

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 have prompted new research in the use of traditional clustering algorithms and the development of new algorithms [7]. 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 bioinformatics problems including promoter identification, mRNA splicing site detection, and regulon prediction [1, 12, 17]. The performance of each gene clustering method depends on the feature vector representation and the clustering algorithm as well as the specific problem [7].

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 [7]. In their review of clustering methods, Jain et al. [7] 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. Different variants of k-means with regard to pattern representation, feature selection and extraction, similarity measures, and initial partition selections have been described [7].

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 [3, 4, 15] and have been successfully used for promoter prediction [5, 9, 11].

The combination of positional weight matrices and neural network clustering algorithms was described initially to predict the O-glycosylation sites in proteins [6]. Murakami and Takagi [12] combined the k-means clustering algorithm and positional weight matrices for the detection of 5' terminus of the splicing sites of mRNA. They report improved performance over using positional weight matrices alone. Recently, Van et al. [17] 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.

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 [2]. 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.

Although our approach does not require a specific clustering algorithm, we have used the kmeans clustering algorithm to partition the data sets. 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.

A graph of the positional weight matrix is used 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 the example in Figure 3, 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*.

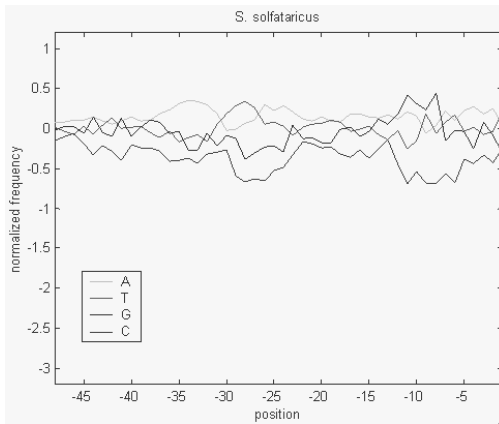


Figure 3. Positional weight matrix of *Sulfolobus solfataricus* gene 5' flanking sequences before clustering.

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 [8] 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 [10].

In both prokaryotes and archaea some genes are transcribed together on one segment of mRNA [13] in a cluster of linked genes 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 [13]. 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. Example results with the promoter region of the genome of *S. solfataricus* (downloaded from Genbank <http://www.ncbi.nlm.nih>) are presented.

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*. 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 4. This result has not been previously reported.

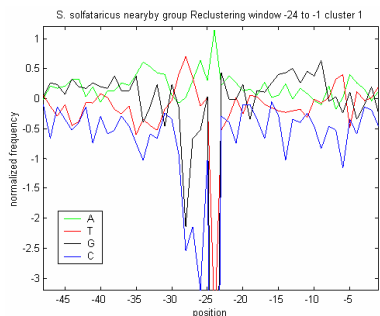


Figure 4 Positional weight matrix of the cluster with a unique gene regulation pattern in *S. solfataricus*.

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 [13, 14, 16]. Our results also show a new pattern with both an A and G box that has not been previously reported. In the current study, we have used the simple k-means clustering algorithm, the positional weight matrix and Euclidean distance. Investigation of other cluster algorithms, 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] 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.
- [2] A. P. Boyle and J. A. Boyle, "Global analysis of microbial translation initiation regions" submitted, 2002.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, 1999, pp. 264-323.
- [8] 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.
- [9] S. Levy, S. Hannenhalli, and C. Workman, "Enrichment of regulatory signals in conserved non-coding genomic sequence," *Bioinformatics*, vol. 17, 2001, pp. 871-877.
- [10] B. Lewin, *Genes VI*, Oxford University Press, 1997.
- [11] 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.
- [12] 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.
- [13] J. Soppa, "Transcription initiation in Archaea: facts, factors and future aspects," *Mol. Microbiol.*, vol. 31, 1999, pp. 1295-1305.
- [14] 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.
- [15] 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.
- [16] 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*," *Extremophiles*, vol. 4, 2000, pp. 175-179.
- [17] 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.