

An *In Silico* Approach to Cluster CAM Kinase Protein Sequences

U.S.N Murty*, Amit Kumar Banerjee, Neelima Arora

Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology, Hyderabad-500607, A.P., India

*Corresponding author: Dr. U.S.N Murty, Deputy Director/ Scientist "F" Head, Biology Division, Indian Institute of Chemical Technology, Hyderabad- 500007, India, Tel: +91 40 27193134; Fax: +91 40 27193227; E-mail: murty_usn@yahoo.com

Received December 12, 2008; Accepted February 20, 2009; Published February 20, 2009

Citation: Murty USN, Amit KB, Neelima A (2009) An *In Silico* Approach to Cluster CAM Kinase Protein Sequences. J Proteomics Bioinform 2: 097-107.

Copyright: © 2009 Murty USN, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

As we are ushering in new age of data driven world, we face an enormous challenge of deriving information from heaps of data available. The amount of data being generated is overwhelming and this calls for exploring novel and effective methods for clustering and classification of such data. CAM kinase family is known to contain many enzymes involved in important physiological processes. In the present study, 13 important physicochemical parameters were calculated for 56 sequences of CAM kinase family *in silico*. Self organizing Maps (SOM) were employed for the classifying and clustering similar sequences and visualization of high dimensional data spaces as they are known for their capability to maintain the essence of topological relationships between the features. SOM effectively yielded 4 clusters which were distinct from each other and marked by characteristic features.

Keywords: Kohonen map; Self Organizing Maps (SOM); CAM Kinase; Bioinformatics; *in silico*; clustering

Introduction

The urge to describe and explore not only the complex phenomena of life but also to seek answers to what lies beyond the realm of current understanding of life processes at molecular level continues to be a major inspiration in modern biology. Human mind is an advance neural cognitive system, a fact exemplified and reinforced by its learning and decision making ability, hence, his endeavor to automate the process of learning and decision making process by devising and employing machine learning techniques should not come as a surprise. In the present data-driven, information-starved world, this gold rush to produce enormous volumes of data would have been of no avail if not empowered by advanced powerful and comprehensive machine learning methods for analysis. The exponential rise in challenges posed to a biologist has propelled a new impetus for development of new and efficient algorithms and

methods for analysis of such data or exploring the existing ones in biological contexts. Proliferation of low cost technology, astonishing growth in computing power and interdisciplinary nature of this field has led to revolution of a sort in recent times. Literature abounds with examples of application of machine learning in biological systems (Tarca *et al.*, 2007). Though both supervised and unsupervised learning methods are being employed in bioinformatics analyses yet unsupervised learning methods are attracting more interest as they offer many advantages like elimination of need of labeling and predefined knowledge of classes and are valuable in gaining an understanding of basic nature of data.

Self Organizing Map (also known as Kohonen Map) is a unsupervised learning algorithm (Kohonen *et al.*, 2001) used for clustering and reducing dimensions of complex data with-

out loosing 'essence' of the data and is capable of organizing data based on the similarity by putting entities geometrically close to each other. SOMs have been applied in diverse fields like assessment of water quality (Walley *et al.*, 2000), classification of communities (Chon *et al.*, 1996, Arab *et al.*, 2004; Tison *et al.*, 2005), gene expression studies (Tamayo *et al.*, 1999), disease diagnosis (Chen *et al.*, 2000; Hoshi *et al.*, 2006), medical imaging (Chuang *et al.*, 2007), biochemical profiling (Kaartinen *et al.*, 1998) and epidemiology (Murty and Arora, 2007). Self organizing maps have been earlier used in classification of families (Andrade *et al.*, 1997), secondary structure determination (Unneberg *et al.*, 2001) and pattern recognition in proteins (Hanke *et al.*, 1996). Owing to its use for multidimensional data visualization, SOM has aptly become the method of choice in bioinformatics studies (Hsu *et al.*, 2003). Previously, data mining techniques have been employed for clustering and classification of Internal Transcribed Spacer sequences in mosquito species (Banerjee *et al.*, 2008, 2009).

The interplay of various inherent sequence and structural features of biological molecules is quite complex and intriguing. Minute and slight variation in physicochemical properties even in the member of same protein family is of common occurrence. Data mining techniques like SOM can be employed to aid the knowledge discovery processes in such instances.

The Ca²⁺/calmodulin-dependent kinases (CaMK) belong to family of structurally related Serine /threonine-specific protein kinase, which are activated in response to elevation of intracellular Ca²⁺, and include CaMKI, CaMKII, CaMKIV and CaMK-kinases (CaMKKs). These are known to play a role in a wide range of activities like regulation of diverse biological events mediated by intracellular calcium like muscle contraction, neurotransmitter release and gene expression (Eto *et al.*, 1999; Nairn *et al.*, 1985; Edelman *et al.*, 1987; Soderling *et al.*, 1996; Braun *et al.*, 1995). This study is an attempt to cluster CaMK kinase sequences belonging to different species on basis of their physicochemical properties by applying Kohonen maps.

Materials and Methods

Sequence Collection and Pre-processing

CAM kinase protein sequences were retrieved from the SWISS-PROT, a public domain protein database (Bairoch and Apweiler, 2000). During the sequence retrieval process,

the keyword 'Calcium/calmodulin-dependent protein kinase' was used which yielded 68 sequences. Sequences representing putative, partial, precursor and fragment of CAM Kinase protein were excluded from the study. Hence, 56 unique proteins were retrieved and considered for this study. The selected CAM kinase protein sequences were retrieved in FASTA format and used for further analysis.

Reconstruction of Phylogeny

All 56 sequences were considered for reconstruction of phylogeny. PHYLIP (Felsenstein, 1982) was used for this purpose. CLUSTALW (Thompson *et al.*, 1994) was employed for the initial multiple sequence alignment. Alignment output was used as input for Seqboot and Protpars program and finally Consense program was used to get the best tree with maximum parsimony method (Felsenstein, 1983) which was visualized with TREEVIEW (Page, 1996) (Fig. 7 in Supplement).

Feature Identified as Parameters for SOM

Physicochemical Characterization

Calculation of physicochemical properties of proteins by traditional experimental methods besides being expensive, is time consuming and cumbersome. The ProtParam is a program used for predicting various physical and chemical properties which may be useful in enhancing our knowledge for experiment design. Physicochemical properties like Length, Molecular Weight, Isoelectric point, Number of negatively charged amino acids, Number of Positively charged amino acids, Extinction coefficient (considering all cysteine residues appear as half cystines), Extinction coefficient *(assuming that no cysteine appears as half cystine), Instability coefficient, aliphatic index and GRAVY were calculated using ProtParam (<http://expasy.org/tools/protscale.html>) (Gasteiger *et al.*, 2005) for these sequences (Table 1 in Supplement). Amino acid composition of the protein sequences can reveal their nature; hence, amino acid composition was also computed (Data not shown).

Secondary Structure Prediction

SOPMA (Self Optimized Prediction Method from Alignment) (Geourjon and Deléage, 1995) was employed for prediction of secondary structure features like alpha helix, extended strand, beta turn and random coils in terms of percentage for all the sequences (Table 2 in Supplement). These

features (except amino acid composition) were considered as input parameters for self organizing maps for further analysis.

Data Mining – Self Organizing Maps

In SOM, the neurons are organized in a lattice, typically a one or two-dimensional array, which is placed in the input space and is spanned over the input distribution. It is feasible to achieve a map of input space where imminence between units or clusters in the map represents closeness of the input data using a two-dimensional SOM network. Processing units in the SOM lattice are associated with weights of the same dimension of the input data. Using the weights of each processing unit as a set of coordinates, the lattice can be positioned in the input space. Throughout the learning stage, the weights of the units change their position and “move” towards the input points. Progress of the movement acquires a gradually slower pace and network is almost “frozen” in the input space at the end of the learning stage. On the completion of the learning stage, the inputs can be associated to the nearest network unit. On visualization, the inputs can be associated to each cell on the map. Cells that evidently contain analogous entities can be considered as a cluster on the map. These clusters are generated during the learning phase without any prior information. The main application of the SOM is the visualization of high-dimensional data in a two dimensional way and the construction of abstractions akin to other clustering techniques.

Steps Involved in the Algorithm

1. **Initialization:** Randomly initialize a weight vector (W_i) for each neuron I $W_i = [w_{i1}; w_{i2}; \dots; w_{in}]$; n denotes the dimension of input data.

2. **Sampling:** Select an input vector $X = [x_1, x_2, \dots, x_n]$

3. **Similarity matching:** Find the winning neuron whose weight vector best matches with the input vector $j(t) = \arg \min \{\|X - W_j\|\}$

4. **Updating:** Update weight vector of winning neuron, such that it becomes still closer to the input vector. Also, update weight vectors of neighbouring neurons-the further the neighbour, the lesser the degree of change.

$$W_i(t+1) = W_i(t) + \alpha(t) X \text{ hij}(t) X [X(t) - W_i(t)]$$

$\alpha(t)$: learning rate that decreases with time t , $0 < \alpha$

$$(t) = 1$$

$$h_{ij}(t) = \exp(-\|r_j - r_i\|^2 / 2 \times \sigma(t)^2)$$

$\|r_j - r_i\|^2$ = distance between winning neuron and other neurons

$\sigma(t)$ = neighbourhood radius that decreases with time t .

5. Continuation: Repeat steps 2–4 until there is no change in weight vectors or up to certain number of iterations. For each input vector, find the best matching weight vector and allot the input vector to the corresponding neuron/cluster.

Data Normalization

Data was normalized linearly such that value in each category ranged between 0 and 1. This is done to get unbiased results while ensuring equal importance to all parameters while clustering.

$$\text{Normalization Formula} = \frac{\text{Original data value} - \text{Minimum Data value}}{\text{Maximum data value} - \text{Minimum Data value}}$$

Results and Discussion

The length of considered sequences varied from 335 to 926 and the molecular weight was found to be in the range of 38163.7-105122.7. The sequences that lie on higher extreme of molecular weight were found to be peripheral Plasma protein belonging to *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. All the sequences possess more negatively charged residues except Q10KY3, Q96NX5, Q91VB2, Q7TNJ7, Q9P7I2, P11730, Q07250, Q13554, Q13555, Q6DGS3 while Q923T9 and Q2HJF7 contains equal number of negatively and positively charged residues.

The pH at which a protein carries no charge and exists as zwitterion is termed as Isoelectric point (pI). The pI value of all considered CAM kinase protein sequences were in the range of 4.83 -9.11 where 13 proteins (understandably those with higher number of negative amino acids except for Q00168) are basic and rest of them are acidic. The instability index which gives clue about the stability of a protein *in vitro* can be calculated using the following formula:

$$i = L - 1$$

$$II = (10/L) * \sum_{i=1}^{L-1} DIWV(x(i)x(i+1))$$

$$i = 1$$

where L denotes length of sequence, DIWV(x(i)x(i+1)) is the instability weight value for the dipeptide starting in position i.

This will be particularly useful in comparing the metabolic stabilities of proteins. All the considered sequences were classified as unstable except Q14012 (37.09), Q9P7I2 (38), Q16566 (31.64) and O42844 (36.67) as a value > 40 indicates an unstable protein. The aliphatic index (AI) which is defined as the relative volume of a protein occupied by aliphatic side chains is regarded as a positive factor for the increase of thermal stability of globular proteins(Ikai, 1980). It can be calculated by the formula:

$$\text{Aliphatic index} = X(\text{Ala}) + a * X(\text{Val}) + b * X(\text{Leu}) + b * X(\text{Ile})$$

where X (Ala), X (Val), X (Ile) and X (Leu) are the amino acid compositional fractions.

Aliphatic index ranged from 76.24- 96.31. From the molar extinction coefficient of tyrosine, tryptophan and cystine (cysteine does not absorb appreciably at wavelengths >260 nm, while cystine does) at a given wavelength, the extinction coefficient of the native protein in water can be com-

puted using the following equation:

$$E(\text{Prot}) = N(\text{Tyr}) * \text{Ext}(\text{Tyr}) + N(\text{Trp}) * \text{Ext}(\text{Trp}) + N(\text{Cystine}) * \text{Ext}(\text{Cystine})$$

where (for proteins in water measured at 280 nm): N= number, Ext(Tyr) = 1490, Ext(Trp) = 5500, Ext(Cystine) = 125.

Extinction coefficients of considered sequences at 280 nm range from 30410 to 98180 M⁻¹ cm⁻¹ assuming all cysteine residues appear as half cystines. High value of extinction coefficients of some sequences connotes incidence of Cys, Trp and Tyr in high concentration. The extinction coefficients are useful in determining protein concentration required for quantitative study of protein-protein and protein-ligand interactions in solutions.

The Grand Average hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence (Kyte and Doolittle, 1982). Low values of GRAVY indices which ranged from -0.571 to -0.214 indicate the possibility of better interaction with water. The secondary structure indicates whether a given amino acid

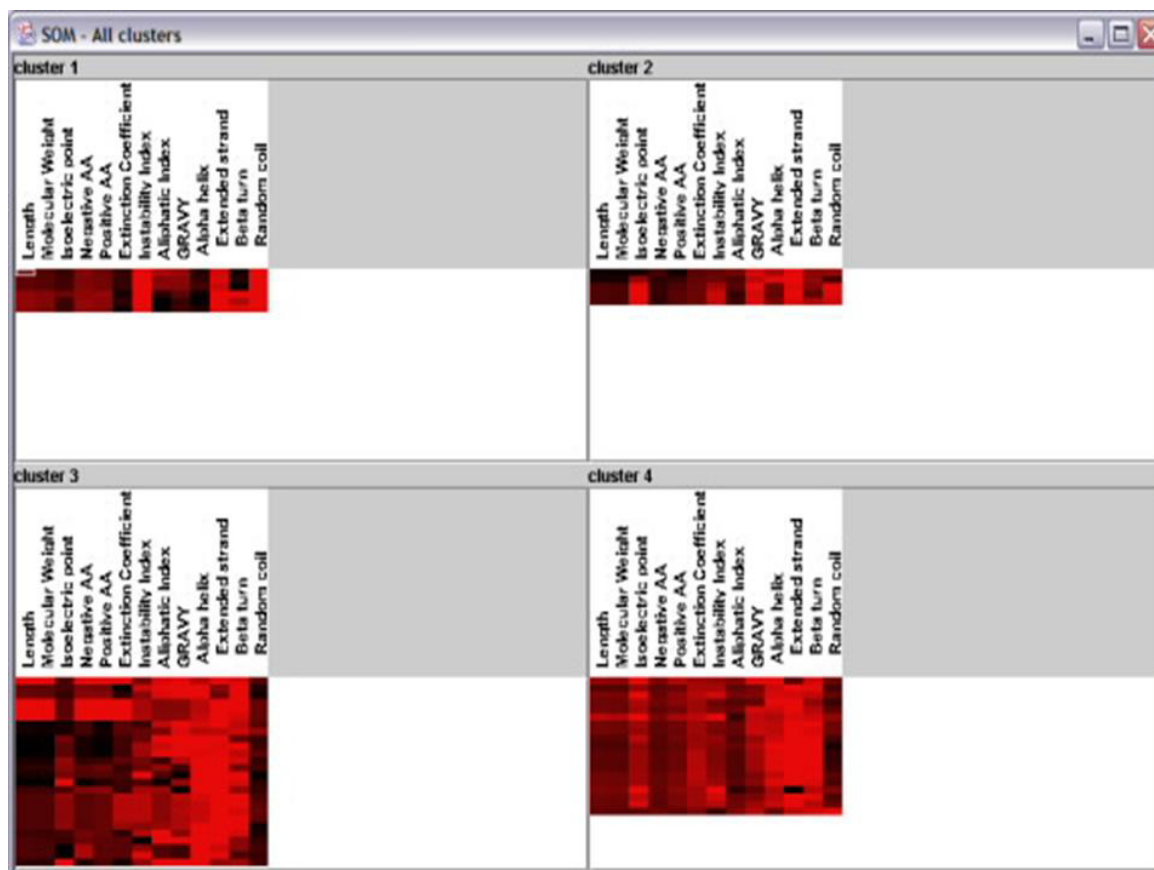


Figure 1: 2*2 grid showing SOM clusters.

lies in a helix, strand or coil. Secondary structure features as predicted using SOPMA are represented in Table 4. The results revealed that alpha helices were found to be predominant followed by random coil, extended strands and beta turns in majority of the sequences while for sequences (Accession number: Q91YS8, Q63450, Q96NX5, Q91VB2, Q7TNJ7, Q13554, Q13555, Q923T9, O42844, Q8N5S9, Q8VBY2, P97756, Q96RR4, Q8C078, O88831), random coils outnumbered other secondary structural features. For Calcium/calmodulin-dependent protein kinase type II beta chain (Protein ID: P28652) belonging to *Mus musculus*, random coils were found to be equal to alpha helices. Normalized data was clustered using SOM on a 2x2 grid (shown in Figure 1). Unsupervised learning was done on the fly using the data using a learning constant of 0.01 and for 10,000 iterations following which the data got clustered based on the neighborhood distance.

In short:

Total no of sequences selected for study =56

Total number of input parameters =13

Total iterations per sequence to form a neuron = 100000

Total iterations to form 4 grid (2X2) structure = 5600000

Successful or winning neurons = 4

Unsuccessful neuron = 0

In short, all 4 neurons were successful and the data got assembled into 4 clusters. The pie chart below (Fig.2) shows the distribution of sequences in the clusters.

Cluster (1, 1): This cluster contains 6 sequences which are exclusively Calcium/calmodulin-dependent protein kinase kinase sequences belonging to *Homo sapiens*, *Mus musculus* and *Rattus norvegicus* and thus, is characterized by very similar trends which make this cluster distinct from all other clusters. This cluster is also marked by lowest values of GRAVY and isoelectric point. At the same time, this cluster shows a distinctly high range of values of instability index and random coils and uniformly low range of extinction coefficient.

Cluster (1, 2): 5 sequences lie in this cluster. Q91YS8

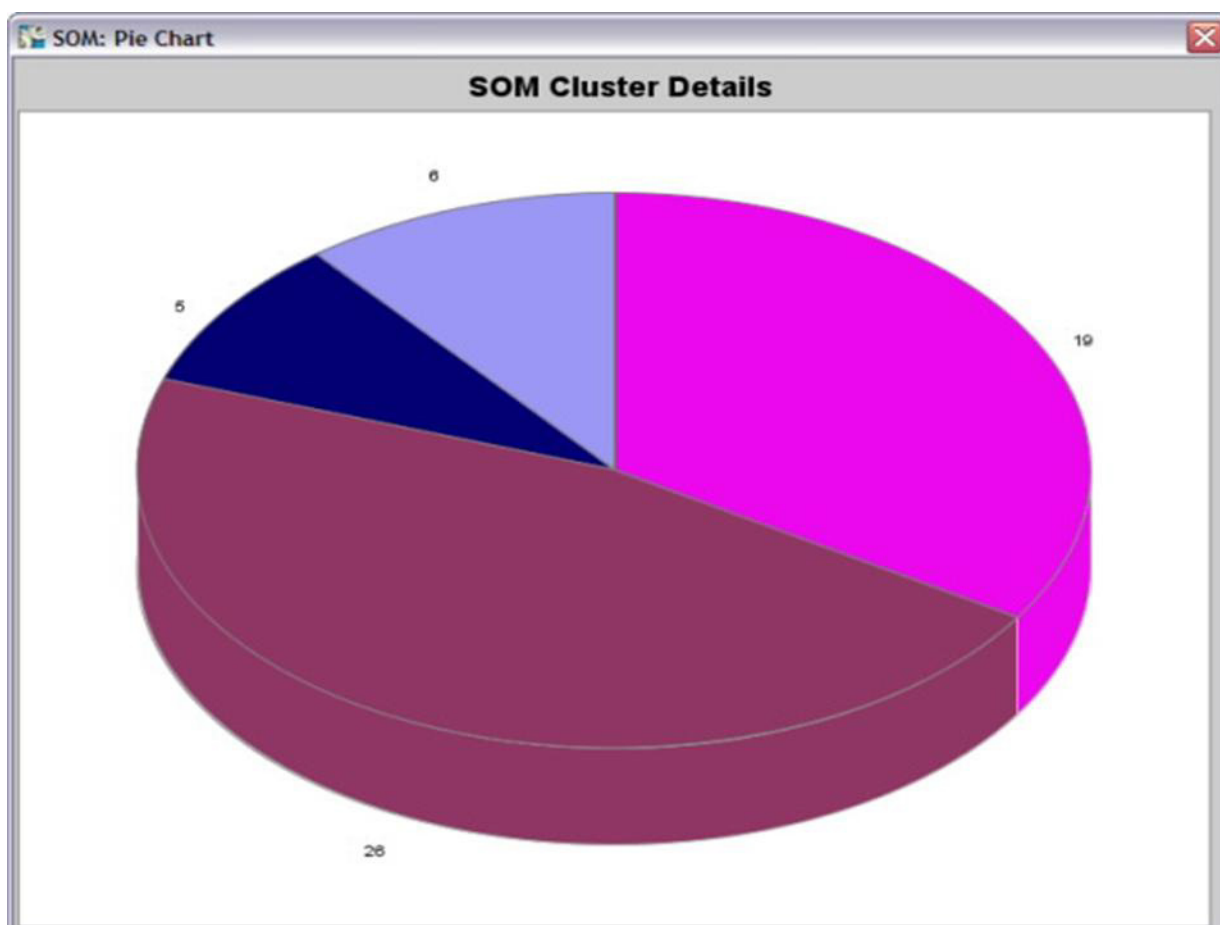


Figure 2: Pie chart showing distribution of sequences in SOM clusters.

SOM Cluster (1,1)

Locations: 6, Parameters: 13 Display Range: 0 to 1



Figure 3: Cluster (1, 1).

SOM Cluster (1,2)

Locations: 5, Parameters: 13 Display Range: 0 to 1



Figure 4: Cluster (1, 2).

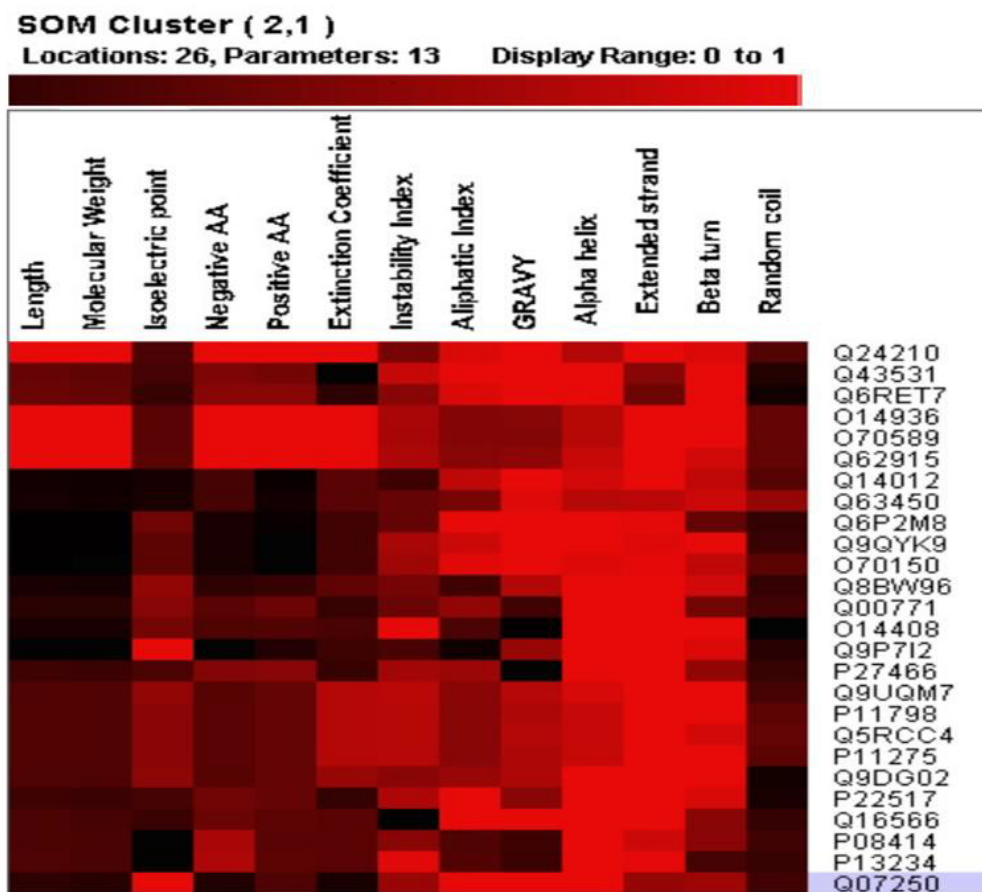


Figure 5: Cluster (2, 1).

and Q8IU85 although similar in length and type of amino acids varied in isoelectric point, GRAVY, alpha helix and random coils. Q96Nx5, Q91VB2, 7TNJ7 which belonged to Calcium/calmodulin dependent protein kinase type 1G got clustered together and showed similar profiles though differing slightly in Instability index, GRAVY and beta turn. This cluster comprised of shortest sequences where random coils were more than alpha helices. This cluster is marked by uniformly high range of extended strands.

Cluster (2, 1): This cluster is constituted by 26 sequences. Except for 4 sequences (Q24210, o14396, O70859, Q62915, this cluster comprises of sequences with low molecular weight and length. O14936, O70589 and Q62915 which belonged to peripheral plasma membrane protein showed nearly identical profiles in SOM cluster and got assembled in neighboring cells. Calcium/calmodulin-dependent serine/threonine-protein kinases sequences also showed similar range of values and were placed at neighboring places. Sequences that belonged to Calcium/calmodulin-dependent protein kinase type II alpha chain also

were lying in proximity in the cluster with similar profiles and differed markedly from next sequence that belonged to Calcium/calmodulin dependent protein kinase type II Delta chain sequence from *Xenopus laevis*.

Cluster (2, 2): 19 sequences that got assembled in this cluster are Calcium/calmodulin dependent protein kinase type II sequences except Q10KY3 which is described as Calcium/calmodulin-dependent serine/threonine-protein kinase 1. All these sequences are longer and are of high molecular weights. In general, the alpha helices were more in number as compared to random coils in the considered sequences. Sequences belonging to Calcium/calmodulin-dependent protein kinase type II beta chain got positioned in vicinity in this cluster and showed similar profiles for all the parameters except for Q13554. 3 sequences that belong to Calcium/calmodulin-dependent protein kinase type II gamma chain also got clustered together with slight variation. Gradient in parameter values can be attributed to the fact that this cluster is assemblage of various types of sequences belonging to Calcium/calmodulin-dependent protein kinase type II α , β

SOM Cluster (2,2)

Locations: 19, Parameters: 13

Display Range: 0 to 1

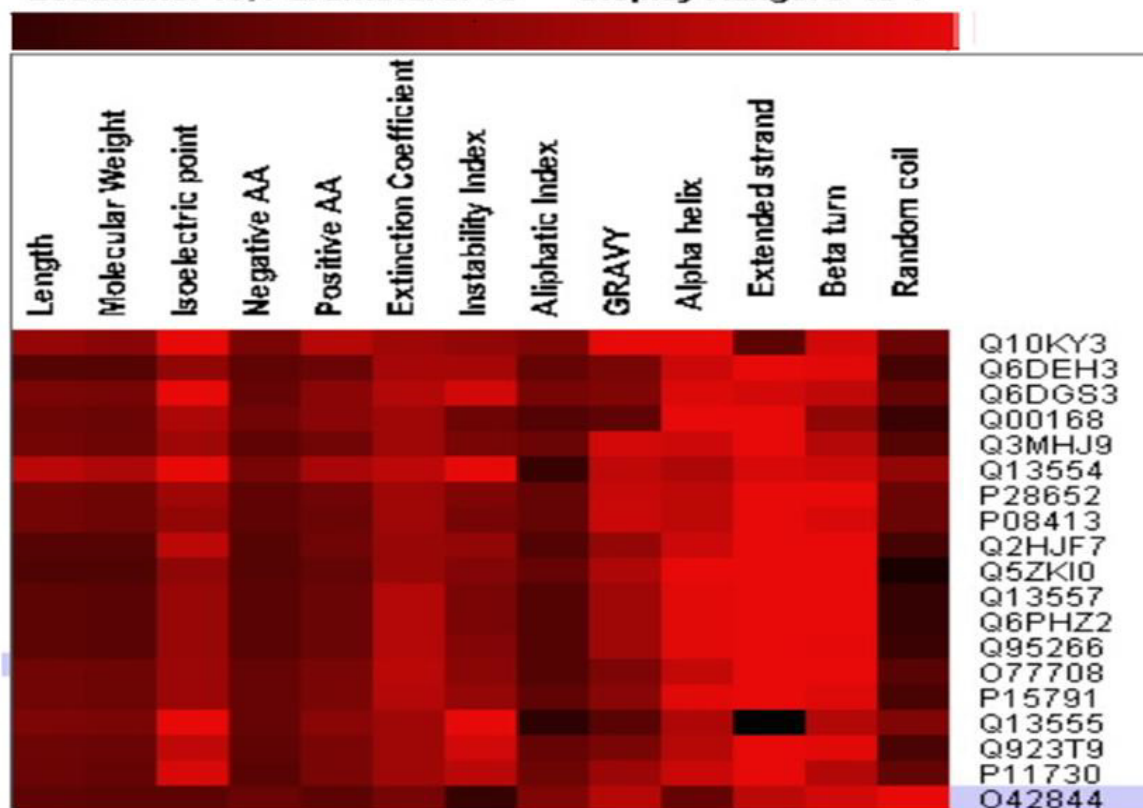


Figure 6: Cluster (2, 2).

chains, ä and ã chains belonging to various species. Even trivial differences at the sequence level and type of chain are clearly reflected in case of sequences from *Danio rerio*.

Future Perspective

A fairly good amount of raw sequence data pertaining to Protein super families and families exist in public domain databases. Conventional methods for defining a protein family rely on signatures, motifs and structural or functional domain information. The method presented in this report allow us to think in a different direction where we can go for further sub-classification of these available large data and this approach may provide a cue for sophisticated intelligent classification and clustering enabling categorization of new subclasses or classes which may aid in new criteria generation for tapping into this wealth of information.

Conclusion

Bioinformatics analyses have been employed by researchers to provide substantial information about the biological macromolecules in shortest span while eliminating to a certain extent, the need of time consuming expensive experiments. With the exponential rise in amount of data being

generated, one can not overlook the need of exploring new methods for clustering and classification of such data. Recently, there have been attempts to employ data mining approaches in biological relevance (Banerjee *et al.*, 2007, 2008). Artificial Neural Networks (ANN) like Self Organizing Maps have innate penchant to learn and can recognize patterns in data without prior information (Lampinen and Oja, 1992). SOM is highly effective sophisticated data clustering tool for visualizing complex data by reducing dimensions. These have been successfully exploited in bioinformatics in chromosome structural studies (Kyan *et al.*, 2001), motif discovery (Mahony *et al.*, 2006, Arrigo *et al.*, 1991), identification of genome signature (Abe *et al.*, 2002), codon usage diversity (Kanaya *et al.*, 2001, Wang *et al.*, 2001), gene prediction (Mahony *et al.*, 2004), identification of transcription binding sites (Mahony *et al.*, 2005), sequence analysis (Oja *et al.*, 2005), nucleic acid classification (Naenna *et al.*, 2003) and gene expression analysis (Ressom *et al.*, 2003; Covell *et al.*, 2003).

In this study, physiochemical properties were calculated for 56 CAM kinase sequences using *in silico* tools. SOMs were employed to segregate data according to variation in properties and group them in separate clusters according to

trend observed in properties. SOMs seem to be a perfect solution for clustering and visualization of such sequence data for easy interpretation owing to its innate simplicity.

Acknowledgement

Authors thank Dr. J.S.Yadav, Director, IICT for his continuous support and encouragement. AKB thanks CSIR for Senior Research Fellowship. NA thanks DST for Research Associate fellowship. Authors thank anonymous reviewers for their valuable suggestions for improvement of manuscript.

References

1. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, et al. (2002) A novel bioinformatics strategy for unveiling hidden genome signatures of eukaryotes: Self organizing map of oligonucleotide frequency. *Genome Informatics*. 13: 12-20.
2. Andrade MA, Casari G, Sander C, Valencia A (1997) Classification of protein families and detection of the determinant residues with an improved self organizing map. *Biological Cybernetics*. 76: 441-450.
3. Arab A, Lek S, Lounaci A, Park YS (2004) Spatial and temporal patterns of benthic invertebrate communities in an intermittent river (North Africa). *Ann De Limnol* 40: 317-327.
4. Arrigo P, Giuliano F, Scalia F, Rapallo A, Damiani G (1991) Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. *Computer Applications in Biosciences* 7: 353-357.
5. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28: 45-48.
6. Banerjee AK, Arora N, Varakantham P, Murty USN (2008) Exploring the Interplay of Sequence and Structural Features in Determining the Flexibility of AGC Kinase Protein Family: A Bioinformatics Approach. *Journal of Proteomics and Bioinformatics* 1: 77-89.
7. Banerjee AK, Arora N, Murty USN (2007) Stability of ITS2 secondary structure in Anopheles: What Lies Beneath? *International Journal of Integrative Biology* 1: 232-238.
8. Braun AP, Schulman H (1995) The multifunctional calcium/calmodulin-dependent protein kinase: From form to function. *Annu Rev Physiol* 57: 417-445.
9. Chen D, Chang RF, Huang YL (2000) Breast Cancer Diagnosis using Self-Organizing Map for Sonography. *Ultrasound in Med & Biol* 26: 405-411.
10. Chon TS, Park YS, Moon KH, Cha EY (1996) Patternizing communities by using an artificial neural network. *Ecol Model* 90: 69-78.
11. Chuang CH, Cheng PE, Liou M, Liou CE, Kuo YT (2007) Application of Self-Organizing Map (SOM) for Cerebral Cortex Reconstruction. *International Journal of Computational Intelligence Research* 3: 26-30.
12. Covell DG, Wallqvist A, Alfred A, Rabow TN (2003) Molecular Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray Data. *Molecular Cancer Therapeutics* 2: 317-332.
13. Edelman AM, Blumenthal DK., Krebs EG (1987) Protein serine/threonine kinases. *Annu Rev Biochem* 56: 567-613.
14. Eto K, Takahashi N, Kimura Y, Masuho Y, Arai K, et al. (1999) Ca²⁺/Calmodulin-dependent Protein Kinase Cascade in *Caenorhabditis elegans*: Implication In Transcriptional Activation. *J Biol Chem* 32: 22556-22562.
15. Felsenstein J (1983) Parsimony in systematics: biological and statistical issues. *Annual Review of Ecology and Systematics*. 14: 313-333.
16. Felsenstein J (1982) Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* 57: 379-404.
17. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, et al. (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) *The proteomics protocols handbook*. Humana New York 571-607.
18. Geourjon C, Deléage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci* 11: 681-684.

19. Hanke J, Beckmann G, Bork P, Reich JG (1996) Self-Organizing hierarchic network for pattern recognition in protein sequence. *Protein Science* 5: 72-82.
20. Hoshi K, Kawakami J, Sato W, Sato K, Sugawara A, et al. (2006) Assisting the Diagnosis of Thyroid Diseases with Bayesian-Type and SOM-Type Neural Networks Making Use of Routine Test Data. *Chemical & Pharmaceutical Bulletin* 54: 1162-1169.
21. Hsu AL, Tang SL, Halgamuge SK (2003) An unsupervised hierarchical dynamic self organizing approach to cancer class discovery and marker gene identification in microarray data. *Bioinformatics* 19: 2131-2140.
22. Ikai AJ (1980) Thermostability and aliphatic index of globular proteins. *J Biochem* 88: 1895-1898.
23. Kaartinen J, Hiltunen Y, Kovanen PT, Ala-Korpela M (1998) Application of self organizing maps for the detection and classification of human blood plasma lipoprotein lipid profiles on the basis of ¹H NMR spectroscopy data. *NMR in Biomedicine*. 11: 168 – 176.
24. Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, et al. (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): Characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. *Gene* 276: 89-99.
25. Kohonen T (2001) *Self-Organizing Maps*. 3rd edition (Berlin, Heidelberg: Springer Press).
26. Kyan MJ, Guan L, Arnison MR, Cogswell CJ (2001) Feature Extraction of Chromosomes From 3-D Confocal Microscope Images. *IEEE Transactions on Biomedical Engineering*, 48: 1306-1318.
27. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157: 105-132.
28. Lampinen J, Oja E (1992) Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*. 2:261-272.
29. Mahony S, McInerney JO, Smith TJ, Golden A (2004) Gene prediction using the Self- Organizing Map: Automatic generation of multiple gene models. *BMC Bioinformatics* 5: 23.
30. Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar D (2005) Transcription factor binding site identification using the self-organizing map. *Bioinformatics* 21:1807-1814.
31. Mahony S, Benos PV, Smith TJ, Golden A (2006) Self-organizing neural networks to support the discovery of DNA-binding motifs. *Neural Networks*. 19: 950-962.
32. Murty USN, Arora N (2007) Application Of Self-Organizing Maps For Prioritization Of Malaria Control Operations In Changlang District, Arunachal Pradesh. *The Internet Journal of Epidemiology* 4(2).
33. Banerjee AK, Arora N, Murty US (2009) Clustering and Classification of *Anopheles* Spacer Sequences using Self Organizing Maps. *The Internet Journal of Genomics and Proteomics* 7 (1).
34. Banerjee AK, Kiran K, Murty US, Venkateswarlu Ch (2008) Classification and identification of mosquito species using artificial neural networks. *Comput Biol Chem*. 32: 442-447.
35. Naenna T, Bress RA, Embrechts MJ (2003) DNA classifications with self-organizing maps (SOMs). In *Proceedings of the IEEE international workshop on soft computing in industrial applications*.
36. Nairn AC, Hemmings HC Jr, Greengard P (1985) Protein kinases in the brain. *Annu Rev Biochem* 54: 931–976.
37. Oja M, Sperber GO, Blomberg J, Kaski S (2005) Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*. 15: 163-179.
38. Page RDM (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* 12: 357-358.
39. Resson H, Wang D, Natarajan P (2003) Clustering gene expression data using adaptive double self-organizing map. *Physiol Genomics* 14: 35-46.
40. Soderling TR (1996) Structure and regulation of calcium/calmodulin-dependent protein kinases II and IV. *Biochim Biophys Acta* 1297: 131-138. (DOI: 10.1021/cr0002386).

41. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96: 2907–2912.
42. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S (2007) Machine Learning and Its Applications to Biology. *PLoS Computational Biology* 3: e116.
43. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22: 4673-80.
44. Tison J, Park YS, Coste M, Wasson JG, Ector L, et al. (2005) Typology of diatom communities and the influence of hydro-ecoregions: A study on the French hydrosystem scale. *Wat Res* 39: 3177–3188.
45. Unneberg P, Merelo JJ, Chacón P, Morán F (2001) SOMCD: Method for evaluating protein secondary structure from UV circular dichroism spectra. *Proteins: Structure, Function, and Bioinformatics*. 42: 460 – 470.
46. Walley WJ, Martin RW, O'Connor MA (2000) Self-organising maps for classification of river quality from biological and environmental data. In: R. Denzer, D.A. Swayne, M. Purvis and G. Schimak, Editors, *Environmental Software Systems: Environmental Information and Decision Support*. IFIP Conference Series, Kluwer Academic Publishers, Boston Hardbound, pp. 27–41.
47. Wang HC, Badger J, Kearney P, Li M (2001) Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Molecular Biology and Evolution*. 18: 792-800.