INTERACTIVE CLUSTERING FOR EXPLORATION OF GENOMIC DATA

XIUFENG WAN
xw6@cs.msstate.edu
Department of Computer Science
Mississippi State University
Box 9637
Mississippi State, MS 39762

SUSAN M. BRIDGES
bridges@cs.msstate.edu
Department of Computer Science
Mississippi State University
Box 9637
Mississippi State, MS 39762

JOHN A. BOYLE
jab@ra.msstate.edu
Department of Biochemistry and
Molecular Biology
Mississippi State University
Box 9650
Mississippi State, MS 39762

ALAN P. BOYLE
apb22@cs.msstate.edu
Department of Computer Science
Mississippi State University
Box 9637
Mississippi State, MS 39762

ABSTRACT

The complete genomic sequences for many organisms, particularly primitive organisms with relatively small genomes (prokaryotes), are now available. We describe an approach that supports interactive exploration of patterns in genomic data by combining use of positional weight matrices, the k-means clustering algorithm, and a visualization tool. Users interact with the system by examining a visualization of the "average" pattern found in each cluster for the sequence under consideration and determine if further clustering or modified clustering is desired. The effectiveness of this approach is demonstrated by a study of promoter sequences in archaea.

INTRODUCTION

Clustering has received renewed attention within the last ten years as a field of study within knowledge discovery and data mining. The massive amounts of data that have become available in a variety of fields has prompted new research in the use of traditional clustering algorithms and the development of new algorithms [8]. Key problems to deal with in cluster analysis are determining which features to use to generate meaningful clusters and interpreting cluster results. Interactive clustering has received substantial attention within the text mining community [1, 16]. We describe an interactive, iterative clustering approach for exploration of genomic data that allows scientists to visualize cluster results and direct the clustering process. This method allows incremental exploration of clusters of sequential patterns.

Gene clustering attempts to partition sequences composed from an alphabet of four nucleotide bases (A,T,C,G) into different groups based on a feature vector representation of each sequence. Gene clustering techniques have been used to explore a variety of problems including expression profile analysis, promoter identification, mRNA splicing site detection, and regulon prediction [2, 6, 15, 24]. The performance of each gene clustering method varies according to feature vector representation and clustering algorithm implementation as well as the specific problem [8, 10].

The method we describe uses a positional weight matrix for comparing sequences, the k-means clustering algorithm (with $k = 2$ at each step) to generate clusters, and a visualization tool that allows the user to analyze clustering results and direct further clustering. Our use of this approach in the study of promoter sequences in several species of archaea has yielded interesting scientific results.

In the remainder of this paper, we present related work, describe the method we have developed, introduce the genomic question that we have studied using the method and present results of one study.

RELATED WORK

Clustering methods have been widely applied in data analysis in many fields [8]. In their review of clustering methods, Jain et al. [10] report that the k-means clustering algorithm is one of the most frequently and successfully used methods. It is popular for clustering large datasets due to its simplicity and its small time and space complexity [10]. Different variants of k-means with regard to pattern representation, feature selection and extraction, similarity measures, and initial partition selections have been described [8, 10]. Recently, k-means clustering has been successfully applied to cluster gene expression data [17].

Positional weight matrices (PWMs) are a statistical model for representing the feature vector using probabilistic values at each position of the data set. Since their introduction into biological sequence analysis, PWMs have been utilized as a standard method to represent the promoter signal [4, 5, 21] and have been successfully used for promoter prediction [7, 12, 14]. This representation has also been used to evaluate the transcription factor bindability of DNA sequences [23].

The combination of positional weight matrices and neural network clustering algorithms was described initially to predict the O-glycosylation sites in proteins [9]. Murakami and Takagi [15] combined the k-means clustering algorithm and positional weight matrices for the detection of 5’ terminus of the splicing sites of mRNA. They report that their method exhibited better performance than using positional weight matrices alone. Recently, Van et al. [24] combined positional weight matrices and Monte Carlo sampling partition methods to successfully identified regulatory sites in the genome sequence of *E. coli*.

INTERACTIVE GENE CLUSTERING

We describe a method for studying groups of genes from a single genome or group of related genomes that allows scientists to explore different sequence patterns. Figure 1 illustrates our interactive, iterative gene clustering approach.

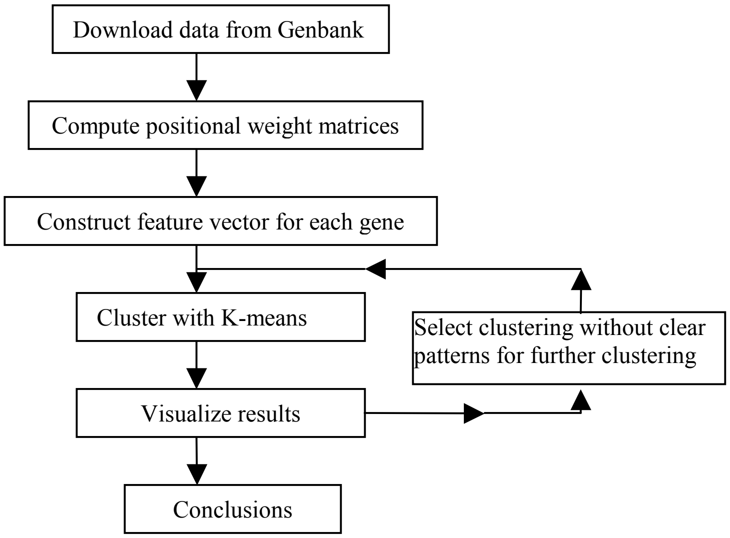
Download data from Genbank
Compute positional weight matrices
Construct feature vector for each gene
Cluster with K-means
Select clustering without clear patterns for further clustering
Visualize results
Conclusions

Figure 1 Interactive gene clustering algorithm.

The sequence data for a genome is downloaded from Genbank or some other repository, the subsequences of interest relative the beginning or end of each gene are extracted and the positional weight matrix is computed. Feature vectors are constructed using the positional weight matrix and these feature vectors are clustered. A visualization of each cluster is produced using a positional weight matrix for each cluster. At this point the user can choose to terminate the clustering process or can invoke the clustering algorithm on one of the clusters. The result is a binary tree of clusters. We have applied this process to the 5' flanking region of genes of species of archaea in order to identify different regulatory patterns.

Biological sequences such as DNA and protein can be thought of as strings of characters from an alphabet. DNA sequences use a four character alphabet (A, T, C, G). Positional weight matrices are used to represent the probability of each character at each position in a sequence. Given a set of sequences S = {S_1, S_2, S_3, …, S_m}, the positional weight matrix of S can be computed as

$$\text{PWM(S)} = \log((\sum_{i=1}^{n}\sum_{j=1}^{m} P(i, x_{i,j}))/f_i)$$

where *n* is the length of the sequence, *m* is the number of sequences, and f_i is the expected frequency. Figure 2 illustrates the process of computing the positional weight matrix for a group of sequences. Each DNA sequence is converted to four binary sequences, one for each base. The binary vectors for each base are summed by position. The value at each position is then divided by the expected frequency of each base. The log of these values is then taken to give the log likelihood of each base occurring in each position. Positive values represent a high likelihood and negative values represent a low likelihood of occurrence.

The feature vector for a particular sequence is constructed by selecting the probability of occurrence of the base in each position in the sequence from the PWM for each position [4].

Although our approach does not require a specific clustering algorithm, we have used the k-means clustering algorithm to partition the data sets. Initial cluster centers are chosen by randomly selecting k feature vectors to serve as cluster centers. Each remaining object is then assigned to the nearest cluster (using Euclidean distance). The centroid of each cluster is then computed and the process is repeated until there is no change in the cluster centers. Our approach uses a k value of 2 to perform a binary search. The advantage of binary search is that it is easy for the user to see differences in two clusters at a time and it produces a hierarchical representation of the clustered data.

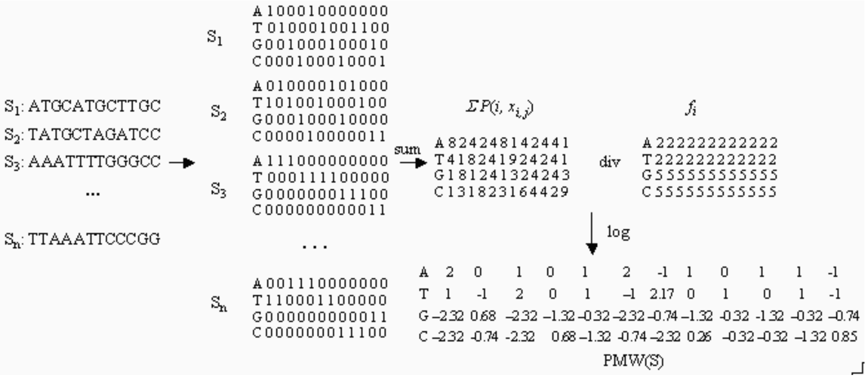

Figure 2. An example of how to compute a position weight matrix.

We use a graph of the positional weight matrix to provide a visualization of the patterns of the entire genome and of the clusters produced. The scientist uses this visualization to direct the clustering process. Figure 3 shows an example of a graph of the positional weight matrix of the 5' flanking region of the genes of *Sulfolobus solfataricus* before clustering. The beginning of the gene is taken as position 0. Five' flanking regions occur "upstream" from the beginning of the gene. A window size is specified to encompass the region of interest to the scientist. In this example, the window is -48 to -1. Initially, two clusters of the initial set of sequences are produced. A graph of the PWM of each of the resulting clusters is generated for examination by the user. For each resulting cluster, the user has several choices:

- Apply the clustering algorithm again to one or both of the clusters
- Redefine the window size before clustering again
- Sort the sequences using domain knowledge before clustering
- Backtrack if the clustering did not add information.

In the next section, we describe how these operations have been applied to cluster the 5' flanking regions of the genes of *Sulfolobus solfataricus.*

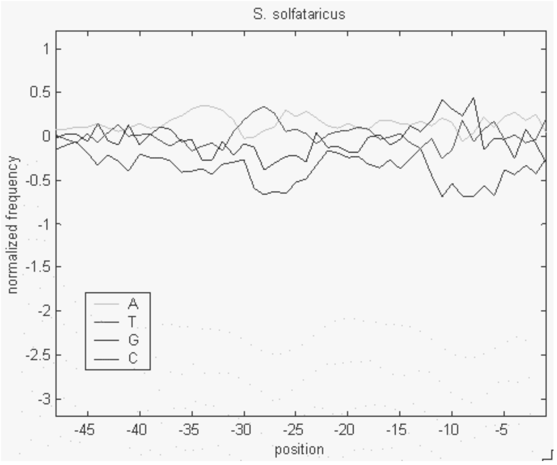

Figure 3. Positional weight matrix of *Sulfolobus solfataricus* gene 5' flanking sequences before clustering.

TRANSCRIPTION AND TRANSLATION SEQUENCES IN ARCHAE

Living organisms can be classified into three groups: prokaryotes, archaea, and eukaryotes. Prokaryotes include the bacteria and other primitive organisms that do not have nuclear membranes. Archaea exhibit features of both prokaryotes and eukaryotes [11] and their position in evolutionary history has long been a subject of debate.

Promoters are involved the initiation of transcription of the gene from DNA to mRNA and are sections of the DNA sequence located "upstream" from genes. Prokaryotes also often have a pattern called a Shine-Darlgarno (SD) sequence that is used as a ribosome binding site for the translation of mRNA to protein that is upstream from the start of the gene [13].

In both prokaryotes and archaea some genes are transcribed together on one segment of mRNA [19]. This cluster of linked genes is called an operon.

Eukaryotes rarely have operons. Previous studies in archaea have shown that the first gene in an operon and single genes have a translation pattern similar to eukaryotes whereas the internal genes of operons have a translation pattern similar to prokaryotes [19, 23]. Graphs of the positional weight matrices of the 5' flanking region of the *Sulfolobus solfataricus* genome (Figure 3) exhibit a mixture of gene regulation patterns. We have used our interactive clustering approach to cluster the promoter regions of genes of several species of archaea into groups that exhibit different gene transcription initiation and translation initiation patterns.

Results with the promoter region of the genome of *S. solfataricus* (downloaded from Genbank http://www.ncbi.nlm.nih) are presented as an example.

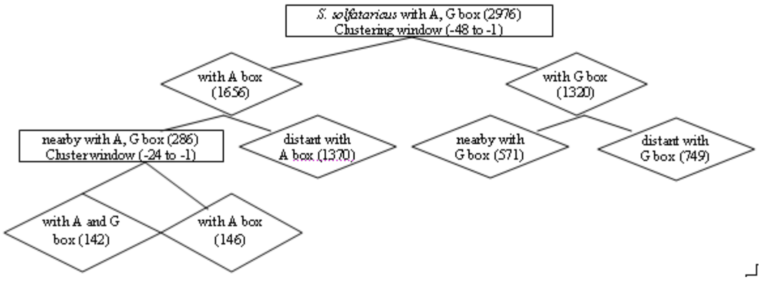
S. solfataricus with A, G box (2976)
Clustering window (-48 to -1)
with A box (1656)
with G box (1320)
nearby with A, G box (286)
Cluster window (-24 to -1)
distant with A box (1370)
nearby with G box (571)
distant with G box (749)
with A and G box (142)
with A box (146)

Figure 4 Clustering results of *S. solfataricus* promoter regions.

CLUSTERING RESULTS

We applied the interactive clustering approach using a window of -48 to -1 for the first round of clustering to the total gene data set of *S. solfataricus*. Figure 4 shows the clustering results obtained using this approach. Subsequent rounds of clustering sometimes used a different window and sometimes were applied to clusters that had been sorted into “nearby” and “distant” groups based on the proximity of neighboring genes. Dr. John Boyle directed the clustering process and “labeled” the resulting clusters. Most patterns were consistent with previous studies. However, there is a small group of nearby genes (142) having both an “A box” and “G box” as illustrated in Figure 5. This result has not been previously reported.

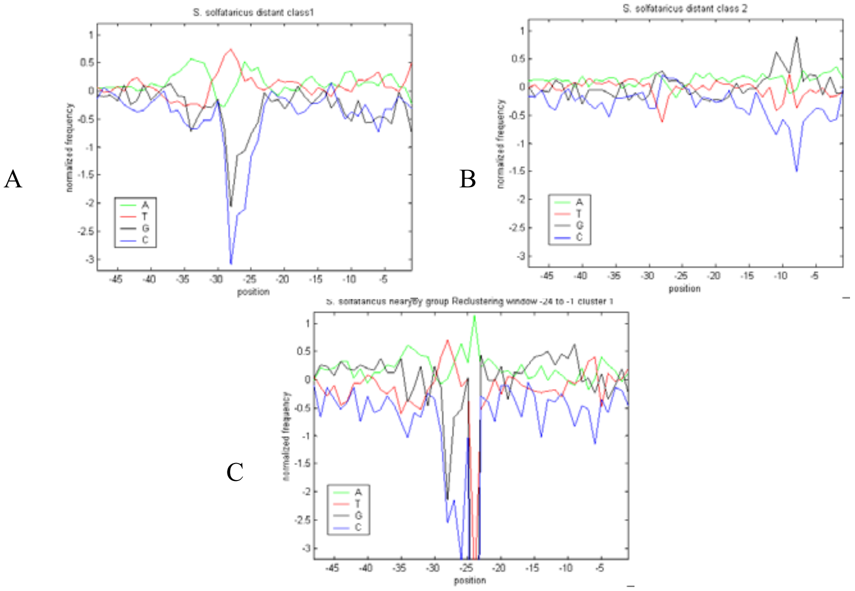

Figure 5 Positional weight matrices of the gene regulation patterns after clustering based on the promoter sequences of *S. sulfotaricus*. A. the cluster with A box; B. the cluster with G box; C. the cluster with weak A and G boxes.

CONCLUSIONS

In summary, we have developed an interactive gene clustering method based on positional weight matrices and the k-means clustering algorithm that allows scientists to visualize and control the clustering process. We have successfully separated the genes of *S. solfataricus* into two classes of genes with different patterns consistent with previous reports [19, 20, 22]. Our results also show a new pattern with both an A and G box that has not been previously reported. This unique combination of translation and transcription initiation patterns requires further experimental investigation. In the current study, we have used the simple k-means clustering algorithm. The effectiveness of other clustering algorithms should also be investigated. We have used the positional weight matrix and Euclidean distance to measure similarity of feature vectors. Investigation of other representations and similarity measures is also planned. The approach that we describe is currently only partially automated. We plan to build a web-based interface to fully automate the approach.

REFERENCES

[1] R. Agrawal, R. Bayardo, and R. Srikant, *Athena: Mining-based Interactive Management of Text Databases.* Research Report RJ 10753, IBM Almaden Research Center, San Jose, CA 95120, 1999.

[2] E. Benitez-Bellon, G. Moreno-Hagelsieb, and J. Collado-Vides, “Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA,” *Genome Biol.,* vol. 3, 2002, pp. RESEARCH0013.

[3] A. P. Boyle and J. A. Boyle, “Global analysis of microbial translation initiation regions,” submitted, 2002.

[4] P. Bucher, “Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences,” *J. Mol. Biol*. vol. 212, 1990, pp. 563-578.

[5] P. Bucher and B. Bryan, “Signal search analysis: a new method to localize and characterize functionally important DNA sequences,” *Nucleic Acids Res*. vol. 12, 1984, pp. 287-305.

[6] M. B. Eisen, P. T. Spellman, P. B. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proc. Natl. Acad. Sci. U S A*, vol. 95, 1998, pp. 14863-14868.

[7] M. Frisch, K. Frech, A. Klingenhoff, K. Cartharius, I. Liebich, and T. Werner, “In silico prediction of scaffold/matrix attachment regions in large genomic sequences,” *Genome Res*. vol. 12, 2002, pp. 349-354.

[8] J. Han and M. Kamber, *Data Mining Concepts and Techniques,* Morgan Kaufmann, New York, 2000, pp 335-391.

[9] J. E. Hansen, O. Lund, J. Engelbrecht, H. Bohr, J. O. Nielsen, and J. E. Hansen, “Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase,” *Biochem J.* vol. 308, 1995, pp. 801-813.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Computing Surveys*, vol. 31, 1999, pp. 264-323.

[11] P. J. Keeling and W. F. Doolittle, “Archaea: narrowing the gap between prokaryotes and eukaryotes,” *Proc. Natl. Acad. Sci.* U S A, vol. 92, 1995, pp. 5761-5764.

[12] S. Levy, S. Hannenhalli, and C. Workman, “Enrichment of regulatory signals in conserved non-coding genomic sequence,” *Bioinformatics,* vol. 17, 2001, pp. 871-877.

[13] B. Lewin, *Genes VI*, Oxford University Press, 1997.

[14] X. Messeguer, R. Escudero, D. Farre, O. Nunez, J. Martinez, and M. M. Alba, “PROMO: detection of known transcription regulatory elements using species-tailored searches,” *Bioinformatics*, vol. 18, 2002, pp. 333-334.

[15] K. Murakami and T. Takagi, “Clustering and detection of 5' splice sites of mRNA by k weight-matrices model,” *Pac. Symp. Biocomput.,* 1999, pp. 171-181.

[16] S. M. Ruger and S. E. Gauch, *Feature Reduction for Document Clustering and Classification.* DTR 2000/8, Dept. of Computing, Imperial College, London, Sept. 2000.

[17] G. Sherlock, “Analysis of large-scale gene expression data,” *Brief Bioinform.* vol. 2, 2001, pp. 350-362.

[18] M. M. Slupska, A. G. King, S. Fitz-Gibbon, J. Besemer, M. Borodovsky, and J. H. Miller, “Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*,” *J. Mol. Biol.* vol. 309, 2001, pp. 347-360.

[19] J. Soppa, “Transcription initiation in Archaea: facts, factors and future aspects,” *Mol. Microbiol.,* vol. 31, 1999, pp. 1295-1305.

[20] J. Soppa, “Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box,” *Mol. Microbiol.* vol. 31, 1999, pp. 1589-1592.

[21] R. Staden, “Measurements of the effects that coding for a protein has on a DNA sequences and their use for finding genes” *Nucleic Acids Res.* vol. 12, 1984, pp. 551-567.

[22] N. Tolstrup, C. W. Sensen, R. A. Garrett, and I. G. Clausen, “Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus,*” *Extremophile*s, vol. 4, 2000, pp. 175-179.

[23] T. Tsunoda and T. Takagi, “Estimating transcription factor bindability on DNA,” *Bioinformatics*, vol. 15, 1999, pp. 622-630.

[24] N. E. Van, M. Zavolan, N. Rajewsky, and E. D. Siggia, “Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics,” to appear in *Proc Natl Acad Sci U S A*, May 2002.