Molecular Phylogeny of Y-Box Proteins and their Cold Shock Domains

A Mani, D Gupta

Purpose: Y-box proteins are a family of highly conserved nucleic acid binding proteins that are conserved from prokaryotes to human. They are supposed to be involved in both transcriptional and translational control. Little is understood about their evolution end genomic diversity among different taxa. The study was performed by combining Bioinformatic and phylogenetic approaches in order to address first cross family evolution of Y-box proteins from different eukaryotic organisms.

Methods: For sequence analysis, inferring the phylogenetic tree and evolutionary characterization the protein sequences of Y box proteins from 17 different species were used. The tree was drawn by Neighbour-joining method while using the bootstrapping as a test of inferred phylogeny.

Result: The phylogenetic trees were constructed from multiple aligned sequences showing bootstrap values on nodes and species codes on leaves. The analysis of data led to a single most consistent tree.

Discussion and conclusion: The results endow with an excellent perception about the evolutionary order of Y-box proteins. The study established an overall framework of information for the family of Y-box proteins, which may facilitate and stimulate the study of this gene family across all organisms.

INTRODUCTION

The Y-box proteins are the most evolutionarily conserved nucleic acid-binding proteins yet described, found in bacteria, plants and animals. All vertebrate Y-box proteins contain a variable N-terminal domain, a Cold Shock Domain, and a C-terminal tail domain. The CSD is a highly conserved nucleic acid binding domain that confers RNA- and single stranded and double stranded DNA binding activities to the Y-box proteins. The most extensively studied Y-box protein, YB-1 has been found to be upregulated during prostate cancer tumor expression [1]. Increased YB-1 expression has been correlated with DNA topoisomerase II α and proliferating cell nuclear antigen expression in human lung cancer and colorectal cancer [2] and linked to markers of cellular proliferation in osteosarcoma [3]. YB-1 has been identified as cell cycle stage specific transcription factor [4]. They have ability to bind Y-box elements [5, 6, 7, 8]. The eukaryotic Y-box proteins were originally identified through their ability to interact with DNA containing a reverse CCAAT box, the Y-box sequence CTGATTGGCCAA [9]. This sequence is found in a variety of promoter regions, including those of the MHC class II genes [9, 10] and genes encoding germ cell-specific functions [5] and in these contexts the Y-box proteins are considered to act as regulators of transcription. Although many Y-box proteins function to regulate transcription and some demonstrate RNA specificity [11], others have an mRNA stabilization role in the cytoplasm [12]. They have been implicated in various cellular processes, including adaptation to low temperatures, cellular growth, nutrient stress and stationary phase [13]. The discovery of a domain, the cold-shock domain that shows strikingly high homology and similar RNA-binding properties to CSPs in a growing number of eukaryotic nucleic-acid-binding proteins suggests that these proteins have an ancient origin.

Despite of being highly conserved and having major role in transcription and translation as well as their nucleic acid binding properties, little is understood about the evolution and genomic diversity of Y-box proteins among different taxa. Objective of this study was to evolutionarily characterize these proteins in higher eukaryotic organisms as well as to carry out comparative analysis of these organisms based on Y-box protein sequences. We also attempted to find out whether these proteins are conserved enough to represent evolutionary order as inferred by using complete genome.
MATERIALS AND METHOD

In order to search Y-box protein family members BLAST [14] was performed by using blastp program in the protein database at NCBI [15]. Mus musculus y-box protein’s gi|2745892|gb|AAB94768.1 amino acid sequence was selected as query. From the hits 17 sequences (Table 1) each from different species were selected for further studies. All the sequences were taken in FASTA format. The sequences were examined individually and aligned using ClustalW [16]. Bioedit version 7.0.9.0 [17] was used for manual editing and analysis of sequences. Kyte J and Doolittle [18] method was used to plot hydrophobicity profile. Entropy was calculated as

\[ H(l) = -\sum f(b,l) \log_2 f(b,l) \]

where \( H(l) \) = the uncertainty, also called entropy at position \( l \), \( b \) represents a residue (out of the allowed choices for the sequence in question), and \( f(b,l) \) is the frequency at which residue \( b \) is found at position \( l \). The information content of a position \( l \), then, is defined as a decrease in uncertainty or entropy at that position. A window of defined size that was 13 is moved along a sequence, the hydropathy scores were summed along the window, and the average (the sum divided by the window size) was taken for each position in the sequence. The mean hydrophobicity value was plotted for the middle residue of the window. Eisenberg et. al. method [19] was used to plot hydrophobic moment profile with a window size of 13 residues having six residues on either side of the current residue and rotation angle, \( \theta = 100 \) degrees.

\[ \mu H = \left( \left[ \sum H_n \sin(\delta n) \right] + \left[ \sum H_n \cos(\delta n) \right] \right)^{(1/2)} \]

Where \( \mu H \) is the hydrophobic moment, \( H_n \) is the hydrophobicity score of the residue \( H \) at position \( n \), \( \delta = 100 \) degrees, \( n \) is position within the segment, and each hydrophobic moment is summed over a segment of the same defined window length.

For a conserved region search within the multiple aligned sequences minimum segment length was set to 15 residues, maximum average entropy was set to 0.4 and the gaps were limited to 2 per segment. Multiple sequence alignment, phylogenetic and molecular evolutionary analyses were conducted using MEGA version 4 [20]. For pair wise and multiple alignment gap open penalty was -7 and gap extension penalty was -1 [21]. BLOSUM weight matrix was used for substitution scoring [22]. Hydrophilic gap penalties were used to increase the chances of a gap within a run (5 or more residues) of hydrophilic amino acids; these are likely to be loop or random coil regions where gaps are more common. The multiple alignments of sequences of Y-box proteins and cold shock domain were used to create phylogenetic trees. The evolutionary history was inferred using the Neighbour-Joining method [23]. All the characters were given equal weights. The bootstrap consensus tree inferred from 10000 replicates was taken to represent the evolutionary history of the taxa analyzed [24]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) were shown next to the branches [25]. The tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the poisson correction method and are in the units of the number of amino acid substitutions per site [25]. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). There were a total of 158 positions in the Y-box proteins’ final dataset, out of which 87 were parsimony informative. Cold shock domains final dataset was having 72 positions out of which 16 were parsimony informative. Phylogenetic analyses were conducted in MEGA4 [20].
RESULTS AND DISCUSSION

MULTIPLE SEQUENCE ALIGNMENT

The Multiple alignment of cold Y-box protein sequences (Fig. 1) resulted into an alignment having 409 positions. The Multiple alignment of cold shock domains (Fig. 2), which vary in the Y box proteins from position from 93 to 164, resulted into alignment with 72 positions suggesting their high conservedness. By statistical analysis of multiple aligned sequences it was observed that glycine, arginine, proline, alanine, glutamine and valine are the most frequently present amino acids with frequency percentage of 12.76, 12.03, 9.52, 8.21, 8.21 and 5.76 respectively. While within conserved sites glycine, valine, alanine, asparagine, glutamine, lysine and serine are the most frequently occurring amino acids with frequency percentage of 19.13, 14.87, 9.25, 8.50, 8.38, 6.37 and 5.12 respectively. These Y-box proteins contain a cold shock domain which is 70 or 71 residues long and highly conserved. It was observed that within the cold shock domain valine, glycine, asparagine, lysine, glutamine and alanine are the most frequently occurring amino acids with frequency percentage of 14.44, 11.37, 9.54, 9.38, 8.30 and 6.06 respectively. The multiple aligned sequence of Y-box protein was found with No. of conserved sites=54, No. of parsimony informative sites=230 and No. of singleton sites=84. While the multiple aligned sequence of cold shock domain of the protein was found with No. of conserved sites=40, No. of parsimony informative sites=16 and No. of singleton sites=15.

Figure 4

Figure 1a: Multiple sequence alignment of Y-box proteins (Position 1-150)
CONSERVED REGION SEARCH

A conserve region search resulted into three regions (Fig. 3a, 3b and 3c) from position 95 to 126 (segment length = 32), 128 to 153 (segment length = 26) and 156 to 180 (segment length = 25) with an average entropy of 0.2151, 0.2142 and 0.3942 respectively. This conservation has already been upheld by minimal entropy shown by positions 95 to 180.
Molecular Phylogeny of Y-Box Proteins and their Cold Shock Domains

Figure 8
Figure 3a: The first conserved domain found in Y-box proteins

Figure 9
Figure 3b: The second conserved domain found in Y-box proteins

Figure 10
Figure 3c: The third conserved domain found in Y-box proteins

ENTROPY PLOT
An entropy plot, measure of the lack of the information content and the amount of variability, was generated for all the aligned positions. The plot shows that entropy rarely touches a scale of two, showing minimal entropy at several positions from position 96 to position 170 where entropy rarely crosses a scale of one, which is a sign of better alignment in the region.

Figure 11
Figure 4: Entropy (Hx) Plot for aligned Y-box protein sequences

HYDROPHOBICITY PROFILE
A hydrophobicity profile plot shows that mean hydrophobicity of the protein for most of the positions is in
all the species is below zero, occasionally it turns to be positive. Maximum hydrophobicity is observed from positions 10 to 100 for Danio rerio and Columba livia and between 140 to 180 positions. Schistosoma japonicum exhibits high increase in hydrophobicity from position 340 to 360 and 380 to 410. These proteins are basically of non-hydrophobic in nature as most of the positions are across showing below mean hydrophobicity in the case of most of the organisms studied here.

**Figure 12**
Figure 5a: Kyte and Dolittle scale mean hydrophobicity profile plot

![Hydrophobicity Profile Plot](image)

**Figure 13**
Figure 5b: Colour legend for Kyte and Dolittle scale mean hydrophobicity profile plot

![Colour Legend](image)

**PHYLOGENETY**
The phylogenetic tree were constructed by using Neighbour–joining method (Figure 6 and figure7). The tree shows different organisms on tree nodes branched on the basis of their Y-box proteins. Schistosoma japonicum makes a totally diverged branch from the main tree among 17 selected proteins. Node for Endopterygotans (Chironomus tetanus, Bombyx mori, Drosophila melanogaster and Aedes aegypti) is supported by lower bootstrap values i.e. 68% while the node for vertebrates is supported by very high bootstrap value i.e. 100%. Node for Teleosteiens (Carassius auratus, Danio rerio and Oryzus laetipus) and mammals’ node is supported by 99% bootstrap value. Node for mammals is
supported by 100% bootstrap value; the only exception is Xenopus tropicalis which is an amphibian. These trees give an idea about the evolutionary order of Y-box proteins. The tree inferred from the cold shock domains (Figure 7) does not show overall consistency with that of whole Y-box tree but gives an idea about their different evolutionary trend. Despite of highly conserved nature of Y-box proteins and cold shock domains this phylogeny does not seem to be completely consistent with the established taxonomy perhaps due to use of a specific protein rather than complete genomes.

**Figure 14**
Figure 6: Bootstrap consensus phylogenetic tree of Y-box proteins created by Neighbour-joining method showing bootstrap support values on the nodes.

**Figure 15**
Figure 7: Bootstrap consensus phylogenetic tree of Y-box proteins’ cold shock domains created by Neighbour-joining method showing bootstrap support values on the nodes.

Though both the trees (Figure 6 and 7) are somewhat similar to the established taxonomy of the organisms but neither the tree made by using whole Y-box protein nor the tree made by only cold shock domain completely comply with established evolutionary order. This inconsistency is higher while using whole Y-box protein rather than only cold shock domain perhaps due to presence of variable N-terminal and C-terminal domains in the protein. Use of only cold shock domain decreases the difference between the order of organisms on the nodes of the phylogenetic tree inferred by NJ method and the tree of established taxonomy since variable N-terminal and C-terminal domains were omitted and inference was based only on the conserved cold shock domain. Molecular phylogeny of various proteins like HSP 90 gene family [26] and coronin gene family [27] have also been inferred with the same inconsistencies, though these studies including the present study furnish an idea about the evolutionary trend in milieu of the protein studied. From figure 7 it can also observed that the within chordates (Birds and mammals) there is greater level of variance in the protein sequence while non chordates, mostly insects, have more conserved Y-box proteins. It appears that the cold shock proteins of prokaryotes which evolved by adding N and C terminal domains to perform diverse functions in eukaryotes by ascending as cold shock domains were probably given selective advantage during the course of evolution for those changes which increased their efficiency for performing more functions which in turn probably increased variations in the amino acid sequences. These variations have become more prevalent among higher organisms with reference to Y-box proteins.

This study presents first comparative proteomic study and evolutionary analysis of the Y-box family of proteins based on molecular phylogeny across different families of 17 eukaryotic organisms. The results endow with an excellent perception about the evolutionary order of Y-box proteins. The study established an overall framework of information for the family of Y-box proteins, which may facilitate and stimulate the study of this gene family across all organisms.

**ACKNOWLEDGEMENTS**

AM is thankful to University Grants Commission, New Delhi for a research fellowship. The work has been supported by a DBT-BIF Grant to DKG under its BTISNet scheme.

**References**

Author Information

Ashutosh Mani, M.Sc.
Center of Bioinformatics, Institute for Interdisciplinary Studies, University of Allahabad

Dwijendra K. Gupta, Ph.D.
Center of Bioinformatics, Institute for Interdisciplinary Studies, University of Allahabad