Clustering and Classification of Anopheline Spacer Sequences using Self Organizing Maps

A Banerjee, N Arora, U Murty

Citation

Abstract
ITS2, a well known phylogenetic marker is widely used in taxonomic studies. This study exploits a novel approach to classify and cluster the Anopheline species based upon their spacer (ITS2) sequences. As secondary structure of ITS2 is crucial for the function, derived parameters based on secondary structure along with sequence composition were considered for this study. Self Organizing Map (SOM), a neural network approach was adopted for classification and clustering of Anopheline sequences. This data mining approach for clustering and classification will aid in unveiling of inherent relationships among the various parameters contributing to ITS2 structure stability.

INTRODUCTION
Malaria is the most devastating parasitic disease of human, exacting an estimated toll of 300–500 million new infections and 1.5–3.0 million deaths annually (World Malaria Report 2005). 41% of population lives in endemic regions in 107 countries under constant threat of malaria (World Malaria Report 2005). The completion of triad of mosquito, parasite and human genome fuelled the effort to see malaria in new light and provided much-needed impetus to studies at molecular level (Aultman et al. 2002). For millennia, the reprehensible mosquito has mined the riches of the human bloodstream, with the availability of Anopheles genome; it is the research community’s turn to mine the molecular riches of the mosquito (Jasny et al. 2002). The quest to answer myriad questions on Anopheles taxonomy necessitated the search of reliable molecular markers to resolve phylogeny and existing ambiguities. But the sheer complexity and colossal magnitude and pace of availability of molecular data being generated in this post genomic era is often overwhelming and perplexing to molecular entomologists.

Nuclear ribosomal RNA genes (rDNA) are organized in clusters containing the 18S, 5.8S and 28S subunits in eukaryotic organisms. Two internal transcribed spacers (ITS) namely, ITS1, separating 18S and 5.8S genes and ITS2 lying between 5.8S and 28S genes are known to occur (Fedoroff 1979). These spacer sequences are used extensively as reliable markers for taxonomic classification across taxa and exploited for phylogenetic reconstruction by virtue of its fast evolution. (Coleman 2003, Alvarez and Wendel 2003).


Spacer sequences often projected as the most efficient weapon in arsenal for resolving phylogenetic relationships at different divergence levels (Hillis and Dixon 1991) do suffer from certain shortcomings. Though ITS2 sequences are used to resolve phylogenies at intra-individual, population and interspecies levels yet owing to their high variability, their use is often restricted to closely related species, finding little usage in phylogeny. The divergence in ITS sequences that stems from recombinant and pseudogenic variants often leads to misleading results and hence, reliance on ITS sequence only can prove costly making this marker a double edged sword. Role of secondary conformations of the ITS regions in defining the cleavage sites to release the ribosomal genes during the maturation process is a well known phenomenon. If not more, secondary structure of RNA is as important as the sequence for the function. ITS2
Secondary structure prediction serves in furnishing additional information for phylogenetic inferences and differentiating in functional and pseudogenic ITSs (Wesson et al. 1992).

Although the tertiary structure of a functional RNA molecule is crucial determinant of its function, but prediction of its three dimensional structure from the sequence is difficult and cumbersome. However, the secondary structure is known to be conserved in functional RNAs and important to the function of the RNA. Secondary structure models can be used for improving alignments at higher systematic levels even with strongly divergent regions such as the ITS, and the framework dictated by the secondary structure is considered as a tool for expanding the preliminary molecular phylogenies. Hence, the secondary structure is usually considered a sufficient approximation of the tertiary structure and several methods for predicting the secondary structures have been developed and implemented.

Since ITS2 secondary structure of numerous eukaryotes has been elucidated in the recent past (Joseph et al. 1999), presence of same overall secondary structure in eukaryotic groups can not be ruled out (Coleman and Vacquier 2002, Mai and Coleman 1997, Michot et al. 1999). Among various parameters known to stabilize the RNA secondary structure, structural energy holds prime importance. This study is an attempt to cluster the sequences based on the inherent information and visualize valuable interrelation among different inherent features of sequence and secondary structure of ITS2 of the Anopheline species.

**MATERIALS AND METHODS**

**DATA COLLECTION AND DATA SET PREPARATION**

ITS2 sequences of Anopheline species were retrieved from NCBI. Sequences were checked for redundancy and filtered. The final curated input dataset contains a total 123 sequences.

Along the parameters known to contribute towards RNA secondary structure stability, the parameters considered for this study are listed in table 1:

<table>
<thead>
<tr>
<th>SL No</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hairpin Loop</td>
</tr>
<tr>
<td>2</td>
<td>Internal Loop</td>
</tr>
<tr>
<td>3</td>
<td>Bulge Loop</td>
</tr>
<tr>
<td>4</td>
<td>Multi Loop</td>
</tr>
<tr>
<td>5</td>
<td>Exterior Loop</td>
</tr>
<tr>
<td>6</td>
<td>Total Bases</td>
</tr>
<tr>
<td>7</td>
<td>G/C Ratio</td>
</tr>
<tr>
<td>8</td>
<td>Energy</td>
</tr>
<tr>
<td>9</td>
<td>GC content</td>
</tr>
</tbody>
</table>

**SECONDARY STRUCTURE PREDICTION**

RNA secondary structure consists of stems and loops. Mainly five types of loops are present in RNA secondary structure, namely, interior, hairpin, exterior, multi and bulge. For in depth analysis, calculation of secondary structure and determination of structural conservation is essential.

**RNA FOLDS ANALYSIS**

Probable target accessibility (loops) was determined using Sfold (statistical Folding and Rational Design of Nucleic Acids) in the Sribo program based on statistical sample of Boltzmann ensemble for secondary structures. Different loops generated from ITS 2 data using Sribo Program were calculated. (fig. 1)

**STRUCTURAL ENERGY CALCULATION**

Structural energy seems to be most important factor influencing the structural stability. The secondary structure with the lowest possible free energy value, the minimum free energy (MFE) structure, is predicted to be the most stable secondary structure for the strand. Among the sub-optimal structures calculated by Sribo program, lowest energy holding stable structures were considered and utilized for data mining analysis to interpret the influence of different factors on secondary structure stabilization.

**GC CONTENT CALCULATION**

GC content is known to influence structural energy. GC percentage was determined using GC calculator (http://www.genomicsplace.com/gc_calc.html). All non-DNA characters except N were stripped before computing.

Besides the above mentioned parameters, other features like total bases were calculated manually.

**DATA MINING ANALYSIS**

We are living in a data rich information poor world where the magnitude of data generated from the high through put methods is overwhelming; Data mining opens a new
Clustering and Classification of Anopheline Spacer Sequences using Self Organizing Maps

window of opportunity in this arena. In the present study, data mining approach was utilized to find out the concealed information inherent in the sequence that finally affects the structural stabilization.

**SELF ORGANIZING MAPS (SOM): ARTIFICIAL NEURAL NETWORKS (ANNS) IS AN ABSTRACT SIMULATION OF A REAL NERVOUS SYSTEM THAT CONTAINS A COLLECTION OF NEURON UNITS COMMUNICATING WITH EACH OTHER VIA AXON CONNECTIONS.**

In SOM, neurons compete with each other to earn the right of representing the input data (Kohonen 2001). As a result, data in the multidimensional attribute space can be abstracted to a much smaller number of latent dimensions organized on a basis of a predefined geometry in a space of lower dimensionality, usually a regular two-dimensional array of neurons. By this way the structures embedded in the input data can be revealed which is the input space and is spanned over the distribution. Using a SOM network, it is possible to obtain a map of input space where closeness between units or clusters in the map represents closeness of the input data. Processing units in the SOM lattice is 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. During the learning stage the weights of the units change their position and “move” towards the input points. This “movement” becomes slower and at the end of the learning stage, the network is “frozen” in the input space. After the learning stage the inputs can be associated to the nearest network unit. When the map is visualized, the inputs can be associated to each cell on the map. One or more cell that clearly contains similar objects can be considered as a cluster on the map. These clusters are generated during the learning phase without any other information. It is not necessary to supply to the network cluster prototypes or examples. SOMs cluster the data in a manner similar to cluster analysis, but have an additional benefit of ordering the clusters and enabling the visualization of large numbers of clusters. These clusters are arranged in a low-dimensional topology-usually a grid structure that preserves the neighborhood relations in the high dimensional data (Kohonen T 1982, Nurnberger A. and Detyniecki 2002, Cuadros-Vargas et al. 2003). The characteristic that distinguishes the SOM net from the other classification algorithms is that not only similar inputs are associated to the same cell but also neighborhood cells contain similar documents. This property together with the easy visualization makes the SOM map a useful tool for visualization and clustering of large data sets.

**PARAMETERS IDENTIFIED FOR SOM:**

Structural parameters like Hairpin Loop, Internal Loop, Bulge Loop, Multi Loop, External Loop, Energy and inherent sequence parameters like total bases, G/C ratio, and GC content% were considered for this study.

**DATA NORMALIZATION:**

Summarized data is normalized linearly such that minimum value in each category is 0 and the maximum 1. This is done to ensure that all the parameters are given equal importance when clustering is done.

**RESULTS AND DISCUSSION**

**Figure 1:** Distribution of various loops in sequences

In short:

- Total no of sequences selected for study = 123
- Total number of input parameters = 9
- Total iterations per sequence to form a neuron = 10,000
- Total iterations to form 4 grid (2X2) structure = 123,000,000
- Successful or winning neurons = 4
- Unsuccessful neuron = 0
Cluster (1, 1): This cluster contains 4 sequences and is characterized by moderate values for all the parameters. External loop shows the least variation while maximum variation was observed in bulge loop.

Cluster (1, 2): This cluster comprises of total 69 sequences. Maximum variation was observed in internal loop followed by hairpin, multi, bulge and external loop. Structural energy is high in all the sequences except for An. homunculus and An. cruzi. Different sibling species of An.crucians showed more variation in loops while maintaining the similar values for GC content and no. of bases and similar value for structural energies. Same trend was observed in An. annulipes, An. fluviatilis and An. rangeli sibling species.

Cluster (2, 1): The sequences falling in this cluster show uniformly high energies and similarly high G/C ratio while GC content% for these sequences is found to be quite low contrary to the popular belief of GC content being the most important parameter in determining the Structural energy. Highest variation is observed in internal loop followed by hairpin, multi, bulge and external loop. Anopheles bancrofti genotype D present in this cluster shows a very high value of external loop. Although two sequences belonging to Anopheles Lesteri differ in all sequences and structural features, the structural energy is quite the same.
Cluster (2, 2): Total 28 sequences fall in this cluster. This cluster is characterized by variation in structural energies which is reflected in the gradient. This cluster has sequences that show differences in base number unlike other cluster. External loop showed lowest variation followed by multi loop. Variation in structural energy for different Anopheles farauti sequences was observed while sequence features did not show variation for these sequences. Variation in structural features possibly has a profound influence in these. Sibling species Anopheles dirus A and An. dirus D showed similar values for all parameters except hairpin loop and multiloop.

DISCUSSION

Molecular taxonomists are generally overwhelmed by complexity of smothering sequence information owing to their number and sibling status of Anopheline species. Dearth of reliable molecular markers has led to an unquenched search to find new ones and utilizing the existing knowledge at different levels to unveil new patterns and phylogenetic inferences. Among them, ITS 2 sequence draws special attention, which is known as a well trusted marker among molecular entomologists but the sheer complexity of phylogeny often hinders and limit its use for reaching to meaningful conclusions and its application in resolving phylogeny across several taxa is debatable (Banerjee et al. 2007). Data mining approaches can be utilized for harnessing of ITS2 secondary structure information of numerous Anopheline ITS2 sequences which remain lying undeciphered in several public domain repositories. Derived secondary structural information is valuable and reliable in this context.

Since its inception by McCulloch and Pitts in 1993, ANN has come a long way and now encompasses a wide range of fields. Application of neural networks within the medical domain for clinical diagnosis, image analysis and interpretation and drug development have been reviewed in past. SOM is a novel approach that belongs to the class of unsupervised neural networks with competitive learning algorithm ability. The SOM approach is useful for extracting implicit, valuable, and interesting data from vast quantities of information. In this approach, neurons compete with each other to earn the right of representing the input data (Oja and Kaski 1999, Kohonen 2001). As a result, data in the multidimensional attribute space can be abstracted to a much smaller number of latent dimensions organized on a basis of a predefined geometry in a space of lower dimensionality, usually a regular two-dimensional array of neurons. Using this approach, the patterns embedded in the input data can be revealed. SOMs cluster the data in a manner similar to cluster analysis, but have an additional benefit of ordering the clusters and enabling the visualization of large numbers of clusters (Bock 2004). This technique is particularly useful for the analysis of large datasets where similarity matching plays a very important role. SOM compresses information while preserving the most important topological and metric relationships of the primary data items (Kirk and Zurada 1999). SOMs have successfully been applied for classification of DNA sequences based on codon usage (Kanaya et al. 2001, Supek and Vlahovicek 2004), nucleotide frequencies (Abe et al. 2003), virtual potentials (Sousa and Sousa 2003), protein sequences analysis (Ferran and Ferrara 1992, Ferran et al. 1994) and clustering of...
microarray data (Toronen et al. 1999) and in various epidemiological studies (Valkonen et al. 2002, Wang et al. 2002).

In the data set considered, GC content ranges from 44.6 % to 70.8% where Anopheles engensensis showed lowest and Anopheles farauti showed highest GC content. G/C ratio also showed a great deal of variation ranging from 0.76 to 1.45 with Anopheles annulieps and Anopheles vaanenedi showing minimum and maximum G/C ratio respectively. Structural energy, the major factor in stabilizing ITS2 secondary structure varied from –368 Kcal to –51.8 Kcal. A great deal of variation was observed in the number of various loops in secondary structure. Among the loops, highest variation was observed in internal loop followed by multi loop, bulge loop and hairpin loop while exterior loop showed least variation according to number of loops generated. Majority of sequences (47.9% of total sequences) did not show the presence of exterior loops while 39.85% showed only 1 exterior loop and only 1 sequence showed 9 exterior loops. Only 1 sequence was found to lack the internal loop while maximum number of internal loops was obtained in Anopheles epiroticus. Lowest number of bulge loop was observed in An. sinensis while highest number was observed in An. dirus D. Anopheles crucians B showed 13 multiloop could not be detected in 3 sequences and highest number observed was 33.

Clustering and visualization of sequence data using SOM according to inherent features enable efficient interpretation and analysis. The relationship of structural energy with sequence composition features and structural parameters can be explained using this technique. SOM reduces the complexity of multidimensional data hence can be effectively used for finding explicit relationships in such cases.

CONCLUDING REMARKS

RNA secondary structure is crucial to three dimensional structure but determination of the correct structure and folding pattern of ITS2 is cumbersome. It is practically unfeasible to calculate the effect of parameters influencing the structural energy of the RNA structure by conventional experimental approaches. With exponential increase in sequences, complexities in deriving interpretation and inferences from the accumulated data will pose an infinite challenge. Data mining approaches can streamline and facilitate in elucidating inherent explicit hidden information in these cases and will empower us in determining not-so-obvious interrelationships. Different RNA folding algorithms also take into account the structural energy as the major determinant in furnishing RNA secondary structure models and conformation. Clustering and visualization of such data will definitely add meaningful dimensions to our understanding of the relationships among the sequence features and structural parameters that come into play in determining the structural energy. This approach can be further fine-tuned in resolving ambiguities using differences at the RNA structural level for identification of sibling species complexes.

ACKNOWLEDGEMENT

The authors are grateful to the Director, Indian Institute of Chemical Technology, Hyderabad for his continuous support and encouragement.

CORRESPONDENCE TO

Dr. Upadhyayula Suryanarayana Murty Scientist “F”/Deputy Director Head, Biology Division, Indian Institute of Chemical Technology, Hyderabad-500007, India. E-mail: murty_usn@yahoo.com Phone: +91 40 27193134; Fax: +91 40 27193227

References


(http://rbm.who.int/wmr2005/html/exsummary_en.htm.)

Author Information

Amit Kumar Banerjee, M.Sc.
Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology

Neelima Arora, M.Sc.
Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology

Upadhyayula Suryanarayana Murty, M.Sc., Ph.D.
Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology