Title: A Hybrid Framework for Protein Sequences Clustering and Classification Using Signature Motif Information

Names of authors: Wei-Bang Chen, Chengcui Zhang*

Full affiliation: Department of Computer and Information Sciences
University of Alabama at Birmingham, Birmingham, AL 35294, USA

Email: {wbc0522, zhang}@cis.uab.edu

* Corresponding Author:

Dr. Chengcui Zhang
CH 127
1530 3rd Ave S
Birmingham AL 35294-1170, USA
Email: zhang@cis.uab.edu
Tel: 1-205-934-8606 Fax: 1-205-934-5473

Abstract

In this paper, we propose an unsupervised hybrid framework for protein sequence clustering and classification which incorporates protein structural motif information. The proposed framework consists of three stages: protein structural motif scan, hybrid clustering, and sequence classification. The incorporation of protein structural motif detected by ScanProsite service provides a better measurement in calculating the sequence similarity. The proposed two-phase hybrid clustering approach combines the strengths of the hierarchical and the partition clustering. Phase I adopts the hierarchical agglomerative clustering to pre-cluster multi-aligned sequences. Phase II performs the partition clustering which initiates its partition based on the result from Phase I and uses profile Hidden Markov Models (HMMs) to represent clusters. The profile HMMs are then stored in the database for unknown sequences classification, which is done by finding the best alignment of a sequence to each existing profile HMM. Our experiments demonstrate the effectiveness and the efficiency of the proposed framework for biological sequence clustering and classification.

1. Introduction

The analysis and prediction of local structures and functions of protein sequences is one of the most interesting topics today in bioinformatics. The functions of a protein are determined by its three dimensional structure [18], i.e., the tertiary and quaternary structure of a protein, along with the environment. By studying the structure of the protein, we can better understand its functions. The three dimensional structure of a protein can be determined by X-ray crystallography, Nuclear Magnetic Resonance (NMR), and Cryo-Electron tomography (CET). However, X-ray crystallography is still the most effective way to determine the high-resolution three dimensional structure of a protein. As of July 21, 2009, the Research Collaboratory for Structural Bioinformatics - Protein Data Bank (RCSB PDB) holds 54,558 protein structures, in which 47,314 (87%) protein structures are determined by X-ray crystallography, whereas only 6,946 (13%) and 171 (<1%) protein structures are determined by NMR and Electron Microscopy, respectively. No matter what approach is used for determining protein structures, these experiments are very expensive and will require complex sample preparation processes. For instance, when X-ray crystallography is performed on protein crystals, it needs to prepare a large amount of highly purified protein (typically more than 10mg) with the concentration of 10mg/ml in order to progressively grow protein crystals. In addition, the process of growing protein crystals is tedious, which usually takes from several days to weeks. Therefore, structural biology researchers often have to rely on cost inefficient and time consuming experiments to study protein structures. Due to the rapid growth of protein databases which makes these traditional approaches not feasible, an alternative way to determine the three dimensional structure of protein and to further infer the functions of protein is through the prediction of the protein structures from its primary sequences.

One way to predict the structures and functions of unknown protein sequences is through sequence comparison. This approach is based on the assumption that if a newly discovered protein sequence is sufficiently similar to the sequence of a protein of known structure, we can infer that the two proteins have similar structures [8]. Past research has shown that similar protein sequences almost always have similar structures and functions and therefore can be categorized into a group called protein family [7, 12]. Within the same protein family, member sequences have similar functions and structures. This idea suggests that it is possible to predict the structures and functions of unknown protein sequences through sequence clustering and classification based on sequence similarity. By clustering related sequences into groups based on their sequence similarity (the clustering step) and then assigning an unknown sequence to its closest cluster (the classification step), we can infer the sequence's functions and structures by analyzing the common functions and structures shared by member sequences in the same cluster.

In this paper, we propose a two-phase hybrid clustering algorithm which takes advantage of both hierarchical clustering algorithms and partitioning clustering algorithms. An efficient bottom-up hierarchical clustering method is proposed as the first phase of the clustering stage. However, a drawback of the hierarchical clustering method is that it cannot perform any adjustment to its previous merging or splitting steps, which may lead the clustering result to a very poor local optima once a bad step of merging or splitting occurs. In addition, the time-complexity of a naïve implementation of the algorithm is $O(n^3)$. To alleviate this problem, we introduce *k*-means clustering algorithm, a partitioning clustering method, as our second phase of the proposed hybrid clustering algorithm to refine the pre-clustering result produced by the first phase. *k*-means clustering algorithms usually involve an iterative process which starts with a set of initial cluster centroids, reassigns sequences to their closest cluster, and then refines the cluster centroid. *k*-means clustering algorithms have a complexity of $O(n^2)$, which is more efficient than the hierarchical clustering. However, problems arise when the initial cluster centroids are randomly selected. With different selections of initial seeds, it may end up with a rather diverse set of clustering results, some of which are associated with very poor local optima. The possibility of seeing empty clusters during the iterative process is another problem. To better guide the *k*-means clustering, in this paper, we use the clustering result from the hierarchical clustering as the initial partition (initial set of clusters) for *k*-means. However, *k*-means clustering algorithms usually require a cluster centroid to represent that cluster. We can use either the similarity based methods or model-based methods to generate the cluster centroids. The similarity-based methods generate a new seed (centroid) by examining the similarity score between each pair of objects in the cluster, while the model-based methods generate a new centroid by building a model to represent that cluster [9]. The model-based methods thus provide better interpretability than do the similarity-based methods since the resulting model for each cluster directly characterizes that cluster. In this paper, we adopt the model-based methods for characterizing clusters. In particular, our approach is based on a probabilistic modeling technique, profile Hidden Markov Model [3, 4] (profile HMM), to represent protein families as clusters.

Hidden Markov Models (HMMs) are statistical modeling methods which are not only broadly used in speech recognition field, but also commonly used for describing biological sequences in computational biology [14]. A hidden Markov model is a stochastic process that considers a system with a finite set of states, each of which is connected by transitions with associated probabilities. The transitions between the states are governed by the transition probabilities, and hence, convey a clear Bayesian semantics. A set of biological sequences can be seamlessly described by a hidden Markov model since we can consider a Markov chain as a protein which is actually a long poly-peptide chain that consists of linearly connected amino acids as a finite set of states. For this reason, in this study, we adopt HMMs as a tool for the profile analysis of a protein family, which statistically characterizes a set of biological sequences by comparing and identifying their common sequence patterns.

It is worth mentioning that the HMM profiles produced by the clustering stage will be further used to classify unknown protein sequences in the classification stage. With the proposed approach, we can effectively and quickly find the optimal alignment of a protein sequence against an existing model of protein families. It not only facilitates the classification stage, but also benefits the future model maintenance. To classify a previously unseen protein sequence, there is no longer a need to compute the distance between the new protein sequence and every other protein sequence in the dataset with this framework. Thus the time complexity of the classification process can be drastically reduced.

In addition, the proposed protein sequence clustering and classification framework incorporates protein structural motif information into the modeling process with the purpose to improve the accuracy of sequence clustering and classification. A protein forms its quaternary structure by assembling several polypeptide chains, each of which is a protein subunit that is known as tertiary structure. A simple protein structural motif is a polypeptide pattern that combines a few secondary structure elements, i.e., local segments of biopolymers such as α helix, β sheets, and loops, to form a three-dimensional structural element of which a complex protein structural motif is composed. Further, structural motifs act as building blocks of functional domain which is a part of a protein sequence that can evolve and function independently of the rest of the protein chain. In general, protein sequences in the same family have several common protein domains [13]. This suggests that the structural motifs associate with the protein domains which correlate to the functions of a protein. Therefore, increasing the weight of structural motifs or domain areas when performing the pair-wise alignment could likely lead to a more accurate evaluation of sequence similarity. For this reason, in this study, we incorporate signature motif information into the proposed protein sequence clustering and classification framework. In practice, we use the ScanProsite (http://www.expasy.org/tools/scanprosite/), a web-based tool provided by PROSITE database which is an annotated collection of structural motif descriptors dedicated to the identification of protein families and domains, to scan protein sequences against the PROSITE database for detecting PROSITE signature motifs in protein sequences [1, 11].

In exploring a proper way of increasing the weight of domains, we propose and compare two different methods: Union and Position Probability methods. Weight normalization by the sequence length is performed prior to the clustering and classification. This normalization step is necessary, which makes the weight increment independent of the lengths of structural motifs. These two proposed weighting schemes mainly address on the protein sequence regions that are covered by structural motifs. The first weighting scheme is called union weighting scheme which considers each position in a protein sequence as “all or none”. This weighting scheme increases the weight if a position in the sequence resides in the structural motif regions. Otherwise, less weight will be given to the non-structural motif regions. The second weighting scheme is the so-called position probability weighting scheme which focuses on analyzing the probability of each position in a protein sequence belonging to structural motif regions. In this weighting scheme, higher weight will be assigned to a position if that position has a higher chance residing on the structure motif regions.

Classification is in its nature a task-oriented step. Therefore, there is no single classification scheme that can meet fit all needs. For protein sequence classification, most approaches adopt supervised machine learning methods which require domain experts to provide the ground-truth according to the purpose of classification. Supervised learning algorithms lack flexibility since they must need the provided ground-truth to train classifiers. In this paper, we proposed to get rid of the ground-truth supplying process by introducing an unsupervised machine learning algorithm. In particular, we tightly integrate the protein clustering and the protein classification in a single framework. The classifiers can be trained by all the sequences in the corresponding clusters, each of which represents a protein family. With this framework, users only need to provide one parameter, i.e., the number of clusters, in order to cluster and classify newly discovered protein sequences, making the proposed framework very easy to use.

In brief, the goal of our research is to develop an unsupervised framework for protein sequence clustering and classification. For clustering, our experimental results demonstrated that the inaccuracy of the hierarchical agglomerative clustering can be compensated by the profile HMM based *k*-means partition clustering since the model-based partition clustering can better describe the dynamic properties of the data in a cluster. On the other hand, the unsupervised initial partition generated in Phase I can prevent the subsequent clustering process from being stuck in a bad local minimum. Our experimental results also exhibit that the weight increment on structural motif areas will increase the accuracy of clustering and classification. For classification, our results show the robustness of the proposed scheme based on HMM profiles.

In the rest of this paper, we describe the proposed protein sequence clustering and classification framework in Section 2. Section 3 demonstrates the experimental results, and Section 4 concludes the paper.

2. The proposed framework

2.1 The overview of the proposed framework

The proposed protein sequence clustering and classification framework comprises three stages. In the first stage, all input sequences are multi-aligned with the EBI ClustalW [6]. In the meanwhile, their signature motifs are predicted by adopting the ScanProsite which scan protein sequences against the PROSITE database for detecting PROSITE signature motifs in protein sequences. In the second stage, an unsupervised hybrid clustering algorithm is applied in order to build a model for each protein family. We measure in the third stage the similarity between an unknown protein sequence and each protein family model produced in the second stage in order to classify the unknown protein sequence into its corresponding family. Our experimental results show that the proposed framework improves the accuracy in both the clustering and the classification stages by increasing the weight of the signature motifs which are conceived as an important structure of functional domain that may appear in the evolutionarily related proteins, i.e. a protein family. Fig. 1 illustrates the overview of the proposed protein sequence clustering and classification framework.

In the following subsections, we first describe the signature motif detection and the weight increasing schemes, followed by the proposed hybrid clustering algorithm and the unknown sequence classification.

2.2 Signature motif detection

As illustrated in Fig. 1, in the proposed framework, the first stage is to detect the signature motifs of a given protein from its primary sequence. To predict the signature motifs of protein sequences, the PROSITE database is used in this study as the signature motifs boundary predictor [11]. The signature motif boundary prediction is achieved by using the ScanProsite tool to search a given sequence against the profiles in the PROSITE database. All signature motif hits in the result returned by ScanProsite are collected as the estimated signature motif boundaries [1] for that sequence.

However, the procured signature motif boundary information cannot be directly used in the proposed framework. This is because that each signature motif boundary returned is marked by its start and stop positions in the raw sequence without any gap in between, while the proposed framework accepts multi-aligned sequences as input which often contain many gaps (due to the multi-alignment). Hence, it is essential to map the signature motif boundary positions from the original sequence to the corresponding multi-aligned sequence.

The position mapping can be achieved by constructing an index mapping table T which is initialized as a vector of length n with the initial values of 1 to n, where n is the length of the multi-aligned sequence. Subsequently, we collect the position indices of all the gaps in the multi-aligned sequence and use the indices to remove the corresponding elements in the vector. To this end, the index mapping table T should have the same length as the original sequence. Assume that a signature motif boundary starts at the position i and stops at the position j in the raw sequence, its corresponding positions in the multi-aligned sequence can be obtained by looking up the i^{th} and j^{th} entries, i.e. the $T[i]$ and $T[j]$, in the index mapping table. Fig. 2 illustrates this idea for signature motif position index mapping.

2.3 Weighting schemes

To select a proper weighting scheme for signature motifs is another important issue in the sequence similarity measurement. Since the location of a signature motif may vary from sequence to sequence, this matters not only in calculating the similarity score of two sequences, but also in computing the similarity score of two clusters. To deal with this issue, we propose two weighting schemes, including the union of signature motifs, and the position probability of signature motifs. The details of the proposed schemes are described as below:

Let X and Y be two clusters whose included sequences are $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$, respectively.

Union of signature motifs

$$UXY = UX \cup UY \tag{1}$$

where UXY is the signature motif union of two clusters X and Y; UX denotes the signature motif union of the cluster X which is defined in Equation 2.

$$UX = Ux_1 \cup \ldots \cup Ux_i \cup \ldots \cup Ux_m \tag{2}$$

where Ux_i is the signature motif union of the i^{th} sequence in the cluster X. The definition of Ux_i is given in Equation 3.

$$Ux = fd_{x1} \cup \ldots \cup fd_{xj} \cup \ldots \cup fd_{xp} \tag{3}$$

where fd_{xj} denotes the j^{th} signature motif in the sequence x and p is the number of signature motifs in x.

Position probability of signature motifs

The position probability of signature motifs is defined as the probability of a position in a sequence belonging to signature motifs. To calculate the position probability of signature motifs, assume that a cluster $X=\{x_1, \ldots, x_m\}$ consists of m protein sequences with the length l. We can create a matrix M with the size $m \times l$ and initiate all the elements in M as 0.

For a sequence x_i in X, if a position j in that sequence belongs to a signature motif, the corresponding entry M_{ij} in M is marked 1. After examining all the sequences, we calculate the sum for each column in M, which indicates the number of sequences at this position belonging to a signature motif. To avoid the zero probability, we add one to each column sum as a pseudo count method. The position probability weighting scheme w is thus a vector of size l which collects $w(j)$ as the probability of a position j belonging to a functional domain in a set of aligned sequences. This vector can be obtained by Equation 4.

$$w(j) = \left[\frac{1 + \sum_{i=1}^{m} M_{ij}}{l + \sum_{i=1}^{m} \sum_{j=1}^{l} M_{ij}} \right]^2, j = 1 : l \tag{4}$$

In the probability weighting scheme, we give much higher weight for the region belonging to the function domain.

In this paper, we test the performance of the proposed framework with or without the use of signature motif information. In addition, we apply the above two weighting schemes with signature motifs and compare their performance with the Equal-weighting scheme in which all positions are equally weighted and no signature motif information is used. Fig. 3 illustrates the idea of the two proposed weighting schemes.

2.4 Hybrid clustering

Hierarchical clustering

The hierarchical clustering step plays an important role in the proposed hybrid clustering algorithm, which supervises subsequent partitioning clustering algorithm though hierarchical clustering itself is an unsupervised method. More specifically, the agglomerative hierarchical clustering algorithm serves as the first step in this stage to provide the initial clusters and guide the subsequent *k*-means clustering.

In the agglomerative hierarchical clustering step, each protein sequence is considered an independent data object which initially forms a cluster of its own. The agglomerative hierarchical clustering algorithm progressively merges two clusters with the highest similarity score. Gradually, the fusions cause the remaining clusters grow larger and larger. This merging process will not stop until the desired number of clusters is formed, and its entire progress can be represented by a dendrogram which illustrates how two clusters are merged at each successive stage of analysis.

Essentially, in each step, the merging of two clusters totally depends on the measurement of the pair-wised similarity score. Hence, the similarity measure is the key factor in the hierarchical clustering. There are several different approaches to determining the similarity between two clusters, including single-link, complete-link, and average-link approaches [9]. In the proposed hybrid algorithm, we adopt the average-link approach in which the similarity value of two clusters is defined as the average similarity score between any two sequences in the two clusters, as formulated in Equation 5.

$$S(X,Y) = \frac{1}{m \cdot n} \times \sum_{i=1:m;\, j=1:n} s(x_i, y_j) \tag{5}$$

where $S(X,Y)$ denotes the similarity score of two clusters $X=\{x_1,\ldots x_i ,\ldots, x_m\}$ and $Y=\{y_1,\ldots y_j ,\ldots, y_n\}$. x_i and y_j are two member sequences in X and Y, respectively. $s(x_i, y_j)$ denotes the similarity score between two sequences x and y, which is defined in Equation 6.

$$s(x_i, y_j) = \sum_{p=1}^{l} sm_p \times w_p \tag{6}$$

where sm_p is the score of the p^{th} pair of amino acids in the two aligned sequences (x_i, y_j) with length l, which can be looked up in the amino acid substitution matrices such as BLOSUM62 [10]. w_p denotes the weight for the p^{th} pair of amino acids in the aligned sequences, which is stored in a weighting vector w obtained by one of the two proposed weighting schemes or by the Equal-weighting scheme. It is worth noting that when aligning sequences with ClustalW, many gaps are inserted into sequences in order to align them. However, the insertion of gaps should be penalized, which is widely known as the 'gap penalty'. In this paper, we set the gap penalty to -9 and extend the BLOSUM62 substitution matrix with the addition of the gap penalty. In this matrix, the similarity score of a gap and an amino acid is -9 (the lowest). The similarity score of two gaps is 0.

While the similarity score of two clusters can be iteratively computed in a pair-wised manner, the two closest clusters are merged in each cycle. This iterative process will continue until the desired number of clusters is obtained, that is, the entire dataset is clustered into *k* clusters. The clustering results are then passed to the partitioning clustering step for further refinement.

Partitioning clustering

A typical *k*-means partitioning clustering algorithm includes the following steps: (1) randomly select *k* seeds as the centroids of *k* clusters, (2) assign a data point to a cluster with which it has the shortest distance; repeat this step for all data points, (3) calculate new centroids of clusters based on the current cluster members, (4) reassess the cluster membership of each data point by their distances to the new centroids, and (5) repeat Steps 2 to 4 till some stopping criterion is met.

In the second step of the proposed hybrid clustering algorithm, we adopt a modified *k*-means partitioning clustering algorithm. First, we use the pre-clustering result from the hierarchical clustering step as the initial partition of sequences. We then use a profile Hidden Markov Model (HMM) to represent the centroids of clusters for *k*-means partitioning clustering. Unlike the traditional *k*-means partitioning clustering algorithm which randomly selects *k* seeds as the initial centroids of clusters, the proposed hybrid clustering algorithm constructs profile Hidden Markov Models to represent the cluster centroids. The initial models are built from the clusters generated from the previous step, i.e. the hierarchical clustering results. The subsequent process is to iteratively evaluate and update the sequence membership and refine the cluster models (as new cluster centroids). We explain the idea of the proposed *k*-means partitioning clustering algorithm as follows.

Assume we need to group *n* sequences into *k* clusters. First, we need to build a profile HMM as the centroid for each cluster *C*. The profile HMM model, a common statistical tool, is widely used for modeling structured sequences composed of symbols. In our previous study [2], we proposed to model a cluster of sequences with position-specific scoring matrix (PSSM) which is actually a concise version of classical profile HMM for modeling sequences since PSSM only models match states. In this study, we adopt profile HMM based on the Plan 7 architecture [3, 4]. We depict a profile HMM model of length 3 as state transition diagrams in Fig. 4, where M, I, and D nodes represent match states, insert states, and delete states, respectively. The flanking states, including S, N, B, E, C, and T, are used for proper modeling of the ends of the sequence. The estimation of profile HMM transition probability is achieved by using pseudo-counts on multi-align protein sequences with the same cluster label.

Once a HMM profile is built and used as the centroid for each cluster, the reassessment of cluster membership for all sequences is performed. For this purpose, we need to measure the similarity between a sequence and a cluster model. The higher the similarity score, the more likely that sequence belongs to the cluster represented by that model. The optimal alignment of a protein sequence to a profile HMM is used to calculate a maximum similarity score using log-odd ratios for emission probabilities and log probabilities for state transitions.

After obtaining the similarity scores of a sequence and all the HMM models, we reassign the sequence to the cluster with which it has the highest similarity score. The subsequent model building and the update of sequences membership are iteratively refined until there is no further change in the clusters [4, 5, 9, 15].

It is worth mentioning that the hybrid clustering algorithm not only produces clustering results, but also generates a set of profile HMMs at the end of the clustering stage. The resulted protein family profiles are saved in the database for the subsequent classification of unknown sequences.

2.5 Classification

Classification is known as a process that systematically arranges objects in groups or categories according to established criteria. At this stage, the proposed framework uses the profile HMMs resulted from the clustering stage as classifiers in order to classify previously unseen sequences into their corresponding protein family. Using profile HMMs as classifiers is similar to the cluster membership assessment process in the *k*-means partitioning clustering stage. The sequence classification can be achieved by calculating the similarity score between an unknown sequence and each profile HMM, and assigning the sequence to the cluster which yields the maximum similarity score.

3. Experimental results

The proposed framework is tested on a protein sequence dataset procured from PANTHER database. In PANTHER database, proteins are classified by expert biologists into families and subfamilies of shared functions, which are then categorized by molecular functions and biological process ontology terms [19]. The dataset used in our experimental study is composed of 10 protein families with various numbers of sequences, including “olfactory receptor” (PTHR11398: 1,071 sequences), “RAS-related GTPase” (PTHR11708: 1,070 sequences), “leucine-rich trans-membrane proteins” (PTHR23154: 765 sequences), “ATP-dependent AMP-binding enzyme” (PTHR11968: 611 sequences), “serine/threonine protein kinase” (PTHR22985: 579 sequences), “myosin” (PTHR13140: 542 sequences), “translation factor” (PTHR23115: 515 sequences), “heat shock protein 70kDa” (PTHR19375: 307 sequences), “chaperonin” (PTHR11353: 281 sequences), and “small heat-shock protein (HSP20) family” (PTHR11527: 202 sequences), with 5,943 protein sequences in total. In addition to protein sequences, the signature motif locations in each sequence are subsequently procured from the PROSITE database. The ScanProsite, a web-based tool provided by PROSITE, is adopted for searching a given sequence against the profiles in the PROSITE database. All signature motif hits in the result returned by the ScanProsite are collected as the estimated signature motif boundaries for that sequence.

We evaluate the performance of the proposed framework with repeated random sub-sampling validation. Specifically, this validation method randomly divides the entire dataset into training data and testing data sets with a ratio of 1:9, i.e., 595 protein sequences for a training data set and 5,348 protein sequences for a testing data set. Within each training data set or each testing data set, the proportions of the protein families reflect the proportions of the protein families in the entire dataset. For each training-testing data pair, the clustering algorithms group the training data for producing clusters each of which contains a group of similar protein sequences. Subsequently, all the training sequences in each cluster are used to build a profile hidden Markov model which represents a protein family. In our experimental setting, 3 rounds of random sub-sampling validation are performed on the entire dataset with non-overlapped training data.

3.1 The clustering evaluation criterion

To evaluate the performance of the proposed hybrid clustering method for protein sequences, we adopt the V-measure as the performance indicator which is reported by Rosenberg and Hirschberg [17]. V-measure is a conditional entropy-based measurement and it provides an objective criterion to describe the quality of cluster results and is independent of the clustering algorithms being used. In this study, we adopt V-measure to compare the performance of the proposed hybrid clustering algorithm with two other clustering algorithms, i.e. the hierarchical clustering algorithm and the *k*-means partitioning clustering algorithm. The *homogeneity* and *completeness* are two essential criteria that determine the quality of clusters in V-measure. The homogeneity implies that ideally each cluster should include only those data points that belong to one class of the ground truth. On the other hand, the completeness measure indicates that all the data points that are members of a class in ground truth should be in one cluster. As a matter of fact, V-measure is the weighted harmonic mean of homogeneity and completeness.

The homogeneity measure is defined as the normalized conditional entropy of the class given the clusters. For example, if a cluster contains only data points of a single class, then its entropy should be zero. It is essential to normalize the raw conditional entropy by the maximum entropy of the class since the raw conditional entropy is not independent of the data size and distribution of classes.

To calculate the homogeneity, we first define a matrix $A= \{a_{ck}\}$, where a_{ck} is the number of elements that are members of both class c and cluster k. Thus the rows of this matrix hold classes/ground truth and the columns contain cluster/experimental results. Let the set of ground truth classes be C and the set of clusters in the experimental result be K, the raw conditional entropy $E(C|K)$, the maximum entropy $E(C)$ and the homogeneity hm, can be defined as follows:

$$E(C) = -\sum_{c=1}^{|C|}\left(\frac{\sum_{k=1}^{|K|} a_{ck}}{N}\log\left(\frac{\sum_{k=1}^{|K|} a_{ck}}{N}\right)\right) \quad (7)$$

$$E(C \mid K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N}\log\frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}} \quad (8)$$

$$hm = 1 - \frac{E(C \mid K)}{E(C)} \quad (9)$$

where N is the total number of data points (e.g., the total # of sequences in our case.)

The completeness can be described as the normalized conditional entropy of K given the ground truth C. For instance, if the result is perfect, i.e. all the data points of a class are in one single cluster, the entropy would be zero. Also to make the calculation symmetrical to homogeneity, we normalize this conditional entropy by the maximum reduction in entropy of clusters. Thus, we define the raw conditional entropy $E(K|C)$, the entropy of clusters $E(K)$, and the completeness cm as follows:

$$E(K) = -\sum_{k=1}^{|K|}\left(\frac{\sum_{c=1}^{|C|} a_{ck}}{N}\log\left(\frac{\sum_{c=1}^{|C|} a_{ck}}{N}\right)\right) \quad (10)$$

$$E(K \mid C) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{a_{ck}}{N}\log\frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}} \quad (11)$$

$$cm = 1 - \frac{E(K \mid C)}{E(K)} \quad (12)$$

After the homogeneity and completeness values have been calculated, we compute the V-measure as the weighted harmonic mean of the homogeneity and completeness. The mathematical expression for V-measure is similar to F-measure, a commonly used evaluation method for information retrieval introduced by Van Rijsbergen in 1979 [16]. It can be written as

$$V_\beta = \frac{(1+\beta^2)\times hm \times cm}{(\beta^2 \times hm) + cm} \quad (13)$$

where β is a constant, which if greater than 1 means that the completeness is weighted β times more strongly; otherwise the homogeneity is weighted more in the calculation. In this study, since homogeneity and completeness are equally important, we use β =1 in computing the V-measure in clustering performance evaluation.

3.2 Clustering algorithm performance evaluation

To demonstrate the effectiveness of the proposed hybrid clustering algorithm, we compare it with three other clustering algorithms, the hierarchical clustering algorithm, the PSSM-based (Position-Specific Scoring Matrix) *k*-means clustering algorithm [2], and the profile HMM-based *k*-means clustering algorithm. We further examine the performance of difference combinations of hierarchical clustering and partition based clustering algorithms (PSSM and Profile HMM *k*-means) in the two steps of a hybrid clustering scheme. To examine the impact of different signature motif weighting schemes on the clustering results, we introduce 2 signature motif weighting schemes as described in Section 2. We present in Table 1 the performance results of fourteen different combinations of hybrid clustering and three one-step (non-hybrid) clustering methods, using different weighting schemes.

In the subsequent experiments, the weighting scheme A denotes the Equal-weighting method, which implies that signature motifs are not emphasized. B indicates the union of signature motifs, and C represents the position probability weighting scheme. In particular, the position probability weighting scheme is represented by a digit followed by a letter (C). This digit *d* indicates the exponential power used in the weight vector construction. For example, in a protein sequences, if position 1 is considered to be on a signature motif, and position 2 does not belong to a signature motif, then the position 1 is 2^d times as important as position 2. In our experiments, we test various *d* values, i.e., *d* =1~5, on the training data in order to find the optimal position probability weighting scheme.

Table 1 compares the proposed profile HMM-based hybrid clustering algorithm with several other clustering algorithms, including: hierarchical clustering algorithm, HMM-based *k*-means clustering with randomly initiated cluster centroids, PSSM-based *k*-means clustering with randomly initiated cluster centroids, and PSSM-based hybrid clustering algorithm. The V-measure values (with *d* = 1) of the hierarchical clustering with various weighting schemes, i.e., A, B, 1C, 2C, 3C, 4C, and 5C, are 0.273, 0.293, 0.278, 0.347, 0.508, 0.524, and 0.540, respectively. The V-measure value of the HMM-based *k*-means clustering with randomly initiated cluster centroids is 0.314. The V-measure value of the PSSM-based *k*-means clustering with randomly initiated cluster centroids is 0.342. The V-measure values of the proposed profile HMM-based hybrid clustering (Hybrid: Hierarchical + HMM) with various weighting schemes, i.e., A, B, 1C, 2C, 3C, 4C, and 5C, are 0.280, 0.305, 0.283, 0.360, 0.526, 0.556, and 0.565, respectively. The V-measure values of the PSSM-based hybrid clustering (Hybrid: Hierarchical + PSSM) with various weighting schemes, i.e., A, B, 1C, 2C, 3C, 4C, and 5C, are 0.214, 0.242, 0.218, 0.276, 0.412, 0.409, and 0.440, respectively.

The experimental results given in Table 1 show that the proposed hybrid clustering algorithm (Hybrid: Hierarchical + Profile HMM) has the best performance among all the methods included for comparison and has the highest V-measure value (with *d* = 1) of 0.565. In addition, there are several other observations that can be gleaned by examining the values more carefully in Table 1. First, it shows that the results of most hybrid clustering methods are better than that of their constituent hierarchical clustering (1st step) or *k*-means clustering (2nd step) alone. Second, the probability weighting scheme is demonstrated better than the union weighting scheme for structural motifs. Third, almost all the clustering methods that incorporate signature motif information are better than their counter parts which do not emphasize motifs (or adopt equal-weighting scheme). In addition, although now shown in Table 1, our experimental results using V-measure with higher *d* values (*d* = 2~5) also demonstrate that the proposed hybrid method outperforms all the others. This is because that the proposed method usually has a higher completeness value than the others, and higher-order V-measures assign higher weights to completeness than homogeneity.

Table 2 demonstrates the effectiveness of adding signature motif information in protein sequence clustering. The performance of all the clustering algorithms in Table 2 can be improved when signature motif information is added and properly weighted. For example, in both hierarchical clustering and PSSM-based *k*-means clustering, the position probability weighting scheme significantly improves the V-measure values (Hierarchical B: 0.293, 1C: 0.278, 2C: 0.347, 3C: 0.508, 4C: 0.524, 5C: 0.540 / A: 0.273; PSSM-based *k*-means clustering: B: 0.366, 1C: 0.386, 2C: 0.428, 3C: 0.427, 4C: 0.353 / A: 0.342). The results again suggest that incorporating the signature motif information can actually improve the clustering results.

It is worth mentioning that from the experimental results we can also observe a correlation between the hierarchical clustering algorithm and the hybrid clustering algorithm. A higher V-measure value in the hierarchical clustering phase usually yields a better overall performance of the hybrid clustering algorithm. However, we do not observe any significant correlation between the *k*-means clustering and the hybrid clustering. This observation may suggest that it is the hierarchical clustering that mainly guides the hybrid clustering algorithm by forming the initial clusters, and the *k*-means clustering algorithm further refines the initial clustering results. However, *k*-means clustering, if performed alone, does not demonstrate the similar correlation as does the hierarchical clustering because its performance is largely affected by the initial partition.

3.3 Classification performance evaluation

In the classification stage, we predict the protein family of the 5,348 testing sequences in the testing data. The testing sequences are tested against all the profile HMMs produced from the clustering stage. Each testing sequence is assigned to the protein family with the highest sequence-against-model score. Ideally, the classification results should be evaluated against the ground truth which is the actual protein family. However, in this unsupervised clustering framework, it is not feasible to map a classifier to its protein family since it is possible that sequences from different protein families are merged together, which makes it sometimes hard to identify the model-family correspondence by just identifying the majority family of sequences in that cluster. One of the reasons is that the number of sequences in each family varies widely. Consequently, what could happen in clustering is that – if we associate a cluster with the family to which most of its cluster members belong, there could be more than one cluster (and their models) associated with one single family. Another reason that we do not explicitly associate a cluster model with a family is that, since the proposed framework is completely unsupervised, the ground truth is supposedly not known in priori. Sequences are clustered solely based on their sequence similarity and signature motifs. The output of unsupervised clustering, including both a set of clusters and their models, represent only groups of similar sequences based on the features (e.g., signature motifs) we choose to use. However, there is still a way to evaluate the classification performance by measuring the homogeneity and completeness of sequences in each cluster after classification against the ground truth family information. In particular, we adopt V-measure as the criterion for performance evaluation which provides a way of combining homogeneity and completeness measures into one single measure.

We run the proposed classification algorithm on 3 training-testing data pairs, each of which contains 5,348 testing sequences. Each testing sequence is tested against all the existing protein family classifiers produced during the hybrid clustering stage. In addition, we compare the classification performance based on the proposed HMM-based hybrid clustering with that based on the PSSM-based hybrid clustering. According to Table 2, PSSM-based k-means method produces its best V-measure (0.428) when the weighting scheme 2C is used. Therefore, 2C is used as the weighting scheme in the second stage of PSSM-based hybrid clustering for its comparison with our proposed method. We show the performance of classification in Table 3. In Table 3, the profile HMM classifiers produced from the proposed hybrid clustering framework, i.e., hierarchical clustering (with weighting scheme: 5C) followed by profile-HMM based *k*-means clustering, produces the best classification result (0.552); meanwhile, the hierarchical clustering (with weighting scheme: 5C) followed by PSSM based *k*-means clustering (with weighting scheme: 2C) has the lowest classification performance (0.187). This indicates that the proposed profile HMM based hybrid clustering algorithm outperforms the PSSM based hybrid clustering algorithm. In addition, the experimental results suggest that the classifiers produced from the proposed hybrid clustering framework also benefit from the use of signature motif information.

4. Conclusions and future work

In this paper, we propose an unsupervised clustering and classification framework for protein sequences. In the experimental results, we compare the proposed two-phase hybrid clustering algorithm with several other clustering algorithms, including the hierarchical clustering algorithm, PSSM-based *k*-means clustering algorithm, and HMM-based *k*-means clustering algorithm. The results demonstrate that the proposed two-phase hybrid clustering algorithm for sequence clustering outperforms its constituent hierarchical clustering (1st step) or *k*-means clustering (2nd step) alone. This suggests that the proposed two-phase hybrid clustering algorithm combines the strength of both the hierarchical agglomerative clustering and the profile HMMs based *k*-means partitioning clustering and avoids the problems arisen when using either of them alone.

In addition, when a new sequence is inserted into a cluster, the proposed framework is very flexible and efficient when it comes to updating the corresponding protein family model that represents a protein family. This is because the profile HMMs can be dynamically updated without recalculating the similarity between the newly inserted sequence and every other sequence in the same cluster, therefore, we can simply add a newly discovered protein sequence into its corresponding protein family and efficiently update the associated family model after successfully predicting its family label by testing the protein sequence against all the existing profile HMM classifiers.

Further, the results suggest that the performance of the sequence clustering can be greatly improved when increasing the weight on the signature motif regions. When signature motif information is added and properly weighted with the proposed probability weighting scheme, the V-measure values are usually higher than that of the union weighting scheme in most of the cases. This indicates that the proposed probability weighting scheme can better represent the common structural motifs in a set of protein sequences.

Moreover, to further improve clustering results, we strongly believe that increasing the weight for the signature motif in the 2nd stage, i.e., the profile HMMs-based *k*-means partitioning clustering, may help bring us closer to the goal since we have shown in our experimental results that introducing signature motif information into hierarchical clustering algorithm or the PSSM-based *k*-means partitioning clustering algorithm is effective. Therefore, it is highly possible that the performance of the profile HMM-based *k*-means clustering can be gained by adding the signature motif information. However, this is not a trivial task and needs further exploration in our future work.

Currently, we obtain the signature motif information by using the ScanProsite to detect existing motif patterns in the protein sequences. However, it is worth noting that not every protein sequence has signature motif information which can be detected by ScanProsite in the PROSITE database. In our case, there are in total 1,015 protein sequences whose signature motif information cannot be obtained in our experimental dataset. For these sequences, the proposed weighting schemes have no impact on differentiating them. This indicates that we must incorporate other structure information rather than signature motif information in order to properly separate those protein sequences when calculating their similarity scores.

5. Acknowledgement

The research of Dr. Chengcui Zhang is supported in part by NSF DBI-0649894 and UAB ADVANCE program.

6. References

[1] E. de Castro, C.J. Sigrist, A. Gattiker, V. Bulliard, P.S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch and N. Hulo, ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, *Nucleic Acids Research* **34** (2006),W362−W365.

[2] W.-B. Chen, C. Zhang and H. Zhong, *An unsupervised protein sequences clustering algorithm using functional domain information*, Proceedings of the 2008 IEEE International Conference on Information Reuse and Integration (IRI 2008), 2008, 76−81.

[3] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis - probabilistic models of proteins and nucleic acids*, Cambridge University Press, Cambridge, UK, 2000.

[4] S.R. Eddy, Profile hidden Markov models, *Bioinformatics* **14** (1998), 755−763.

[5] S.R. Eddy, Hidden Markov models, *Current Opinion in Structural Biology* **6** (1996), 361−365.

[6] The EMBL-EBI Tools. (http://www.ebi.ac.uk/Tools/clustalw2/index.html)

[7] A.J. Enright and C.A. Ouzounis, GeneRAGE: a robust algorithm for sequence clustering and domain detection, *Bioinformatics* **16** (2000), 451−457.

[8] M. Gribskov, A.D. McLachlan and D. Eisenberg, *Profile analysis: detection of distantly related proteins*, Proceedings of the National Academy of Sciences of the United States of America (PNAS) **84** (1987), 4355−4358.

[9] J. Han and M. Kamber, *Data mining: concepts and techniques*, San Francisco, CA, USA: Morgan Kaufmann Publishers, 2001.

[10] S. Henikoff and J.G. Henikoff, *Amino acid substitution matrices from protein blocks*, Proceedings of the National Academy of Sciences of the United States of America (PNAS) **89** (1992), 10915−10919.

[11] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B. Cuche, E. De Castro, C. Lachaize, P.S. Langendijk-Genevaux and C.J.A. Sigrist, The 20 years of PROSITE, *Nucleic Acids Research* **36** (2007), D245−D249.

[12] E.V. Kriventseva, M. Biswas and R. Apweiler, Clustering and analysis of protein families, *Current Opinion in Structural Biology* **11** (2001), 334−339.

[13] C.J. Ku and G. Yona, *Domain-based protein hierarchy and detection of semantically significant domain architectures*, Proceedings of the 13th Annual International Conference on Intelligent Systems for Molecular Biology, 2005.

[14] W. Mahdi, S. Werda and A.B. Hamadou, A hybrid approach for automatic lip localization and viseme classification to enhance visual speech recognition, *Integrated Computer-Aided Engineering* (*ICAE*), **15** (2008), 253−266.

[15] I.S. Main and I. Dubchak, Representing and reasoning about protein families using generative and discriminative methods, *Journal of computational biology: a journal of computational molecular cell biology* **7** (2000), 849−863.

[16] C.J. van Rijsbergen. *Information Retrieval*, Butterworths, London, 1979.

[17] A. Rosenberg and J. Hirschberg, *V-measure: a conditional entropy-based external cluster evaluation measure*, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, 410−420.

[18] I. Rudolfová and J. Zendulka, *Clustering of protein sequences*, Proceedings of 1st International Workshop WFM'06, 2006, 71−78.

[19] P.D. Thomas, M.J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan and A. Narechania, PANTHER: a library of protein families and subfamilies indexed by function, *Genome Research* **13** (2003), 2129−2141.

Wei-Bang Chen and Chengcui Zhang – 2nd Revision

Table Captions:

Table 1. Performance comparison of the proposed framework with other clustering algorithms

Table 2. Effectiveness of adding signature motif in sequence clustering

Table 3. Performance comparison of sequence classification

Table 1. Comparison of the proposed framework with other clustering algorithms

Clustering Method	**Weighting Scheme**	**Homogeneity**	**Completeness**	**V-measure**
Hierarchical	A	0.186	0.522	0.273
	B	0.211	0.515	0.293
	1C	0.187	0.545	0.278
	2C	0.252	0.558	0.347
	3C	0.444	0.604	0.508
	4C	0.477	0.582	0.524
	5C	0.504	0.581	0.540
HMM-based *k*-means (random initial)	A	0.234	0.488	0.314
PSSM-based *k*-means (random initial)	A	0.257	0.512	0.342
Hybrid (Hierarchical + HMM)	A-A	0.191	0.535	0.280
	B-A	0.220	0.526	0.305
	1C-A	0.191	0.551	0.283
	2C-A	0.262	0.579	0.360
	3C-A	0.437	0.667	0.526
	4C-A	0.471	0.682	0.556
	5C-A	0.493	0.663	**0.565**
Hybrid (Hierarchical + PSSM)	A-A	0.139	0.561	0.214
	B-A	0.167	0.536	0.242
	1C-A	0.142	0.561	0.218
	2C-A	0.186	0.542	0.276
	3C-A	0.316	0.616	0.412
	4C-A	0.320	0.584	0.409
	5C-A	0.347	0.610	0.440

Weighting Scheme:
A: Equal weighting scheme
B: Union weighting scheme
C: Probability weighting scheme

Note:
In hybrid clustering methods, the weighting schemes used at both steps are shown in the second column. For example, 2C-A indicates that the weighting scheme used in the first step is Scheme C with power 2, while the scheme used in the second step is Scheme A.

Table 2. Effectiveness of adding signature motif in sequence clustering

Clustering Method	**Weighting Scheme**	**Homogeneity**	**Completeness**	**V-measure**
Hierarchical	A	0.186	0.522	0.273
	B	0.211	0.515	0.293
	1C	0.187	0.545	0.278
	2C	0.252	0.558	0.347
	3C	0.444	0.604	0.508
	4C	0.477	0.582	0.524
	5C	0.504	0.581	0.540
PSSM-based *k*-means (random initial)	A	0.257	0.512	0.342
	B	0.280	0.531	0.366
	1C	0.292	0.579	0.386
	2C	0.335	0.606	0.428
	3C	0.336	0.593	0.427
	4C	0.267	0.523	0.353
	5C	0.248	0.509	0.334

Weighting Scheme:
A: Equal weighting scheme
B: Union weighting scheme
C: Probability weighting scheme

Table 3. Performance comparison of sequence classification

Classifiers					**Homogeneity**	**Completeness**	**V-measure**
ID	**Level 1 Clustering**		**Level 2 Clustering**				
Hybrid A-A	Hierarchical	A	HMM-based *k*-means	A	0.166	0.581	0.256
Hybrid B-A	Hierarchical	B	HMM-based *k*-means	A	0.209	0.572	0.298
Hybrid 1C-A	Hierarchical	1C	HMM-based *k*-means	A	0.163	0.580	0.254
Hybrid 2C-A	Hierarchical	2C	HMM-based *k*-means	A	0.245	0.613	0.349
Hybrid 3C-A	Hierarchical	3C	HMM-based *k*-means	A	0.410	0.705	0.517
Hybrid 4C-A	Hierarchical	4C	HMM-based *k*-means	A	0.436	0.724	0.543
Hybrid 5C-A	Hierarchical	5C	HMM-based *k*-means	A	0.458	0.698	0.552
Hybrid 5C-2C	Hierarchical	5C	PSSM-based *k*-means	2C	0.133	0.344	0.187

Figure captions

Fig. 1 The overview of the proposed protein sequence clustering and classification framework

Fig. 2 Signature motif position index mapping

Fig. 3 Two signature motif weighting schemes

Fig. 4 The state diagrams of the profile HMM

Fig. 1 The overview of the proposed protein sequence clustering and classification framework

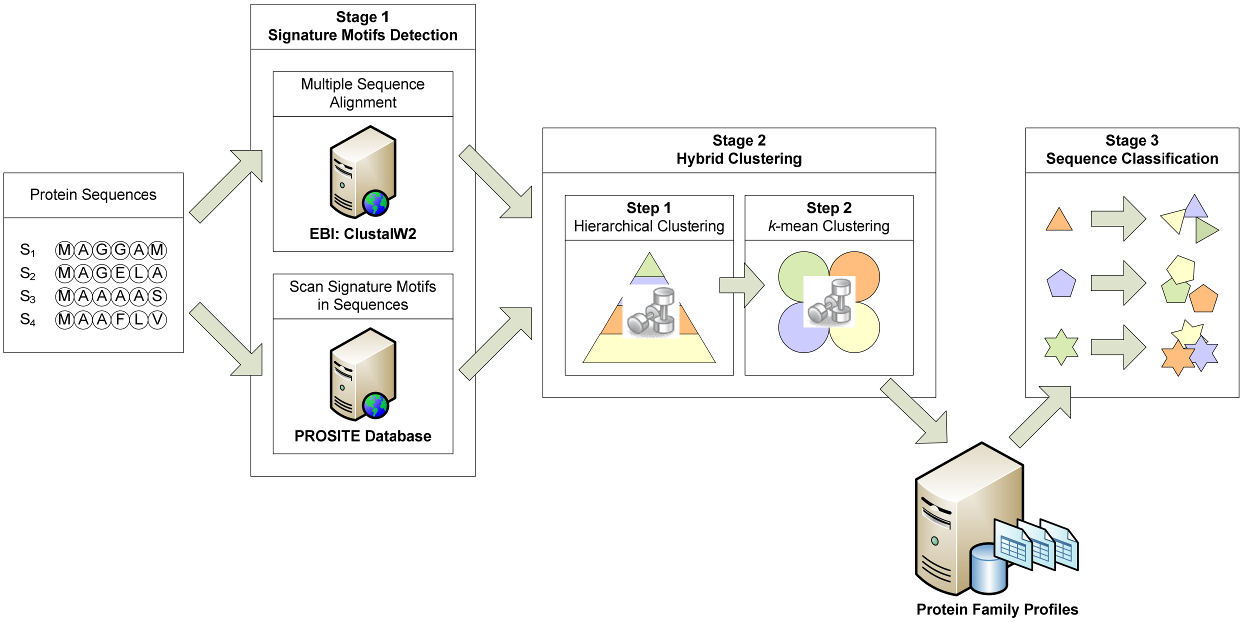


Fig. 2 Signature motif position index mapping

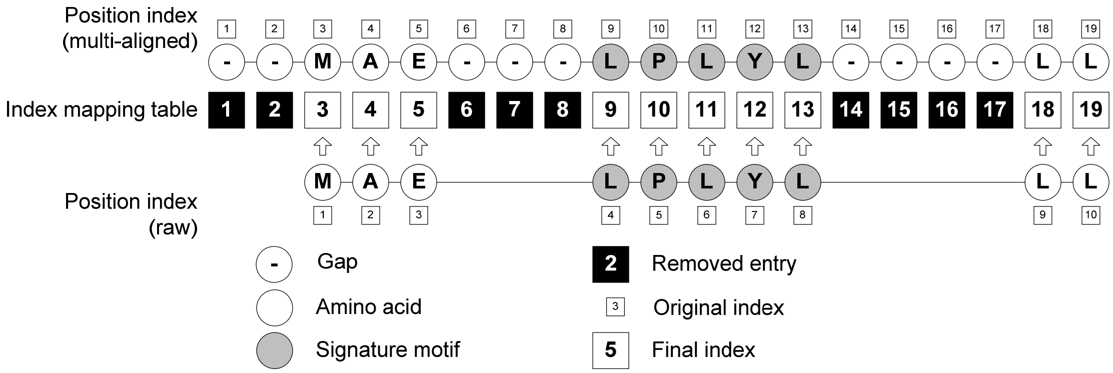


Fig. 3 Two signature motif weighting schemes

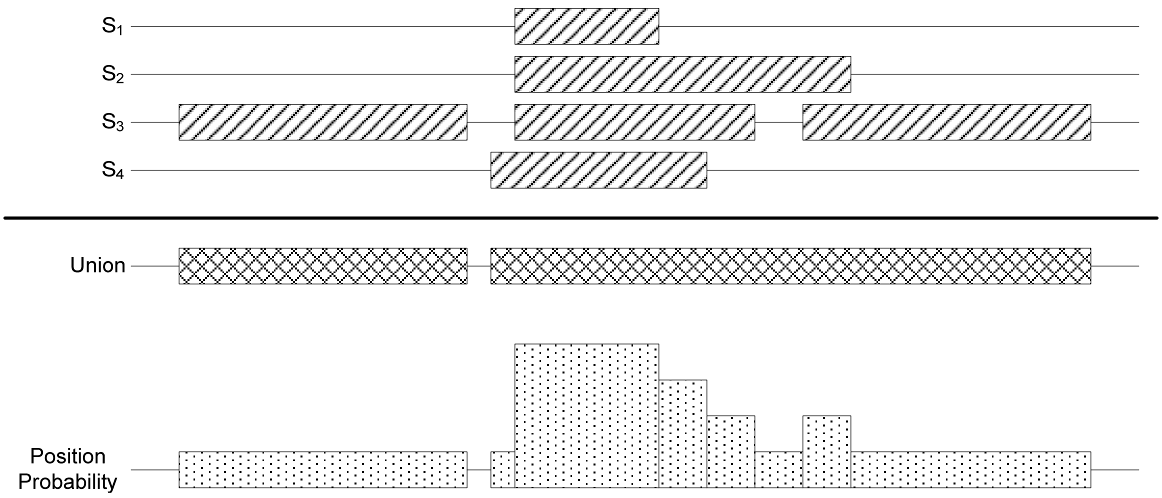
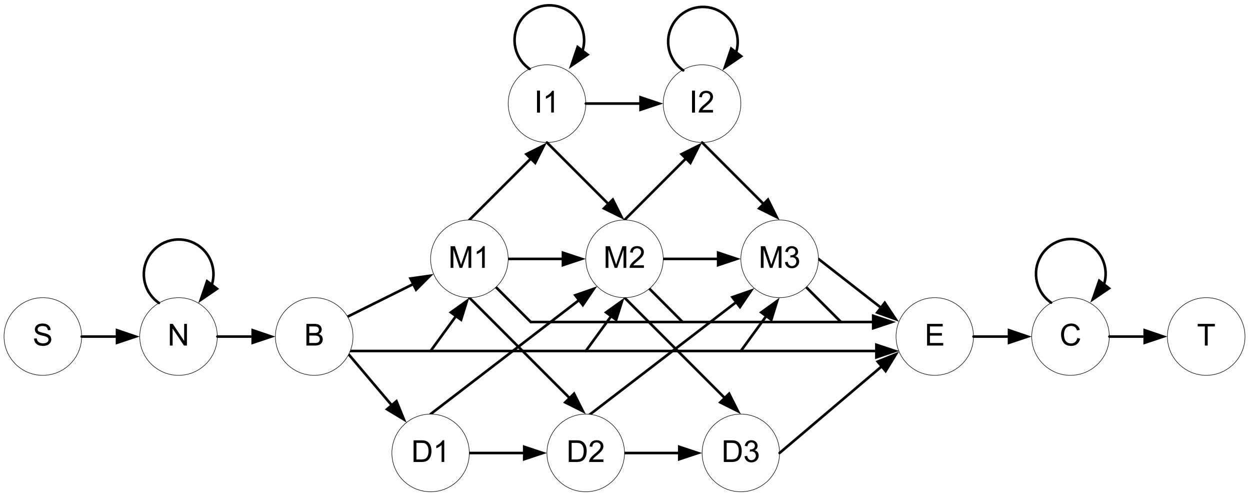


Fig. 4 The state diagrams of the profile HMM


I1
I2
M1
M2
M3
S
N
B
E
C
T
D1
D2
D3