Identification of Regulatory Elements in Archaea using Self-Organizing Maps

Alan P. Boyle[1], John A. Boyle[2], Susan M. Bridges[3]

Keywords: self-organizing maps, regulatory elements, archaea, operon, *Sulfolobus solfataricus*

1 Introduction.

An explosion in the amount of available genomic data has changed the ways in which these data are analyzed. It has become possible to locate transcription and translation regulatory regions based on large scale comparisons of regions upstream of Open Reading Frames (ORFs). The use of self-organizing maps (SOMs) can aid in this search for cis regulatory elements in an organism by segregating similar patterns in different parts of the map.

A SOM is a basic neural network algorithm based on unsupervised learning. It implements reduced dimensionality mapping of the training set to produce a map that follows the probability density function of the data.[1] This unsupervised training system provides a relatively fast clustering that is, in many ways, better than traditional clustering models. The process of clustering upstream regions attempts to divide sequences of DNA into different groups based on feature vector values derived from a a positional weight matrix.[2] The use of this approach in the study of 5' flanking regions of *Sulfolobus solfataricus* ORFs has produced specific clustering that reveals different regulatory features associated with sets of ORFs.

It has been previously found that archaea use both eukaryotic and eubacterial means of transcription and translation.[3,4]. The use of SOM clustering has enabled us to identify a Shine-Dalgarno (S-D) region associated with some ORFs is complimentary to the 3' end of 16S ribosomal RNA in *Sulfolobus solfataricus*. We have observe an A box and a B box in a relatively fixed position upstream of the start of translation. By analysis of known data sets, we show that ORFs that are internal members of operons cluster together and generally lack the transcriptional feature we have identified. Conversely, first ORFs in operons cluster and have the A and B boxes.

2 Approach.

We used an approach that supports dynamic exploration of regulatory patterns in clusters of ORFs in *Sulfolobus solfataricus* by use of positional weight matrices and Kohonen's self-organizing map (SOM) algorithm. We have explored the data using different dimensions in the SOM lattice. We have also varied the extent of the windows to be used in scrutinizing the 5' flanks. ORFs are seen to cluster into groups with and without the regular TATA feature (A and B boxes) and with and without the Shine-Dalgarno sequence.

[1] Department of Computer Science and Engineering and Department of Biochemistry and Molecular Biology, Mississippi State University, Box 9637, Mississippi State, MS 39759, E-mail: apb22@cse.msstate.edu

[2] Department of Biochemistry and Molecular Biology, Mississippi State University, Box 9650, Mississippi State, MS 39759, E-mail: jab@ra.msstate.edu

[3] Department of Computer Science and Engineering, Mississippi State University, Box 9637, Mississippi State, MS 39759,. E-mail: bridges@cse.msstate.edu

The implications of these regulatory features with respect to the location of the ORFs in operons are considered.

3 Results.

ORFs were classified as either Distant or Nearby depending on the location of the their translation start sites relative to the stop codon of the nearest ORF. ORFs with starts located within ±25 nucleotides of the nearest stop codon were classed as Nearby. First ORFs in operons and Internal ORFs represent sets of about 100 ORFs each identified by inspection of the genome.

	First ORFs in Operons			Internal ORFs in Operons		
Features	S-D	TATA	Mixed	S-D	TATA	Mixed
	21%	60%	19%	62%	21%	17%
	Distant ORFs			Nearby ORFs		
	30%	63%	7%	57%	33%	10%

Table 1: Percentage of ORFs in clusters with identifiable cis elements. Window was -5 to -40 from translation start codon. Dimensions of SOM were 2 X 5.

Changes in window size and location had little impact on the clustering. Use of higher dimensions in the SOM allowed finer discrimination of cluster features but past some number of dimensions, the weight matrices become too noisy to analyze.

4 Conclusions

SOM used in conjunction with positional weight matrices allows for visualization of patterns of cis regulatory elements in genomes. Here we use *Sulfolobus solfataricus* as an example and show that the preponderance of ORFs internal to operons have only an S-D element as a recognizable feature. The preponderance of first genes in operons lack an S-D sequence but have a TATA box in a relatively fixed location. Other ORFs may lack this element or, more likely, have it in another location relative to the start codon. They may also have significant deviation from the recognizable consensus sequence. Division of ORFs into Nearby and Distant gives a similar clustering of cis elements as seen in the operon data. It would be expected that Nearby ORFs are much more likely to be members of an operon as compared to Distant ORFs.[5]

References

[1] Vesanto, J. (1999) SOM-based data visualization methods. Intelligent-Data-Analysis, 3: 111–126.

[2] Staden, R. (1984) Measurements of the effects that coding for a protein has on a DNA sequences and their use for finding genes. Nucleic Acids Res 12: 551-567.

[3]Kyrpides NC and Ouzounis CA (1999) Transcription in archaea. Proc Natl Acad Sci USA 96: 8545-8550

[4]Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. Gene 234: 187-208

[5] Wan, Xiufeng, Susan M. Bridges, and John A. Boyle. (2004) Revealing gene transcription and translation initiation patterns in arcaea using an interactive clustering model, Under review.