

Classification algorithms and analyzing the functionality of protein families

L. Gao^{1,2} & D. K. Y. Chiu²

¹*Department of Computer Science and Engineering,
Xidian University, China*

²*Department of Computing and Information Science,
University of Guelph, Canada*

Abstract

The rapid growth of bio-sequence data has resulted in an increasing demand for reliable algorithms that group proteins in a meaningful way. Many traditional classification and clustering algorithms have been adapted or directly applied to protein sequences or structures. To capture protein functionality, new algorithms have recently been proposed specifically aiming at incorporating protein functions. In this paper, we review some of the classification and clustering algorithms for proteins. We divide algorithms into four categories based on their use of dissimilarity measure, density characteristics, computational modeling and information of evidence. The algorithm based on information of evidence analyzes the biomolecular sequences as discrete-valued n-tuples such that discrete values rather than their variables are selected as evidence for the final groupings. The advantage of this approach is that the configuration of the final clusters does not depend on a reliable distance measure, a predefined computational model of the clusters, reliability of the adaptive learning method or a measure of the density function. Finally, the methods are reviewed with respect to the quality of reflecting functionality of the protein family.

Keywords: protein families, protein functionality, classification algorithm, clustering algorithm.

1 Introduction

Proteins are building blocks of organisms and fundamental substances of life that play an important role in executing and regulating many biological processes.



There are only 20 different types of natural amino acids of a protein, but they can be combined to generate a large number of protein sequences. In the study of proteins, there are three important attributes: sequence, structure, and biological function [1]. The sequence is essentially the string of amino acids which comprises the protein. The structure of the protein is the way the protein is outlined, perhaps in the three dimensional space. Most important, yet elusive, is the protein function. The protein function reflects its actual role in the specific organism in which it exists. Understanding the protein function is critical for most applications, such as drug design, genetic engineering, or pure biological research.

The number of protein sequences in public databases grows rapidly with the progress of experimental technologies in molecular biology and large bioinformatics projects. The large size of sequence data makes it desirable to know the inter-relationships between proteins. Thus, cluster the sequence data into meaningful or manageable groups is widely used [2]. It is one of the most important subjects in bioinformatics to provide meaningful suggestions as to what the functional relationships between proteins are. In the past, clustering and classification algorithms have been widely recognized as basic techniques in statistics and computer science, and have been studied extensively. The major goal of clustering is to create a partition of objects such that objects in each group have similar characteristics. The resulting groupings can potentially reveal unknown groups or categories that may lead to a better understanding of the nature. For example, a common way to tackle the complexity of protein function prediction is to use database searches to find proteins that are similar, thus inferring the protein function. This method can be generalized to searching within the context of protein clustering or classification, and thus in principle, a more reliable analysis [3].

This paper is organized as follows. We start by describing the mathematical formulation and the challenges of clustering on protein families. The following sections describe classification or clustering algorithms based on the methodology used: similarity-based clustering, model-based clustering, density-based clustering and evidence-based clustering.

It is important to note that vast amount of research were made in the field of proteomics involving interactions between proteins. Herein, we focus on algorithms based on clustering and classifying protein family.

2 Mathematical formulations and background of protein clustering

Let $N = \{e_1, e_2, \dots, e_n\}$ be a set of n elements, and let $C = \{c_1, c_2, \dots, c_m\}$ be a partition of N into subsets. That is, the subsets are disjoint and their union is N . Each subset is called a cluster, and C is called a clustering solution, or simply a clustering [4]. The goal in a clustering problem is to partition the set of elements N into homogeneous, well-separated or discontinuous clusters. That is, we require that elements from the same cluster will be highly similar to each other, while elements from different clusters will have lower similarity to each other.



Note that this formulation does not define a single optimization problem: homogeneity and separation can be defined in various ways, leading to a variety of optimization problems. Even when the homogeneity and separation are precisely defined, those are two objectives which are typically conflicting: the higher the homogeneity, the lower the separation, and vice versa.

Clustering problems and algorithms are often represented in graph-theoretic terms. We therefore include some basic definitions on graphs.

For a vertex $v \in V$, define the weight of v to be the sum of weight of the edges incident on v . A cut C in G is a subset of its edges, whose removal disconnects G . The weight of C is the sum of the weights of its edges. A minimum weight cut is a cut in G with minimum weight. In case of non-negative edge weights, a minimum weight cut C partitions the vertices of G into two dis-joint non-empty subsets $A, B \subset V, A \cup B = V$, such that $E \cap \{(u, v) : u \in A, v \in B\} = C$. For a pair of vertices $u, v \in V$, the distance between u and v is the length of the shortest path which connects them.

Clustering and analysis of protein families in large database is one of the principal objectives in structural and functional genomics. Protein family clustering can significantly contribute to the depiction of functional diversity of homologous family. So far, there are many different databases based on sequences, structure, domain, motif, and phylogenetic profile, as well as the combination of sequence and structure. In analyzing the proteins, algorithms have to be devised corresponding to different database type according to how it is constructed [5,6].

3 Clustering algorithms

Many methods are currently available for the clustering of proteins into families, including hierarchical clustering, partitioning clustering, graph theoretic approaches, statistical approach, and neural network approach. Full-sequence analysis helps in classifying proteins based on the known sequences. However, protein function often stems from its protein structure. This gives rise to the idea of classifying protein based on structure [39]. The drawback of structured-based clustering is that the number of sequence for which structure is available is relatively small. The other aspect is to incorporate the protein functionality into the analysis. The difficulty in this case is that its precise nature may not be known. The analysis may then be handicapped by an objective that is circular. To put the algorithms into perspective, we review them based on their use of data attributes.

3.1 Similarity based clustering

One of the most common approaches towards classifying proteins is using sequence similarity. Sequence similarity is a well-studied subject, and numerous software packages for biological sequence are available. Often similarity in their sequence, structure and biological functions also implies a common ancestral relationship. However, biological functions of a protein often are treated

separately from their sequences or structures, the clusters generated may not convey this information

3.1.1 Hierarchical clustering

Hierarchical algorithms create a hierarchical decomposition of set N . Clustering solutions are typically represented by a dendrogram, a tree that iteratively split N into smaller subsets until each subset consists of only one subject. The dendrogram can either be created from the leaves up to the root (agglomerative approach) or from the root down to the leaves (divisive approach) by merging or dividing clusters at each step. For protein sequence clustering, the typical algorithms of hierarchical clustering are single linkage (SL) clustering [7–11], SL-KL clustering [2], and p -quasi complete linkage [8,12].

SL clustering is one of the simplest clustering methods and is widely used for the clustering of protein sequences. Firstly constructing a linkage graph: given a set of protein sequences S and a cutoff threshold C for similarity, a linkage graph $G=(V,E)$ is a simple graph such that each vertex $v \in V$ represents a protein sequence (i.e. $|V|=|S|$ and there exist an edge $e = \{v_p, v_q\}$ ($e \in E, v_p, v_q \in V$ if $H(P, Q) \geq c$ in any two of protein sequences P and Q ($P, Q \in S$) and v_p and v_q represent P and Q respectively. A set of protein sequence S is to divide S into $G_1, G_2, \dots, G_n$ such that each $G_i (1 \leq i \leq n)$ is any of maximal connected subgraphs or an isolated vertex with no incident edge in the linkage graph of S .

Although this method is simple, fast and widely used, it has difficulty in detecting an appropriate cutoff score for sequence similarity. Similarities between proteins sharing a signature vary greatly in terms of the length of the signature and the conservation of their residues. In addition, a protein is sometimes similar to another at the sequence level even though they do not share any signatures. Thus, its relationship to functional similarity may not be immediate.

Currently protein-clustering applications use either a rigid cutoff score or a set of cutoff scores with different levels [13]. An SL-based clustering method, named SL-KL classifies proteins with different levels that are selected automatically [2]. It uses a graph representation of a given set of proteins, whose vertices denote proteins and edges denote pairs of proteins having sequence similarities above a certain cutoff score.

To find distant relationships, a p -quasi complete linkage algorithm, also use the graph representation, is proposed. A p -quasi complete graph $QC=(V,E)$ is a graph such that $\deg(v) \geq \lceil p(|V|-1) \rceil$, for all $v \in V$, where $\deg(v)$, p and $|V|$ denote the degree of a vertex v , connectivity ration of QC ($0 \leq p \leq 1$) and the number of all vertices in V , respectively. Clearly a 1-quasi complete graph is a complete graph. Thus a connectivity ration p means how close to a complete graph. The goal is to search for the minimum number of the covers of given protein sequences such that each cover is represented as a maximal p -quasi complete subgraph (or an isolated vertex) in the linkage graph of the protein sequences. This method allows some proteins to overlap between two or more

groups. This overlap is essential for clustering a protein set containing multi-domain structure proteins that have two or more signatures (referred to motif or domain). However, its application to a large number of proteins is difficult to its computational cost.

3.1.2 Graph theoretical approaches

A number of methods on graph decomposition based on the connectivity among vertices have been proposed to identify protein families. Enright et al. [11] adopts a Markov cluster algorithm based on graph representation, which can be applicable for finding distant relationships by changing a parameter affecting its cluster granularity. The method proposed by Abascal and Valencia [14] adopts a normalized cut algorithm. The method differs from the other methods in that it focuses on protein carefully retrieved by iterative homology searches to avoid spurious homologs. In most case, proteins with a significant sequence similarity are related functionally, though many exceptions exist. Such similarity is widely used to infer protein functions. However, retrieved proteins with apparent similarities are often uninformative, such as similarities to hypothetical proteins. Similarities between the queries and well-characterized proteins are often modest at best. It is not easy to predict its function from such modest similarities. To cope with this problem, it is important to incorporate alternative approaches, such as signature identification and sequence clustering. Kawaji et al. [13] developed a time and space efficient clustering algorithm. It uses a graph representation where its vertices and edges denote proteins and their sequence similarity above a certain cutoff score. It repeatedly partitions the graph by removing edges that have small weights, which correspond to lower sequence similarities. To find the appropriate partitions, they introduced a score combining the normalized cut and a locally minimal cut capacities. For comparison with other clustering methods, such as NCUT, SL-KL and SL, they executed those algorithms using their experimental data.

It is concluded that the normalized cut criterion is more appropriate for modeling distantly-related protein groups. On the other hand, graph partitioning does not find all the relationships correctly for multi-domain proteins. A scalable clustering method according to such structure could be more desirable.

3.1.3 *K*-means clustering

K-means is another commonly used clustering algorithm and is a typical partitioning algorithm. Partitioning algorithms construct a partition of a database *D* of *n* objects into a set of *k* cluster. *k* is an input parameter for these algorithms, which typically starts with an initial partition of *D* and then uses an iterative control strategy to optimize an objective function.

We have seen many methods have been developed for clustering proteins according to their sequence similarity. However, most of these methods tend to have a computational complexity that is at least quadratic on the number of sequences, as they require an all-against-all initial analysis. Furthermore, these methods require a reliable distance measure. When different attributes from sequence, structure and functionality are considered, a meaningful distance measure may be difficult to define.



Guralnik and Karypis [15] present an entirely different approach to protein clustering that does not require all against-all analysis and uses a near-linear complexity k -means based clustering algorithm, which is still based on similarity. The key idea is to find a set of feature that capture the sequential nature of the various proteins, project each protein into a new space whose dimensions are these features, the features refer to frequent motifs, and then use a traditional vector-space k -means based clustering algorithm to find the protein clusters in that space. Their experimental evaluation shows this approach appears promising and leads to reasonably good clusters. However, some of the resulting clusters contain proteins from different families. The proteins in those families do not share long conserved motifs. On the other hand, it loses some of the important information, such as the order of the motifs and their position in each sequence. To further improve performance of clustering algorithm, they have developed a centroid-based approach and have had a preliminary result.

3.2 Model based clustering

3.2.1 Support vector machine (SVM)

Support vector machine (SVM) is a classification algorithm that provides good performance in many application domains. SVM is a special kind of feed forward neural network which approximately implements structural risk minimization. This is different from general neural network using empirical risk minimization. SVM uses a set of training data to build a classifier by weighing each training item by how much it contributes to the overall classification. During the past three years, SVM has been applied very broadly within the field of computational biology, to pattern recognition problems including protein remote homology detection, microarray gene expression analysis, functional classification of promoter regions, prediction of protein-protein interactions [17]. The advantage of this approach is that the problem of scaling in combining different features can be by-passed using an adaptive training method.

Standard SVM (or binary SVM) classify objects into two classes by finding a hyperplane that separates a given training set according to these classes in a best way. Since SVM is only capable of making binary decisions, it is necessary to enhance them to distinguish between more than two classes and set-valued predictions, which are one-versus-rest, one-versus-one and voting vector [18]. A drawback of all of the mentioned approaches of multi-class SVM is that the extra effort for introducing another class leads to the use of additional binary SVM distinguishing the new class. To avoid this overhead, Krieger et al. [19] proposed a new idea. They do not extend the class set at once. Instead they refine the post processing of the voting vectors gained from the one-versus-one approach. A thorough experimental evaluation of their algorithm for classifying proteins from SWISS-PROT is tested, demonstrating that their method is capable to classify multi-class with high accuracy and an efficient adequate for real world application.

SVM is a good technique for pattern recognition. The complex part is the optimization of the variable. A good feature of SVM is that no problem domain

knowledge is required. The key in designing a SVM is in choosing the correct kernel function.

3.2.2 Hidden Markov models (HMMs)

A hidden Markov model is a stochastic process which attempts to model a family of sequences. With the model, it is able to provide the probability of other test sequences belonging to that family. A HMM uses the concept of a finite state machine where each transition between a state has a probability of being taken and each state has an output probability. These models can then be compared with a test sequence and a similarity score produced. It has been shown that the state machines have a direct equivalence with regular expression. Therefore hidden Markov models can be used to model sequences of characters. Protein sequences can be represented as a linear sequence of characters, so some researcher studied how HMM could be used to model protein sequence and their study produced a promising result ^[16]. Since then a great deal of research has gone into refining HMM for protein sequence analysis [21–23].

The problem of detecting and classifying multi-domain proteins has been address by a number of approaches, which rely on the detection of individual domains using BLAST reports, domain database dictionaries or iterative sequence comparison. Some of these methods rely on manual intervention for family assignment of multi-domain proteins. All the above mentioned techniques suffer from a number of drawbacks: they can be too computationally intensive, somewhat inaccurate or not fully automatic. Enright et al. [11] presented a novel algorithm that generates accurate protein families using MCL (Markov Cluster) formalism for graph clustering by flow simulation. The actual implementation of the algorithm allows the efficient and rapid clustering of any arbitrary set of protein sequences. This method does not operate directly on sequence but on a graph that contains similarity information, it avoids the expensive steps of sequence alignment. Instead, global patterns of sequence similarity are detected and used to partition the similarity graph into protein families.

Using HMM to model protein sequence has a fundamental flaw [16]. Protein sequence have three-dimension shape which is based roughly on amino acid sequence, only certain three-dimensional shapes are stable as there are bonds between the different amino acid sequences. This is important as the three-dimensional structure of a protein relates to its function and consequently the family it belongs too. The main flaw using a generative HMM is that it may not be flexible enough to accounts for the bonds between the amino acids that are not next to each other. Following on from this, one of the ways of making the implementation of an HMM easier is using the assumption that the output of one state does have bearing on the output of another state, in another words they have independent probabilities. Unfortunately, in biology, this is not the case and the output of one state has a large bearing on the output of the next state. To get around these problems a much richer model is needed [16].

3.2.3 Combining SVM and HMMs

Several statistical models have been created to aid in the classification but so far they have only had partial success. They can give suggestions to the protein's



superfamily, but may not be a definite answer. Jaakkola et al. [24] proposed a method combine a hidden Markov model with the discriminate technique of a SVM. This method builds on the previous work of using a HMM but uses a different approach to decode the model. It is claimed that this method is able to provide better results than just using a HMM. This method is known as SVM-Fisher method for detecting remote protein homologies. The SVM-fisher method is able to provide superior results than just using a generative HMM, so it relies on a good HMM model, but can still improve the results from a poor HMM model. One major problem with the current method is that it does not analyze how the amino acids relate to each other in different parts of the amino acid sequences. This is a major drawback with the current theory behind the SVM-fisher method and in particular with the HMM being used to model the data [16]. Although there are some drawbacks about this method, it is still a very promising method for the classification of protein sequence.

3.2.4 Neural networks

Neural networks provide a new suite of nonlinear algorithm for classification. Neural networks have been chosen as technical tools for the protein sequence classification due to the following two reasons [25]: (1) the extracted features of protein sequences are distributed in a high dimensional space with complex characteristics which is difficult to satisfactorily model using some statistical or parameterized approaches; and (2) neural networks are able to use the raw continuous values as system input. Basically, there are two types of neural models applicable for protein sequence classification task, i.e., supervised feed-forward networks (FNNs), which includes MuLtilayer Perceptron (MLP) and Radial-Basis Function (RBF) networks [27,28], and unsupervised Self-Organizing Mapping (SOM) [26], which is mainly used for data clustering. The use of SOM is to discover relationships within a set of protein sequence by clustering them into different groups. In contrast, FNN based classification system emphasizes on matching patterns through supervised learning.

Radial-Basis Function (RBF) networks have received much attention due to their learning efficiency. Comparative studies showed that the RBF network outperforms the MLP model in both classification performance and learning efficiency [25]. D.H. Wang et al. have used generalized radial basis function (GRBF) [25] and modular radial basis function (MRBF) [28] for protein sequence classification. For GRBF approach, a set of fuzzy classification rules with confidence factors can be extracted from this network. The initial fuzzy rule set is refined using a new objective function, which compromises between misclassification rate and generalization capability. For MRBF approach, it is designed to combine with a compensational combination at the transition output layer. The connection weights between the final output layer and the transition output layer are optimized to enhance the classification ability. For the two methods, it is required to increase the recognition rate for super-families with low-probability items, reliability and robustness.

Self-Organizing Mapping (SOM) has been used as a divisive clustering approach in many areas. The most prominent disadvantage of SOM approach is

that it is difficult to know when to stop the algorithm. If the map is allowed to grow infinitely, the size of SOM is gradually increased to a point when clearly different sets of expression patterns are identified. Therefore, as with k-means clustering, the user has to rely on other source of information to determine the number of clustering that best present the available data. There are many varieties to SOM, among which the self-organizing trees are a constructive neural network methods that develop into a tree (binary tree) topology during learning [30]. Wang et al. [26] developed self-organizing tree algorithm (SOTA) to analyze pre-aligned sequences of genes (polygenetic reconstruction) in order to classify protein sequences. SOTA is a hierarchical neural network that grows into a binary tree topology. For this reason, SOTA can be considered as a hierarchical clustering algorithm.

Bayesian Neural Network (BNN) is another technique that has been recently used for protein sequence analysis. J.T.L Wang et al. [29] have used the BNN with structural learning for extracting features from protein sequences. The BNN is an important additional to neural network solutions that have been offered to the problem at hand, as they represent a large group of hybrid neural network that combine classical neural network with statistical classification and predication theories [30].

3.3 Density based clustering

Most partitioning methods cluster objects based on a defined distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulties at discovering clusters of arbitrary “shapes” or configurations. In this case, a method called density-based was developed [31,35]. Density-based methods can be used to filter out noise (outliers). Some density-based methods may be able to detect clusters that are not “spherically”-configured.

There are two major approaches for density-based methods [32]. The first approach pins density to a training data point and is reviewed in the sub-section density-based connectivity. The typical algorithm is the DBSCAN algorithm [33]. This algorithm is the first density-based algorithm which judges the density around the neighborhood of an object to be sufficiently dense. The second approach pins density to a point in the attribute space and is explained in the sub-section density functions. Representative algorithm is DENCLUE [34], which models the overall density of a point analytically as the sum of influence functions of a data points around it.

Clustering algorithm based on density aims at detecting high-density regions forming clustering when separated by a low-density inter-cluster region. In this approach, a reliable estimation of density such as reliable probability estimation on a discrete-valued n-tuple is required. Normally the sample size of the data set may not be sufficient for a reliable estimation. For a large n value of n-tuple, that is, data with a large number of components, the sample requirement may be unattainable. Although estimations based on attribute independence of the data component and low-order approximation estimation of n-tuple have been proposed, these assumptions may not be realistic. The drawback of these approaches is that density based on all attribute values of the data component

may not be a good indicator of functionality in many applications including biomolecular data due to the presence of irrelevant attribute values [36].

3.4 Evidence based clustering

One important goal of the evidence-based approach to clustering is to capture the functionality of the biomolecule into the clustering process. However, the functionality that may affect the groupings may be unknown. Evidence-based algorithms [36] are usually divided into several phases. The initial phase identifies the initial clusters using tentative, but reasonable criteria. Then values of the attributes that are relevant to the initial clusters are statistically identified. The final phase refines the clusters based evidences as reflected from selected values of the attributes. The initial clusters are only temporary, defined in such a way so that values selected are reliable, possibly ignoring the uncertain cases.

A method has been proposed previously for protein clustering [36]. Amino acid residues of the sequences are selected using statistical analysis based on significant associations between the residues and the initial clusters, and the residues with each other. Then the selected values are combined using an information measure to infer the final clusters. The information measure, known as weight of evidence, is defined as: $W(\text{class } C \text{ against other classes} \mid \text{selected amino acids in the sequence } \{A_1, A_2, \dots, A_m\}) = \sum_i W(\text{class } C \text{ against other classes} \mid A_i)$. The term to be summed indicates the calculated evidence from each selected amino acid and is defined as: $W(\text{class } C \text{ against other classes} \mid A_i) = \log P(A_i \mid \text{class } C) / P(A_i \mid \text{class not-}C)$. where P is the estimated probability possibly from the sequences to be clustered. The sequence is classified into a class C if W is maximized.

The advantage of this approach is that the configuration of the final clusters does not depend on a reliable distance measure, a predefined computational model of the clusters, reliability of the adaptive learning method or a measure of the density function. In fact the evidence aims at the underlying functionality of the proteins based on the detected significant associations, as reasonable indication even when the functionality is unknown.

4 Conclusions

Grouping proteins into families is important both for biological and computational reasons. The comprehensive classification of proteins into similarity groups is an important but a difficult challenge. So far, diversely different algorithms have been used to generate different results, some algorithms based on sequence similarity, some based on the defined domain and motif, some based on full sequence, and some based on the protein structure. Nevertheless, the most important question is how the characteristics of the groups and the grouping configuration enhance our understanding of the biological nature. Often it means incorporating biological functionality into the analysis either before or after the groups are generated, explicitly from domains or motifs, or implicitly in the algorithm. On the other hand, the relevant functionality may not be known.



While similarity-based algorithm is the most widely used, it is not likely to be sufficient in all cases. One problem is the reliability of the similarity measure and the choice of threshold in generating the groups. Another problem is the detection of multi-domain structure of many protein families. Proteins containing multiple domains can confound these methods and result in incorrect groupings. Proteins with similar domain are associated with each other, even if they have lower similarity score. However, domain-based classification may not be reliable in the sense that many proteins have several domain appearances and some proteins do not have any domain [37,40]. The basic idea of most full-sequence based classification is that of homology transitivity [3,38]. Similarity is not necessarily a transitive relation. Furthermore, using all the amino acids in the sequence without considering their relevance and as equal weight may not be appropriate. After all, the ultimate goal in understanding these biomolecules in nature requires a complementary consideration of the different functional aspects of the individual protein, whether they are known or unknown, in the bigger context of the protein family.

References

- [1] Setubal & Meidanis, Introduction to computational molecular biology, Brooks/Cole Publishing Company, 1999.
- [2] H. Kawaji, Y. Yamaguchi, H. Matsuda & A. Hashimoto, A graph-based clustering method for a large set of sequence using a graph partitioning algorithm. *Genome Informatics*, 12, pp. 93-102, 2001.
- [3] Ori Sasson & Michal Linial, Protein clustering and classification. www.ls.huji.ac.il/~michall/papers/protein_clustering_and_classification.pdf.
- [4] Anil K. Jain & Richard C. Dubes, Algorithms for clustering data. Prentice Hall, 1988.
- [5] R.L. Tatusov, E.V. Koonin & D.J. Lipman, A genomic perspective on protein families. *Science*, 278(24), pp.631~637, 1997.
- [6] M. Pellegrini, Computational methods for protein function analysis. *Curr Opin Chem Biol*, 5(1), pp. 46-50, 2001.
- [7] Watanabe H & Otsuka J, A comprehensive representation of extensive similarity linkage between large numbers of proteins. *Comput Appl Biosci.*, 11(2), pp.159-66, 1995.
- [8] H. Matsuda, T. Ishihara & A. Hashimoto, Classifying molecular sequences using a linkage graph with their pairwise similarities. *Theoretical Computer Science*, 210(2), pp.305-325, 1999.
- [9] Kriventseva EV, Biswas M & Apweiler R, Clustering and analysis of protein families. *Curr. Opin.Struct. Biol.*, 11(3), pp.334-339, 2001.
- [10] A.J. Enright & C.A. Ouzounis, GeneRAGE: A robust algorithm for sequence clustering and domain detection. *Bioinformatics*, 16(5), pp. 451-457, 2000.
- [11] A.J. Enright, S.V. Dongen & C.A. Ouzounis, An efficient algorithm for a large-scale detection of protein families. *Nucleic Acids Research*, 30, (7), pp. 1575-1584, 2002.



- [12] S. Seno, R. Teramoto & H. Matsuda, P-quasi complete linkage analysis for gene-expression data. Proceeding of the IEEE Computer Society Bioinformatics Conference (CSB'02), August 14-16, Stanford, California. pp. 342, 2002.
- [13] H. Kawaji, Y. Takenaka & H. Matsuda, Graph-based clustering for finding distant relationships in a large set of protein sequences. Bioinformatics, 20, (2), pp.243-252, 2004.
- [14] F. Abascal & A. Valencia, Clustering of proximal sequence space for the identification of protein families. Bioinformatics, 18 (7), pp. 908-921, 2002.
- [15] Valerie Guralnik & George Karypis, A scalable algorithm for clustering protein sequences. Workshop on data mining in Bioinformatics (with SIGKDD01 Conference), pp.73-80, 2001.
- [16] Simon Payne. Classification of protein sequences into homogenous families. MSc. Thesis, University of Sheffield, 2001.
- [17] W.S. Noble, Support vector machine applications in computational biology. Kernel Methods in Computational biology, MIT Press, Cambridge, MA, 2004.
- [18] J. Platt, N. Cristianini & J. Shawe-Taylor. Margin DAGs for Multiclass Classification. Proceedings of Neural Information Processing Systems, NIPS'99, MIT Press, pp. 547-553, 2000.
- [19] H.P. Kriegel, P.Kröger, A. Pryakhin & M. Schubert, Using support vector machines for classifying large sets of multi-represented objects. Proc. 4th SIAM Int. Conf. on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, pp.102-114, 2004.
- [20] P. Mitra, C.A. Murthy & S.K. Pal, A probabilistic active support vector-learning algorithm. IEEE Trans. On Pattern Analysis and Machine Intelligence, 26(3), pp. 413-418, 2004.
- [21] A. Krogh, M. Brown, I.S. Mian, K. Sjolander & D. Haussler, Hidden Markov models in computational biology: applications to protein modelling. Journal of Molecular Biology, 235(5), pp. 1501-1531, 1994.
- [22] R. Hughey & A. Krogh. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Computer Applications in the Bioscience, 12(2), pp.: 95-107, 1996.
- [23] S. Eddy, G. Mitchison & R. Durbin, Maximum discrimination hidden Markov models of sequence consensus. Journal of Computational Biology, 2(1), pp.9-23, 1995.
- [24] T. Jaakkola, M. Diekhans & D. Haussler, A discriminative framework for detecting remote protein homologies. Journal of Computational Biology, 7(1), pp. 95-114, 2000.
- [25] D.H. Wang, N.K. Lee & T.S. Dillon, Extraction and optimization of fuzzy protein sequence classification rules using GBRF neural networks. Neural Information Processing, 1(1), pp. 53-58, 2003.
- [26] H. C. Wang, J. Dopazo, L.G. Delafraga, Y. P. Zhu & J. M. Carazo, Self-organizing tree-growing network for the classification of protein sequences. Protein Science, 7(12), pp. 2613 ~ 2622, 1998.

- [27] C. Wu, G. Whitson, J. Mclarty, A. Ermongkonchai & T. C. Chang, Protein classification artificial neural system. *Protein Science*, 1(5), pp. 667 - 677, 1992.
- [28] D.H. Wang, N.K. Lee, T.S. Dillon & N.J. Hoogenraad, Protein sequences classification using radial basis function (RBF) neural networks. R.I. McKay, J. Slaney (Eds.), 15th Australian Joint Conf. on Artificial Intelligence, Canberra, Australia, December 2-6, pp. 477-486, Springer 2002.
- [29] J.T.L Wang, Q. Ma, D. Shasha & C.H. Wu, New techniques for extracting features from protein sequence, *IMB System Journal*, 40(2), pp. 426-441, 2001.
- [30] F. Valafar, Neural network applications in biological sequencing, rohan.sdsu.edu/~faramarz/papers/NNreport.pdf.
- [31] J. Han, M. Kamber & A. Tung, Spatial Clustering Methods in Data Mining: A Survey. H. Miller and J. Han (Eds.), *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis, 2001.
- [32] P. Berkhin, Survey of clustering data mining techniques, umdrive.memphis.edu/vphan/public/berkhin-survey.pdf
- [33] M. Ester, H.-P. Kriegel, J. Sander & X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceeding of the 2nd ACM SIGKDD*, pp.226-231, Portland, Oregon, 1996.
- [34] A. Hinneburg & D.A. Keim, An efficient approaches to clustering in multimedia database with noise. *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pp.58-65, 1998.
- [35] C. Domeniconi & D. Gunopulos, An efficient density-based approach for data mining task. *Knowledge and Information System*, 6(6), pp. 750-770, 2004.
- [36] A.K.C. Wang, D.K.Y. Chiu & W.H. Huang. A discrete-valued clustering algorithm with applications to biomolecular data. *Information Science*, 139(1), pp. 97~112, 2001.
- [37] Jinfeng Liu & B. Rost, Domains, motifs and clusters in the protein universe. *Current Opinion in Chemical Biology*, 7(1), pp. 5-11, 2003.
- [38] P. Pipenbacher, A.S. Chilep, S. Schneckener, A. Schö nhuth, D. Schomburg & R. Schrader, ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformaticc*, 1(1) , pp.1-10, 2002.
- [39] Eidhammer, I. Jonassen & W.R.I. Taylor, *Protein bioinformatics: an algorithmic approach to sequence and structure analysis*, Wiley, 2004.
- [40] L. R. Mayor, K. P. Fleming, A.E. Müller , D. J. Balding & M. J.E. Sternberg, Clustering of Protein Domains in the Human Genome, *Journal of Molecular Biology*, 340(5) , pp.991-1004, 2004.

