JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# INSIGHTS INTO HOX GENES AND ITS EFFECTS ON SARCOMA

**[1]Veena Kumara Adi and [2]Swarnagouri Naganathanhalli**

[1]Associate Professor, [2] 4th year Student B E Biotechnology

Department of Biotechnology,

Bapuji Institute of Engineering and Technology, Davangere 577004, INDIA

drveena.adi@gmail.com

***Abstract :*** Sarcoma is a rare heterogeneous malignant tumor that occurs frequently in connective tissues. They contribute to 1% of the adult population contributing around 100 different types which are differentiated normally by the similarity of the mesenchymal tissue. Of late, some studies have shown the involvement of the HoxC and HoxD genes in sarcomas. The HOX genes (in humans) are organized in four clusters – HOXA, HOXB, HOXC, and HOXD. Each cluster has genes numbered 1-13. In the present study, the structure of the human Hox gene from NCBI Entrez was used to study its conservation and structure. The paralogue Hox13 genes showed high-level conservation in the results sequence alignment and conserved domain search near the 5' end. RNAfold web server conveyed a high negative free energy and PSIPRED, AphaFold Structure Database analysis resulted in 3 helix structures in the form of helix-coil-helix. These 3 helices form the specific binding site in Hox. Despite conservation, there were variations in the expression of Hox 13 paralogous gene in sarcoma tissues. Mutation in HoxC13 and HoxD13 caused changes in stability. The prognosis values from the overall survival analysis in GEPIA2 showed the worst prognosis in HoxC13, and HoxB13 than in HoxD13, and HoxA13. Functional expression by Gene Ontology analysis in STRING showed the immune response functions and related pathways. From GO results, mRNA expression of Hox 13 against immune cells in TIMER, HoxC13 has a larger p-value than HoxB13, HoxD13, and HoxA13. Hox genes have a role in sarcoma tumorigenesis and affect immune-related genes.

***IndexTerms* - HOX Genes, Sarcoma, GEPIA 2, GO Analysis, Bioinformatics**

## I. INTRODUCTION

Genome is the vital part that holds the information about the life we lead on. The Human Genome Project (HGP) is one of history's great feats of exploration. It was an inward voyage of discovery that allowed me to read nature's complete print for building a human being. Several facts were unfolded and had a strong impact on research.

The protein-coding sequence is 1.5% of the genome. Identifying and sequencing these protein-coding genes was done with the assistance of comparative analysis of the genome with other organisms. In the comparative analysis of chromatin-associated proteins and transcription factors, a significant revealed amount of domain architecture was shared between humans and the *Drosophila melanogaster* (Consortium, 2001). Additionally, a systematic analysis of *melanogaster* and human showed that 77% of the distant human disease gene is similar to the *Drosophila* genes related to human disease (Lawrence T. Reiter, 2001). In the disorders such as appendicular skeleton, and limb/feet disorder the HOX gene expressions were identified. Thus, Hox genes are considered one of the common transcription factors which are commonly expressed in developmental disorders (Kornak, 2003). The Hox genes are a small part of the Homeobox genes first recorded in *melanogaster* after two mutation observations – the antennapedia mutation (change of antenna to legs) and bithorax mutation (change of haltere to wing) (Lewis, 1978).

In evolution, the homeobox gene is duplicated into two genes forming the protohox gene as such sponges and nematodes have mapped out five or more genes in the same cluster (Simona Santini, 2003). Evolution after *Drosophila* which consists of ten hox in a cluster, two duplication events occurred in early vertebrates. The fishes have been mapped with more than 5 clusters of hox genes. After duplication, many of the old genes were lost but, a similar number of ancestral genes were conserved. As a result, mammals have four clusters of Hox genes. After HGP, the 230 homeobox genes consisting of 257 sequences were discovered by Nam and Nei (Jongmin Nam, 2005). Hox gene families in humans are evolutionarily preserved. They determine the embryonic development and cell memory gene program. These genes are classified into eleven Homeobox classes – ANTP, PRD, LIM, POU, HNF, SINE, TALE, CUT, PROS, ZF, and CRES. The ANTP class is subdivided into NKL (NK-Like) and HOXL (HOX-Like). The HOX genes (in humans) are organized in four clusters – HOXA, HOXB, HOXC, and HOXD. Each cluster has genes numbered 1-13 which are divided into 7 families – Hox1, Hox2, Hox3, Hox4, Hox5, Hox6-8, and Hox9-13. The 39 Hox genes in 4 clusters are organized on four chromosomal loci aligned in 13 paralogous groups based on sequence homology. (Sylvie Forlani, 2003). Off late studies focusing on Hox genes and their expression are surfacing.

The first Hox gene expression was observed during mesoderm formation in the early gastrulation process. These Hox genes further regulate stem cell differentiation, especially in Mesenchymal stem cells (Laurie K Svoboda, 2014) (Seema Bhatlekar, 2018). Its dysregulation occurs in cancer. Recent studies have shown the involvement of the HoxC and HoxD genes in sarcomas (Sarver, 2015). Soft tissue sarcomas (STS) are rare and heterogeneous malignant tumors forming around 1% of adult tumors. In STS there are at least 100 different molecular and histologic types that exhibit different clinical features (Gamboa, 2020). The sarcomas are differentiated normally by the similarity of the mesenchymal tissue. All the types are broadly genetically differentiated into 2 types: the simple karyotypes with simple genetic alteration and malignancies with the complex karyotypes.

In the Pan-Cancer Atlas study of TCGA, the 6 major sarcomas were studied, including 5 complex and simple types of sarcoma. Those types are 1) dedifferentiated liposarcoma (DDLS), 2) leiomyosarcoma (LMS), 3) undifferentiated pleomorphic sarcoma (UPS), 4) myxofibrosarcoma (MFS), 5) malignant peripheral nerve sheath tumor (MPNST) and 6) synovial sarcoma (SS), where the SS is the simple karyotypic tumor (Network, 2017).

Deregulation of Hox genes is associated with developmental abnormalities and human diseases. Paralogous Hox genes 13(HoxA13, HoxaB13, HoxC13, HoxD13) play a relevant role in tumor development and prognosis. In this study, we try to understand the structure of the paralog HOX-13 (HoxA13, HoxaB13, HoxC13, HoxD13) gene. The paralog Hox gene was further studied for the pathway in the sarcoma.

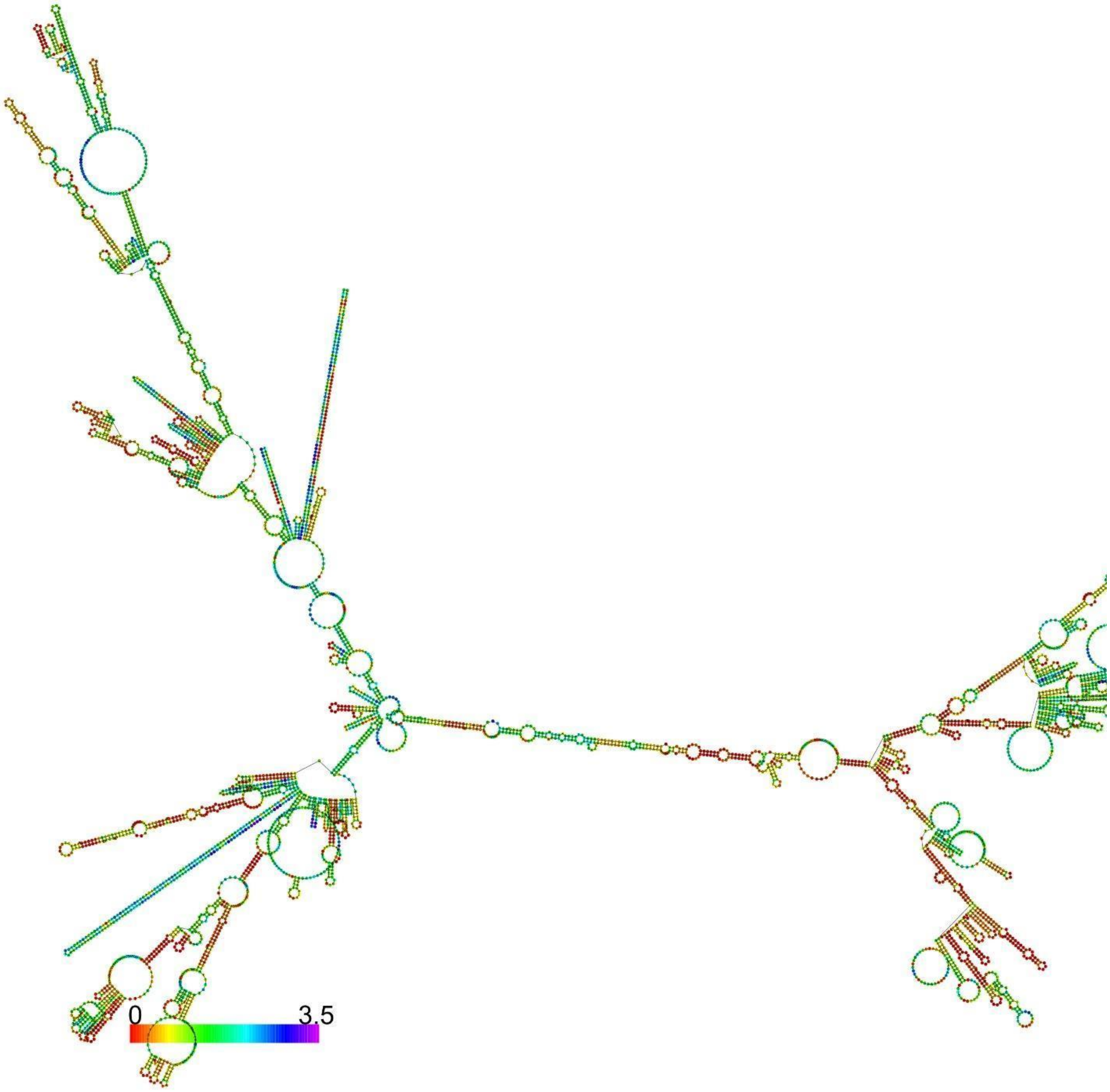
## II. METHODS

In this study, the Hox gene sequences (both mRNA and protein) and their accession numbers were found using features such as advanced search and filters in NCBI Entrez (https://www.ncbi.nlm.nih.gov/) which is the primary text search and data retrieval system. T-Coffee (https://www.ebi.ac.uk/Tools/msa/tcoffee/) for proteins, structural alignment (Expresso) provides simple alignment for the paralog homeobox proteins proving the conversion of the structure. PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) is a secondary structure prediction tool for protein sequence(s) where PSI Blasts the sequence to get the protein hits and the position-specific scoring matrix is iteratively performed to get the secondary structure prediction. The PSIPRED 1.0 (Predict Secondary Structure) was used in understanding the paralog homeobox protein. The Ramachandran plot was obtained for all the sequences to understand the distribution of energies. RNAfold web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) was utilized for the prediction of the structure of mRNA and thus consequently the free energy values of the structure and its stability analysis. AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) was further exerted to predict the protein's tertiary structure based on its AI system. The NCBI CD (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) search allows the search of queries against the Conserved Domain Database (CDD). For most of the searches, the output is provided by the RPS-BLAST with the specified E-value threshold. The paralog Hox 13 genes conserved domain was observed. To understand the interactive explanation of Hox 13 genes in sarcoma, the cBioPortal (http://www.cbioportal.org/) was validated for Hox gene alteration frequencies in sarcoma cancer. Studying of the mutations was done on the I-Mutant 2.0 ( https://folding.biofold.org/i-mutant/i-mutant2.0.html). TCGA, Pan Atlas study of 255 sample study was used. Gene Expression Profiling Interactive Analysis 2 (http://gepia2.ancer-pku.cn/) is a web server for gene expression analysis based on the TCGA and the GTEx databases. Therefore, the GEPIA2 enabled the differential expression of HoxA13, Hoxb13, Hoxc13, and HoxD13 in the sarcoma and normal tissue. Gene enrichment analysis was done to understand the particular pathway expression of Hox gene in sarcoma using STRING (https://string-db.org/). Based on results obtained from the STRING the analysis of tumor-infiltrating cells was done with help of TIMER ( http://cistrome.org/TIMER/ ). GO survival module is used for analyzing each of the HoxA13, Hoxb13, HoxC13, and HoxD13 genes against immune cells.

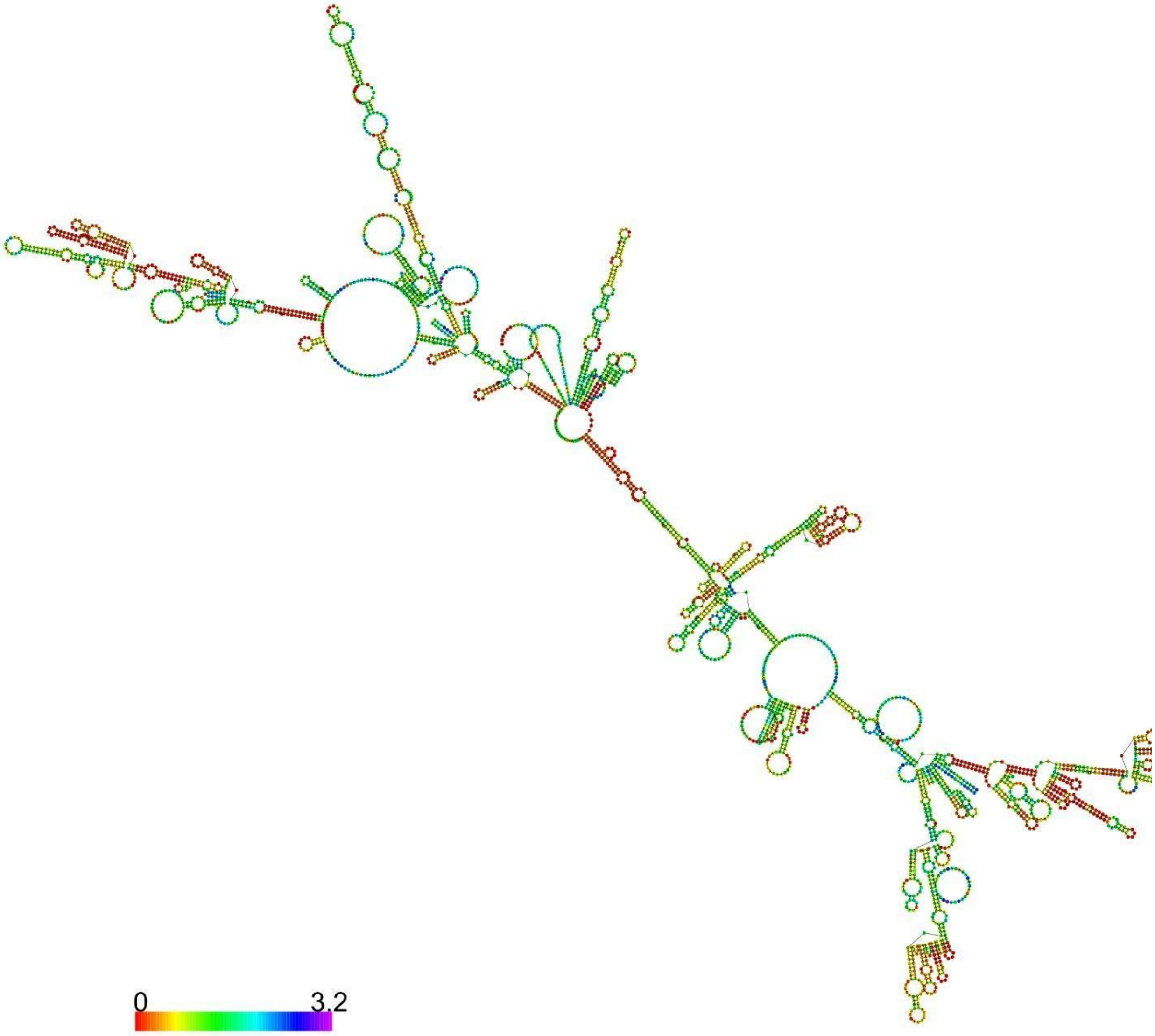
## III. RESULT

The accession numbers NM_000522, NM_006361, NM_017410, and NM_000523 of HoxA13, HoxB13, HoxC13, and HoxD13 mRNA sequence was obtained respectively. The protein sequence of accession numbers NP_000513, NP_006352, NP_059106, and NP_000514 concerning HoxA13, HoxB13, HoxC13, and HoxD13 were derived. The mRNA structure prediction results from the RNA fold web server are shown in fig (1) (2). The MFE (Minimum Free Energy) structure of an RNA sequence is the secondary structure that contributes a minimum of free energy. The structure is predicted based on the loop-based energy model treats the free energy of an RNA secondary structure as the sum of the contributing free energies of the loops. Whereas, the Partition Folding Function (PF) is the probability of the occurrence of the secondary structure contained in the whole Boltzmann ensemble. The Centroid Structure is the secondary structure with minimal base pair distance to all secondary structures in the Boltzmann ensemble.

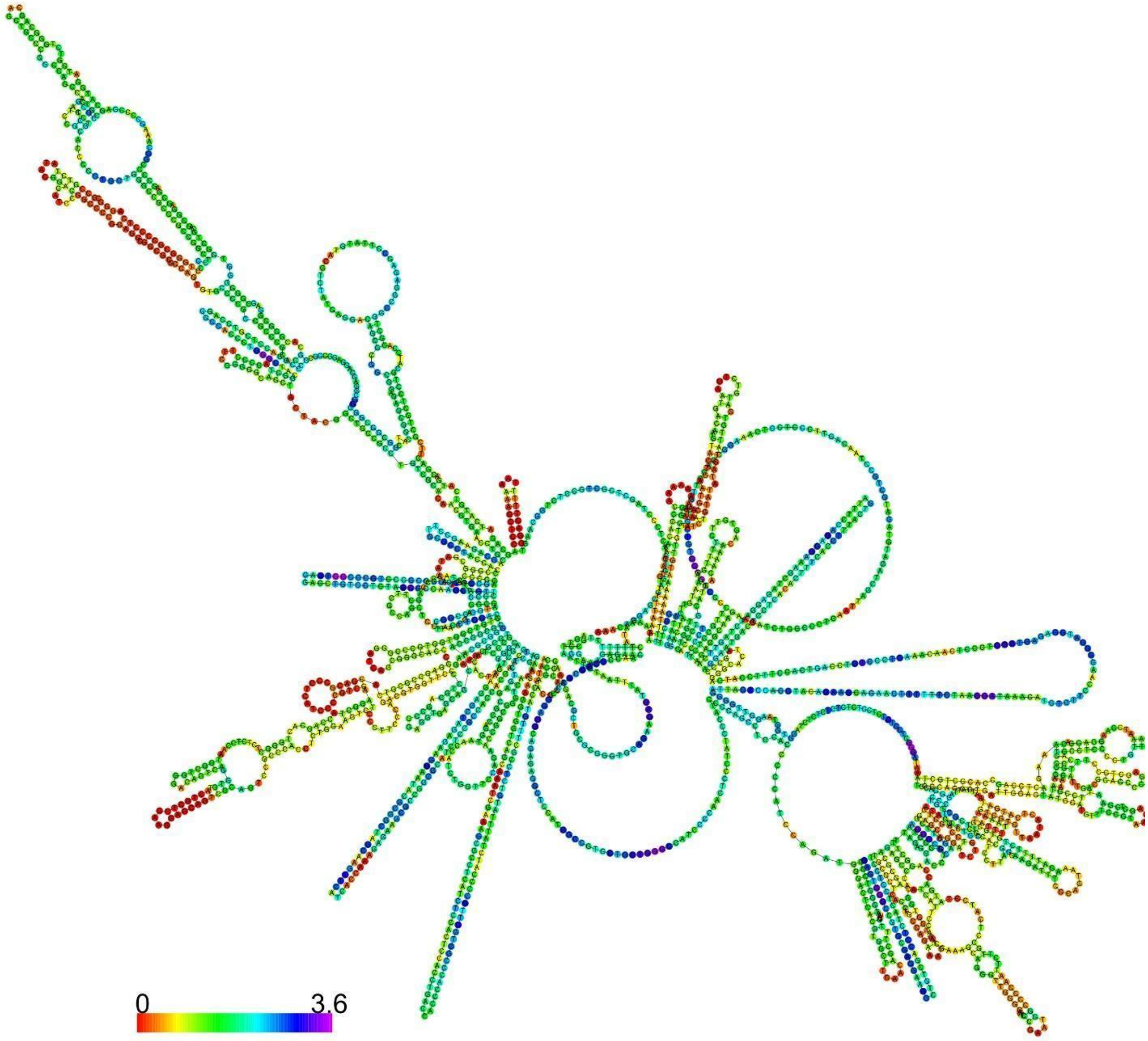
For HoxA13 mRNA the optimal structure with MFE is -1525.30 kcal/mol, PF probability percentage is 0.0, and centroid structure with free energy is -1183.32 kcal/mol. For HoxB13 mRNA the optimal structure with MFE is -1215.30 kcal/mol, PF probability percentage is 0.0, and centroid structure with free energy is -1023.71 kcal/mol. For HoxC13 mRNA the optimal structure with MFE is -888.60 kcal/mol, PF probability percentage is 0.0, and centroid structure with free energy is -532.42 kcal/mol. For HoxD13 mRNA the optimal structure with MFE is -889.00 kcal/mol, PF probability percentage is 0.0, and centroid structure with free energy is -731.80 kcal/mol.



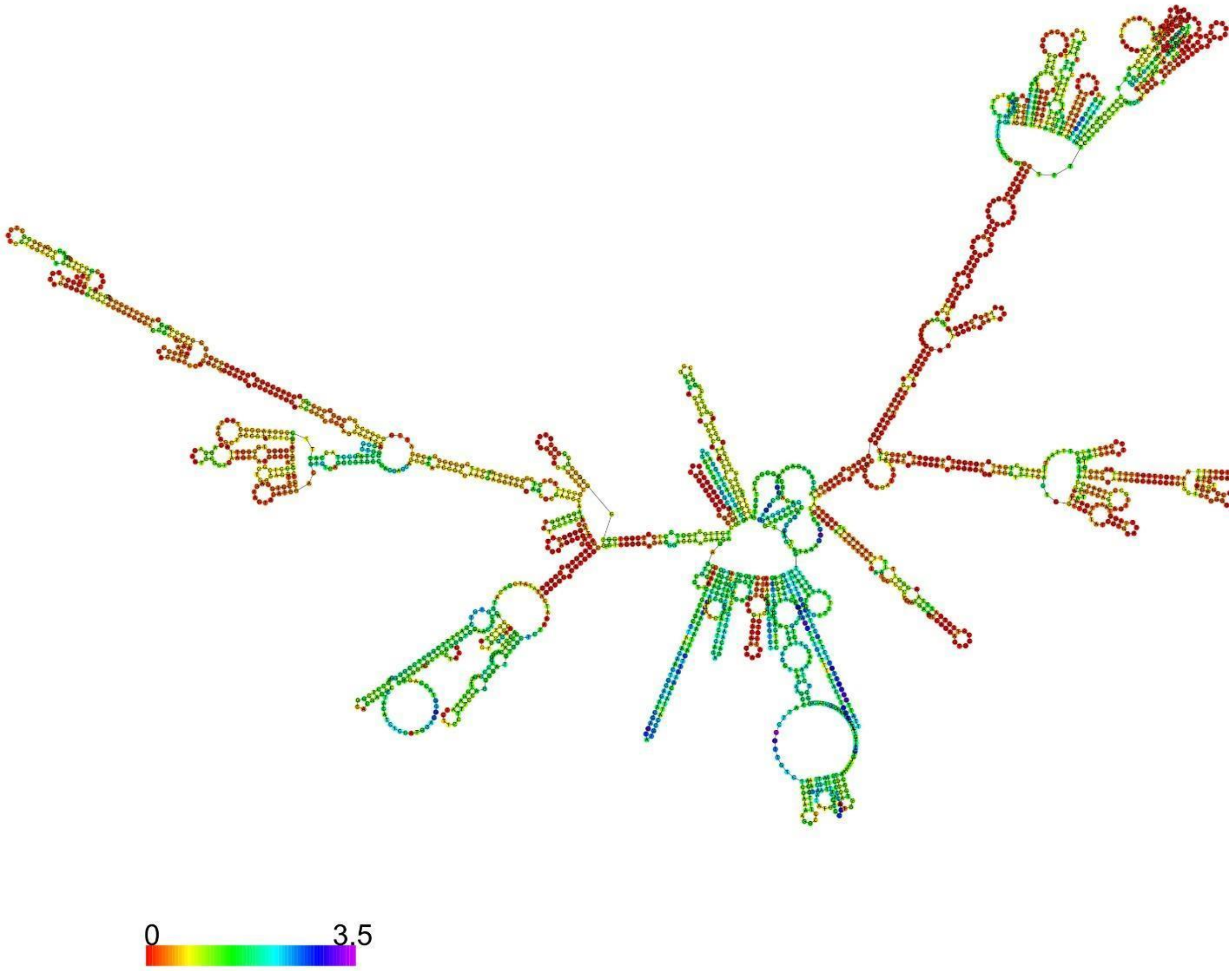
A



B

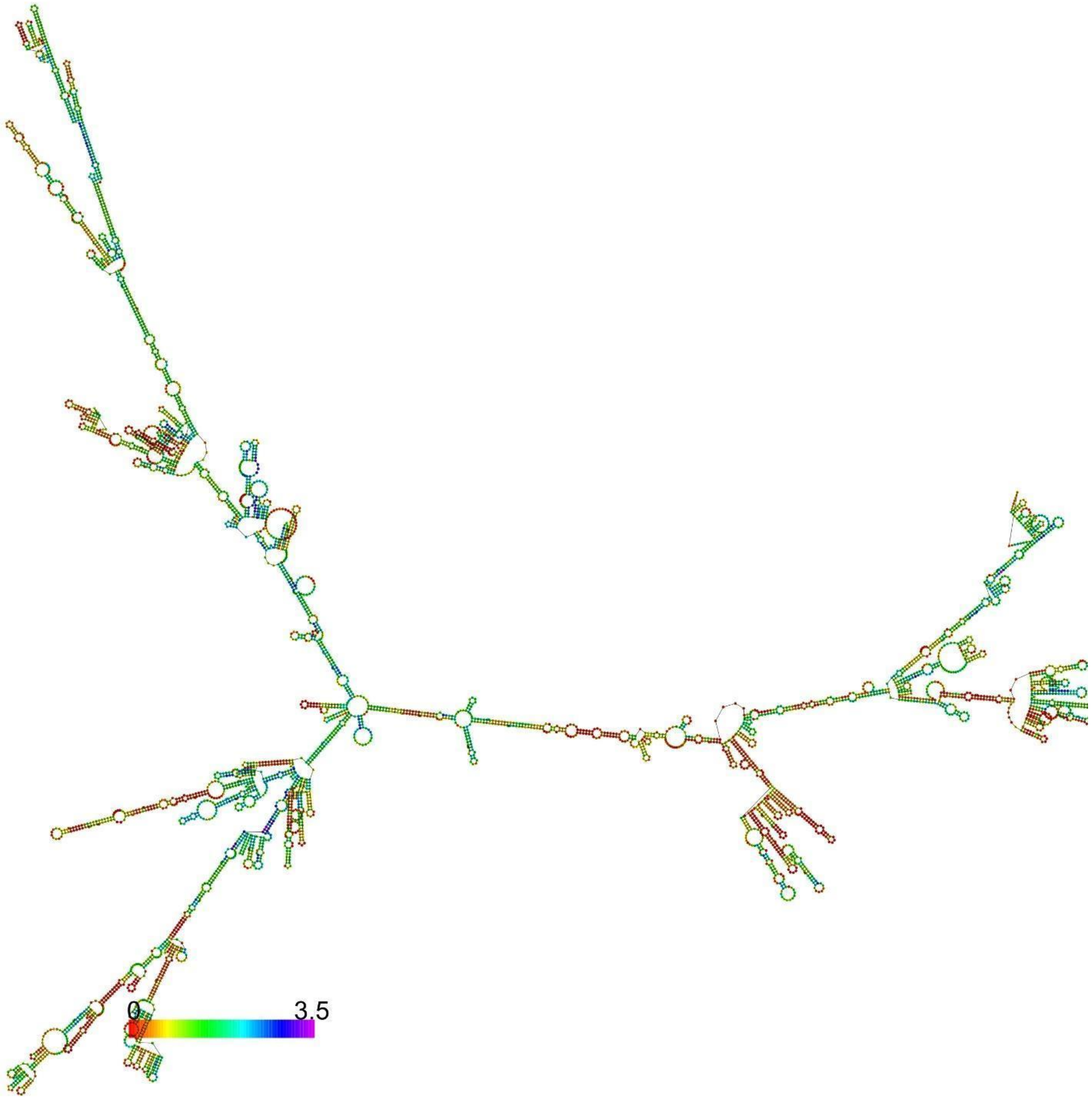


C

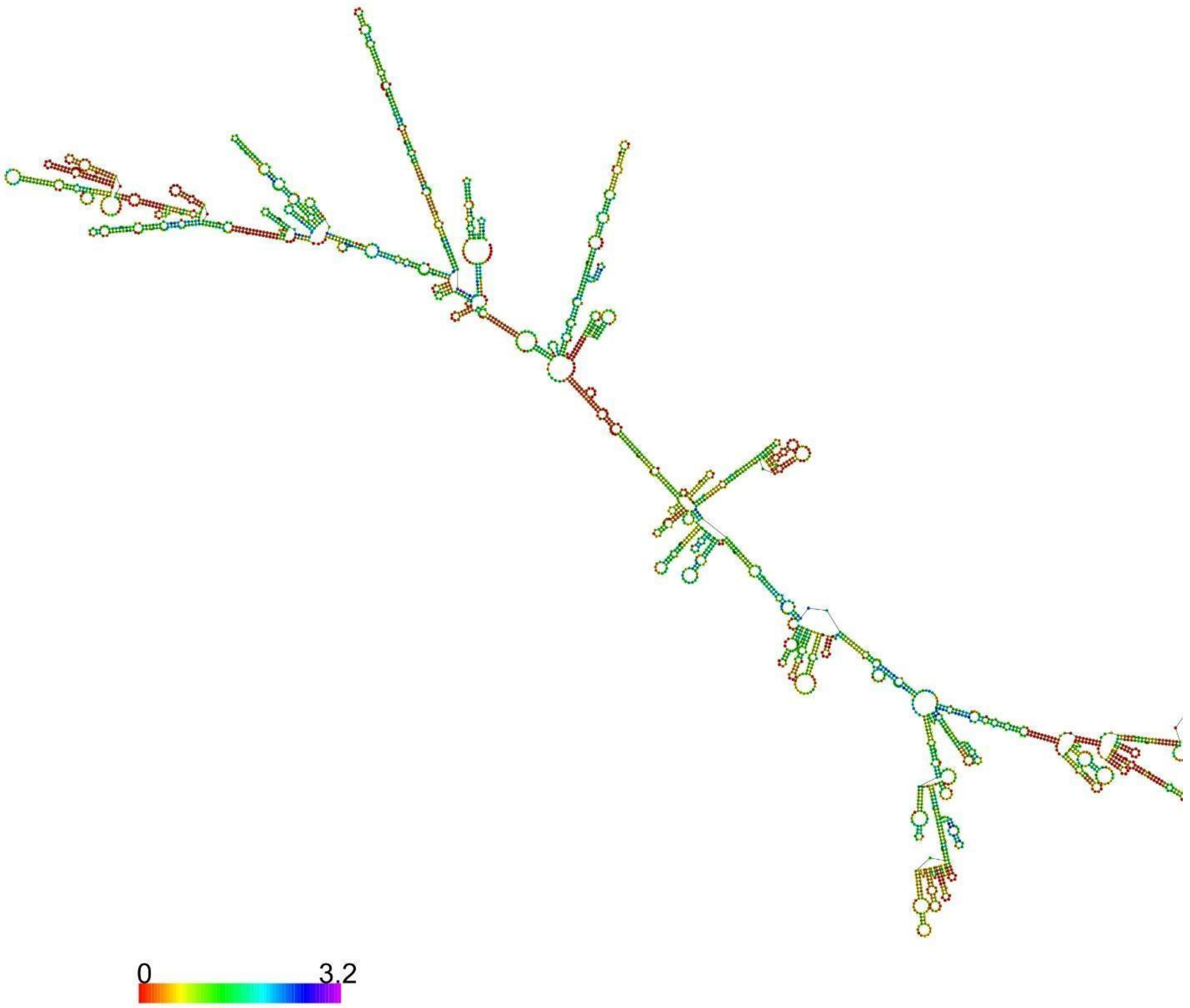


D

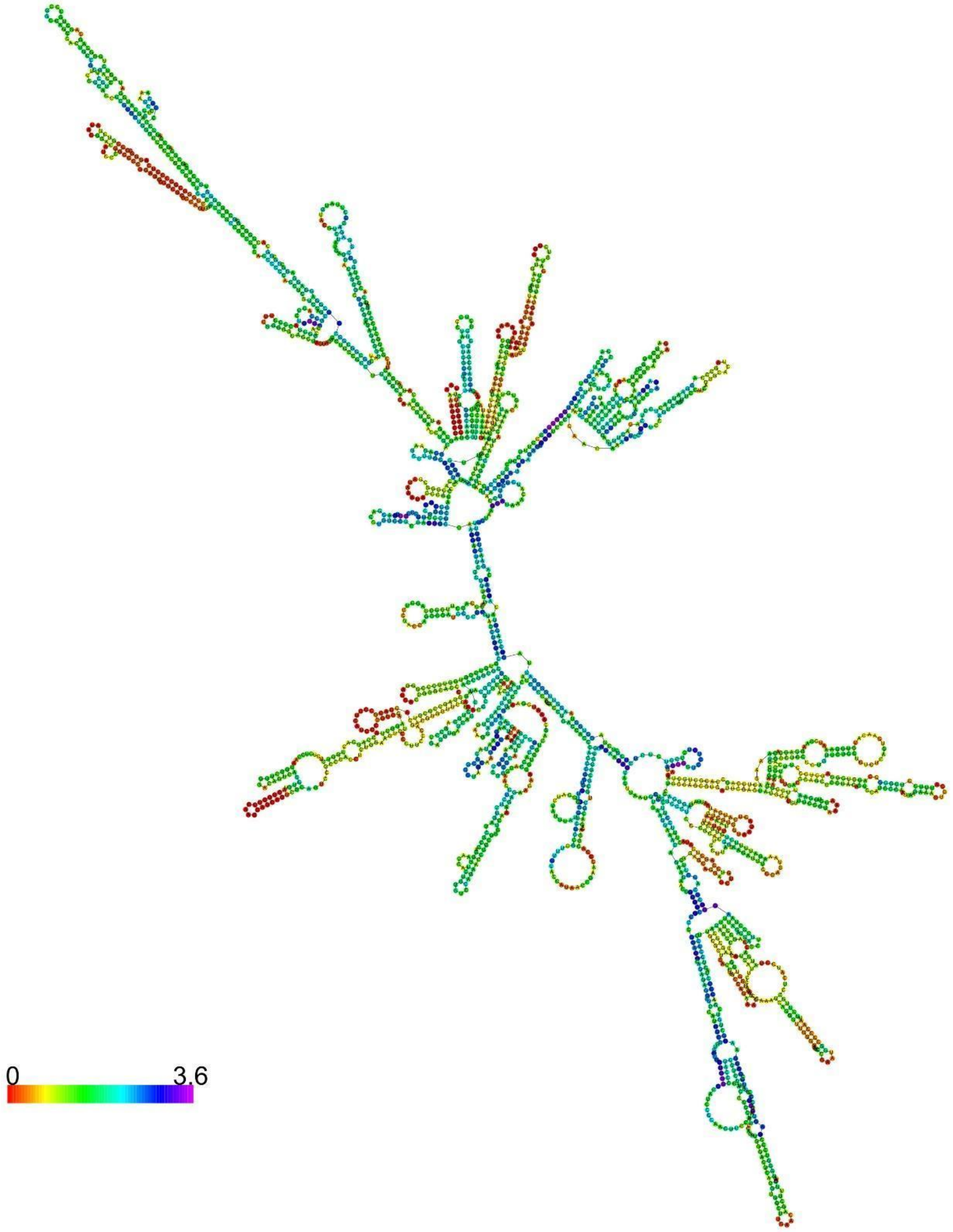
Figure 1: A: Centroid mRNA prediction of HOX A13 Gene, B: Centroid mRNA prediction of HOX B13 Gene, C: Centroid mRNA prediction of HOX C13 Gene, D: Centroid mRNA prediction of HOX D13 Gene.



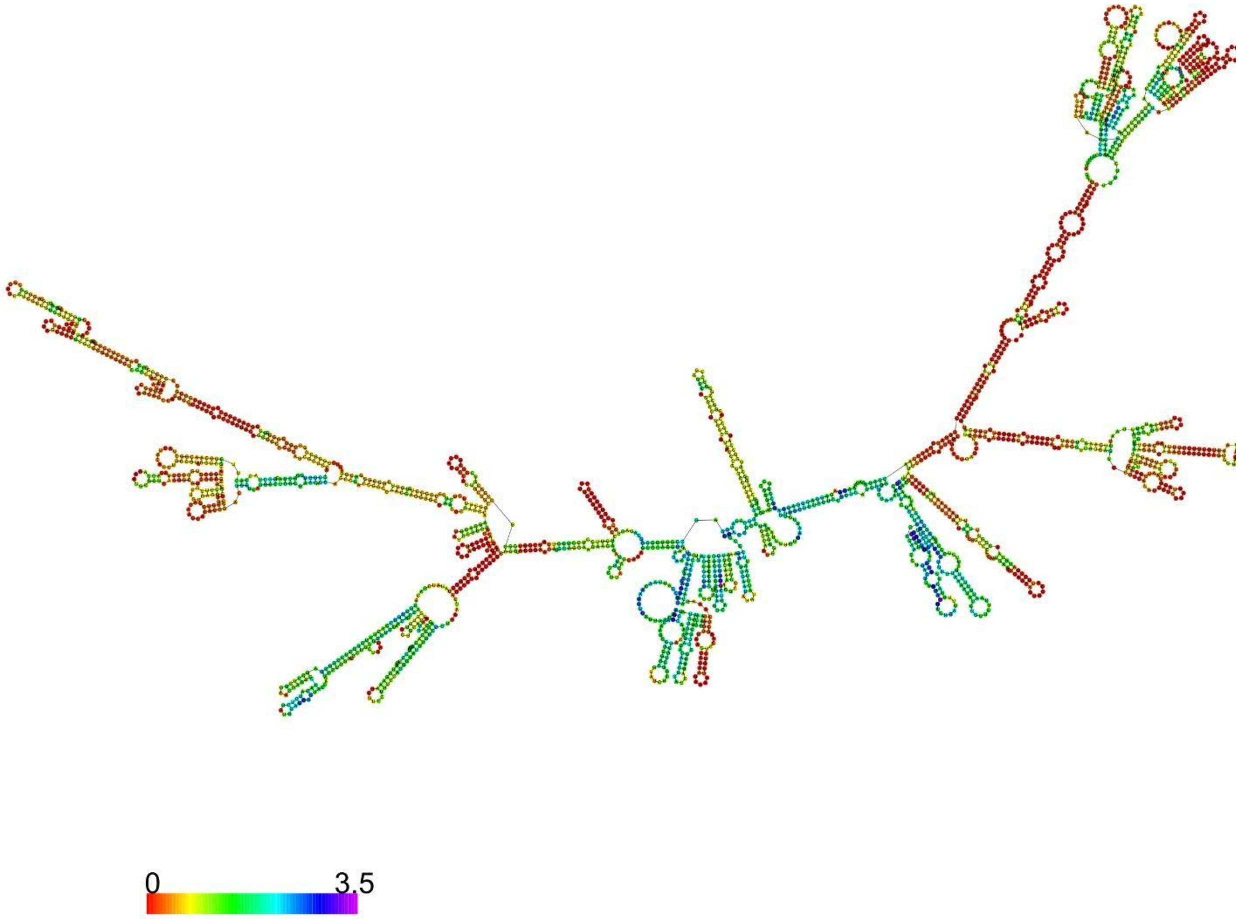
A



B



C



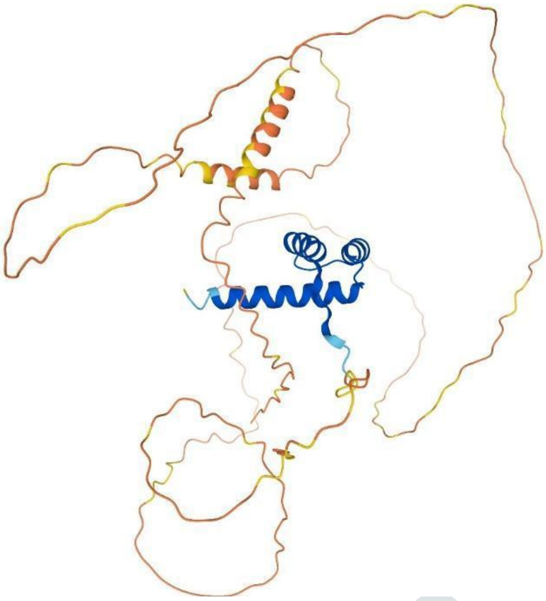
D

Figure 2: A: MFE mRNA prediction of HOX A13 Gene, B: MFE mRNA prediction of HOX B13 Gene,
C: MFE mRNA prediction of HOX C13 Gene, D: MFE mRNA prediction of HOX D13 Gene.

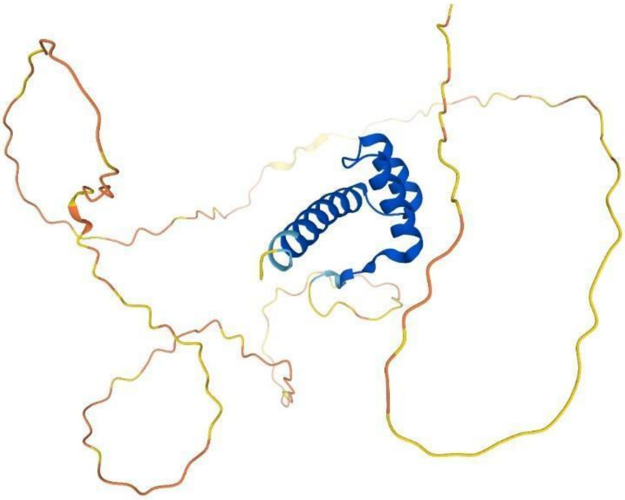
The secondary structure from PSIPRED is shown in (3). The 3 α-helices shaded in the pink are connected with 2 coils in the grey are present near the 5’ end representing the domain. The Helices I and Helices II lie parallel to each other and across them, the third Helices III (Recognition helix) is positioned in a way that it interacts with the DNA in Fig 3. (Sharmila Banerjee-Basu, 2001).

Figure 3: A: HoxA13 secondary structure, B: HoxB13 secondary structure, C: HoxC13 secondary structure, D: HoxD13 secondary structure. All four conserve the Helix-Coil-Helix structure which mainly interacts with the DNA.

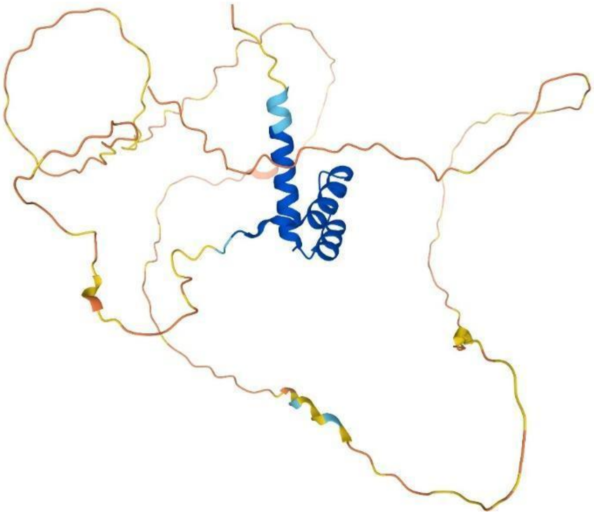
The AlphaFold protein structure database predicted HoxA13, HoxB13, HoxC13, and HoxD13 tertiary structures with three α – helixes with very high (>90%) model confidence. The 3 α-helixes lies in the region predicted in the secondary structure as Fig. 4.



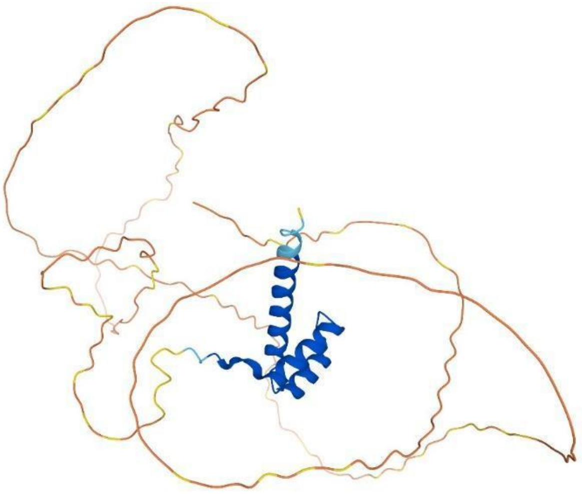
A



B



C

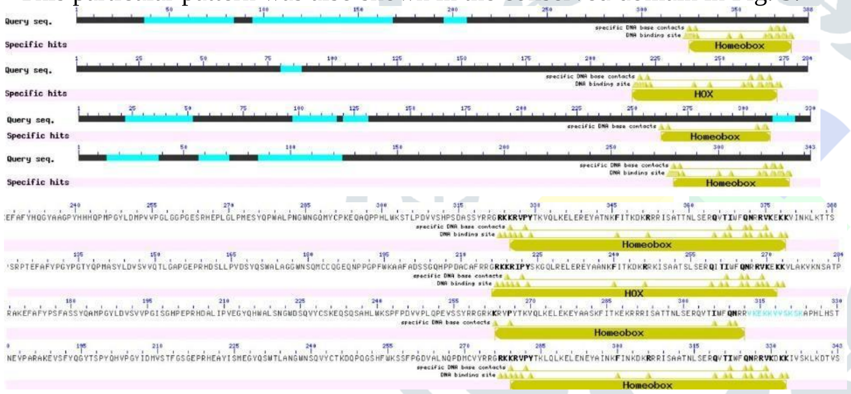


D

Figure 4: A: Tertiary structure of HoxA13 from AlphaFold Protein Database, B: Tertiary structure of HoxB13 from AlphaFold Protein Database, C: Tertiary structure of HoxC13 from AlphaFold Protein Database, D: Tertiary structure of HoxD13 from AlphaFold Protein Database.

From the conservation analysis of the paralog Hox genes, there is absolute conservation of identical amino acids of Helix I (Lue/L and Phe/F), Helix III (Trp/W, Phe/F, Asn/N, Arg/R, Lys/K) (Sharmila Banerjee-Basu, 2001).

This particular pattern was also shown in the conserved domain in Fig. 5.



A

B

Figure: 5 –

A – The domain position in all HoxA13, HoxaB13, HoxC13, and HoxD13 amino acid chains respectively. Homeodomain is in the interval between 325-379 in HoxA13. Homeodomain is in the interval between 216-372 in HoxB13. Homeodomain is in the interval between 263-312 in HoxC13. Homeodomain is in the interval between 279-333 in HoxD13.

B – The amino acids in the homeodomain for DNA binding (i.e; specific DNA binding sites) in the HoxA13, HoxaB13, HoxC13, and HoxD13 respectively.

The specific DNA binding site in the conserved homeodomain shows similar conservation of amino acids near the end of the sequence as is shown in the sequence alignment. The DNA binding domain, therefore, involved in the transcriptional regulation may bind to DNA as a monomer or as homo- and/or heterodimers, in a sequence-specific manner.

The alteration in the gene sequence led to dysregulation which causes cancer. One of the studies has shown the Hox gene involvement in sarcoma (Sharmila Banerjee-Basu, 2001). To analyze alteration in the hox gene cBioPortal tool was enabled. In the cBioPortal the cancer study was queried against HoxA13, Hoxb13, HoxC13, and HoxD13 for profiles for mutations, structural variants, and mRNA expression z-score (threshold ± 2.0) relative to the diploid sample (RNAseq vs RSEM), and protein expression z-score (threshold ± 2.0) RPPA. The alteration frequency in HoxA13, Hoxb13, HoxC13, and HoxD13 were 5%, 7%, 6%, and 4% respectively Fig. 6.

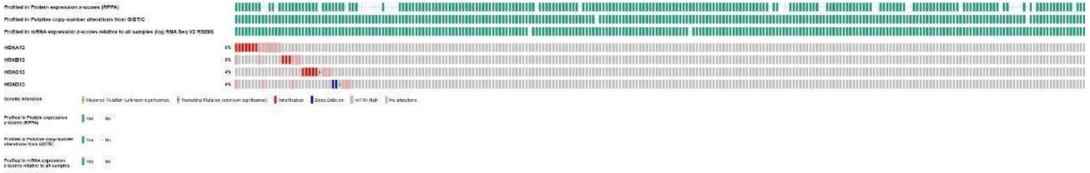


Figure: 6 – The genetic alteration in HoxA13, Hoxb13, HoxC13, and HoxD13 genes for sarcoma.

HoxC13 and HoxD13 expressed missense and truncating mutations in mRNA expression in sarcoma tissues. At position 183 in an amino acid sequence of HoxC13 Glutamine (Q) is substituted. The stability changes due to mutation in HoxC13 protein were obtained from I-Mutant 2.0.

Table 1: Mutations at position 183

| Position | WT | NEW | STABILITY | RI |
|---|---|---|---|---|
| 183 | Q | V | Increase | 1 |
| 183 | Q | L | Increase | 4 |
| 183 | Q | I | Increase | 4 |
| 183 | Q | M | Decrease | 2 |
| 183 | Q | F | Decrease | 1 |
| 183 | Q | W | Decrease | 2 |
| 183 | Q | Y | Increase | 1 |
| 183 | Q | G | Decrease | 6 |
| 183 | Q | A | Decrease | 2 |
| 183 | Q | P | Decrease | 1 |
| 183 | Q | S | Decrease | 6 |
| 183 | Q | T | Decrease | 6 |
| 183 | Q | C | Increase | 3 |
| 183 | Q | H | Decrease | 5 |
| 183 | Q | R | Decrease | 3 |
| 183 | Q | K | Increase | 2 |
| 183 | Q | E | Increase | 3 |
| 183 | Q | N | Decrease | 3 |
| 183 | Q | D | Decrease | 3 |

**WT**: Amino acid in Wild-Type Protein
**NEW**: New Amino acid after Mutation
**RI:** Reliability Index

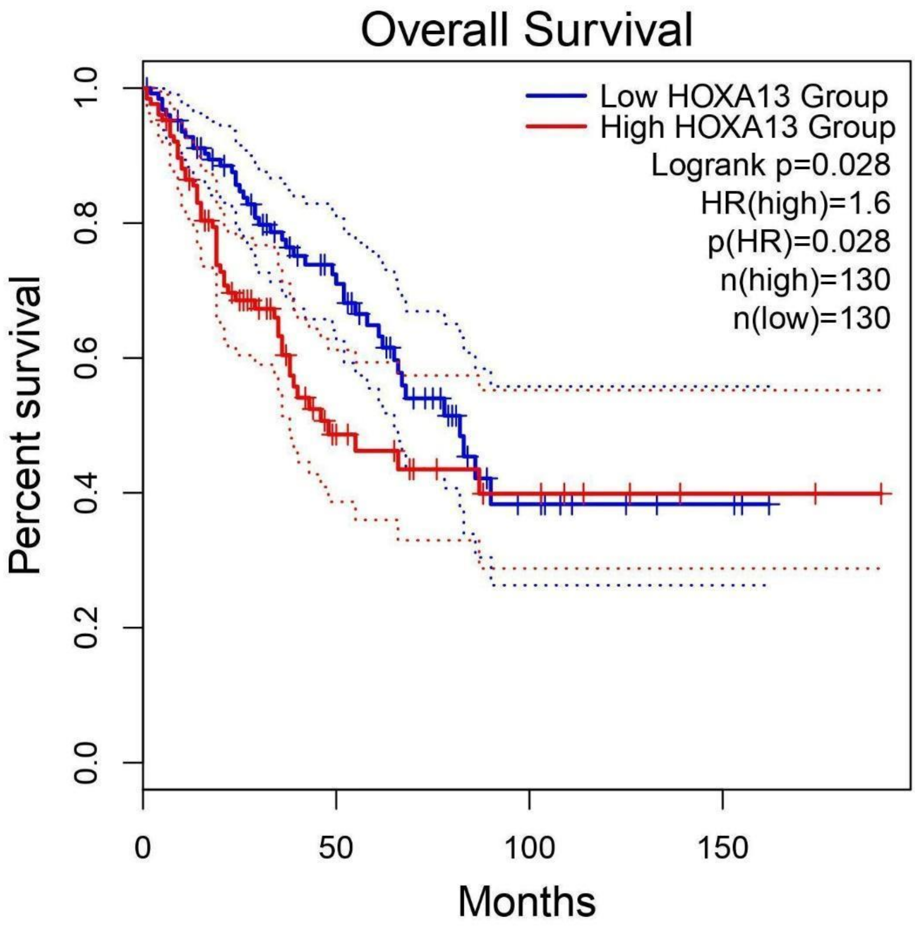
Similarly, the mutation in 185 amino acid positions on HoxD13 in sarcoma tissues was analyzed. In cBioportal the mutation of Alanine (A) to serine (S) was known.

Table 2: Mutations at position 185

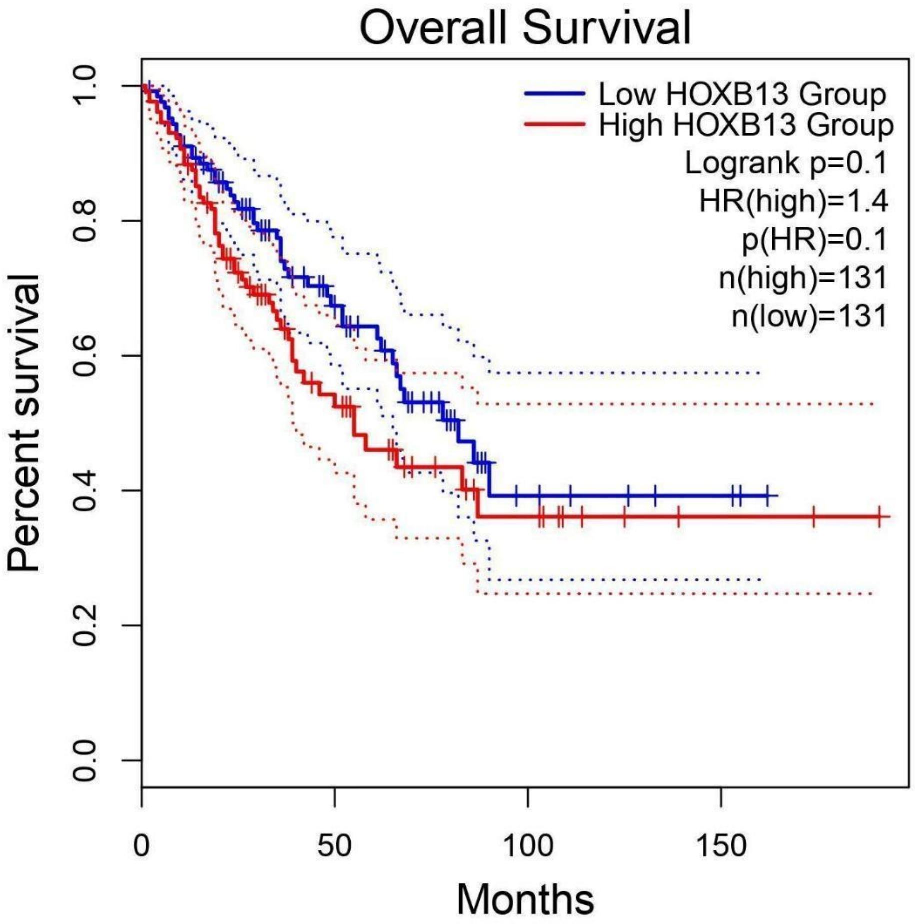
| Position | WT | NEW | STABILITY | RI |
|---|---|---|---|---|
| 185 | A | S | Decrease | 9 |

**WT**: Amino acid in Wild-Type Protein
**NEW**: New Amino acid after Mutation
**RI:** Reliability Index

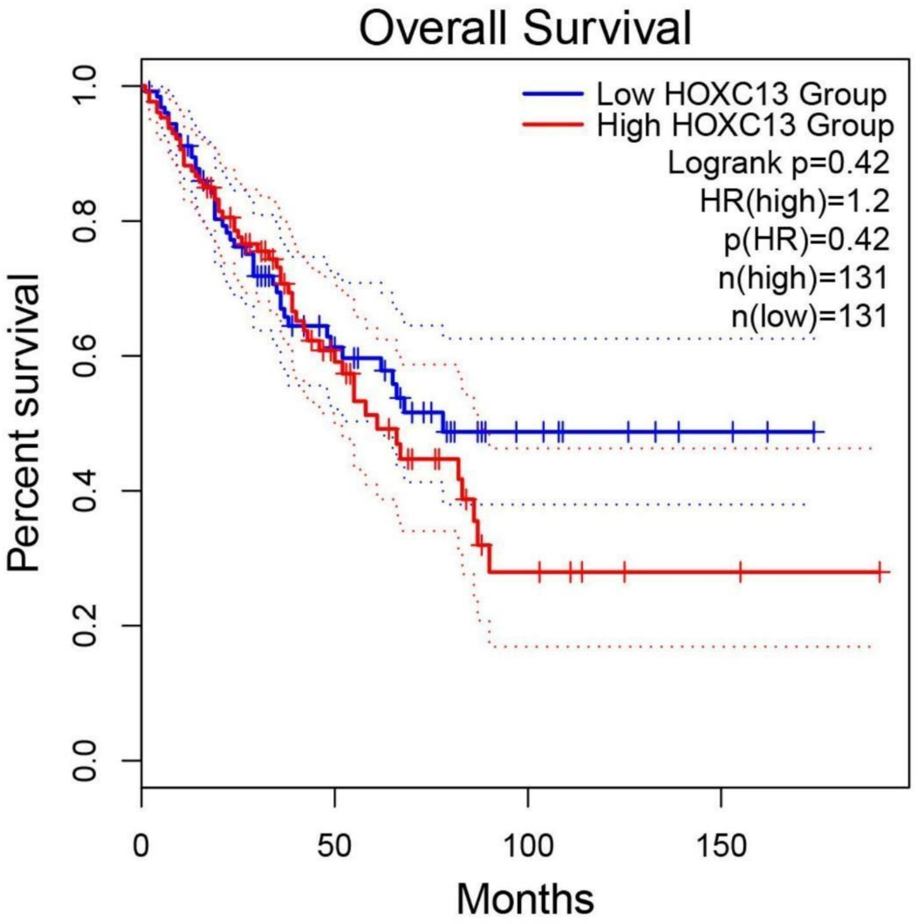
Further, the analysis of mRNA expression of HoxA13, Hoxb13, HoxC13, and HoxD13 concerning sarcoma and normal tissues was done to get the prognosis of a sarcoma patient. It is observed the shorter OS time with worse prognosis in higher expression levels of HoxC13 and HoxB13 compared with lower expression levels of HoxD13 and HoxA13. (Fig. 7, the p>00.5 are shorter OS time and worse prognosis)


Overall Survival
Low HOXA13 Group
High HOXA13 Group
Logrank p=0.028
HR(high)=1.6
p(HR)=0.028
n(high)=130
n(low)=130
Percent survival
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Months

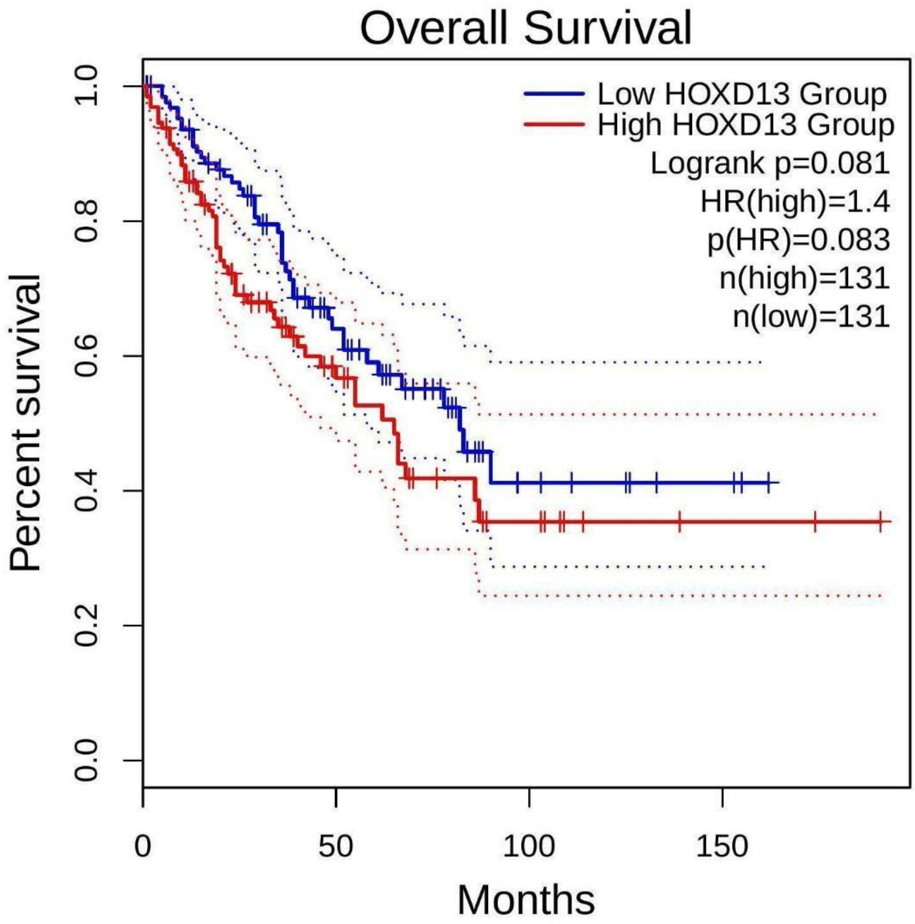
A


Overall Survival
Low HOXB13 Group
High HOXB13 Group
Logrank p=0.1
HR(high)=1.4
p(HR)=0.1
n(high)=131
n(low)=131
Percent survival
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Months

B


Overall Survival
Low HOXC13 Group
High HOXC13 Group
Logrank p=0.42
HR(high)=1.2
p(HR)=0.42
n(high)=131
n(low)=131
Percent survival
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Months

C



D

Figure 7: A: OS time between Hox A13 paralog gene expression and sarcoma cancer with its worse prognosis if p>0.05, B: OS time between Hox B13 paralog gene expression and sarcoma cancer with its worse prognosis if p>0.05, C: OS time between Hox C13 paralog gene expression and sarcoma cancer with its worse prognosis if p>0.05, D: OS time between Hox D13 paralog gene expression and sarcoma cancer with its worse prognosis if p>0.05.
The red line – cases with high expression, the blue line – cases with low expression, HR – hazard ratio.

To analyze the functional expression of the HoxA13, Hoxb13, HoxC13, and HoxD13, some of the closely altered and unaltered genes mRNA expression with HoxA13, Hoxb13, HoxC13, and HoxD13 expression were analyzed in the STRING database. Surprisingly, the immune response functions and related pathways were observed in Fig. 8.

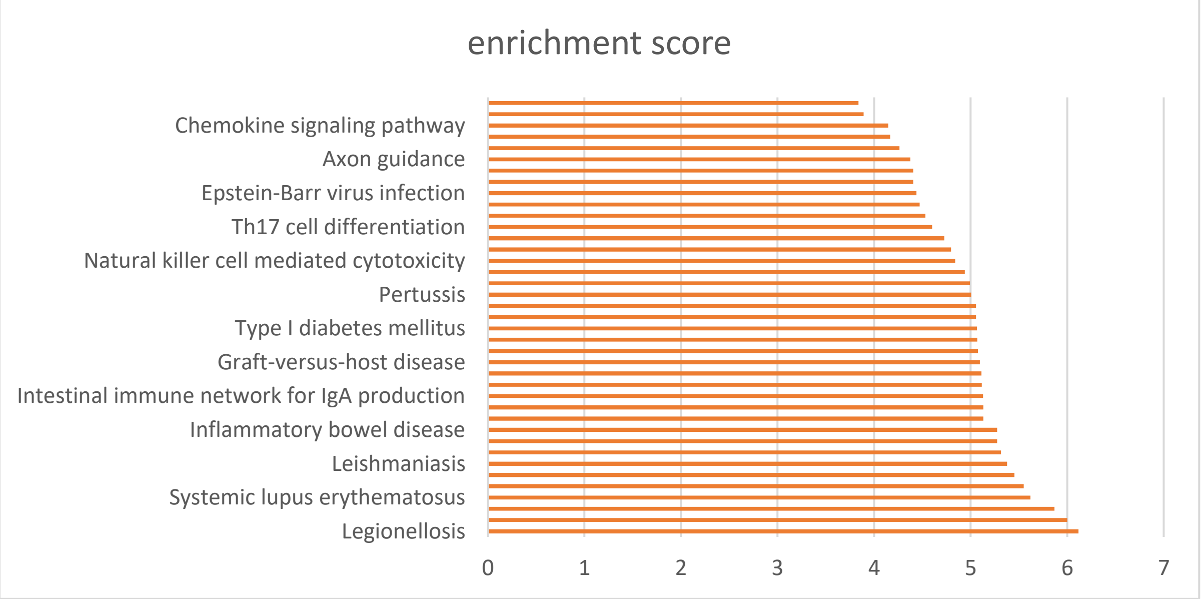


Figure: 8 – GO enrichment analysis of top 500 altered and unaltered similar genes from Hox paralog gene.

Based on the results of the GO enrichment, the mRNA expression against immune cells was analyzed with help of the TIMER tool. The Hox13 paralog gene and similar genes were obtained from the cBioPortal shown in Fig. 9.

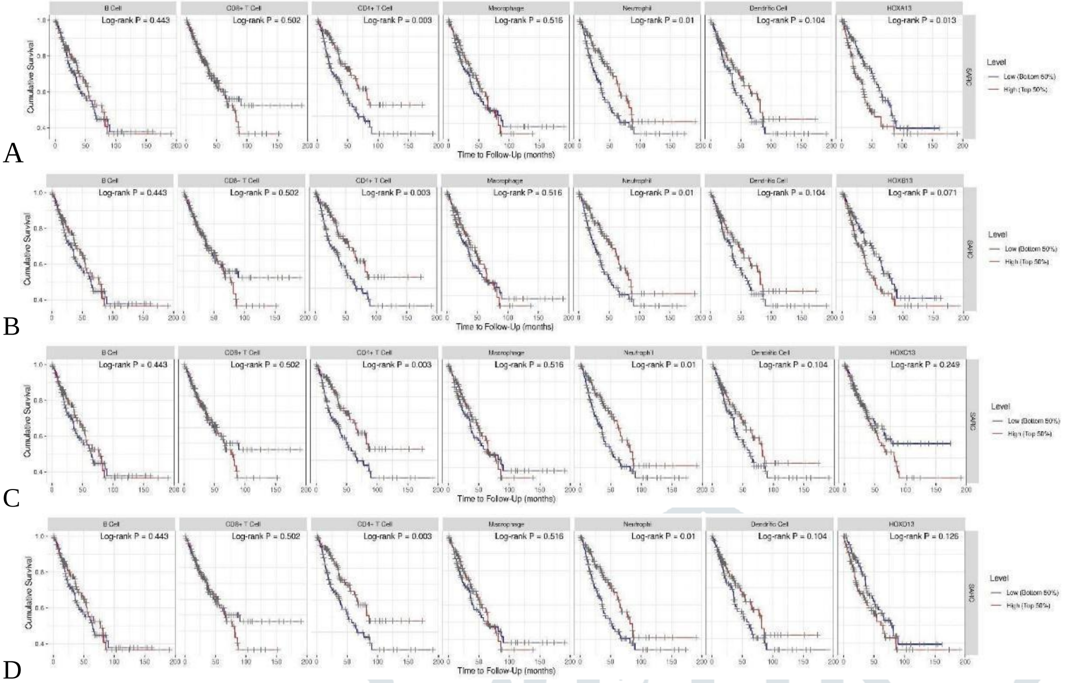
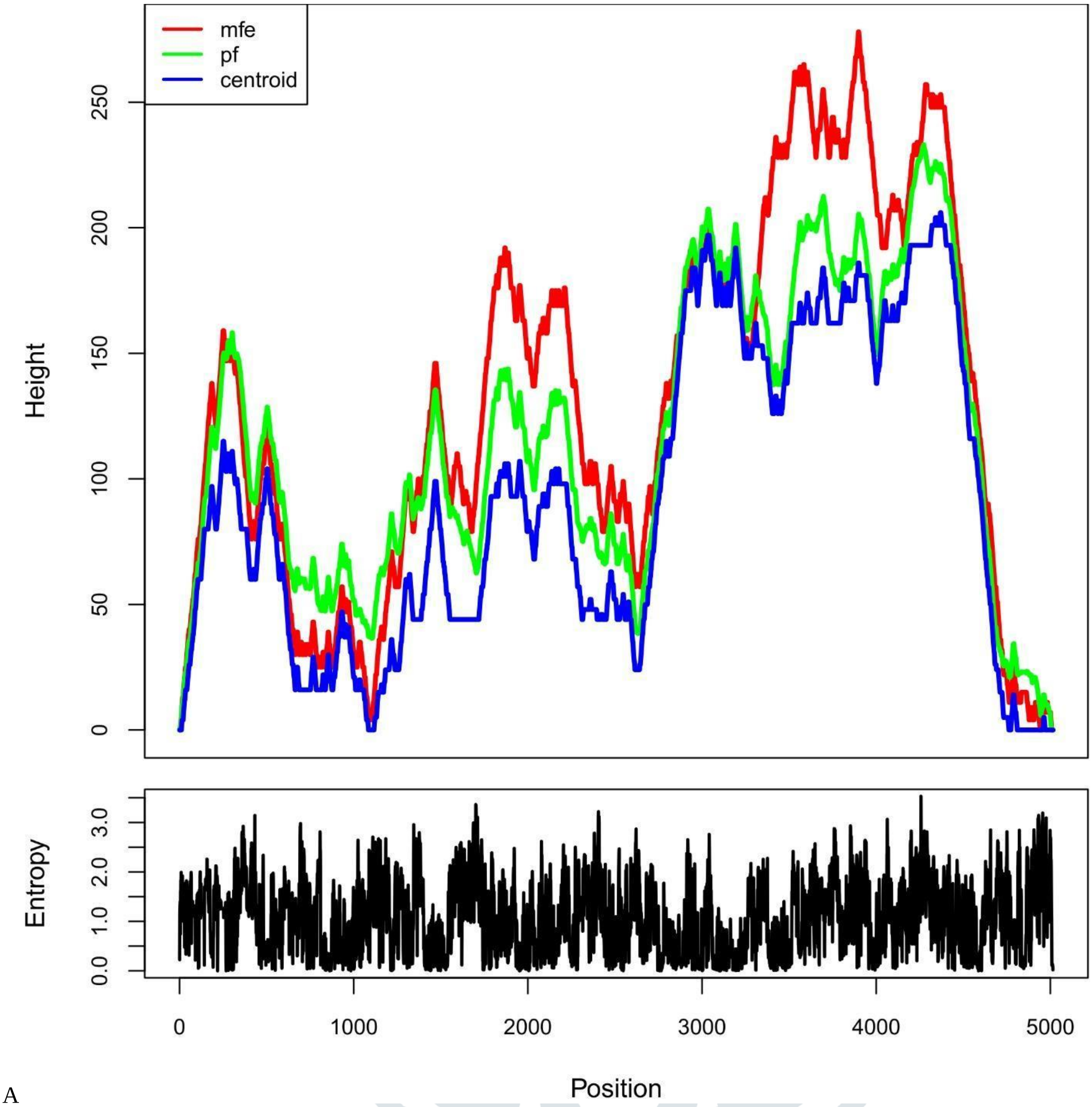


Figure: 9 – A – The mRNA expression in case of HoxA13 (p=0.013) and other immune cells such as B cell (p=0.443), CD8+ (p=0.502), CD4+ (p=0.003), Macrophages (p=0.516), Neutrophils (p=0.01), and Dendritic cells (p=0.104).
B – The mRNA expression in case of HoxB13 (p=0.07) and other immune cells such as B cell (p=0.443), CD8+ (p=0.502), CD4+ (p=0.003), Macrophages (p=0.516), Neutrophils (p=0.01), and Dendritic cells (p=0.104).
C – The mRNA expression in case of HoxC13 (p=0.249) and other immune cells such as B cell (p=0.443), CD8+ (p=0.502), CD4+ (p=0.003), Macrophages (p=0.516), Neutrophils (p=0.01), and Dendritic cells (p=0.104).
D – The mRNA expression in case of HoxD13 (p=0.126) and other immune cells such as B cell (p=0.443), CD8+ (p=0.502), CD4+ (p=0.003), Macrophages (p=0.516), Neutrophils (p=0.01), and Dendritic cells (p=0.104).

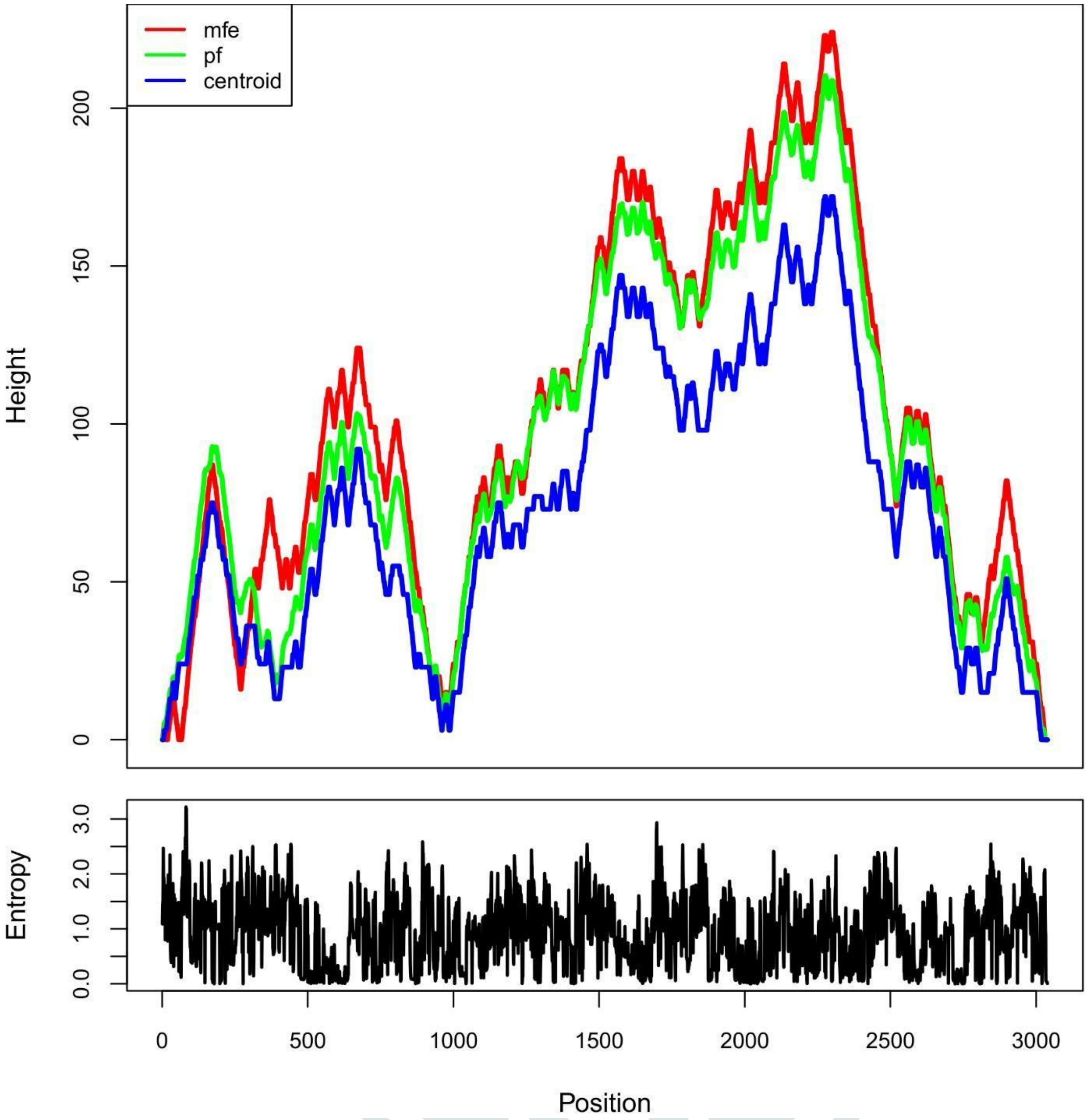
## IV. DISCUSSION

The HOX genes are the transcription that regulates the expression from the embryonic stage. The development of the anterior-posterior axis of the embryo is based on the expression of the Hox genes. These are the least repeating regions in the genome which assist its highly complex nature (Consortium, 2001) (Simona Santini, 2003) (Jongmin Nam, 2005). The Hox gene does not bind to any with high affinity. There binding of the Hox to DNA is done such that the specificity of binding depends on the multiple binding sites that are spatially separated but are in the neighborhood. The two highest affinity binding sites share a 5’-ATCATTA-3’ consensus sequence. The low-affinity binding sites are spatially separated to direct the co-factor binding (Bony De Kumar, 2021) (Gruschus, 1997). The binding of the DNA is further directed by the amino acid sequence in the Homeodomain. The N-terminal arm makes the specific binding sites with the hexapeptides. Much Hox protein has conserved hexapeptide motif consisting of amino acids – tryptophan (W) and methionine (M) (Piper, 1999) (Gruschus, 1997).

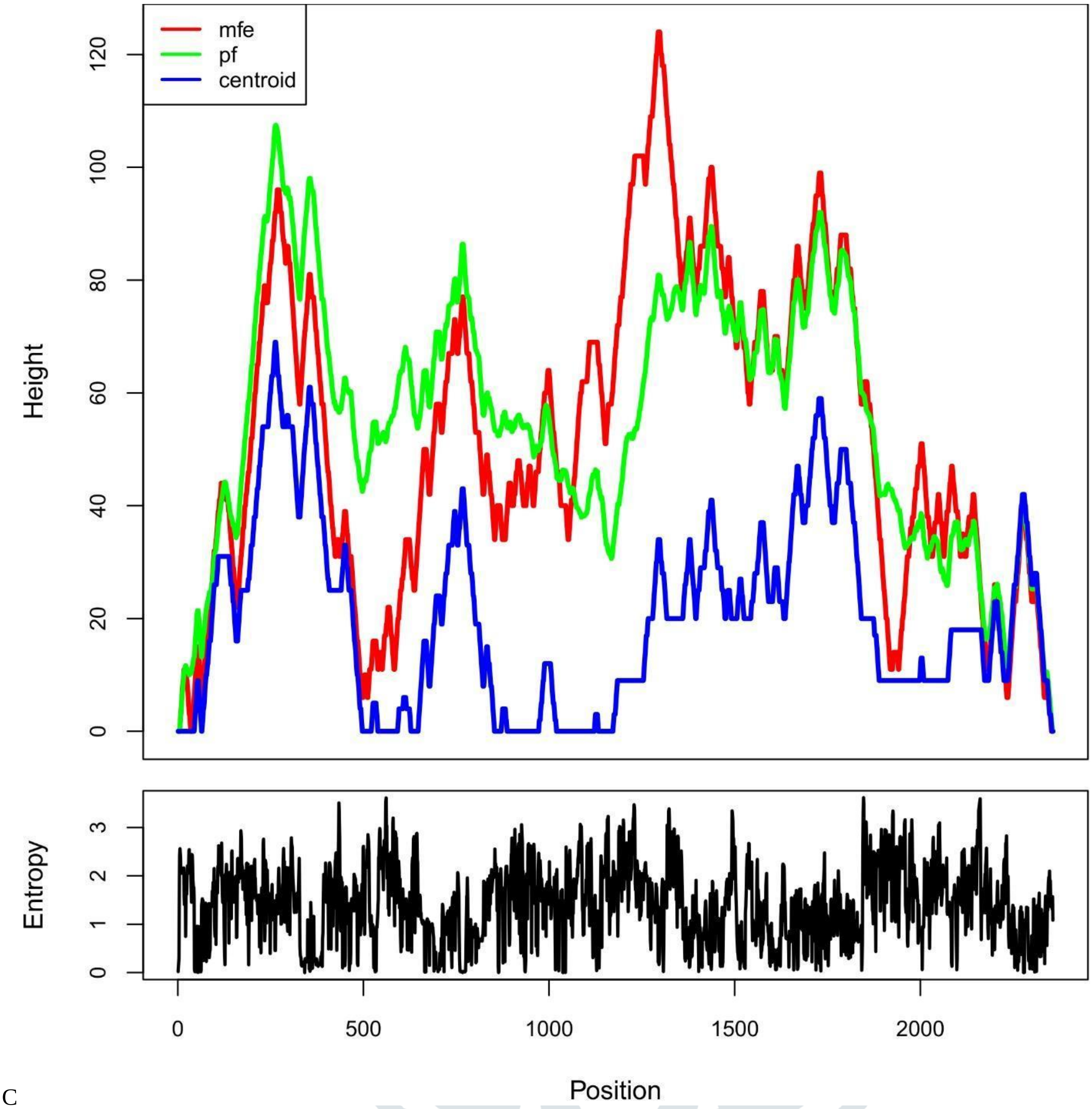
From the mRNAs structure prediction, we observed that HoxA13 and HoxB13 have high free energy and with a large number of base pairs concentrated at the 5’end region. For the HoxC13 mRNA, the plot represents the height distribution of MFE high in the central and 5’ end and for HoxD13, the height is a peek at the 3’end. However, all four paralog genes have high negative free energy shown in Fig. 10.


mfe
pf
centroid
Height
0
50
100
150
200
250
Entropy
0.0
1.0
2.0
3.0
0
1000
2000
3000
4000
5000
Position

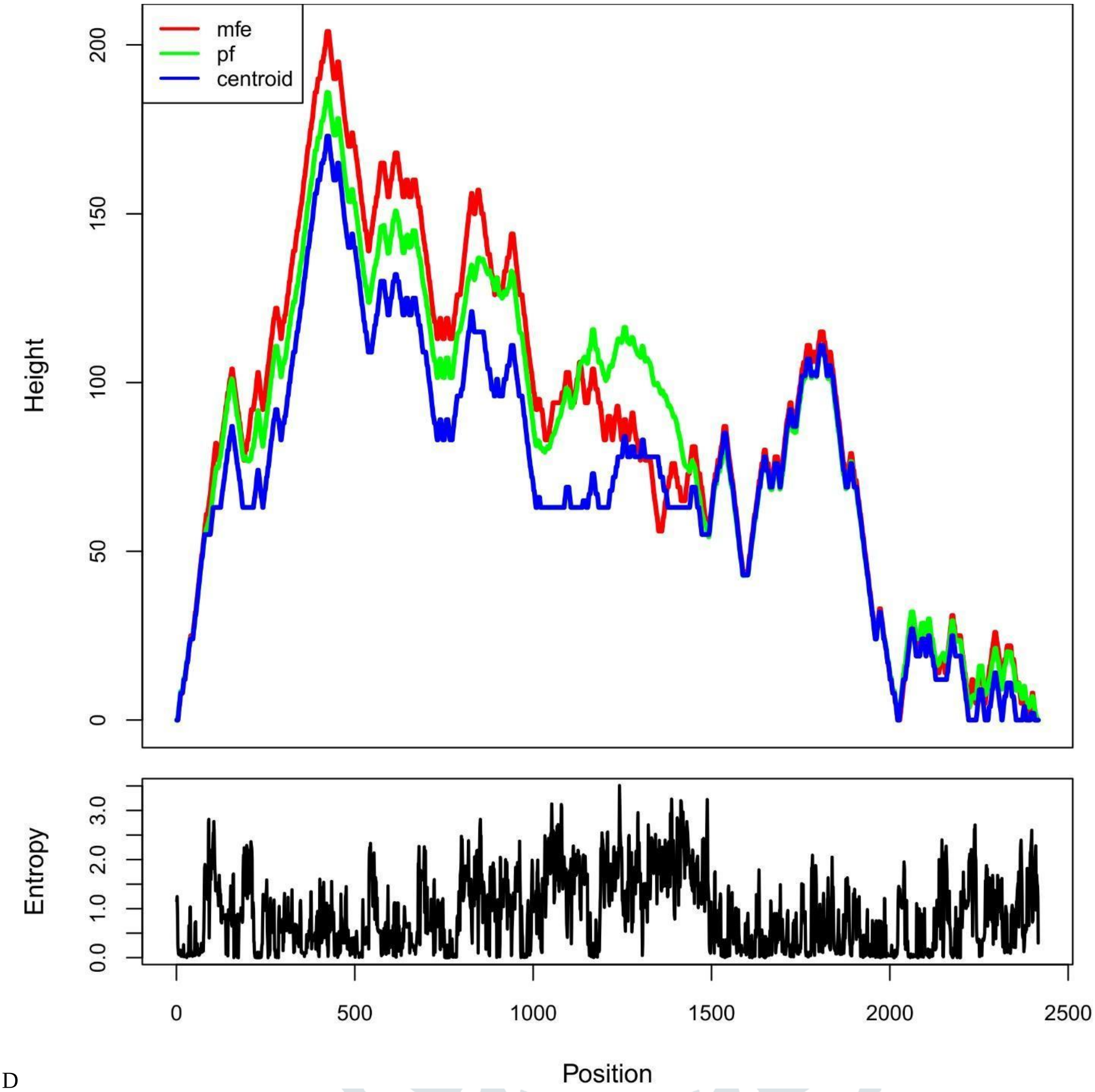
A


mfe
pf
centroid
Height
0
50
100
150
200
Entropy
0.0
1.0
2.0
3.0
0
500
1000
1500
2000
2500
3000
Position

B


mfe
pf
centroid
Height
0
20
40
60
80
100
120
Entropy
0
1
2
3
0
500
1000
1500
2000
Position

C

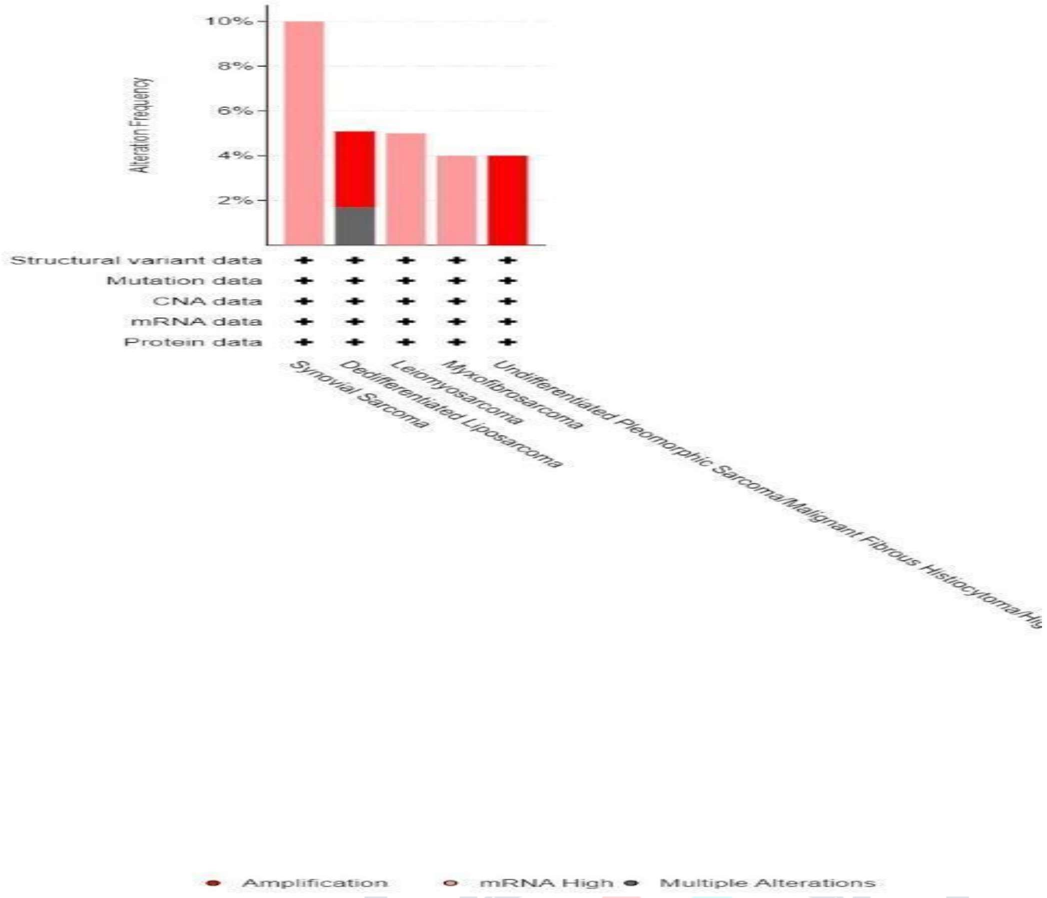


D

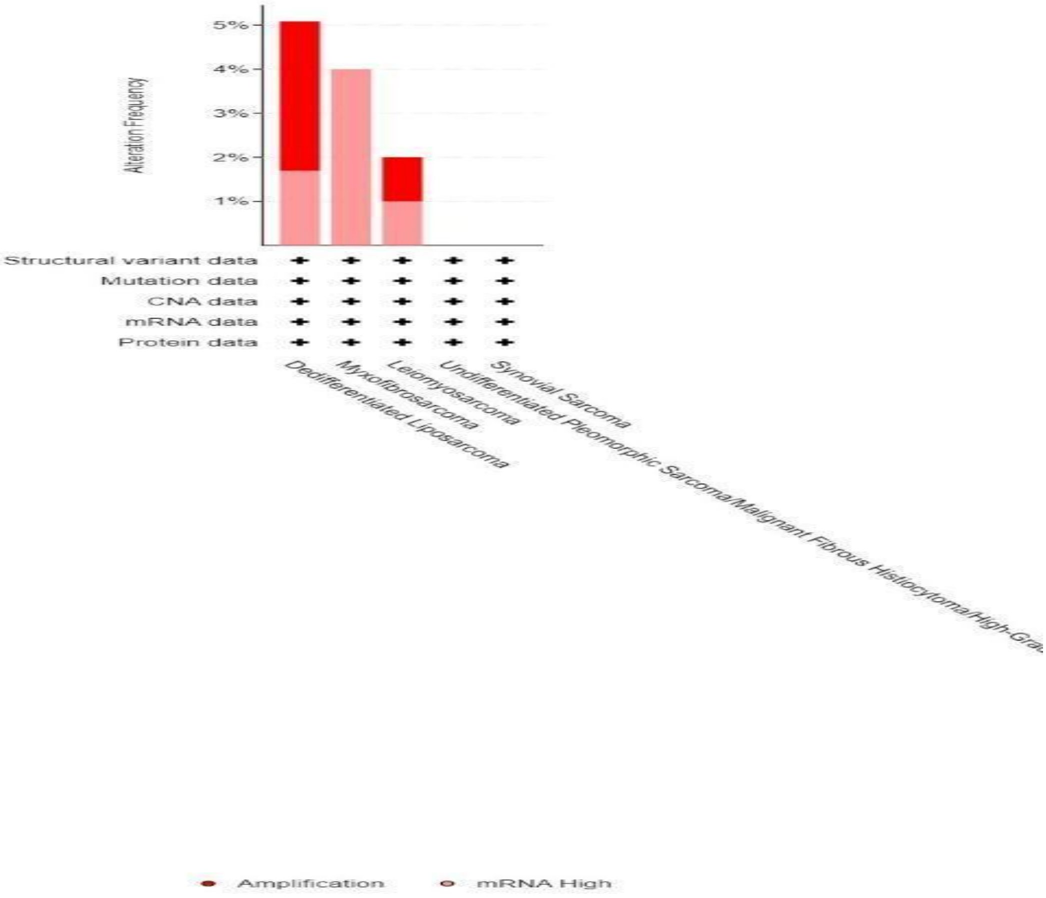
Figure 10: A: The graph of MFE, PF, and centroid of HoxA13, B: The graph of MFE, PF, and centroid of HoxB13, C: The graph of MFE, PF, and centroid of HoxC13, D: The graph of MFE, PF, and centroid of HoxD13.

The prediction of the secondary structure and tertiary structure conveyed a two-helix-coil-helix structure. The helix I and helix II lie parallel to each other whereas, the helix III also called the recognition helix lies across them. The major groove of DNA binds with the recognition helix. The main conservation in the Hox gene occurs in the recognition helix. In the helix III the conservation of 5 amino acids (Trp, Phe, Asn, Arg, Lys) is seen. The N-terminal arm makes contact with a minor groove and hence, the helix III with the n-terminal arm confers the DNA binding specificity in the Homeobox gene (Sharmila Banerjee-Basu, 2001).

The amino acid sequence alignment and conserved domain showed a high level of similarity between all the Hox13 paralog genes. Despite the high conservation in the paralog Hox 13 gene, the difference in the functional expression is seen. The unique characteristics in each of the Hox13 gene assists in the differential expression which is in the case of sarcoma tumor subtypes as shown in Fig. 11. The mutations analyzed showed both an increase and decrease in stability. However, the mutations' ability to bring change in the expression was very less when protein was expressed as a whole.


Alteration Frequency
10%
8%
6%
4%
2%
Structural variant data
Mutation data
CNA data
mRNA data
Protein data
Synovial Sarcoma
Dedifferentiated Liposarcoma
Leiomyosarcoma
Myxofibrosarcoma
Undifferentiated Pleomorphic Sarcoma/Malignant Fibrous Histiocytoma/Hig
● Amplification
○ mRNA High
● Multiple Alterations

A


Alteration Frequency
5%
4%
3%
2%
1%
Structural variant data
Mutation data
CNA data
mRNA data
Protein data
Dedifferentiated Liposarcoma
Myxofibrosarcoma
Leiomyosarcoma
Undifferentiated Pleomorphic Sarcoma/Malignant Fibrous Histiocytoma/High-Gra
Synovial Sarcoma
● Amplification
○ mRNA High

B

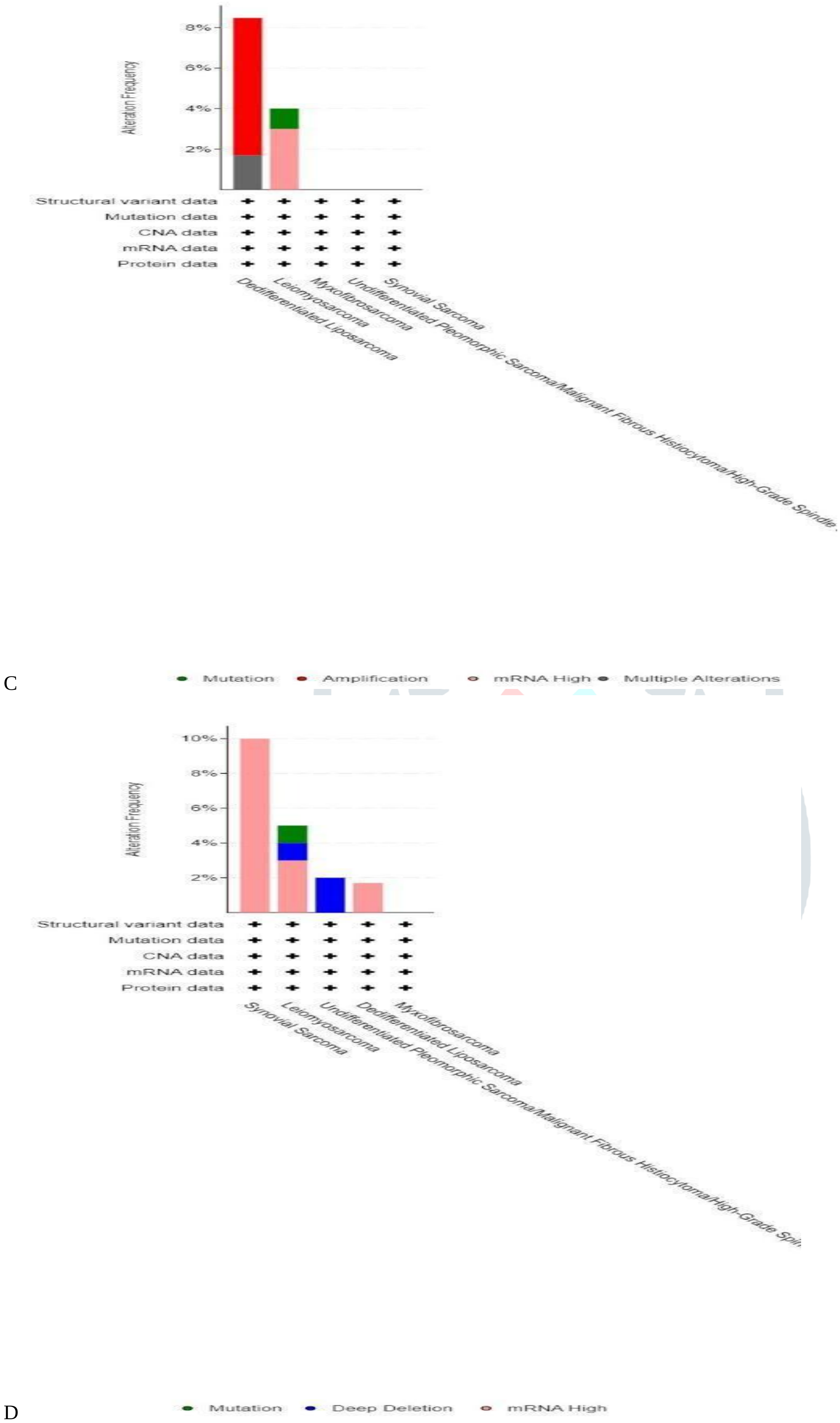


Figure: 11 –

A – HoxA13 expression in different tumor subtypes. Multiple Alteration, Amplification, and high mRNA expression are a predominant ways in sarcoma.

B – HoxB13 expression in different tumor subtypes. Amplification and high mRNA expression are predominant in sarcoma.

C – HoxC13 expression in different tumor subtypes. The Amplification, high mRNA expression, and Mutations are dominant in sarcoma.

D – HoxD13 expression in tumor subtypes. The high mRNA expression, Mutations, and Deep deletion are predominant in sarcoma.

The sarcoma was analyzed for the alterations in the cBioPortal showed a change in the DNA copy number. But upon closer look at tumor subtypes showed varying causes in the Hox13 paralog gene. When we further analyzed the prognosis value of the sarcoma concerning HoxA13, HoxB13, HoxC13, and HoxD13 genes. The worse prognosis value was HoxC13 followed by HoxB13, HoxD13, and HoxA13.

Sarcomas are the most diverse group of malignant tumors, with only mesenchymal cells common characteristics between them as they differentiate from them. Mesenchymal cells isolated from the umbilical cord showed expression of various Hox genes. HoxC 13 and HoxD 13 expression were observed in bone marrow-derived mesenchymal cells. Moreover, the mesenchymal cells themselves are heterogenous and hence, the origin of sarcoma cells is important. For simple and stable karyotypes, the principle is to modify transcriptome and interfere with regulation like in Liposarcoma, Ewing sarcoma, and synovial sarcoma. The overexpression of HoxD13 was observed in the Ewing sarcoma. Both BMI-1 and EZH2 are always over-regulated in sarcoma. The fusion of both leads to dysregulation of posterior HoxD i.e.; HoxD10, HoxD11, and HoxD13 12 (von Heyking, 2017) (Laurie K Svoboda, 2014). In dedifferentiated liposarcoma, the histochemical studies show overexpression of HoxC13. Liposarcoma is directly related to the blocking of adipocyte differentiation and Hox genes are translation regulators in adipogenesis. Therefore, the HoxC gene is highly considered to be an interfering gene in liposarcoma (Cantile, 2013).

GO enrichment was performed to analyze the Hox 13 gene's function in sarcoma. The GO enrichment resulted in the biological pathways of the immune response. When profiling of the immune-related genes in soft tissue sarcoma was done, HoxB13, HoxC13, and HoxD13 showed significant p-value and HoxA13 showed moderate p-value. It has been observed that complex sarcomas work with immune-related genes. In the case of nerve sheath fiber tumors, the Wnt pathway plays a critical role in maintaining the tumor (Pridgeon, 2017). Immune-related genes have been confirmed role in tumorigenesis.

Nevertheless, the lack of systematic information on mesenchymal cells, sarcoma, and the lack of systematic analysis of immune-related genes and their clinical significance needs further study. Also, the analysis of the Hox genes in the immune-related gene in sarcoma related to carcinogenesis merits further evaluation.

## V. CONCLUSION

The Hox genes are less repetitive, most complex sequences that code for specific binding sites in the Hox protein. From the structure prediction, we observed all the 3 α-helixes help in forming a high-affinity binding site which is highly conserved as can be seen in the results of the sequence alignment and conserved domain search. The results of mRNA structure fold prediction revealed secondary structure with high negative minimum free energy in all four paralog genes affirming the stability and low repetition region for mRNA itself. The Ramachandran plot results show the conservation of some particular amino acids in the helix I and helix III. Despite the high degree of conversion, the expression of Hox 13 paralogous have some uniqueness to it. The expression of the Hox 13 paralogous genes in the sarcoma varied significantly. High mRNA expression, amplification, and mutations are predominantly visible in the sarcoma subtypes. Further studying of mutations studied, their effect on the development of sarcoma was of less significance. The over-survival analysis in GEPIA2 was done to get prognosis values which showed, that HoxC13 and HoxB13 showed worse prognosis values compared to the HoxD13 and HoxA13.

GO analysis of the Hox13 and closely related genes showed pathways related to immune cell signaling, differentiation, and pathogen attacking systems from the STRING database. From the list of closely related genes, the top 500 altered and unaltered genes were selected, and with Hox13 genes, mRNA expression was analyzed with immune cells in TIMER showed similar results to that of OS analysis. The p-value of HOXC13 against all the tumor cells was highest in all Hox13 genes. In conclusion, the Hox genes with sarcoma need further study.

## REFERENCES

[1] Banerjee-Basu, S., & Baxevanis, A. D. (2001). Molecular evolution of the homeodomain family of transcription factors. Nucleic acids research, 29(15), 3258–3269. https://doi.org/10.1093/nar/29.15.3258

[2] Bony De Kumar, Diane C. Darland (2021). The Hox protein conundrum: The “specifics” of DNA binding for Hox proteins and their partners. Developmental Biology, 477, 284-292. https://doi.org/10.1016/j.ydbio.2021.06.002

[3] Cancer Genome Atlas Research Network. Electronic address: elizabeth.demicco@sinaihealthsystem.ca, & Cancer Genome Atlas Research Network (2017). Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell, 171(4), 950–965.e28. https://doi.org/10.1016/j.cell.2017.10.014

[4] Cantile, M., Galletta, F., Franco, R., Aquino, G., Scognamiglio, G., Marra, L. ... De Chiara, A. (2013). Hyperexpression of HOXC13, located in the 12q13 chromosomal region, in well-differentiated and dedifferentiated human liposarcomas. Oncology Reports, 30, 2579-2586. https://doi.org/10.3892/or.2013.2760

[5] Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic acids research, 33(Web Server issue), W306–W310. https://doi.org/10.1093/nar/gki375

[6] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery, 2(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

[7] Daniel W A Buchan, David T Jones, The PSIPRED Protein Analysis Workbench: 20 years on, Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W402–W407, https://doi.org/10.1093/nar/gkz297

[8] Daniel W A Buchan, David T Jones, The PSIPRED Protein Analysis Workbench: 20 years on, Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W402–W407, https://doi.org/10.1093/nar/gkz297

[9] Fábio Madeira, Young mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, Rodrigo Lopez, The EMBL-EBI search and sequence analysis tools APIs in 2019, Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W636–W641, https://doi.org/10.1093/nar/gkz268

[10] Forlani, S., Lawson, K. A., & Deschamps, J. (2003). Acquisition of Hox codes during gastrulation and axial elongation in the mouse embryo. Development (Cambridge, England), 130(16), 3807–3819. https://doi.org/10.1242/dev.00573

[11] Gamboa, A.C., Gronchi, A. and Cardona, K. (2020), Soft-tissue sarcoma in adults: An update on the current state of histotype-

specific management in an era of personalized medicine. CA A Cancer J Clin, 70: 200-229. https://doi.org/10.3322/caac.21605

[12] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., & Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, *6*(269), pl1. https://doi.org/10.1126/scisignal.2004088

[13] Garth R. Brown, Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D. Pruitt, Donna R. Maglott, Terence D. Murphy, Gene: a gene-centered information resource at NCBI, Nucleic Acids Research, Volume 43, Issue D1, 28 January 2015, Pages D36–D42, https://doi.org/10.1093/nar/gku1055

[14] Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., & Hofacker, I. L. (2008). The Vienna RNA websuite. Nucleic acids research, 36(Web Server issue), W70–W74. https://doi.org/10.1093/nar/gkn188

[15] Gruschus, J. M., Tsao, D. H., Wang, L. H., Nirenberg, M., & Ferretti, J. A. (1997). Interactions of the vnd/NK-2 homeodomain with DNA by nuclear magnetic resonance spectroscopy: basis of binding specificity. Biochemistry, 36(18), 5372–5380. https://doi.org/10.1021/bi9620060

[16] Holland, P.W., Booth, H.A.F. & Bruford, E.A. Classification and nomenclature of all human homeobox genes. BMC Biol 5, 47 (2007). https://doi.org/10.1186/1741-7007-5-47

[17] Hu, C., Chen, B., Huang, Z. et al. Comprehensive profiling of immune-related genes in soft tissue sarcoma patients. J Transl Med 18, 337 (2020). https://doi.org/10.1186/s12967-020-02512-8

[18] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001). https://doi.org/10.1038/35057062

[19] Jongmin Nam, Masatoshi Nei, Evolutionary Change of the Numbers of Homeobox Genes in Bilateral Animals, Molecular Biology and Evolution, Volume 22, Issue 12, December 2005, Pages 2386– 2394, https://doi.org/10.1093/molbev/msi229

[20] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[21] Kornak, U., & Mundlos, S. (2003). Genetic disorders of the skeleton: a developmental approach. American journal of human genetics, 73(3), 447–474. https://doi.org/10.1086/377110

[22] Kristina von Heyking, K., Roth, L., Ertl, M., Schmidt, O., Calzada-Wack, J., Neff, F., Lawlor, E. R., Burdach, S., & Richter, G. H. (2016). The posterior HOXD locus: Its contribution to phenotype and malignancy of Ewing sarcoma. Oncotarget, 7(27), 41767–41780. https://doi.org/10.18632/oncotarget.9702

[23] Lappin, T. R., Grier, D. G., Thompson, A., & Halliday, H. L. (2006). HOX genes: seductive science, mysterious mechanisms. The Ulster medical journal, 75(1), 23–31.

[24] Laskowski, R.A., Jabłońska, J., Pravda, L., Vařeková, R.S. and Thornton, J.M. (2018), PDBsum: Structural summaries of PDB entries. Protein Science, 27: 129-134. https://doi.org/10.1002/pro.3289

[25] Lewis E. B. (1978). A gene complex controlling segmentation in Drosophila. Nature, 276(5688), 565–570. https://doi.org/10.1038/276565a0

[26] Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J. C., Rodig, S., Signoretti, S., Liu, J. S., & Liu, X. S. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome biology*, *17*(1), 174. https://doi.org/10.1186/s13059-016-1028-7

[27] Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., Li, B., & Liu, X. S. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer research, 77(21), e108–e110. https://doi.org/10.1158/0008-5472.CAN-17-0307

[28] Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Apr:gkac240. DOI: 10.1093/nar/gkac240. PMID: 35412617; PMCID: PMC9252731.

[29] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, Sameer Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, *Nucleic Acids Research*, Volume 50, Issue D1, 7 January 2022, Pages D439–D444, https://doi.org/10.1093/nar/gkab1061

[30] Paolo Di Tommaso, Sebastien Moretti, Ioannis Xenarios, Miquel Orobitg, Alberto Montanyola, Jia-Ming Chang, Jean-François Taly, Cedric Notredame, T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension, Nucleic Acids Research, Volume 39, Issue suppl_2, 1 July 2011, Pages W13–W17, https://doi.org/10.1093/nar/gkr245

[31] Piper, D. E., Batchelor, A. H., Chang, C. P., Cleary, M. L., & Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. Cell, 96(4), 587–597. https://doi.org/10.1016/s0092-8674(00)80662-5

[32] Pridgeon, M.G., Grohar, P.J., Steensma, M.R. et al. Wnt Signaling in Ewing Sarcoma, Osteosarcoma, and Malignant Peripheral Nerve Sheath Tumors. Curr Osteoporos Rep 15, 239–246 (2017). https://doi.org/10.1007/s11914-017-0377-9

[33] Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., & Bier, E. (2001). A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. Genome research, 11(6), 1114–1125. https://doi.org/10.1101/gr.169101

[34] Santini, S., Boore, J. L., & Meyer, A. (2003). Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. Genome research, 13(6A), 1111–1122. https://doi.org/10.1101/gr.700503

[35] Sarver, A., Sarver, A., Thayanithy, V. et al. Identification, by systematic RNA sequencing, of novel candidate biomarkers and therapeutic targets in human soft tissue tumors. Lab Invest 95, 1077–1088 (2015). https://doi.org/10.1038/labinvest.2015.80

[36] Seema Bhatlekar, Jeremy Z. Fields, Bruce M. Boman. Role of HOX Genes in Stem Cell Differentiation and Cancer. Stem Cells International, 2018. https://doi.org/10.1155/2018/3569493

[37] Svoboda, L. K., Harris, A., Bailey, N. J., Schwentner, R., Tomazou, E., von Levetzow, C., Magnuson, B., Ljungman, M., Kovar, H., & Lawlor, E. R. (2014). Overexpression of HOX genes is prevalent in Ewing sarcoma and is associated with altered epigenetic regulation of developmental transcription programs. Epigenetics, 9(12), 1613–1625. https://doi.org/10.4161/15592294.2014.988048

**[38]** Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, *49*(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074

**[39]** Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., & von Mering, C. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic acids research, 45(D1), D362–D368. https://doi.org/10.1093/nar/gkw937

**[40]** Szymon Chojnacki, Andrew Cowley, Joon Lee, Anna Foix, Rodrigo Lopez, Programmatic access to bioinformatics tools from EMBL-EBI update: 2017, Nucleic Acids Research, Volume 45, Issue W1, 3 July 2017, Pages W550–W553, https://doi.org/10.1093/nar/gkx273

**[41]** Tang, Z., Kang, B., Li, C., Chen, T., & Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic acids research, 47(W1), W556–W560. https://doi.org/10.1093/nar/gkz430

**[42]** Webb, B. and Sali, A. 2016. Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinform. 54: 5.6.1- 5.6.37. https://doi.org/10.1002/cpbi.3

**[43]** Yang, M., Derbyshire, M. K., Yamashita, R. A., & Marchler-Bauer, A. (2020). NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. Current protocols in bioinformatics, 69(1), e90. https://doi.org/10.1002/cpbi.90