Mesirov, J. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences, 101*(12), 4164–4169.

Chen, B., Tai, P., Harrison, R., Pan, Y. (2006). FGK model: A efficient granular computing model for protein sequence motifs information discovery. In *The IASTED international conference on computational and systems biology* (pp. 56–61).

Chen, B., Tai, P., Harrison, R., Pan, Y. (2006). FIK model: A novel efficient granular computing model for protein sequence motifs and structure information discovery. In *The IEEE symposium on bioinformatics and bioengineering* (pp. 20–26).

Chen, B., & Johnson, M. (2009). Protein local 3d structure prediction by super granule support vector machines (super gsvm). *BMC Bioinformatics, 10*(Suppl 11), S15.

Chen, B., Pellicer, S., Tai, P. C., Harrison, R., & Pan, Y. (2008). Efficient super granular svm feature elimination (super gsvm-fe) model for protein sequence motif information extraction. *International Journal of Functional Informatics and Personalised Medicine*, 8–25.

Chou, P. Y., & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry, 13*(2), 222–245.

Chou, P. Y., & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas Molecular Biology, 47*, 45–148.

Crooks, G., Hon, G., Chandonia, J., & Brenner, S. (2004). Weblogo: a sequence logo generator. *Genome Research, 14*, 1188–1190.

Davies, D., Bouldin, D. (1979). A cluster separation measure. In *IEEE Transactions on pattern analysis of machine intelligence, Vol. 1* (pp. 224–227).

Devarajan, K. (2008). Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Comput Biology, 4*(7), e1000029.

Donoho, D., & Stodden, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts. *Advances in Neural Information Processing Systems, 16*.

Dunn, J. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Cybernetics, 3*, 32–57.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics, 21*, 768–769.

Gao, Y., & Church, G. (2005). Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics, 21*(21), 3970–3975.

Han, K., & Baker, D. (1983). Recurring local sequence motifs in proteins. *Molecular Biology, 251*, 2577–2637.

Henikoff, S., Henikoff, J., & Pietrokovski, S. (1999). New features of the blocks database servers. *Nucleic Acid Research, 27*, 226–228.

Henikoff, S., Henikoff, J., & Pietrokovski, S. (1999). BLOCKS++: a non redundant database of protein alignment blocks derived from multiple compilation. *Bioinformatics, 15*(6), 417–479.

Hoyer, P.O. (2002). Non-negative sparse coding. In *Proceedings of ieee workshop on neural networks for signal processing* (pp. 557–565).

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research, 5*, 1457–1469.

Hulo, N., Sigrist, C., Saux, L., Langendijk-Genevaux, P., Bordoli, L., Gattiker, A., et al. (2004). Recent improvements to the PROSITE database. *Nucleic Acid Research, 32*, 134–137.

Kabsh, W., & Sander, C. (1979). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Bioploymers, 22*, 2577–2637.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Kim, J., Park, H. (2008). Sparse nonnegative matrix factorization for clustering, Technical report. GT-CSE-08-01, Computational Science and Engineering, Georgia Institute of Technology.

Kim, H., & Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics, 23*(12), 1495–1502.

Kim, H., & Park, H. (2008). Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications, 30*(2), 713–730.

Lee, D. D., & Seung, H. S. (1997). Unsupervised learning by convex and conic coding. *Advances in neural information processing systems 9* (Vol. 9, pp. 515–521). MIT Press.

Lee, D., & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791.

Li, S.Z., Hou, X., Zhang, H., Cheng, Q. (2001). Learning spatially localized, parts-based representation. In *CVPR '01: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, Vol. 1* (pp. 207–212).

Lin, T. Y. T. Y. (2000). Data mining and machine oriented modeling: A granular computing approach. *Applied Intelligence, 13*(2), 113–124.

Macqueen, J.B. (1967). Some methods of classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (pp. 281–297).

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics, 5*(2), 111–126.

Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications, 416*(1), 29–47.

Pavesi, G., Zambelli, F., & Pesole, G. (2007). Weederh: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. *BMC Bioinformatics, 8*(1), 46.

Peña, J. M., Lozano, J. A., & Larrañaga, P. (1999). An empirical comparison of four initialization methods for the *k*-means algorithm. *Pattern Recognition Letters, 20*(10), 1027–1040.

Ross, D. A., & Zemel, R. S. (2006). Learning parts-based representations of data. *Journal of Machine Learning Research, 7*, 2369–2397.

Sander, C., & Schneider, R. (1991). Database of similarity derived protein structures and the structure meaning of sequence alignment. *Proteins: Structural and Functional Genetics, 9*(1), 56–68.

Siddharthan, R., Siggia, E. D., & van Nimwegen, E. (2005). Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biology, 1*(7), e67.

Wang, G., & Dunbrack, J. R. L. (2003). Pisces: a protein sequence-culling server. *Bioinformatics, 19*(12), 1589–1591.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 267–273). New York, NY, USA: ACM Press.

Yao, Y. (2001). On modeling data mining with granular computing. In *COMPAC* (pp. 638–643).

Zhong, W., Altun, G., Harrison, R., Tai, P., Pan, Y. (2005). Improved *k*-means clustering algorithm for exploring local protein sequence motifs representing common structural property. In *IEEE transactions on nanobioscience, Vol. 14* (pp. 255–265).

Zhou, Q., Wong, W.H. (2004). CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. In *Proceedings of the national academy of sciences of the United States of America, Vol. 101 (33)* (pp. 12114–121190).