

**Comparative *in-silico* analysis of the interactions
between Ultrabithorax and its putative cofactors
across different insect groups.**



***Thesis Submitted towards the Partial Fulfillment of
BS-MS dual degree programme***

By

Vaibhav Prakash Wagh

To

The Department of Biology,

Indian Institute of Science Education and Research, Pune

Under the Supervision of

Prof. L. S. Shashidhara

Indian Institute of Science Education and Research, Pune

For the academic year 2017-18

Certificate

This is to certify that this dissertation entitled "**Comparative in-silico analysis of the interactions between Ultrabithorax and its putative cofactors across different insect groups.**" towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by **Valbhav Prakash Wagh**, at IISER Pune under the supervision of **Prof. L. S. Shashidhara**, Department of Biology, IISER Pune during the academic year 2017-18.



Prof. L.S.Shashidhara

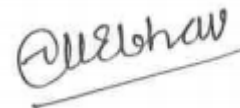
Department of Biology,

IISER Pune

Date:

Declaration

*I hereby declare that the matter embodied in the report entitled “**Comparative in-silico analysis of the interactions between Ubx and its putative cofactors across different insect groups**” are the results of the work carried out by me at the Department of Biology at IISER Pune, under the supervision of **Prof. L.S. Shashidhara**, Department of Biology, IISER Pune and the same has not been submitted elsewhere for any other degree.*



Vaibhav Prakash Wagh

Student, IISER Pune,

Date:

Abstract

In dipterans, the Hox protein Ultrabithorax (Ubx) specifies haltere fate in the third thoracic segment by repressing wing fate. While orthologues of Ubx are expressed in other insects such as *Apis* (Hymenoptera), *Bombyx* (Lepidoptera) and *Tribolium* (Coleoptera), it specifies haltere only in flies. The underlying cause behind this is differential regulation of target genes. Previous studies have shown that *in-vitro* all Hox proteins bind to similar 'AT' rich DNA sequence with similar affinity but *in-vivo* are very specific towards their target genes. Interaction with co-transcription factors can achieve this in-vivo specificity. Ubx is known to be interacting with other partner proteins to regulate the expression of its target genes differentially, and the interaction can be different in different insect species. There are various ways by which Ubx can regulate its target genes one of which is by direct physical interaction with the co-factor proteins.

The current study aims at the direct physical interaction between Ubx and its cofactors in species of *Drosophila* in comparison with other species like *Tribolium castaneum*, *Apis mellifera* and *Bombyx mori*. This study includes prediction of interactions between Ubx and cofactors and identification of structural constraints on their binding sites using homology modeling. Later Based on the separation of binding sites of Ubx and a putative co-transcription factor in various species, we can comment on the evolution of Ubx-cofactor interactions that mediate gene regulation across different insect species. This can help in understanding the evolution of Ubx- cofactor interaction.

Table of contents-

Abstract.....	4
Table of contents-	5
List of figures-.....	6
List of Tables-	7
Acknowledgments.....	8
1 Introduction	9
1.1 Development and role of Homeotic genes-.....	10
1.2 Ultrabithorax (Ubx) –	12
1.3 Morphological differences between T2 and T3 flight appendages-	13
1.4 Structure of Ultrabithorax (Ubx)-.....	15
1.5 Objectives-	17
2 Materials and methods-.....	18
2.1 Sequence Alignments-	18
2.2 Structure modeling-	21
2.3 Molecular Docking-	23
2.4 Transcription Factor Binding Sites (TFBS) search-.....	24
3 Results.....	26
3.1 Identification of evolutionary changes in Ubx and its cofactors across species -.....	26
3.2 Identification of structural constraints on the binding sites-	31
3.3 Predicting the hexapeptide binding sites –.....	38
3.4 Identification of Putative target genes regulated by Ubx and its cofactors-.....	40
4 Discussion-	45
4.1 Evolutionary changes in Ubx and its cofactors-	45
4.2 Hexapeptide bindings with Cofactors -.....	46
4.3 Identification of the target genes based on optimum TFBS separations-	47
4.4 Future Directions-	48
5. References-	49
6 Annexure-1.....	55
6.1 Divergence of Ubx in comparison with its cofactors-	55
6.2 Sequence alignments-.....	63

List of figures-

Figure 1 Divergence of arthropods.....	9
Figure 2: Expression patterns of Hox genes in <i>Drosophila melanogaster</i>	11
Figure 3 Ubx Mutations in <i>Drosophila melanogaster</i> -.....	13
Figure 4 The difference in morphology of flight appendages.....	14
Figure 5 Ubx- Representation of sequence	15
Figure 6 Ubx-Exd-DNA complex	16
Figure 7 Sample BLAST output of Ubx	19
Figure 8 Comparison of Ubx and other cofactor proteins	27
Figure 9 Alignment of Homeodomain region of Ubx.....	28
Figure 10 DNA binding homeodomain of Exd	29
Figure 11 DNA binding domain of MAD	30
Figure 12 Ubx-Exd DNA complex	31
Figure 13 Interacting residues of Exd with Ubx	31
Figure 14 Ubx-MAD and DNA complex 1.....	32
Figure 15 Ubx –MAD and DNA complex 2.....	34
Figure 16 Ubx –MAD and DNA complex 3.....	35
Figure 17 Ubx –MAD and DNA complex.4.....	36
Figure 18 Ubx –MAD and DNA complex 5.....	37
Figure 19 Hexapeptide(FYPWMA) docked over MAD,.....	38
Figure 20 Hexapeptide(FYPWMA) docked over Exd	39
Figure 21 Hexapeptide(FYPWMA) docked over En	40
Figure 22 Frequency distribution of Ubx and Exd binding sites.....	41
Figure 23 Frequency distribution of Ubx and MAD binding sites.....	42
Figure 24 Frequency distribution of Ubx and Exd binding sites.....	43
Figure 25 Frequency distribution of Ubx and Trl	44

List of Tables-

Table 1 Proteins and their PDB IDs obtained from RCSB-PDB.	22
Table 2 Length and sequence difference in Ubx Linker region in four species	32
Table 3 Ubx_MAD different Combinations	33
Table 4 Number of Putative target genes-.....	44

Acknowledgments

I would like to express my sincere and deepest gratitude to my project supervisor Prof. L. S. Shashidhara. I am indebted to him for his invaluable guidance during this project work.

I would like to acknowledge Dr. M.S. Madhusudhan for his constant support and guidance during the project. I thank him for his valuable inputs towards many critical aspects of analysis.

I wish to thank Soumen, Ankit and all the LSS and COSPI lab members for their cooperation and help. My thanks and appreciations also go to my friends in developing the project and people who have willingly helped me out with their abilities.

I thank our former Director Dr. K. N. Ganesh and current Director Dr. Jayant Udgaonkar and Dr. Sanjeev Galande for providing such a wonderful research atmosphere and facilities at IISER Pune.

I would like to thank my family and friends for their constant guidance and moral support to complete the project.

Vaibhav Prakash Wagh

Chapter 1

1 Introduction

Over the course of evolution in arthropods, body plan has undergone modification and resulted in a divergence of appendages. They have the largest diversity of species among which insects have evolved flight over time. There are differences in flight appendages across insect species. These differences arise from segmental diversity, and the underlying cause of this is the evolution of Homeotic (Hox) genes (E. B. Lewis 1982). Hox genes are expressed along the anteroposterior (AP) axis forming the segmental boundary of the appendage morphology. They are transcription factors which bind to DNA and regulate expression of a specific set of genes in specific segments.

Dipterans have diverged from these primitive four-winged insects. The diversity in the flight appendages ranges from two pairs of identical wings in primitive insects to dipterans with a pair of wings and a pair of halteres. The divergence of species over 450 million years is shown in Figure 1. Dipterans have a set of genes which suppress the formation of the second pair of wings. (Weatherbee et al. 1998a) (Weatherbee and Carroll 1999). These genes are regulated by a specific hox gene, but the mechanisms by which they are regulated are not well understood.

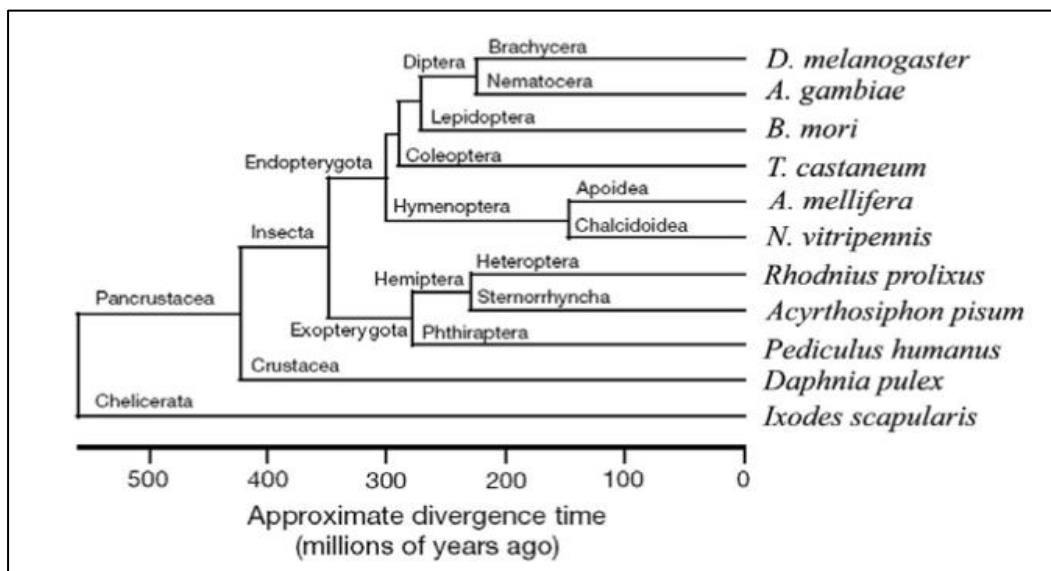


Figure 1 Divergence of arthropods over the course of 450 million years ago (adapted from Honey bee genome sequencing consortia, 2006)

1.1 Development and role of Homeotic genes-

During the process of development of an organism, the basic mechanisms of axis formation and body plan organization are well conserved across animals with bilateral symmetry. Insects have the characteristic segmented body plan, and the sequential expression of different sets of genes establish the body plan along the anteroposterior (A-P) axis. Hox genes which are spatially expressed along AP axis plays a central role in specifying segment identity. Hox genes are functionally conserved across animals because they are formed by series of duplication and divergence events over time.(Maconochie et al. 1996)

Hox genes were first identified based on the phenotypes exhibited by mutations in them. Mutations in a typical hox gene causing gain or loss of function can result in the conversion of one body part along body axis into another and the transformations caused are known as homeotic transformations. For example loss of function mutations in Ubx in *Drosophila* causes haltere to wing transformation in the third thoracic segment. (E. B. Lewis 1982). Hox genes encode for transcription factors which have homeodomain as their DNA binding motif. Which regulate genes involved in the different developmental process such as AP axis segmentation, cell-cycle regulation, differentiation, etc.

Drosophila has eight Hox genes which are classified into two major complexes namely the Antennapedia complex (ANT-C) and the Bithorax complex (BX-C). ANT-C complex includes labial, proboscipedia, deformed, sex combs reduced and Antennapedia which control the identity of the head to the second thoracic segment. BX-C includes Ultrabithorax (Ubx), abdominal-A and abdominal-B they are expressed from part of second thoracic segments to the abdominal segments.(Hughes and Kaufman 2002). The expression patterns of Hox proteins are shown in Figure 2 from embryo to adult stage.

Hox proteins contain a conserved domain called as Homeodomain which acts as DNA binding domain of these transcription factors. They bind and regulate the specific subset of target genes in an organism. Studies have shown that homeodomains cannot dictate DNA-binding specificity on their own. Indeed they bind to similar "AT" - rich DNA binding sites. *In-vitro* experiments have shown that homeodomains bind to similar degenerate TAAT core DNA sequence.. However, *in-*

vivo they are very specific towards their target selection. (Mann, Lelli, and Joshi 2009)
 This is the so-called Hox paradox which is still unanswered.

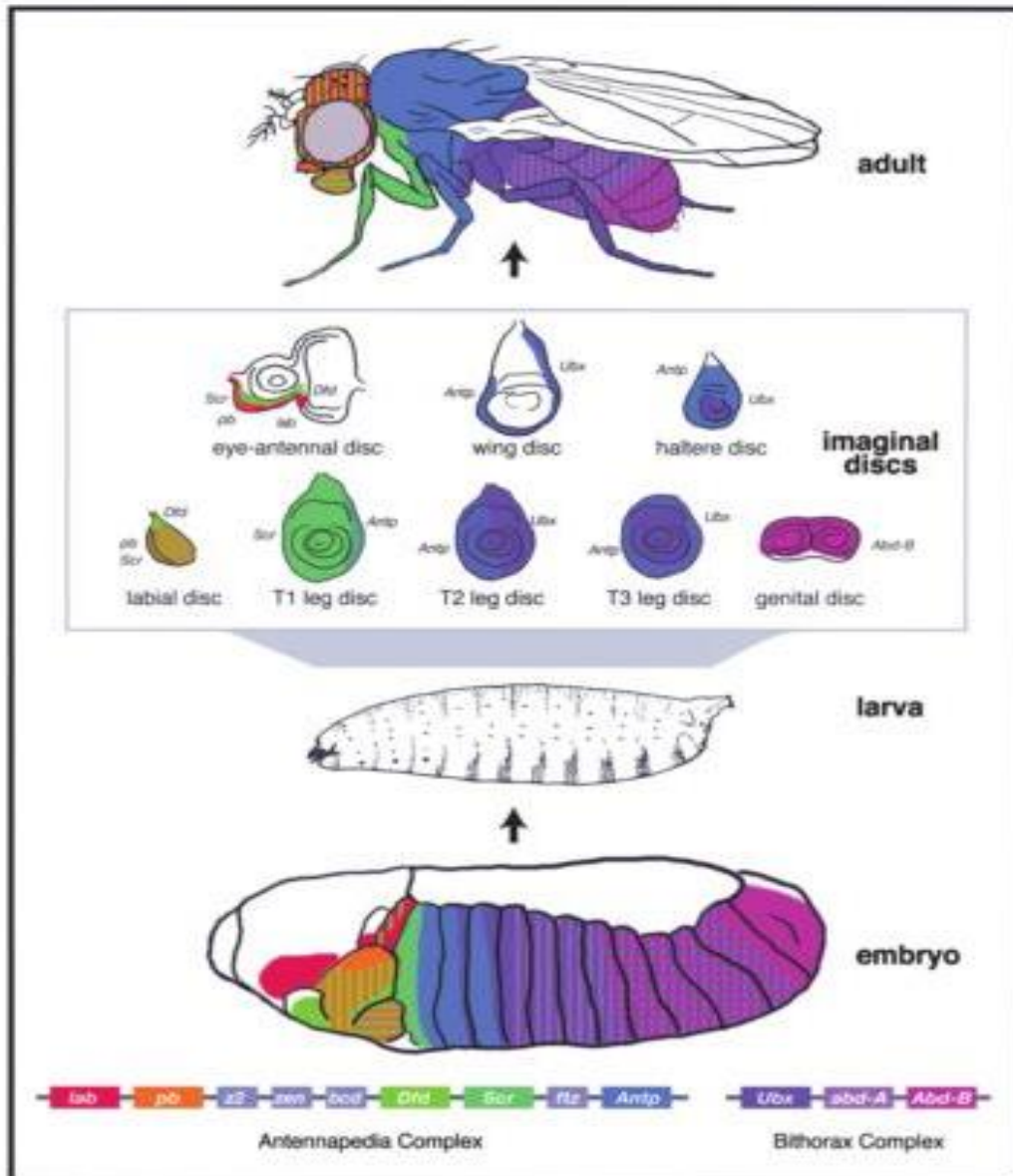


Figure 2: Expression patterns of Hox genes in *Drosophila melanogaster* embryo to adult along anteroposterior axis and in the imaginal discs. (Courtesy: Hughes et.al., 2002)

1.2 Ultrabithorax (Ubx) –

One of these hox proteins Ultrabithorax which are part of Bithorax complex is expressed in posterior part of T2 and T3 in the thorax In *Drosophila* embryos as well in larval stages. (Akam and Martinez-Arias 1985; Weatherbee et al. 1998b). In case of *Drosophila*, Ubx in the T3 segment is responsible for the development of halteres by suppressing the default wing fate. Because overexpression of Ubx in T2 leads to the formation of haltere and removal of Ubx from T3 can alone cause haltere to wing transformation leading to a four-winged fly. Figure 3 shows different phenotypes with the absence of Ubx in T3 showing four-winged fly whereas overexpression of Ubx in T2 showing zero winged fly. This shows that Ubx is necessary and sufficient to cause the transformation.(Akam and Martinez-Arias 1985)

Ubx in other insects species shows more or less similar expression pattern (Figure 4). In *Apis mellifera* embryos the expression pattern is similar to that of Ubx in *Drosophila* embryos.(Walldorf, Binner, and Fleig 2000) Ubx expression in *Tribolium* at the end completion of germ band elongation stage is detected in from parasegment five to parasegment sixteen with highest in the first abdominal segment. Moreover, in *Bombyxmori* Ubx expression is from third thoracic segment to ninth abdominal segment.(Masumoto and Yaginuma 2009)

The expression of Ubx in *Drosophila* in wing imaginal discs is limited to peripodial membrane and in the halteres it is almost everywhere. In case of *Apis mellifera*, more Ubx is present in hind wings compared to forewings. In *Tribolium*, Ubx expression is higher in the wing compared to that of elytron.(Tomoyasu, Wheeler, and Denell 2005; D. L. Lewis, DeCamillis, and Bennett 2000)

Ubx functions at multiple levels in the development of wing disc by suppressing various wing patterning genes(Shashidhara et al. 1999; Weatherbee et al. 1998a). Functionally Ubx is conserved across species because when Ubx derived from *Apis*, *Tribolium* and *Bombyxis* overexpressed in transgenic *Drosophila* it can induce wing to haltere transformation which is identical to changes caused by the overexpression of *Drosophila* Ubx.(Prasad et al. 2016) This suggests that though the function of Ubx it is conserved, but it is regulating different genes in different insect species. The ways by which Ubx is regulating different targets genes is not well understood.

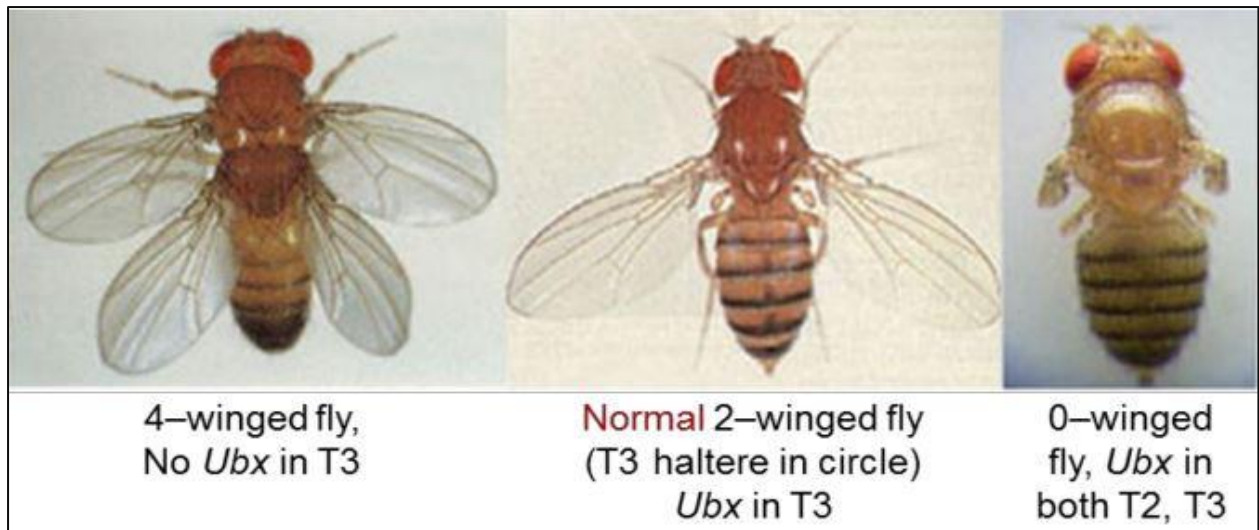


Figure 3 *Ubx* Mutations in *Drosophila melanogaster*- left is the absence of *Ubx* in T3, the middle is wildtype two-winged fly and right panel shows overexpression of *Ubx* in the T2 segment. (Adapted from Lewis 1978)

1.3 Morphological differences between T2 and T3 flight appendages-

In *Apis mellifera*, they have wings on both second and third thoracic segments with smaller differences in venation patterns, and also the orientation of bristles is opposite which helps in interlocking wings during flight. (Walldorf, Binner, and Fleig 2000)

In *Tribolium castaneum* second thoracic segment comprise Elytra which is thick and pigmented which act as a protective shell and third thoracic segment has a pair of wings which helps in flight. (Tomoyasu, Wheeler, and Denell 2005). In *Bombyxmori* the size of the forewing and hindwing differs, the hindwing relatively smaller in size than forewing. (Masumoto and Yaginuma 2009)

In *D. Melanogaster* wing in the T2 segment is a flattened structure composed of veins and interveins. Whereas The T3 segment harbors a modified wing structure, the haltere. It is a bulbous structure and acts as a balancing organ for flies. (Weatherbee et al. 1998a). In *Apis* and *Bombyx* there are not significant differences in forewing and hindwing, but in *Tribolium*, *Ubx* suppresses elytra in T3 and retains wing

identity which is unlike in *Drosophila*. Figure 4 shows the differences in the wing morphologies across four major types of endopterygote insects.

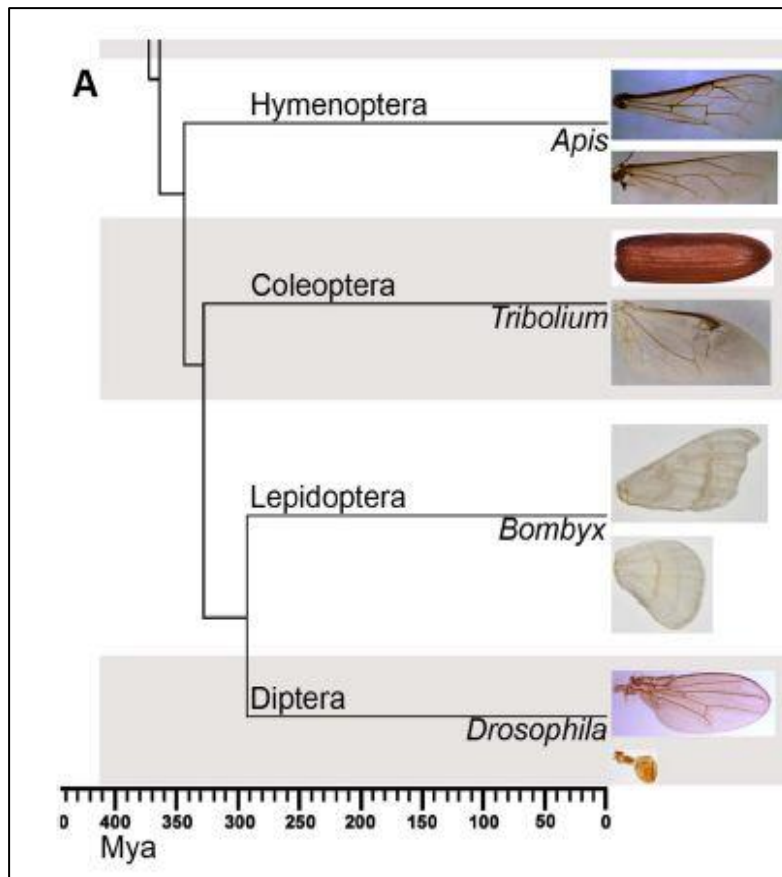


Figure 4 The difference in morphology of flight appendages on T2 and T3 segments. The above image is forewing and below is hindwing across species. In *Tribolium* the forewing is modified into elytra and in *Drosophila* hindwing is modified into haltere. (adapted from Zdobnov)

Previous studies to understand the Ubx regulated gene networks showed that Ubx targets genes at multiple levels of signaling pathways. Moreover, Ubx is regulating its target very specifically in a different organism. (Mohit et al. 2006; Shashidhara et al. 1999; Agrawal et al. 2011a) The differential regulation of target genes by Ubx can be because of the evolution of following mechanisms acting together or independently and may explain the Ubx mediated specification of haltere fate by suppression of wing fate in *Drosophila*.

- 1) Evolutionary changes at cis-regulatory sequences of targets of Ubx.
- 2) Evolutionary changes in the protein sequences of cofactors of Ubx.
- 3) Evolutionary changes in the protein sequences of targets of Ubx.

1.4 Structure of Ultrabithorax (Ubx)-

Ubx belongs to homeodomain family of proteins. Homeodomains are helix-turn-helix motifs with three alpha helices with evolutionary highly conserved DNA binding residues. (Passner et al. 1999). Ubx consists of structured as well as unstructured regions. Figure 5 represents the structural domains of Ubx, starting from N-terminal of the protein, amino-acids 1-102 consists of short structured regions separated by unstructured sequences, from 103-216 is a disordered region which consists of activation domain. Amino acids 250-303 consists of intrinsically disordered regions which consists of FYPWMA motif which is known to interact in the case of Ubx-Exd interaction and alternately spliced microexons. The C-terminal of Ubx includes homeodomain and a repression domain. (Liu, Matthews, and Bondos 2008)

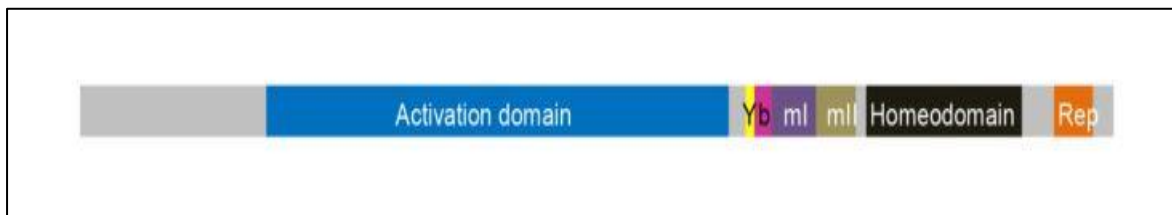


Figure 5 Representation of Ubx Sequence –Blue: transcription activation domain, Yellow: YPWM Exd interaction motif, Black: homeodomain , (Liu, Matthews, and Bondos 2008)

The homeodomain of Ubx consists of the structured region with 60 amino acid forming the helix-loop-helix structure and hexapeptide (HX) FYPWMA towards the N terminal of the Homeodomain and UbdA motif towards the C terminal of the Homeodomain. Hexapeptide and the homeodomain are connected by linker region whose length varies across orthologs as well as between isoforms of Ubx. (de Navas et al. 2011; Liu, Matthews, and Bondos 2008)

Ubx is regulating its targets specifically which suggests that Ubx is achieving this specificity by interacting with other co-transcription factors. Studies have shown that Ubx is regulating spalt by collaborating with MAD and their binding sites are next to each other. (Walsh and Carroll 2007)

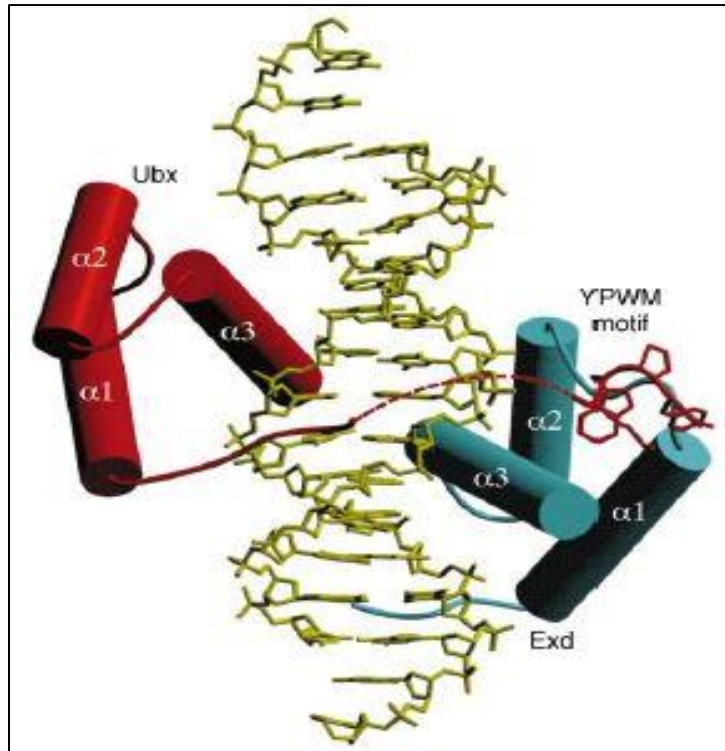


Figure 6 Ubx-Exd-DNA complex, Homeodomains of Ubx and Exd binding on opposite sides of DNA and the YPWM motif interacting with hydrophobic pocket of Exd with the help of a linker region of variable length. (adapted from Knoepfler and Kamps 1995)

One such attempt to find the direct physical interaction between Ubx and its cofactors has identified a short amino-acid sequence called the hexapeptide(HX) forming protein complex between Ubx and binding to DNA. (Figure 6) The HX is required for the interaction with partners because mutations in HX fails to form a complex with extradenticle (exd).(Knoepfler and Kamps 1995; Foos et al. 2015). Also, other studies to explore the possibility of HX interacting with other cofactors such as Biomolecular Fluorescence Complementation experiments have shown that Hox proteins can bind different cofactors and HX is essential for this interaction.(Hudry et al. 2012; Merabet and Lohmann 2015)

To explore the contribution of Ubx interaction with cofactors, previous studies have identified different cofactors or collaborator protein interacting with Ubx(Hudry et al. 2012), and also putative cofactors of Ubx were identified using ChIP-chip experiments performed previously in the lab.(Agrawal et al. 2011b)

This project aims at testing hypothesis that interaction of Ubx with different cofactors is mediating specificity of target gene regulation.

1.5 Objectives-

The objectives are summarised as follows:

- 1) Identification of evolutionary changes in the protein sequences of known cofactors of Ubx in *Drosophila* species and compare it with *Apis mellifera*, *Bombyx mori*, and *Tribolium castaneum*.**
- 2) Identification of structural constraints on the binding sites because of interaction between Ubx and its cofactors using homology modeling.**
- 3) Identification of patterns of binding sites of Ubx and its cofactors on ChIP pulled down sequences.**
- 4) Identification of functionally relevant targets of Ubx to study Ubx-cofactors interactions.**

This report contains series of analysis starting from identification of evolutionary changes in the protein sequences of cofactors using sequence alignments, which provided information about the conservation of proteins and their DNA binding domains across species. Using this information we found structural constraints on the binding sites using tools of homology modeling and molecular docking. With the known constraints on the Ubx and cofactor binding sites, we searched for the occurrences of transcription factor binding sites and putative target genes downstream to these binding sites. Later comparison of this data with ChIP-seq helped us identify functionally relevant targets of Ubx to study Ubx-cofactor interaction.

Chapter 2

2 Materials and methods-

2.1 Sequence Alignments-

Finding orthologues of proteins-

The cofactors of Ubx are known in *Drosophila melanogaster*, to find their orthologs in other insect species PSI-BLAST was used.

Sequence alignments are done to find sequences of significant similarity in a genome or proteome database. Here sequence alignment is used to find orthologues sequences of proteins in protein database of NCBI. We ran PSI-BLAST to find out different orthologues of proteins.(Altschup et al. 1990; Z. Zhang et al. 2000). PSI-BLAST is a Position Specific iterated BLAST where the Protein sequence from *D. melanogaster* is used as a query sequence, and then non-redundant protein database was searched for *Apis mellifera*, *Bombyxmori*, and *Tribolium castaneum*. Here we ran three iterations with a PSI-BLAST threshold of 0.005 and conditional compositional score matrix adjustment. The matrix used here is BLOSUM 62. After three iterations sequences with lower E-values were selected. Most of the times the proteins were listed as orthologues. The sequences were collected and then used for further analysis.

Pairwise Sequence Alignments-

Pairwise sequence alignments are used to find the conservation of protein sequence between *D. Melanogaster* and *Drosophila* species as well as between *Apis mellifera*, *Bombyxmori*, and *Tribolium castaneum*. This can be done using two different ways namely Global sequence alignment and Local sequence alignment. Global Sequence Alignments-

Global sequence alignment is used to find optimal alignment over the entire length of the protein sequences being compared. Global alignment is preferred when sequences to be aligned are of comparable lengths. Global alignments are helpful when two sequences have detectable sequence similarity over the entire length. Cases where there is the insertion of the unrelated motif in a sequence the global alignments are not helpful.

For doing global sequence alignments, EMBOSS Needle (Rice, Longden, and Bleasby 2000; McWilliam et al. 2013) was used. The algorithm used by EMBOSS Needle is Needleman-Wunsch with BLOSUM 62 matrix. The other settings were used as the default.

Local Sequence Alignments-

Local sequence alignments find the domain or short regions of similarities between a pair of sequences. It is helpful when we are interested in finding best alignments between subsequences, finding conserved domains, etc.

For Local sequence alignments, NCBI-BLAST was used.(Altschul et al. 1997; Altschup et al. 1990; Altschul 1991). NCBI-BLAST uses Smith-Waterman Algorithm for the alignments. Here we used BLOSUM 62 matrix with expected threshold 10 and word-size was 3. The following figure shows a sample blast output-

BLAST run gave us multiple scores based on the alignments. They are listed below.

Score	Expect	Method	Identities	Positives	Gaps
256 bits(653)	2e-87	Compositional matrix adjust.	175/406(43%)	200/406(49%)	146/406(35%)
Query 1	MNSYFEQ-ASGFYGHPHQATGMAMGSGGHHDDQTASAAAAAYRGFPLSLGMSPYA----NH				55
Sbjct 1	MNSYFEQ A GFYG H TG A HHD A AAAYR FPL LGMSPYA +H				54
Query 56	HLQ-----RTTQDSPYDASITAACNKIY-----GDGAGAYK----QDCLN				91
Sbjct 55	H R QDSPYDAS+ AC K+Y G +Y +DC				113
Query 92	IKADAV--NGY-----KDIWNTGGNSGGGGGGGGGGGAGGTGGAGNANGGNAAN				140
Sbjct 114	+ NGY KD+W + S				143
Query 141	ANGQNNPAGGMPVRPSACTPD-SRVGGYLDTSGGSPVSHRGGGAGGNVSVSGGNAGGV				199
Sbjct 144	AN Q+N VRPSACTP+ +RVG Y GG GG+ + S GN ++				185
Query 200	QSGVGVAGAGTAWNANCTISGAAAQTAASSLHQASNHTFYPWMAIAGECPEDPTKSKIR				259
Sbjct 186	++WN C+++ +A+Q A Q +NHTFYPWMAIAG				226
Query 260	SDLTQYGGISTDMGKRYSESLAGSLPDWLGTNGLRRRGRQTYTRYQTLELEKEFHNTHY				319
Sbjct 227	NG+RRRGRQTYTRYQTLELEKEFHNTHY				255
Query 320	LTRRRRIEMAHALCLTERQIKIWFQNRRLKKEIQAIKELNEQEK				365
Sbjct 256	LTRRRRIEMAH+LCLTERQIKIWFQNRRLKKEIQAIKELNEQEK				301

Figure 7 Sample BLAST output of Ubx

Following entities were provided by the BLAST (Figure 7) which are listed here

Bit score describes the overall quality of alignment; the score depends upon the use of the scoring system. This takes into account the variable positive scores for identities and positives and the penalty for the gaps. E – Value is the number of different

alignment with score better or equal to the bit score that is expected to occur by chance in a database search. Lower e-value, indicates more significant score and the alignment. Identities are the extent to which two amino acid sequences have the same residues at the same position in an alignment. Whereas positives are changes in the amino acids which conserves their physiochemical properties. Gaps are spaces introduced into alignment to compensate for insertions or deletions in one sequence relative to another. Aligned length is total aligned positions computed by the BLAST, and minimal length is the sequence length of smaller of the two polypeptides.

For quantification of the similarities and difference, the following parameters were used-

1. **Bit score/aligned length-** To assess the bit score per unit length of aligned positions.
2. **Bit score/ minimal length-** To assess the bit score per unit length of smallest polypeptide among the two.
3. **Percentage Identities-**Percentage score of having same amino acid at same positions.
4. **Percentage Positives-** Percentage score of having a conservative substitution of amino acids at same positions

The rationale behind using bit score/ minimal length-

In case of two protein sequences of different lengths are to be compared. Where first few residues are identical, and rest is changed. Here BLAST will align first few positions. In case of parameters bit score/ aligned length, we will end up getting a higher score because of aligning only first few positions. Thus using bit score/minimal length, we get more realistic estimation ensuring verity in case of the proteins in which some protein motifs have remained identical. For example, in case of two proteins of varying lengths 100 and 150 amino acids respectively, the first 50 residues are identical, and rest are diverged thus using Bit score/aligned length we might get 100% identical as the aligned length shall be first 50 amino acids. Using Bit score/minimal length, we get realistic estimation which is 50% identical because here minimal length of 100 and 150 is 100. Thus this comparison parameter would be more useful.

Multiple Sequence Alignments-

We used multiple sequence alignments to find the changes in the DNA binding domains of cofactors of Ubx across species.

Multiple sequence alignments are used to find regions of similarities across multiple sequences. It helps to identify structural and functional domains.

Multiple sequence alignments were done to compare the orthologous sequences across different species. Also, the different protein sequences from an organism were aligned in search of conserved motifs. Clustal Omega (Sievers et al. 2011; McWilliam et al. 2013) was used for multiple sequence alignments here, Input sequence was provided in the form of multiple protein sequences in Fasta file, and other parameters were used as default.

2.2 Structure modeling-

To find out structural constraints on the Ubx-cofactor interactions and their comparison across different species we used homology modeling.

Homology modeling of protein is used to construct a protein structure model of a target protein from its amino acid sequences and experimentally determined structures of a related homologous protein. If the proteins are homologs, then they are likely to have similar structures. The target sequence is Protein sequence to be modeled in this case we use Ubx, Exd, and MAD protein sequences as Target sequences; Templates are homologous protein structures which are experimentally determined.

The proteins which have similar structure are more likely to perform similar functions. Thus predicting protein structure helps in predicting the function of the protein. There are different steps involved which are template search, target-template alignment, model-building, and model evaluation. The reliability of the model depends upon sequence identity between target and templates. Good sequence identity gives more accurate and reliable models. The models with low sequence identity have a higher chance of errors and can be improved using advanced modeling tools.

For modeling the structure of proteins, three different tools were used.

MODELLER:

MODELLER is used for comparative protein structure modeling of protein structure using satisfaction of spatial restraints. (Andrej Sali et al. 1993; Marti et al. 2000; Etheve, Martin, and Lavery 2015)

MODELLER allows the user to choose parameters at each step of modeling and many other things such as multiple template modeling, comparison of structure and modeling ligand in the binding site, etc. The stepwise process while using MODELLER is as follows

Template search- The target sequence was searched in RCSB-PDB database using BLAST. For some of the cofactor, 3D crystal structures were obtained from the PDB database. The proteins with name and PDB ID are listed here-

Protein Name	PDB-ID
Ultrabithorax	4CYC,4UUT,4UUS
Extradenticle (Exd)	4CYC,4UUS
SMAD	1MHD
Engrailed	2P81

Table 1 Proteins and their PDB IDs obtained from RCSB-PDB.

Target –template alignment-

After that target-template alignment was performed using align2d() command. For visual inspection of the alignment files in the .pap format was used.

Model-Building –

Model- building command takes a target-template alignment file and template structure as input and gives out the number of possible models specified.

For few of the cases, multiple templates were used to model proteins, in that case, a step is included before model building where structurally the templates were aligned using salign().

Model Evaluation-

After model building, we get multiple models with different scores mainly MODELLER objective function (molpdf), and discrete optimized potential energy (DOPE) score was considered. Out of these models we selected structures with the lowest DOPE score. As in the above proteins, all proteins were well aligned, and there was not much variation in the DOPE score. After modeling the protein structures-

To look for interaction between Ubx and cofactors when bound to DNA for that DNA bound structures of proteins are needed. For modeling proteins with DNA, we have DNA bound templates of proteins from which DNA was used as a ligand.

Using ligand modeling in binding sites, the extra residues for DNA were added in the alignment file, and further steps were performed. For proteins for which there was no template available or the alignment was poor I-TASSER AND Mod-Web were used,

I-TASSER -Iterative Threading Assembly Refinement (Y. Zhang 2008)-

By providing a sequence of a protein to be modeled, it identifies structural templates from PDB database and using multiple threading approaches and constructs atomic models using iterative template fragment assembly simulations.

ModWeb:

ModWeb is an online version of MODELLER (Pieper et al. 2004). ModWeb is an automated tool which takes protein sequence as input and builds structural models. Out of which best scoring model is selected. The amount of sampling of templates while calculating models can be controlled by assigning different fold assignment methods available. After modeling different proteins with their DNA bound structure, the analysis of structure needs to be done for that UCSF-Chimera was used

UCSF-Chimera-

Chimera is a Protein structure visualization tool. After building protein models, using this different analysis were performed. (Pettersen et al. 2004)

Structure analysis-

Matchmaker used for comparison of two protein structure by superimposing onto each other. Finding clashes/contacts finds unfavorable interaction cases where atoms are too close together and all kinds of direct interactions. The coulombic surface coloring is used to visualize the charge on the surface of protein molecules. Using build structure, B-DNA double helix were built with required binding sites of two proteins with different separation of bases.

2.3 Molecular Docking-

Here homology modeling provided us with possible orientation in which Ubx and cofactor can interact but to find out potential binding sites of hexapeptide from Ubx on cofactors we used molecular docking.

Molecular docking is used to predict possible binding sites of a peptide of interest by algorithms of conformational samplings and minimizing the scoring function. The general steps involved in the docking are Target (Proteins) selection ligand selection and preparation, docking and evaluation of docking results respectively.

Molecular docking was used to find HX interacting with different cofactors

Autodock-Vina

It is an automated molecular docking tool. The steps are as follows: Ligand and Protein preparation by adding polar hydrogen atoms. Then minimizing the rotatable bonds and setting up the grid for search space of ligand binding. (Morris and Huey 2009). After running autodock-vina multiple outputs with all possible binding sites were generated, out of them, we selected models with lowest binding energy. Here our protein is DNA binding, so the sites which overlap with DNA binding regions of proteins were blocked.

CABS-dock-

It is an online tool which takes protein and ligand sequence as inputs and provides possible ligand binding sites on a protein. (Blaszczyk et al. 2016; Kurcinski et al. 2015). CABS-dock here was used with default settings.

2.4 Transcription Factor Binding Sites (TFBS) search-

Structural constraints obtained on the binding sites from Homology modeling for Ubx and cofactors can help in finding the putative target genes. To find the putative target genes we used TFBS search and further listed the downstream target genes.

Transcription factors binding sites were collected from different kinds of literature available.

TRANSFAC - Contains database for eukaryotic transcription factors their experimentally proven binding sites, consensus binding sites and regulated genes. (Matys 2006)

JASPAR is an open-access database of curated and non-redundant transcription factor binding profiles.(Sandelin 2004)

All these tools provide us with either single DNA binding motif of a protein or position weight matrix. Which are later used to find the binding sites of the protein over DNA sequences.

For finding a binding site python programme was written, for scanning PWM over genome of an organism PWM tool (<http://ccg.vital-it.ch/pwmtools/pwmscan.php>) which scans a particular genome database when given a position weight matrix of a transcription factor of interest.

We used matrix constructed from TRANSFAC and JASPAR for TFs like Ubx, Exd, MAD, etc. with individually to scan genome.

Also, two matrices were combined with different distance separation up to ten bases to find out the instances with two Transcription factors binding sites present in the proximity. The occurrences were plotted against the base separation.

Chapter 3

3 Results

3.1 Identification of evolutionary changes in Ubx and its cofactors across species -

Using Pairwise sequence alignments-

Pairwise sequence alignment is used between protein sequences of *Drosophila melanogaster* and other *Drosophila* species as well as between *Apis mellifera*, *Bombyxmori*, and *Tribolium castaneum*. Then the obtained homology score is compared with that of Ubx using four different parameters. Here higher homology score indicates a slower rate of change of proteins.

The parameters used for the analysis are Percentage identity, percentage positives, Bit score per aligned length and bit score per minimal length. The scores are normalized with the scores from *Drosophila melanogaster*. Also, comparison with Cytochrome P450 reductase was performed to check scores calculated when the parameters are compared to an unrelated protein sequence.

While comparing the parameters from Ubx with Exd (figure 8), the parameters show that in comparison to Ubx, Exd is more conserved across all *Drosophila* species as well as in *Apis mellifera*, *Bombyxmori*, and *Tribolium castaneum*. In Exd the percentage identity and percentage positives are almost 100% in species of *Drosophila*, but Ubx has variation in a score ranging from 74% to 100 %. For *Apis*, *Bombyx* and *Tribolium* the scores are higher in Exd in comparison with Ubx. Also in bit score per minimal length and bit score per aligned length scores are close to one in species of *Drosophila* and much higher in *Apis*, *Bombyx*, and *Tribolium* in Exd but lower for Ubx.

The comparison points out that irrespective of the parameters used Ubx is evolving faster in comparison with Exd across species. The high conservation rate of Exd hints at its conserved function as well.

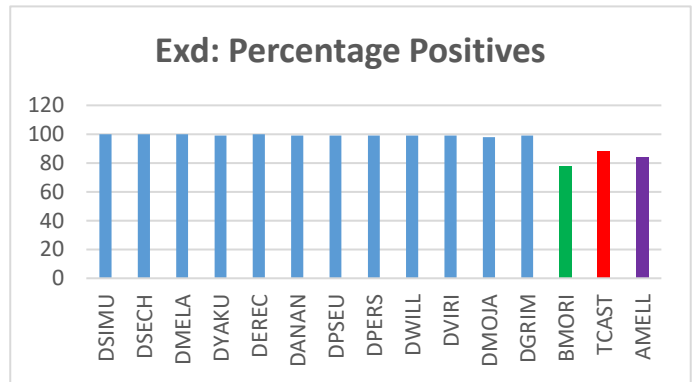
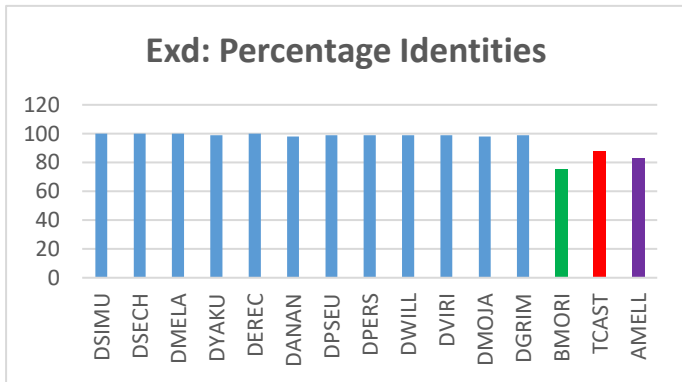
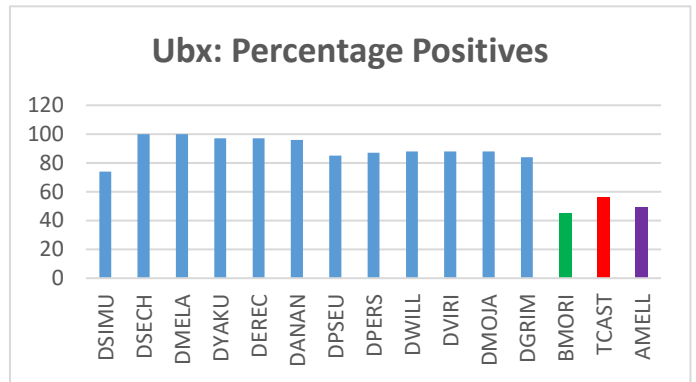
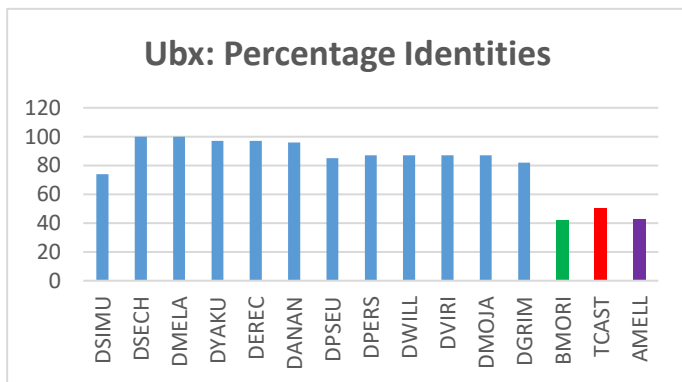
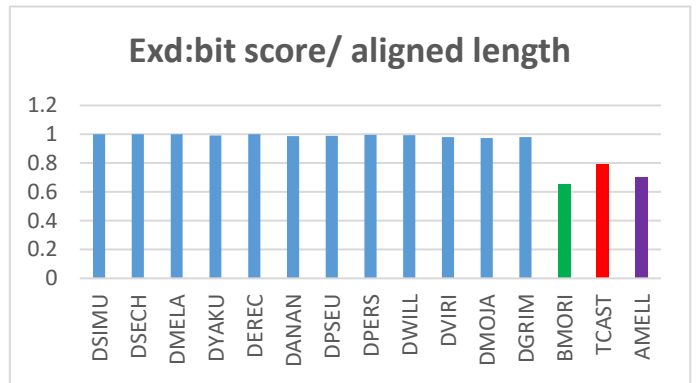
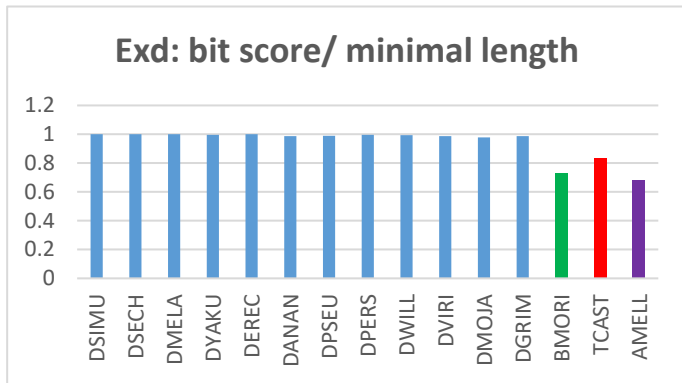
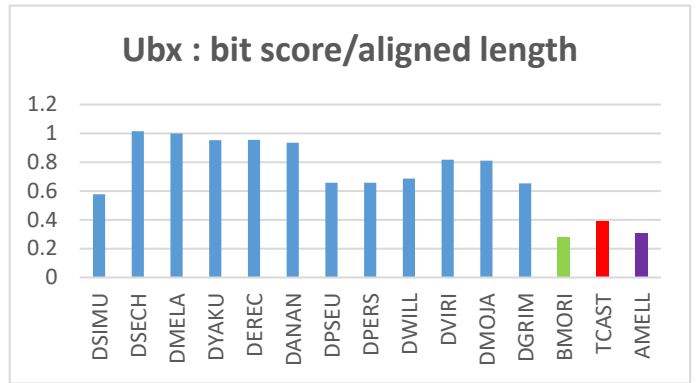
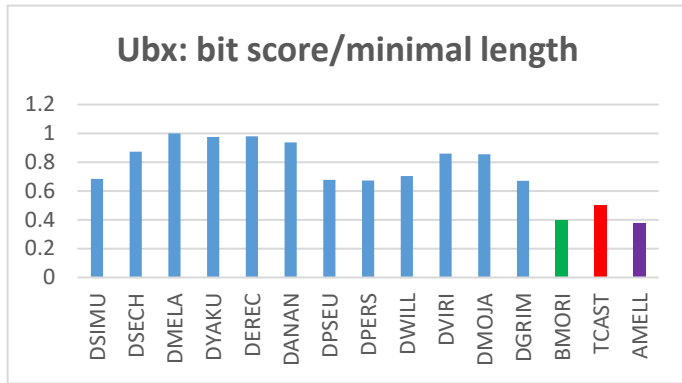


Figure 8 Comparison of Ubx and other cofactor proteins based on four different homology parameters across twelve *Drosophila* species and *Bombyxmori*, *Apis mellifera* and *Tribolium castaneum*. Blue-*Drosophila* Species, Green-*Bombyxmori* (BMORI), Red- *Tribolium castaneum* (TCAST), Purple-*Apis mellifera* (AMELL).

A similar comparison was carried out for other cofactors (please refer to Annexure 1) which shows that in comparison with Ubx the homology indices of MAD, Elf-1, E2f1, Hairy, Adf-1, Hth, and CPR450 are higher. Which indicates the higher rate of change of Ubx. In case of GAF, the protein database in *Apis* and *Bombyx* does not result in orthologous sequence because of very low query cover and no characterization in the database. This can also be because of the absence of protein orthologs.

Identification of changes in the DNA binding domains-

i) Ultrabithorax (Ubx)-

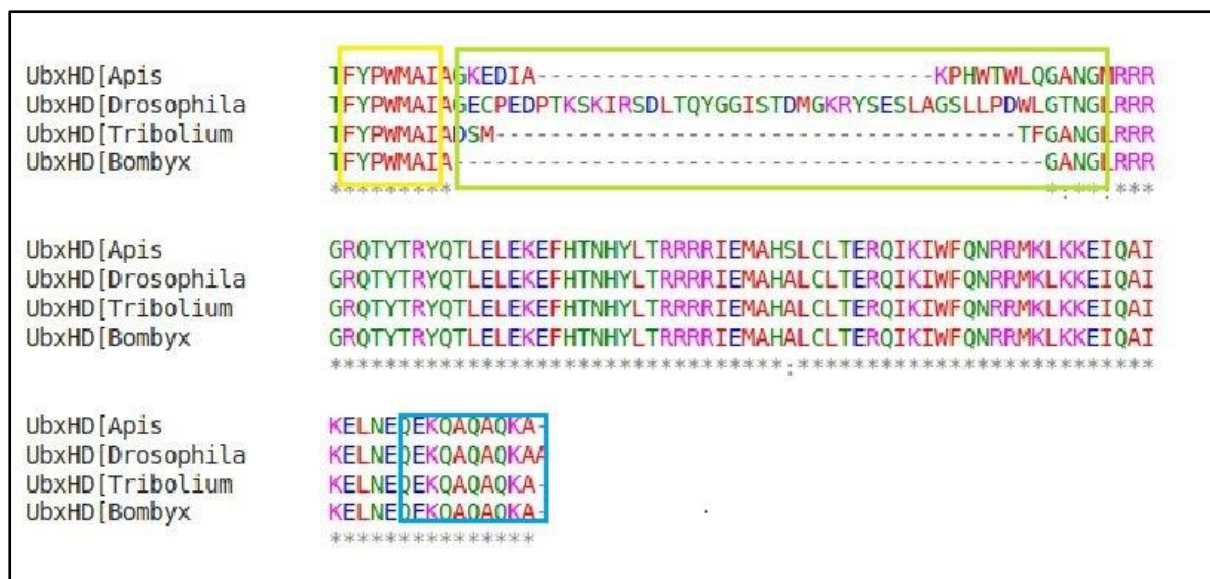


Figure 9 Alignment of Homeodomain region of Ubx. Here each row indicates sequence from species mentioned towards left. The yellow box indicates conserved hexapeptide(FYPWMA) and the green colored box shows the difference in the linker region(LR). The unboxed sequences are homeodomains which are conserved, and the blue colored box shows the UbdA motif. The color codes for amino acids are as follow: Red: Hydrophobic, Blue: Acidic, Magenta: Basic-H, Green: Hydroxyl, sulfhydryl, amine, and glycine.

The alignment of Ubx with its orthologs from the other insect groups showed that Ubx sequences differ in the unstructured (disordered) part of the protein which is towards the N-terminal of the Homeodomain (HD).(Figure 9) However, looking at the hexapeptide(HX), HD and UbdA motif are conserved across above alignments. The major difference between above alignments is in the length of linker region joining HX and HD.

ii) Extradenticle (Exd)-

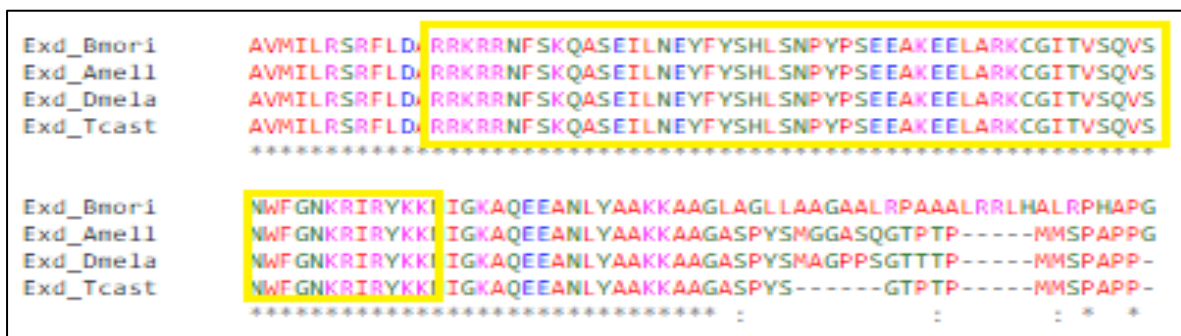


Figure 10 DNA binding homeodomain of Exd in Apis (Amell), Bombyx(Bmori), Tribolium (Tcast) and Drosophila (Dmela), the yellow box highlights the homeodomain (DNA binding domain) of Exd . The color scheme is as mentioned in the figure 9.

This alignment shows the aligned protein sequences of homeodomain from Exd from *Bombyxmori*, *Apis mellifera*, *Drosophila melanogaster*, *Tribolium castaneum* respectively.(Figure 10) When bound to DNA Exd can interact with Hexapeptide(Hx) from Ubx, and the interacting residues are conserved across all of them. The DNA binding domain is highly conserved which implies that across species their DNA binding sites could be identical.

iii) Mothers Against Dpp (MAD)-

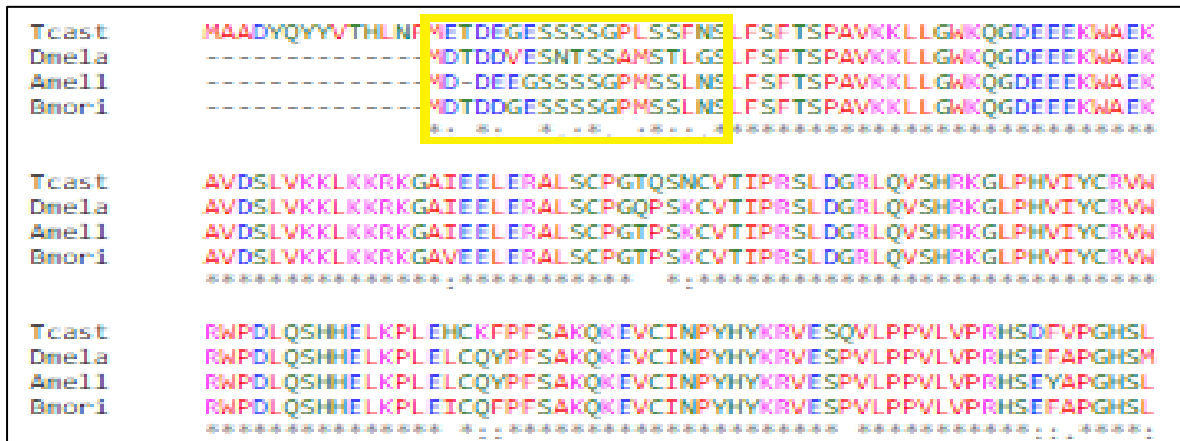


Figure 11 DNA binding domain of MAD (MH1) in *Apis* (*Amell*), *Bombyx*(*Bmori*), *Tribolium* (*Tcast*) and *Drosophila* (*Dmela*).The yellow box indicates the N- terminal residue change in the MH1 domain of MAD.

This is aligned MH1 domain of protein MAD which is the DNA binding domain of the protein(Figure 11) As the alignment shows the MH1 domain is conserved across them. There are few changes in the amino acids in the *Drosophila melanogaster* towards the N-terminal of the protein. These few changes in the DNA binding residues of a protein are crucial for their DNA binding specificity.

Overall, multiples sequence alignments of Ubx, Exd, MAD show that the DNA binding domains of all these proteins are conserved across *Apis*, *Bombyx*, *Tribolium*. Which indicates that the DNA binding sites of these proteins are similar across species, but there can be preferences for one site over another in MAD because of changes in the N-terminal residues of the domain.

Another alignment with different proteins of same species is performed to check whether there are similar motifs which are common in all Ubx interacting cofactors. In that case, proteins were pair wised aligned with each other. The rationale behind performing these alignments was, as all cofactors are interacting with Ubx then there might be residues which are similar in all the cofactors. However, after aligning them using BLAST, evaluated using e-values and aligned lengths which were poor except for the protein E2f1 aligned with others. The tables for the aligned length, e-values and percentage identity are in Annexure I.

3.2 Identification of structural constraints on the binding sites-

i) Ubx-Exd interactions-

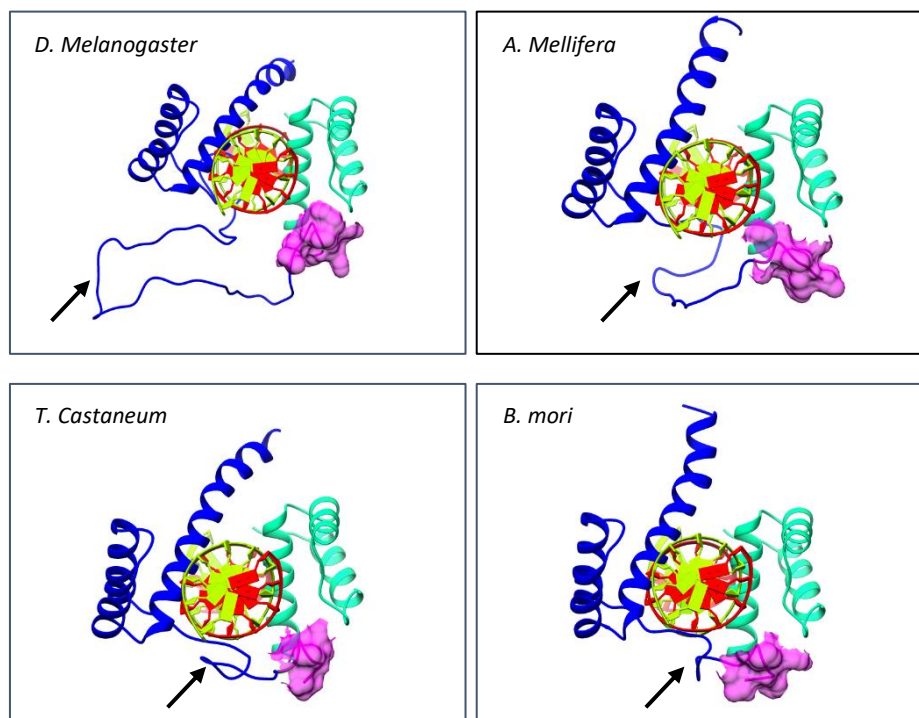


Figure 12 Ubx-Exd DNA complex in *A. mellifera*, *D.Melanogaster*, *B.mori*, *T.castaneum*. Here Red-yellow colored DNA double helix, Ubx in Blue and Exd in Cyan and Hexapeptide(HX) surface in Magenta. Different lengths of Linker region(LR) are pointed by the arrow. Ubx is interacting with Exd with the help of HX mediated by a linker.

We have the experimentally determined structure of Ubx and Exd (4CYC), where a small peptide of length six amino acid is interacting with the hydrophobic pocket formed by residues of extradenticle (highlighted in Figure 13.)

```
GARRKR RNF SKQASE I LNEY FYSH LSNPY PSEEAKEELARKCGI TVSQVS
NWFGNK R IRYKKN I GKAQEEANLYAA
```

Figure 13 The sequences show residues highlighted in green interacting with Hexapeptide (HX) from Ubx,

The modeled structures show Homeodomains of Ubx and Exd bound to DNA and Hexapeptide (HX) from Ubx interacting with Exd with the help of linker region(LR).(Figure12).The lengths of linker region vary across them with *D.melanogaster* having most extended linker whereas *B.mori* with the shortest linker. The difference in the linker region can be seen in the above structures. (Table2).

The length of LR is variable across these four species, which can play a crucial role in determining the difference in interactions.

Species Name	Linker Region(LR) sequence	LR length
<i>Drosophila melanogaster</i>	IAGECPEDPTKSKIRSDLTQYGGISTDMG KRYSESLAGSLLPDWLGTNGL	50
<i>Apis mellifera</i>	IAGKEDIAKPHWTWLQGANGM	21
<i>Tribolium castaneum</i>	IADSMTFGANGL	12
<i>Bombyx mori</i>	IAGANGL	7

Table 2 Length and sequence difference in Ubx Linker region in four species

ii) Ubx-MAD interaction-

MAD is another protein which is shown to be collaborating with Ubx. The MH1 domain of MAD is a DNA binding domain, and here MH1 domain structure is used for structural analysis. After aligning sequences of the MH1 domain, it was observed that it is well conserved across *Apis*, *Bombyx* and *Tribolium*. MAD is binding to DNA with the help of beta-hairpin structured motif. Here from two separate DNA bound structures of Ubx and MAD, we modeled the Ubx and MAD bound DNA complex. MAD is binding to DNA with the help of beta-hairpin loop. (Figure 14)

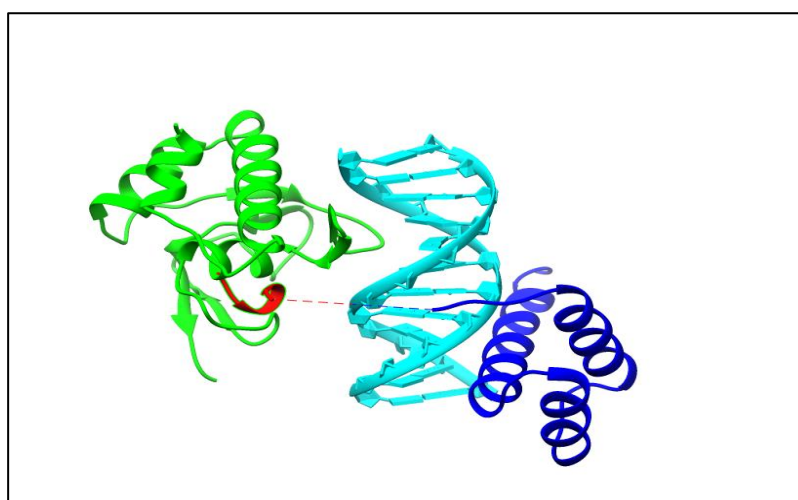


Figure 14 Ubx-MAD DNA bound complex Ubx-blue ,DNA double helix-Cyan,MAD-green and Hexapeptide in Red

To check for all possible orientations by which HX from Ubx can interact with MAD, we used different combinations of binding sites either present on the same strand of DNA or the opposite, also with different distance separation between binding sites of Ubx and MAD. We have considered here binding residues on single strand which are DNA binding sequences obtained from the crystal structures are TTTAT for Ubx and TAGAC for MAD.

Ubx_MAD	MAD_C_Ubx
Ubx_C_MAD	MAD_Ubx

Table 3 Ubx_MAD different combinations of binding sites

Here MAD_C and Ubx_C indicate binding on the opposite strand compared to the previous binding. Later the binding sites of Ubx and MAD were separated to by adding nucleotides in between. Captions for images are in the form of their binding site sequence and separation between them ex. Ubx_MAD_1 indicates Ubx and MAD binding sites separated by 1 Base.

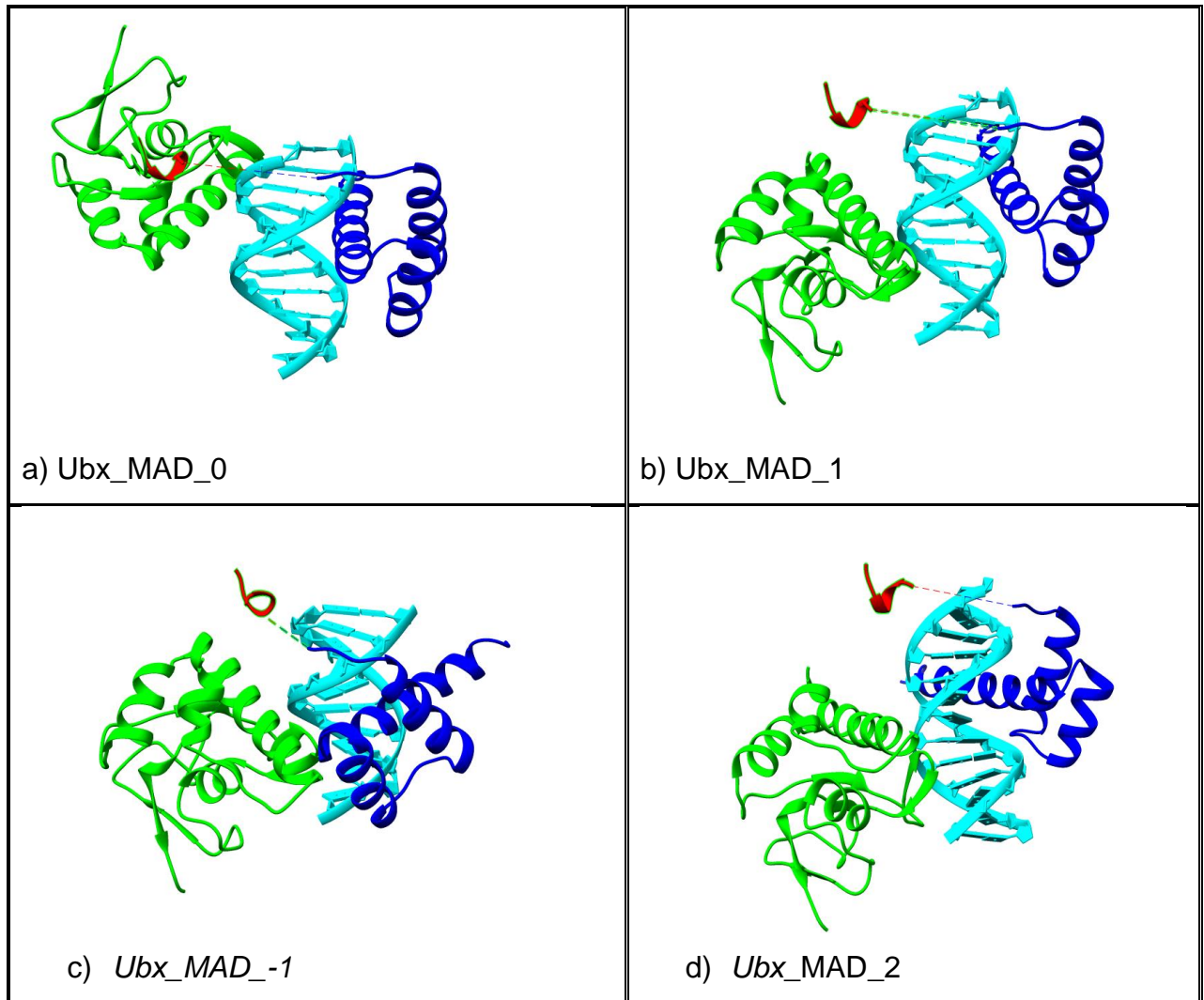


Figure 15 Ubx –MAD and DNA complex Green: MAD, Blue: Ubx, Cyan: DNA double helix, Red: Hexapeptide(HX) image a) is Ubx and MAD binding sites on Same strands with 0 base separation. Similarly, the legends c), d), e) reads Ubx and MAD binding sites on the same strand with separation of 1,-1 and two respectively/

Ubx and MAD binding sites are on the same strand of DNA, When these binding sites are too close there are clashes between proteins. In figure15 case a) MAD binding site is immediately after Ubx, and there is no clash between Ubx and MAD DNA binding regions compared to case c) where Ubx and MAD binding sites have an overlap of 1 base which is causing the hindrance. In the cases b) and d) are with separation of 1 and two bases respectively, proteins are far apart and hence are not causing any clashes.

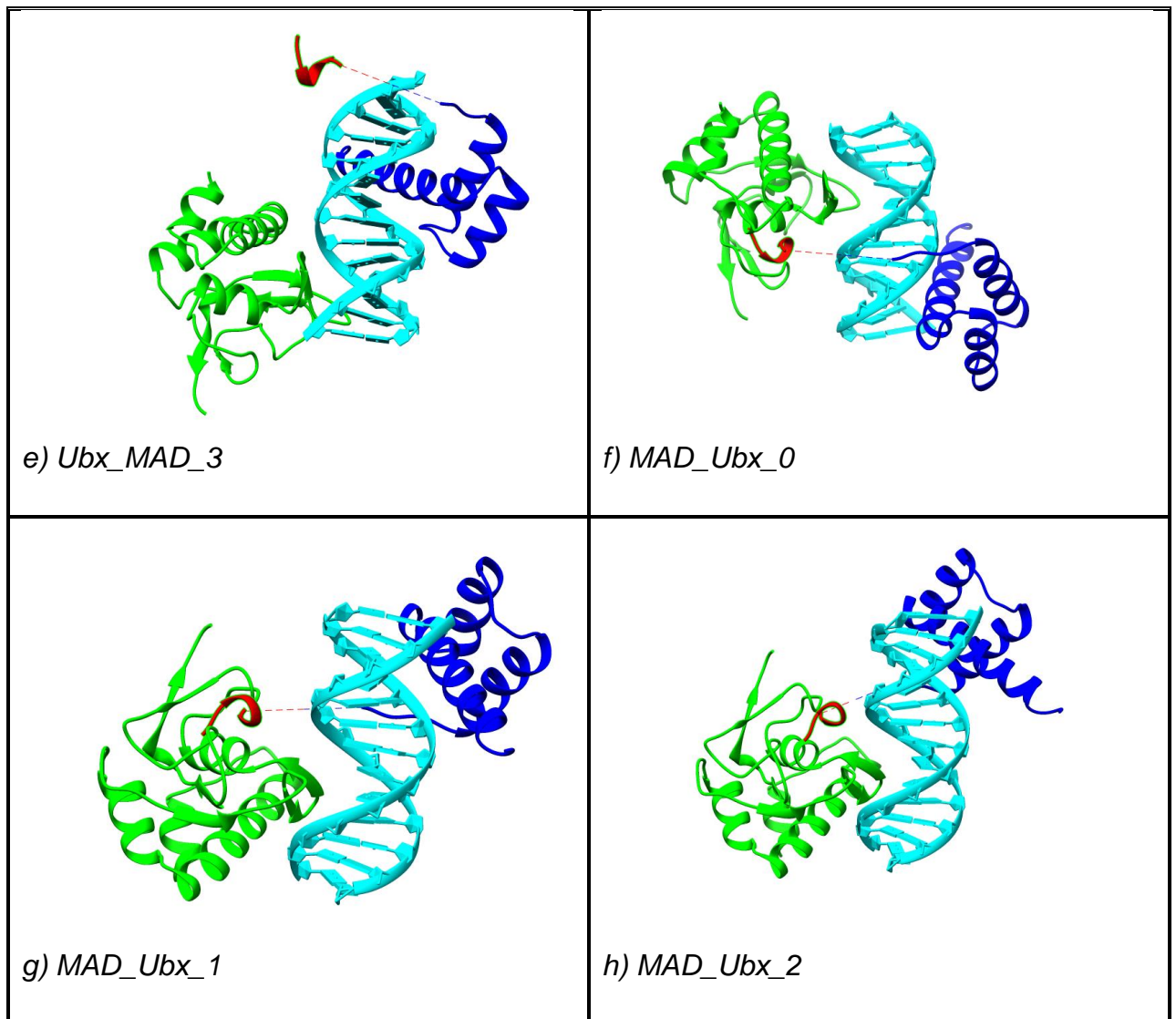


Figure 16 Ubx –MAD and DNA complex Green: MAD, Blue:Ubx,Cyan:DNA double helix, Red:Hexapeptide(HX) image e) is Ubx and MAD binding sites on Same strands with 3 base separation. Figure f), g), h) reads MAD and Ubx binding sites on same strand with separation of 0, 1, 2 bases respectively .

Here in figure 16 case e) the separation between the binding sites of Ubx and MAD is three bases, and there is no hindrance caused by the DNA binding residues. However, in case of f), g), h) Ubx binding site is after MAD binding site, and this positioning allows HX to access a different part of MAD to interact.

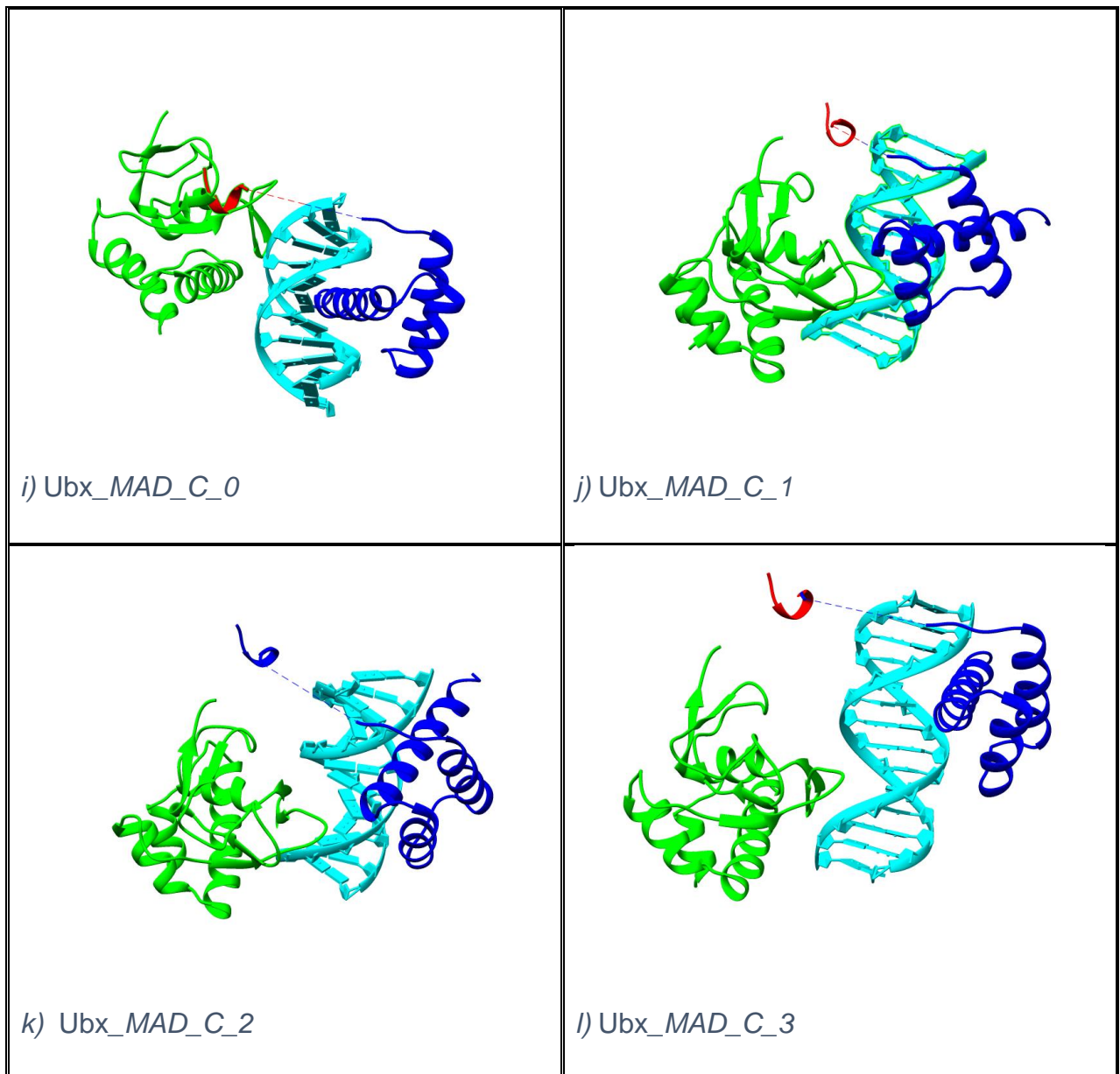


Figure 17 Ubx –MAD and DNA complex Green: MAD, Blue: Ubx, Cyan: DNA double helix, Red: Hexapeptide(HX) image i) is Ubx and MAD binding sites on complementary strands with 0 base separation. Similarly, the legends j), k), l) are Ubx and MAD binding sites on complementary strands with 1,2,3 base separation respectively.

In this figure 17, the MAD binding site is on the opposite strand that of Ubx binding site. This allows interaction with the different side of the MAD which is same as the accessible site of MAD in case MAD_Ubx (figure 16). However, there are clashes between DNA binding residues of Ubx and MAD in case j) which is one base separation, which lowers the possibility of binding Ubx and MAD on opposite strands

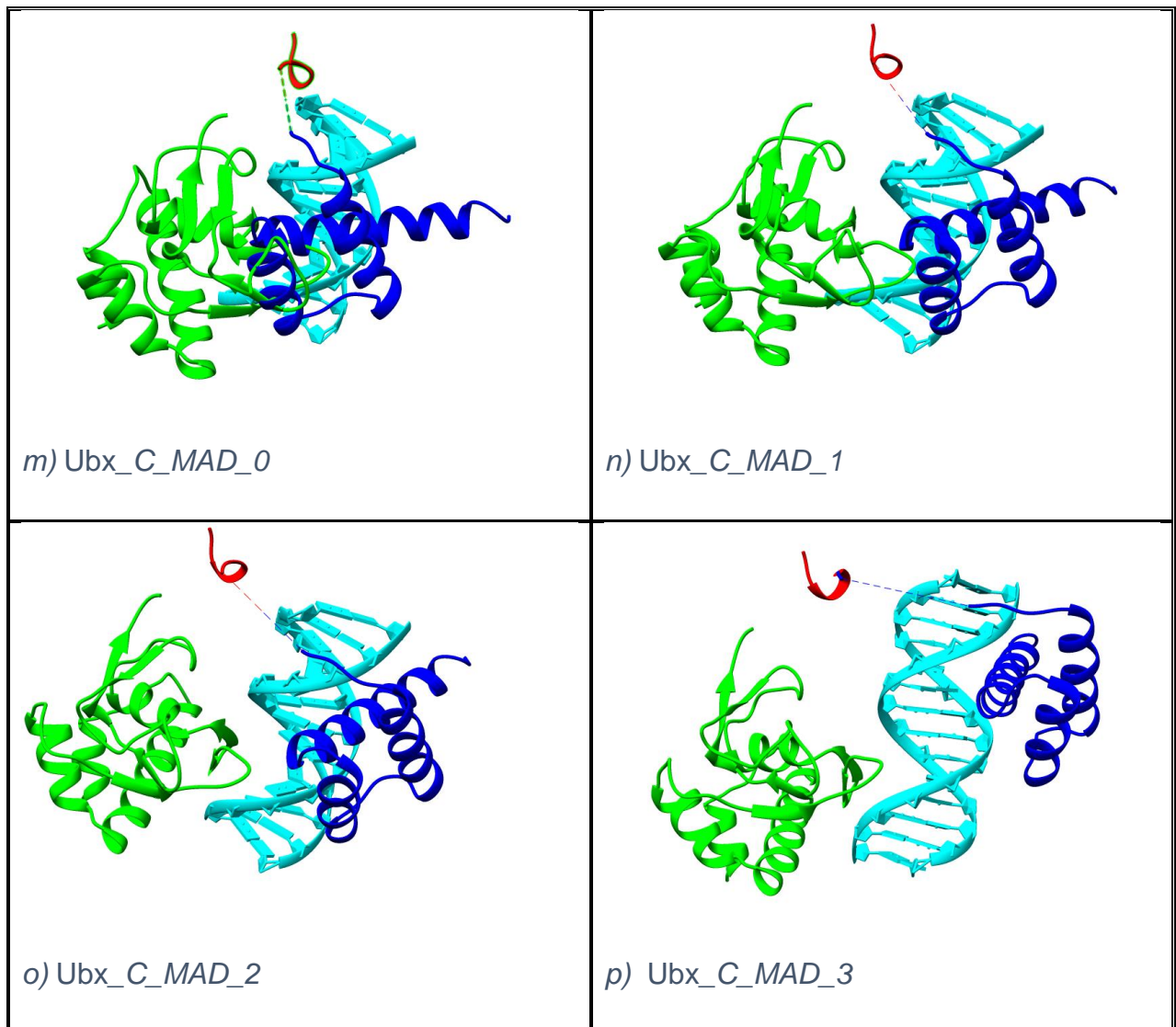


Figure 18 Ubx –MAD and DNA complex Green: MAD, Blue:Ubx, Cyan:DNA double helix, Red:Hexapeptide(HX) image m) is Ubx and MAD binding sites on complementary strands with 0 base separation. Similarly the legends n),o),p) are Ubx and MAD binding sites on complementary strands with separation of 1,2,3 bases respectively.

with one base separation. In cases k) and l) both transcription factors are well separated but still can interact with the help of linker region.

In this figure 18, we have Ubx binding sites on the opposite strand which allows a similar region of MAD accessible to HX as that of Ubx and MAD_C cases.

Overall all the above images show the possible conformations in which HX from Ubx can interact with MAD with their binding sites located on same or different strands. There are two areas in which HX can interact with MAD to test the likelihood of interaction further molecular docking was performed.

Also, there is another protein from the MAD family known as Medea (Med) has similar sequence and structure as MAD, except for its DNA binding residues which

change its DNA binding sites. Otherwise, the MH1 domain of Med is also accessible to HX in the same way as of MAD. Med can interact with HX in the same way as of MAD.

3.3 Predicting the hexapeptide binding sites –

To check for the likelihood of hexapeptide (HX) of Ubx binding to different cofactors, molecular docking was performed using Protein-peptide docking methods.

Molecular docking of Hexapeptide(HX) on MAD, Exd, En gave the probable binding sites of HX. Here small peptide HX is docked over proteins to predict the binding conformation using auto-dock and Cabs dock tools.

MAD and HX Docking-

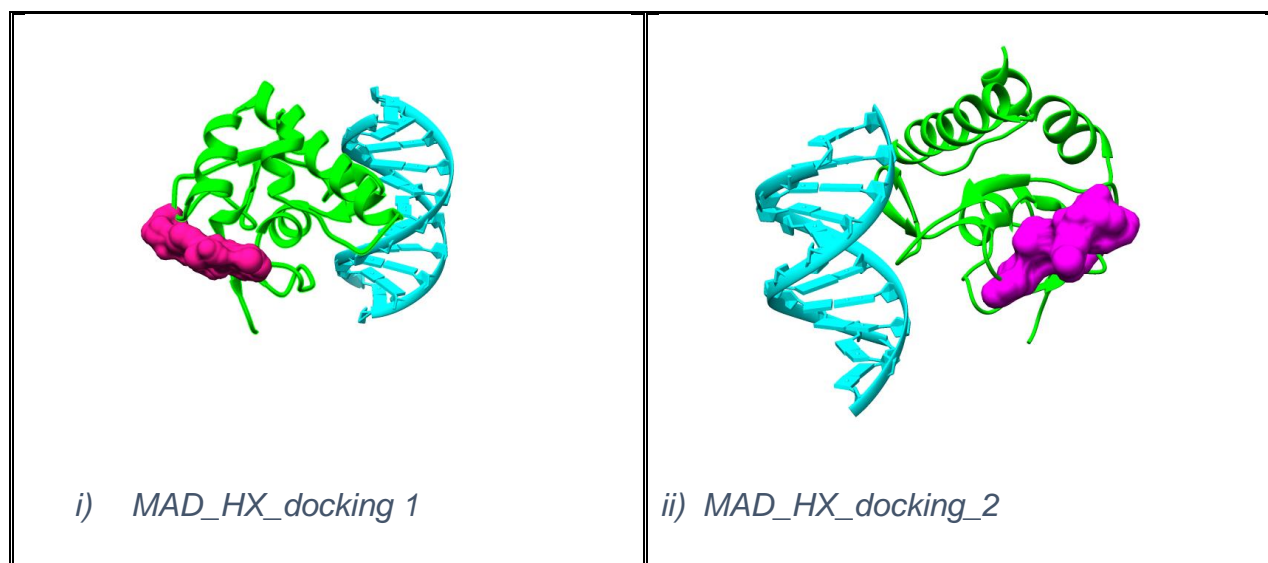


Figure 19 Hexapeptide(FYPWMA) docked over MAD, Green: MAD, Cyan: DNA double helix, Magenta: Hexapeptide

Figure 19 shows Hexapeptide in magenta cloured surface docked over MAD at two different positions. These are considered possible orientations in which HX can bind to MAD in the presence of DNA because other binding sites were present on the sites which were not likely to be accessible by HX when HX is linked to Ubx via linker region.

Extradenticle (Exd) and HX docking-

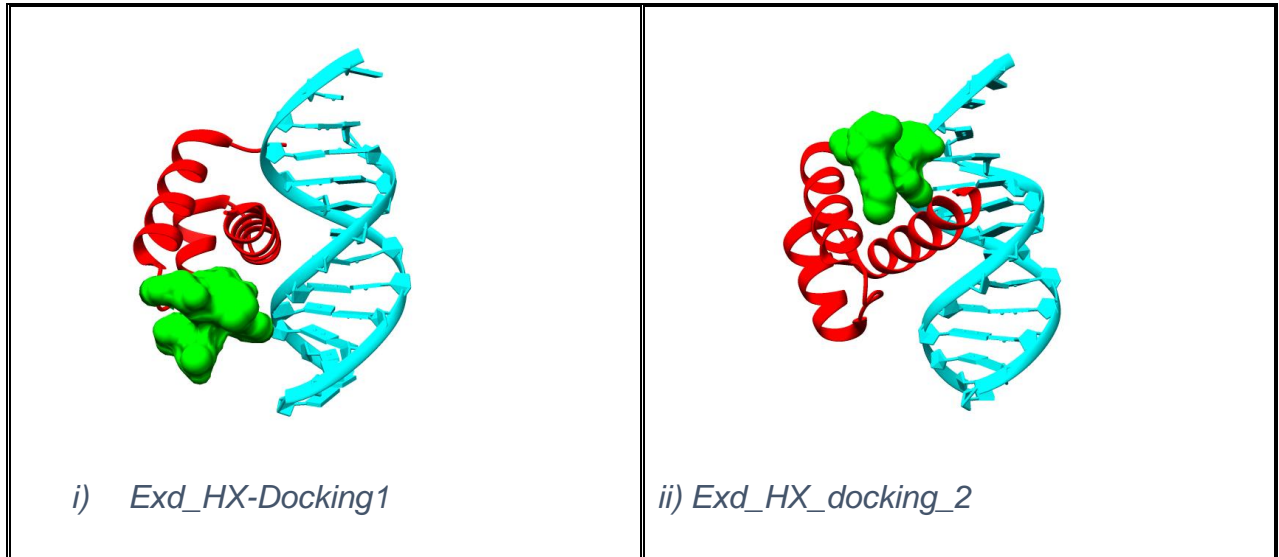


Figure 20 Hexapeptide(FYPWMA) docked in Exd, Red: Extradenticle, Cyan: DNA double helix, Green: Hexapeptide

To test the results obtained from the molecular docking, we checked for the known HX binding site on Exd. In the left panel, the binding of HX is same as mentioned the Ubx-Exd complex.(4CYC) Which added the confidence in the results. And towards the right is another region of Exd to which HX can bind obtained from molecular docking.

Engrailed (En) and HX docking-

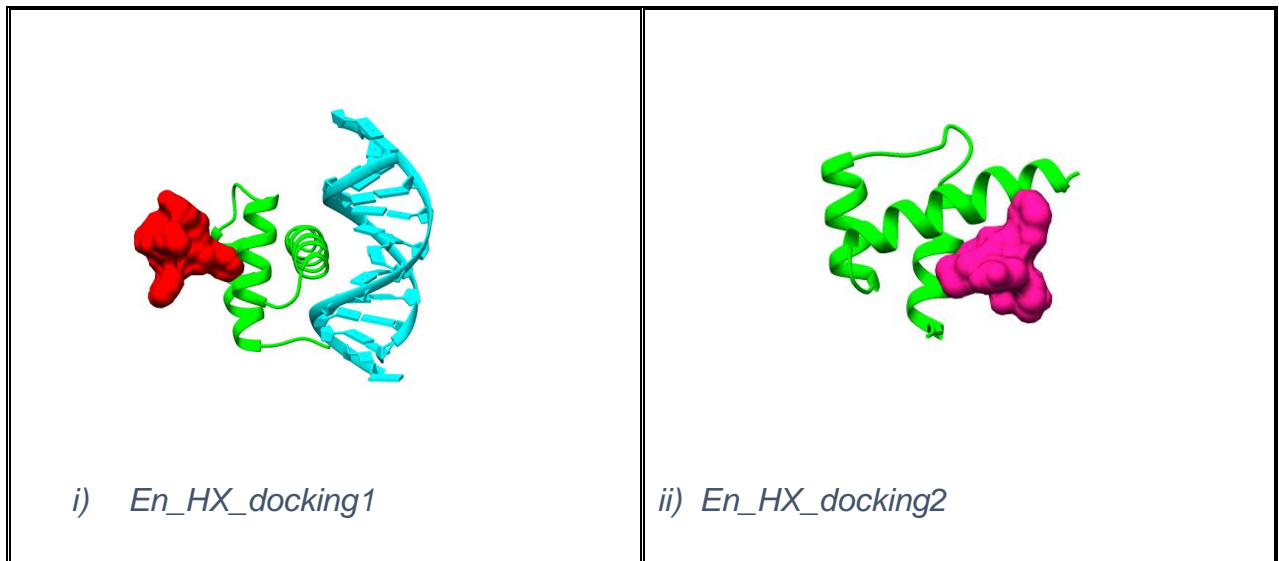


Figure 21 Hexapeptide(FYPWMA) docked over En, Green: engrailed(en), Cyan: DNA double helix, Red: Hexapeptide

Further to check for HX interacting with cofactors with the known structure we used En which is another homeodomain-containing protein. But in this case, only one conformation in which HX can interact with En which is shown in the left panel. Other results of Molecular docking (right panel) lies on the DNA binding residues of engrailed. This indicates fewer chances of Ubx and En interactions mediated by HX. Overall Molecular docking has added more confidence in the results obtained by structure modeling.

3.4 Identification of Putative target genes regulated by Ubx and its cofactors-

The presence of binding sites near each other can allow the proteins to interact. To find the instances in which the binding sites of Ubx and its cofactors like Exd, MAD, and Trithorax-like (Trl) are close to each other we searched for the binding sites over the entire genome. There are two ways by which this is done, first we searched for instances of occurrences of proteins and then calculated the distance between them.

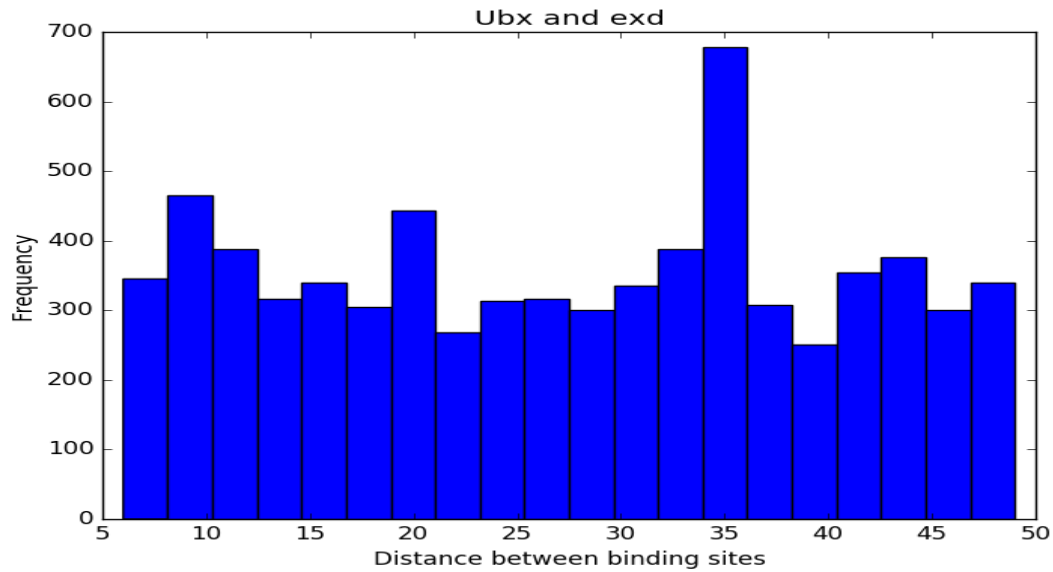


Figure 22 Frequency distribution of Ubx and Exd binding sites searched using single motif up to 50 bases. Y-axis represents the frequency of occurrence of binding sites and X-axis is the base separation between the binding sites.

This plot in figure22 shows Frequency of binding sites of Ubx and Exd separated up to 50 bases. Here single DNA binding sites were searched over the genome. The graph shows a uniform distribution of frequency except for slight increase at 35 base separation. Later we moved on to scanning entire Position weight matrix (PWM) of transcription factors over genome to look for the all possible occurrences. Position weight matrix of Ubx and cofactors together scanned over entire *Drosophila* genome with separation between them increased up to ten bases.

Ultrabithorax and Mothers against DPP-

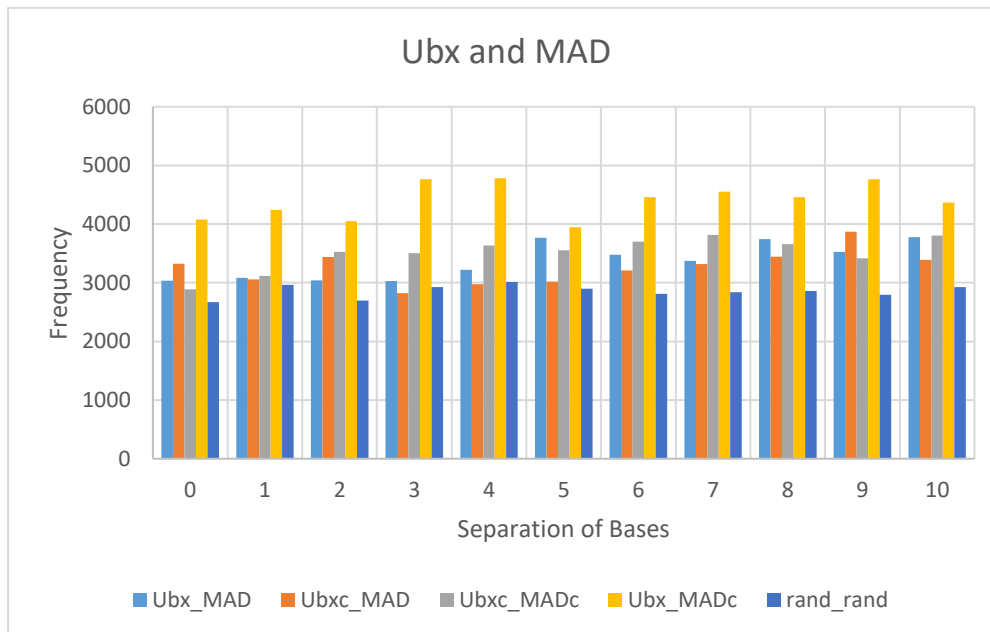


Figure 23 Ubx and MAD binding sites searched using PWM; Y-axis represents the frequency of occurrences of Ubx and MAD binding sites with the separation between them on X-axis increased up to ten bases. Here in the legends suffix 'c' indicates complimentary binding sequence.

Here Ubx is on one strand of DNA and MAD on the opposite which is Ubx_MADc indicated here in yellow have higher occurrences compared to that of other combinations of binding sites. The incidents of the randomly selected matrix are lower compared to the different combinations.

Ultrabithorax and Extradenticle -

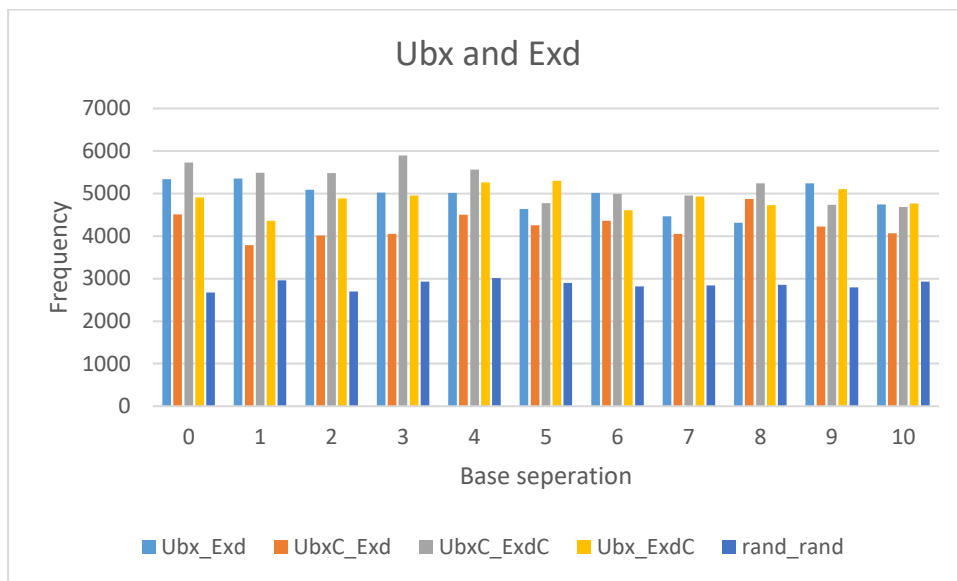


Figure 24 Ubx and Exd binding sites searched using PWM; Y-axis represents the frequency of occurrences of Ubx and Exd binding sites with the separation between them on X-axis increased up to ten bases. Here in the legends suffix 'c' indicates complimentary binding sequence.

The Ubx and Exd show much higher occurrence together which can be because of their similar binding sites. The above graph shows that up to four bases there are higher occurrences of Ubx_ext and UbxC_extC compared to other combinations of binding sites.

Ultrabithorax and Trithorax-like (Trl)-

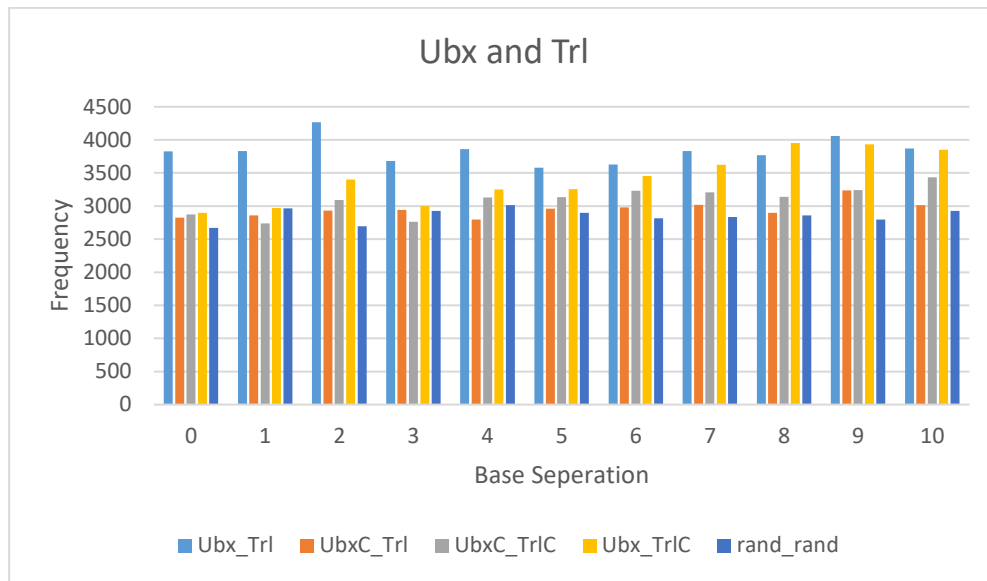


Figure 25 Ubx and Trl binding sites searched using PWM; Y-axis represents the frequency of occurrences of Ubx and Trl binding sites with the separation between them on X-axis increased up to ten bases. Here in the legends suffix 'c' indicates complimentary binding sequence.

In figure 25, there is a higher number of occurrences of Ubx_TrI binding sites on the same strands compared to the other combinations of binding sites indicated here in Blue. This suggests that these binding site combinations are preferred over others.

Based on above occurrences of binding sites we listed the target genes whose transcription start sites are downstream of these binding sites up to 1 Kb. Then we compared these list of genes with the genes obtained from Ubx ChIP-seq in the haltere tissue (performed by Soumen). The genes which are common to ChIP-seq data and above data can be the possible targets of Ubx regulated by Ubx and above cofactors in the context of haltere development.

Name of the Cofactor	No of genes identified using TFBS analysis	No of genes identified using ChIP-seq data.	No of common target genes
MAD	2801	833	278
Exd	3562	833	318

Table 4 Number of Putative target genes

Chapter 4

4 Discussion-

4.1 Evolutionary changes in Ubx and its cofactors-

Sequence alignments allow the comparison of Ubx divergence with that of its cofactors. It can be interpreted that the Ubx is evolving faster as compared to the cofactors. The changes in the Ubx sequences are in the non-homeodomain part of the protein as the homeodomain is conserved. Proteins like Exd, MAD, Pan, Hth have higher homology scores which suggest their relatively low rate of change of proteins. Here bit score per aligned length of Hth and MAD across *Drosophila* species are close to one whereas in case of Ubx it ranges from 0.57 to 1. Also, comparison with unrelated protein CPR450 also showed higher conservation than Ubx. This suggests that Ubx is diverging faster in comparison with other proteins.

Despite changes in the sequence the function of Ubx is conserved which can be because of the conserved Homeodomain and the changes in the non-homeodomain part of Ubx can allow Ubx to interact with different cofactors in different species. There is large region of Ubx which is structurally disordered which might be allowing Ubx to interact with wide range of cofactors as well.

Multiple sequence alignments of Ubx, Exd, MAD, and En show that the DNA binding domains of all these proteins are conserved across *Apis*, *Bombyx*, *Tribolium*. Thus the DNA binding sites would also be conserved if there are no changes in the DNA binding residues of the proteins. But there are slight changes in DNA binding residues of MAD, and that can give rise to differential preferences towards the DNA binding sites. For other cases, we can use same DNA binding sites while doing the T.F. binding sites analysis.

The MSA of Ubx showed us that the length of linker region is longest in case of *Drosophila* which could be the reason behind Ubx interacting with different cofactors. The longer linker region in *Drosophila* is allowing it to interact with cofactors with greater distance separation between binding sites.

Also, there are differences in the isoforms of Ubx in *Drosophila*, and the prevalent isoform in Haltere (Ubx IA) has longer linker region compared to other Ubx isoforms. Previous studies have shown that expression of another isoform in T3

segment causes partial wing like phenotype, this might explain the significance of the length of the linker region in the Ubx mediated regulation of target genes in the context of haltere. The linker can act as a restraining factor while HX is interacting with cofactors, more extended linker region suggests the possibility of interaction over long range.

4.2 Hexapeptide bindings with Cofactors -

Ubx and *exd* in *Drosophila* are known to interact, and the same interaction is expected in other insects. In cases where the binding sites separation is larger, then as the linker region in *Bombyx* is just seven residues the likelihood of interaction decreases. In *Drosophila melanogaster* the length of the linker is 50 residues in comparison with seven residues in *Bombyx mori*. The length of linker region can act as a constraint during the interaction. How far the linker can reach to mediate the interaction of HX cannot be predicted accurately but based on the relative difference in the length of linker among different species, we can comment upon the separation between binding sites. Also, it should be considered that the ability of linker to mediate HX interaction depends on the secondary structure of it, bending of DNA and charge on neighboring residues.

In case of MAD, we tried to find novel interaction as there is evidence for Ubx and MAD proteins together regulating downstream genes. While finding these novel interaction mediated by HX. We came across potential cases in the interaction can happen which includes the cases when Ubx and MAD are binding on opposite strands of DNA. The molecular docking results have given a likely binding region of HX which increases the likelihood of Ubx interacting with MAD when bound in the specific orientation. Similarly, another protein from MAD family Med which has a similar MH1 domain as of MAD can also interact with Ubx with the help of HX the same way as MAD is interacting. The difference in the MAD and Med are in the DNA binding sites. MAD and Med are known to form dimer so to find potential target genes regulated by these proteins one can also look for the presence of all three bind sites together and genes present downstream to these binding sites.

The homeodomain-containing protein such as Exd interacting with Ubx using HX also suggests the possibility of Ubx forming homodimer. The challenge in the case of homology modeling is that we can only predict whether the HX is interacting with the DNA binding domain of another protein because of unavailability of homologous structures. There are possibilities of interaction of HX with non-DNA binding domains of proteins as well.

Molecular docking provides plausible binding regions of HX over cofactors, In the case of MAD and HX molecular docking, resulted into two conformations multiple times and that can imply these interactions to be most likely, and combining that result with the modeled structures increases confidence in the model. But there are limitations to the accuracy of molecular docking in case of peptides. To support above results we again used same tools for HX and Exd which resulted in some models with binding sites same as experimentally determined structures of the Ubx-Exd complex. Molecular docking with Engrailed (En) gave us just one binding site which was not part of the DNA binding domain of En. This suggests that HX can only bind to the region given the orientation of both the proteins allows to bind to DNA which is less likely. In case of En, this is also likely that HX can restrict En from binding to DNA. Molecular docking has helped in finding the plausible binding sites of HX and the cofactors.

4.3 Identification of the target genes based on optimum TFBS separations-

For interaction of Ubx and Cofactors, their binding sites should be close, which can allow them to interact and regulate transcription of downstream genes. The occurrence of binding sites of T.Fs which are collaborating with each other should be higher than compared to random TFBS.

In case of Ubx and MAD, we scanned PWM's over entire genome and the results provided with Ubx and MAD compliment combination with a higher number of occurrences which hints at the higher possibility of Ubx and MAD interacting when bound in that arrangement.

A similar comparison of Ubx and Exd also provided with certain arrangement of binding sites which are more frequent than others. Scanning of binding sites of Ubx and Trl also have a higher frequency for the adjacent arrangement of Ubx and Trl.

The binding sites analysis has given us putative target genes. Presence of binding sites may not necessary have downstream gene as the target gene. To find out target genes in the context of haltere development, ChIP-seq data provides the list of genes of interest. The common target genes between above two datasets are the genes of interest to study Ubx-cofactor interactions.

4.4 Future Directions-

The series of analysis of the interactions between Ubx and cofactor can help us in understanding their combined role in the development of haltere. This analysis can further be extended to another protein which are expected to interact with Ubx. Finding the interactions will help in finding putative genes which are being regulated by Ubx and its cofactors.

In the context of Ubx and its cofactor mediated regulation of target genes, to further find out target genes in case of the development of haltere. The list of genes provided by the analysis of binding sites can further narrowed down using genes which are expressed during the haltere development by using ChIP-seq data. Also by extending the similar analysis in *Apis*, *Bombyx* and *Tribolium* will provide us with a list of genes regulated by the pair of transcription factors and ChIP-seq data for these insects will narrow down the list of potential target genes in a similar way.

Comparison of these sets of genes from *Apis*, *Bombyx*, and *Tribolium* with that of *Drosophila* will give us genes which are differentially expressed and more likely to be regulated by Ubx and the cofactors. The difference in cofactor binding signatures across species can help in classifying the differentially regulated genes. The genes which are different between *Drosophila* and other species can be directly related to haltere specification.

Further, we can check the expression patterns of these potential target genes in comparison with Ubx to find out functionally relevant genes. These functionally relevant genes obtained can be used to study Ubx-cofactor interactions. To further validate the biological significance of these cofactors genetic and biochemical studies can be performed. This will help in providing novel insights into the mechanism of regulation of genes by Ubx.

Chapter- 5

5. References-

Agrawal, Pavan, Farhat Habib, Ramesh Yelagandula, and L. S. Shashidhara. 2011a. "Genome-Level Identification of Targets of Hox Protein Ultrabithorax in *Drosophila*: Novel Mechanisms for Target Selection." *Scientific Reports* 1 (1): 205. doi:10.1038/srep00205.

———. 2011b. "Genome-Level Identification of Targets of Hox Protein Ultrabithorax in *Drosophila*: Novel Mechanisms for Target Selection." *Scientific Reports* 1 (i): 1–10. doi:10.1038/srep00205.

Akam, M E, and a Martinez-Arias. 1985. "The Distribution of Ultrabithorax Transcripts in *Drosophila* Embryos." *The EMBO Journal* 4 (7): 1689–1700. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=554405&tool=pmcentrez&rendertype=abstract>.

Altschul, Stephen F. 1991. "Amino Acid Substitution Matrices from an Information Theoretic Perspective," 555–65.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. doi:10.1093/nar/25.17.3389.

Altschup, Stephen F, Warren Gish, The Pennsylvania, and University Park. 1990. "Basic Local Alignment Search Tool 2Department of Computer Science," 403–10.

Błaszczuk, Maciej, Mateusz Kurcinski, Maksim Kouza, Lukasz Wieteska, Aleksander Debinski, Andrzej Kolinski, and Sebastian Kmiecik. 2016. "Modeling of Protein-Peptide Interactions Using the CABS-Dock Web Server for Binding Site Search and Flexible Docking." *Methods* 93. Elsevier Inc.: 72–83. doi:10.1016/j.ymeth.2015.07.004.

Etheve, Löic, Juliette Martin, and Richard Lavery. 2015. "Dynamics and Recognition

- within a Protein-DNA Complex: A Molecular Dynamics Study of the SKN-1/DNA Interaction." *Nucleic Acids Research* 44 (3): 1440–48. doi:10.1093/nar/gkv1511.
- Foos, Nicolas, Corinne Maurel-Zaffran, María Jesús Maté, Renaud Vincentelli, Matthieu Hainaut, Hélène Berenger, Jacques Pradel, Andrew J. Saurin, Miguel Ortiz-Lombardía, and Yacine Graba. 2015. "A Flexible Extension of the Drosophila Ultrabithorax Homeodomain Defines a Novel Hox/PBC Interaction Mode." *Structure* 23 (2): 270–79. doi:10.1016/j.str.2014.12.011.
- Hudry, Bruno, Sophie Remacle, Marie Claire Delfini, René Rezsöhazi, Yacine Graba, and Samir Merabet. 2012. "Hox Proteins Display a Common and Ancestral Ability to Diversify Their Interaction Mode with the Pbc Class Cofactors." *PLoS Biology* 10 (6). doi:10.1371/journal.pbio.1001351.
- Hughes, C L, and T C Kaufman. 2002. "Hox Genes and the Evolution of the Arthropod Body Plan." *Evolution & Development* 4: 459–99. doi:DOI 10.1046/j.1525-142X.2002.02034.x.
- Knoepfler, P S, and M P Kamps. 1995. "The Pentapeptide Motif of Hox Proteins Is Required for Cooperative DNA Binding with Pbx1, Physically Contacts Pbx1, and Enhances DNA Binding by Pbx1." *Molecular and Cellular Biology* 15 (10): 5811–19.
- Kurcinski, Mateusz, Michal Jamroz, Maciej Blaszczyk, Andrzej Kolinski, and Sebastian Kmiecik. 2015. "CABS-Dock Web Server for the Flexible Docking of Peptides to Proteins without Prior Knowledge of the Binding Site." *Nucleic Acids Research* 43 (W1): W419–24. doi:10.1093/nar/gkv456.
- Lewis, D L, M DeCamillis, and R L Bennett. 2000. "Distinct Roles of the Homeotic Genes Ubx and Abd-A in Beetle Embryonic Abdominal Appendage Development." *Proceedings of the National Academy of Sciences of the United States of America* 97 (9): 4504–9. doi:10.1073/pnas.97.9.4504.
- Lewis, E B. 1982. "Control of Body Segment Differentiation in Drosophila by the Bithorax Gene Complex." *Progress in Clinical and Biological Research* 85 Pt A: 269–88. <http://www.ncbi.nlm.nih.gov/pubmed/7111279>.
- Liu, Ying, Kathleen S. Matthews, and Sarah E. Bondos. 2008. "Multiple Intrinsically

- Disordered Sequences Alter DNA Binding by the Homeodomain of the *Drosophila* Hox Protein Ultrabithorax.” *Journal of Biological Chemistry* 283 (30): 20874–87. doi:10.1074/jbc.M800375200.
- Maconochie, M, S Nonchev, a Morrison, and R Krumlauf. 1996. “Paralogous Hox Genes: Function and Regulation.” *Annual Review of Genetics* 30: 529–56. doi:10.1146/annurev.genet.30.1.529.
- Mann, Richard S., Katherine M. Lelli, and Rohit Joshi. 2009. *Chapter 3 Hox Specificity. Unique Roles for Cofactors and Collaborators. Current Topics in Developmental Biology*. 1sted. Vol. 88. Elsevier Inc. doi:10.1016/S0070-2153(09)88003-4.
- Mart, Marc A, Ashley C Stuart, S Roberto, Francisco Melo, and S Andrej. 2000. “C P S M G G,” 291–325.
- Masumoto, Mika, and Toshinobu Yaginuma. 2009. “Functional Analysis of Ultrabithorax in the Silkworm , *Bombyx Mori* , Using RNAi,” 437–44. doi:10.1007/s00427-009-0305-9.
- Matys, V. 2006. “TRANSFAC(R) and Its Module TRANSCompel(R): Transcriptional Gene Regulation in Eukaryotes.” *Nucleic Acids Research* 34 (90001): D108–10. doi:10.1093/nar/gkj143.
- McWilliam, Hamish, Weizhong Li, Mahmut Uludag, Silvano Squizzato, Young Mi Park, Nicola Buso, Andrew Peter Cowley, and Rodrigo Lopez. 2013. “Analysis Tool Web Services from the EMBL-EBI.” *Nucleic Acids Research* 41 (Web Server issue): 597–600. doi:10.1093/nar/gkt376.
- Merabet, Samir, and Ingrid Lohmann. 2015. “Toward a New Twist in Hox and TALE DNA-Binding Specificity.” *Developmental Cell* 32 (3). Elsevier Inc.: 259–61. doi:10.1016/j.devcel.2015.01.030.
- Mohit, Prasad, Kalpana Makhijani, M. B. Madhavi, V. Bharathi, Ashish Lal, Gururaj Sirdesai, V. Ram Reddy, et al. 2006. “Modulation of AP and DV Signaling Pathways by the Homeotic Gene Ultrabithorax during Haltere Development in *Drosophila*.” *Developmental Biology* 291 (2): 356–67. doi:10.1016/j.ydbio.2005.12.022.

- Morris, Gm, and Ruth Huey. 2009. "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility." *Journal of ...* 30 (16): 2785–91. doi:10.1002/jcc.21256.AutoDock4.
- Navas, Luis F de, Hilary Reed, Michael Akam, Rosa Barrio, Claudio R Alonso, and Ernesto Sánchez-Herrero. 2011. "Integration of RNA Processing and Expression Level Control Modulates the Function of the *Drosophila* Hox Gene Ultrabithorax during Adult Development." *Development (Cambridge, England)* 138 (1): 107–16. doi:10.1242/dev.051409.
- Passner, Jonathan M, Hyung Don Ryoo, Leyi Shen, Richard S Mann, and Aneel K Aggarwal. 1999. "Letters to Nature Structure of a DNA-Bound Ultrabithorax ± Extradenticle Homeodomain Complex" 397 (February).
- Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. "UCSF Chimera - A Visualization System for Exploratory Research and Analysis." *Journal of Computational Chemistry* 25 (13): 1605–12. doi:10.1002/jcc.20084.
- Pieper, Ursula, Narayanan Eswar, Hannes Braberg, M S Madhusudhan, Fred P Davis, Ashley C Stuart, Nebojsa Mirkovic, et al. 2004. "MODBASE , a Database of Annotated Comparative Protein Structure Models , and Associated Resources" 32. doi:10.1093/nar/gkh095.
- Prasad, Naveen, Shreeharsha Tarikere, Dhanashree Khanale, Farhat Habib, and L S Shashidhara. 2016. "A Comparative Genomic Analysis of Targets of Hox Protein Ultrabithorax amongst Distant Insect Species." *Scientific Reports* 6 (May). Nature Publishing Group: 27885. doi:10.1038/srep27885.
- Rice, Peter, Lan Longden, and Alan Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics* 16 (6): 276–77. doi:10.1016/S0168-9525(00)02024-2.
- Sandelin, A. 2004. "JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles." *Nucleic Acids Research* 32 (90001): 91D–94. doi:10.1093/nar/gkh012.
- Shashidhara, L S, N Agrawal, R Bajpai, V Bharathi, and P Sinha. 1999. "Negative

Regulation of Dorsoventral Signaling by the Homeotic Gene Ultrabithorax during Haltere Development in *Drosophila*.” *Developmental Biology* 212: 491–502. doi:10.1006/dbio.1999.9341.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (539). doi:10.1038/msb.2011.75.

Tomoyasu, Yoshinori, Scott R. Wheeler, and Robin E. Denell. 2005. “Ultrabithorax Is Required for Membranous Wing Identity in the Beetle *Tribolium Castaneum*.” *Nature* 433 (7026): 643–47. doi:10.1038/nature03272.

Walldorf, U, P Binner, and R Fleig. 2000. “Hox Genes in the Honey Bee *Apis Mellifera*.” *Development Genes and Evolution* 210 (10): 483–92. doi:10.1007/s004270000091.

Walsh, C. M., and S. B. Carroll. 2007. “Collaboration between Smads and a Hox Protein in Target Gene Repression.” *Development* 134 (20): 3585–92. doi:10.1242/dev.009522.

Weatherbee, Scott D., and Sean B. Carroll. 1999. “Selector Genes and Limb Identity in Arthropods and Vertebrates.” *Cell* 97 (3): 283–86. doi:10.1016/S0092-8674(00)80737-0.

Weatherbee, Scott D., Georg Halder, Jaeseob Kim, Angela Hudson, and Sean Carroll. 1998a. “Ultrabithorax Regulates Genes at Several Levels of the Wing-Patterning Hierarchy to Shape the Development of the *Drosophila* Haltere.” *Genes and Development* 12 (10): 1474–82. doi:10.1101/gad.12.10.1474.

Weatherbee, Scott D, Georg Halder, Jaeseob Kim, Angela Hudson, and Sean Carroll. 1998b. “Ultrabithorax Regulates Genes at Several Levels of the Wing-Patterning Hierarchy to Shape the Development of the *Drosophila* Haltere,” 1474–82.

Zhang, Yang. 2008. “I-TASSER Server for Protein 3D Structure Prediction.” *BMC Bioinformatics* 9: 1–8. doi:10.1186/1471-2105-9-40.

Zhang, Zheng, Scott Schwartz, Lukas Wagner, and Webb Miller. 2000. “A Greedy

Algorithm for Aligning DNA Sequences.” *Journal of Computational Biology* 7 (1–2): 203–14. doi:10.1089/10665270050081478.

Blue- *Drosophila* Species

Green-*Bombyxmori* (BMORI)

Red- *Tribolium castaneum* (TCAST)

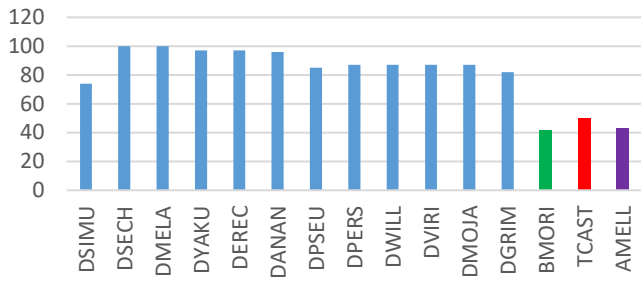
Purple-*Apis mellifera* (AMELL)

6 Annexure-1

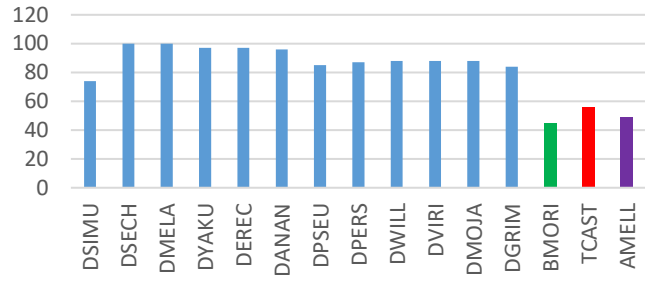
6.1 Divergence of Ubx in comparison with its cofactors-



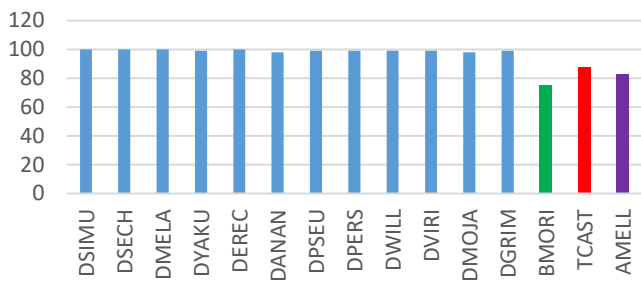
Ubx: Percentage Identities



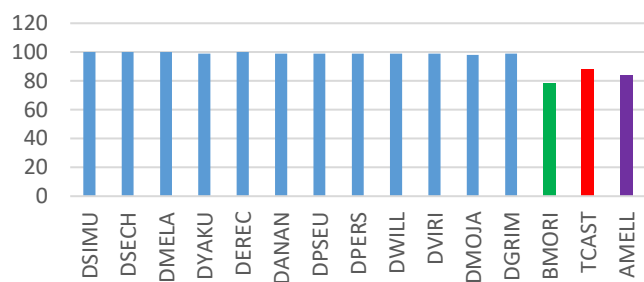
Ubx: Percentage Positives



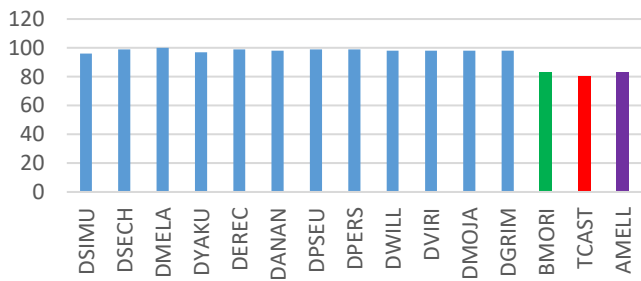
Exd: Percentage Identities



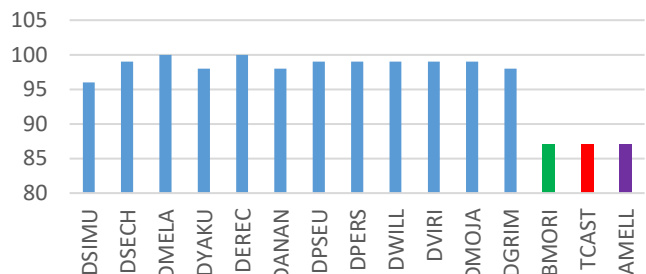
Exd: Percentage Positives



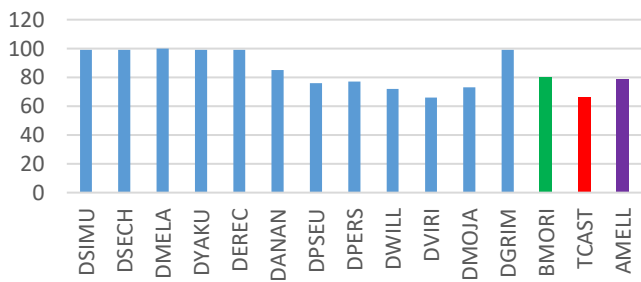
Mad: Percentage identities



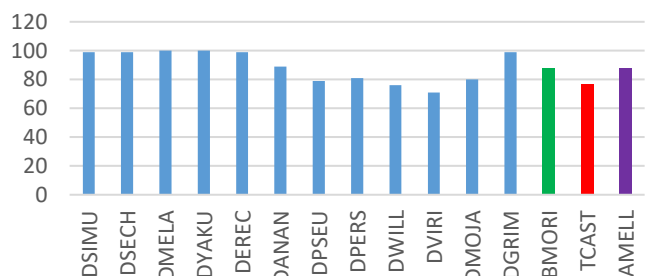
Mad: Percentage Positives



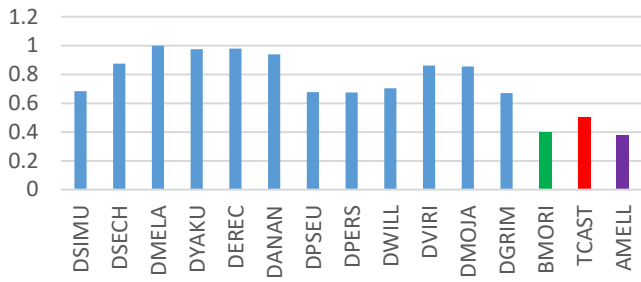
En: Percentage Identities



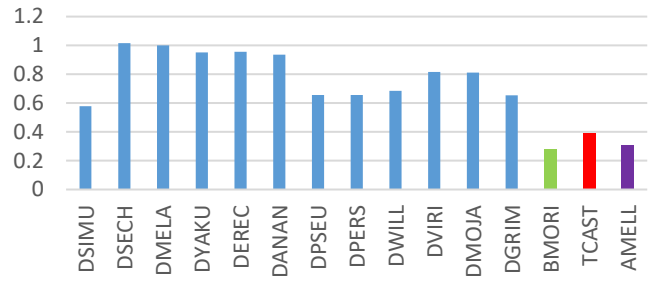
En: Percentage Positives



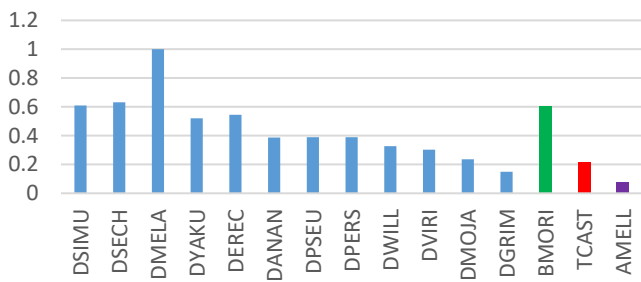
Ubx: bit score/minimal length



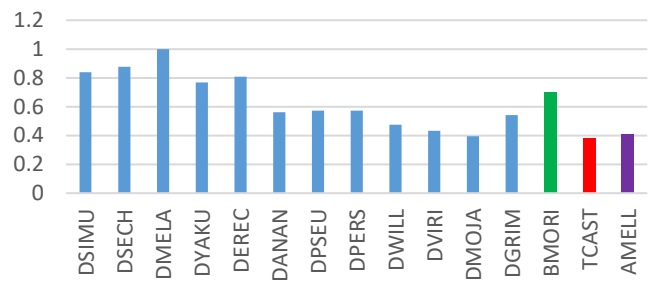
Ubx : bit score/aligned length



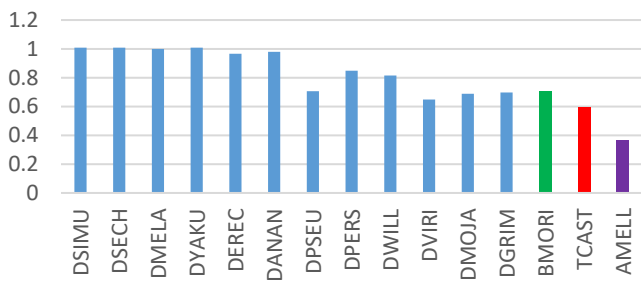
Pan: bit score/ minimal length



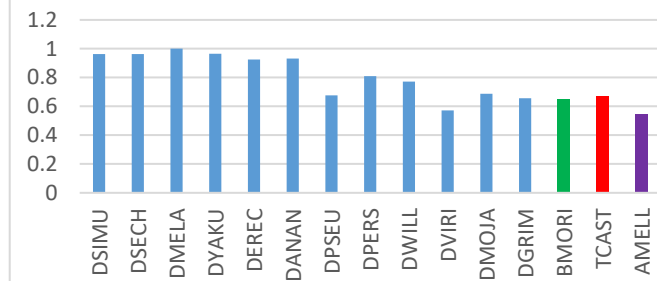
Pan :bit score/aligned length



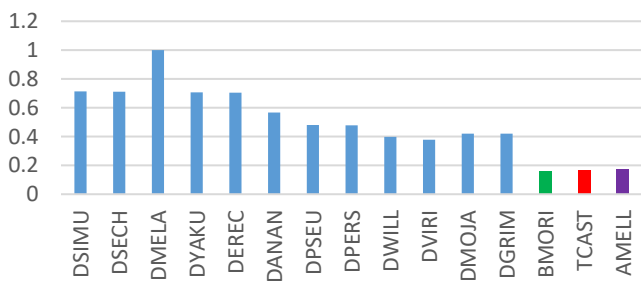
Elf-1: bit score/ minimal length



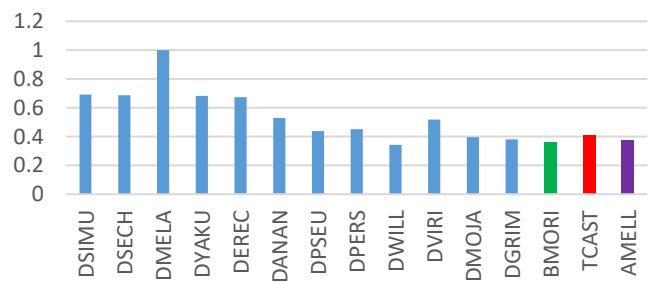
Elf-1: bit score/aligned length



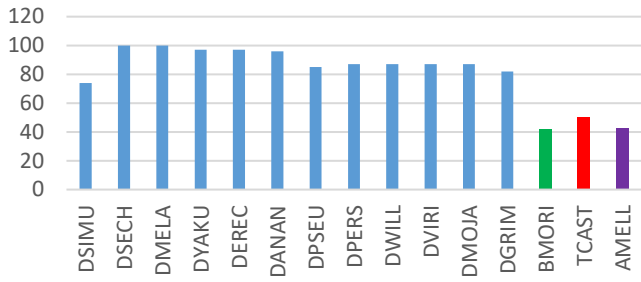
E2f1-bit score/ minimal length



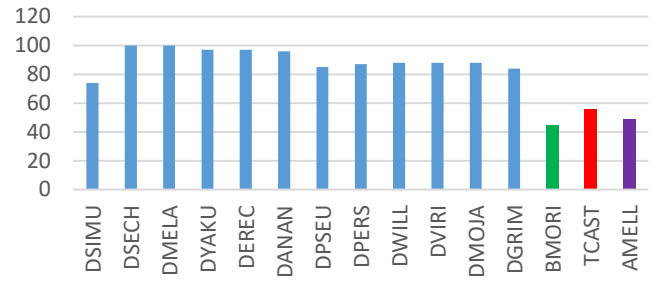
E2f-1: bit score/aligned length



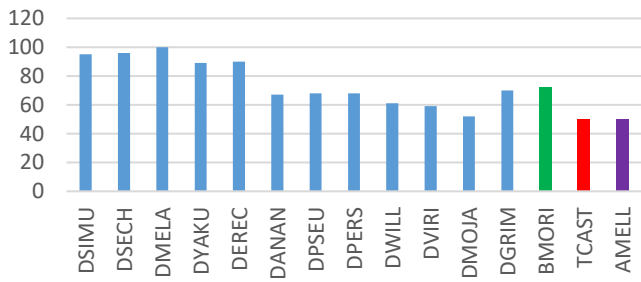
Ubx: Percentage Identities



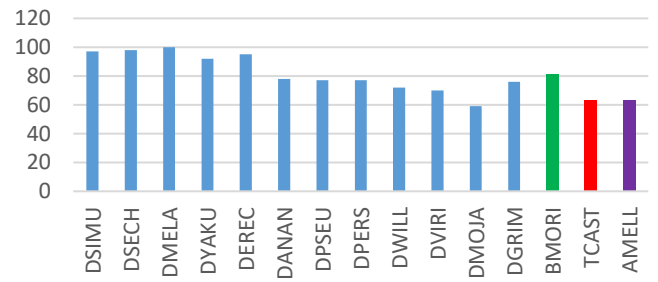
Ubx: Percentage Positives



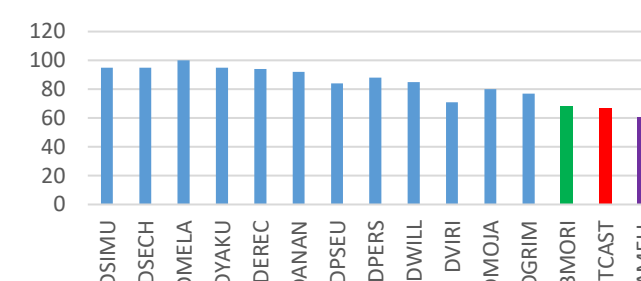
Pan: Percentage Identities



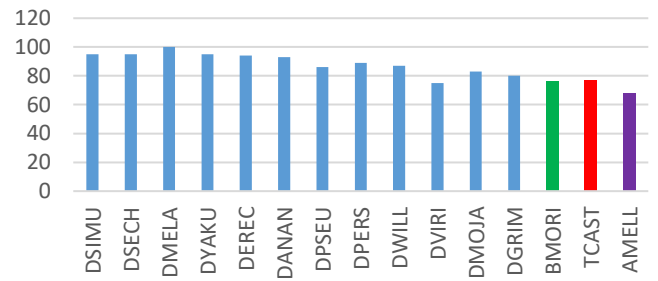
Pan: Percentage Positives



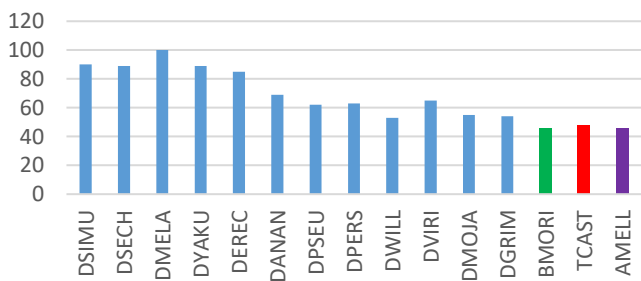
Elf-1: Percentage Identities



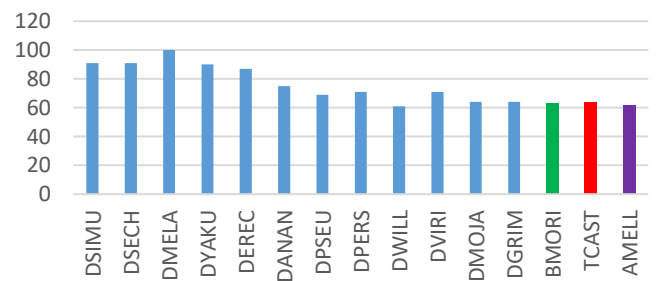
Elf-1: Percentage Positives



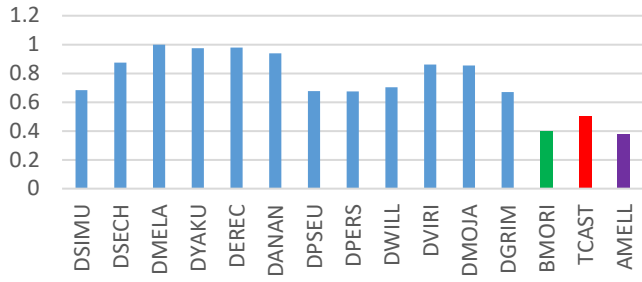
E2f-1: Percentage Identities



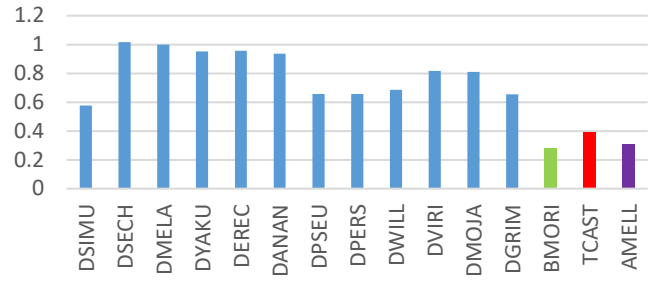
E2f-1: Percentage Positives



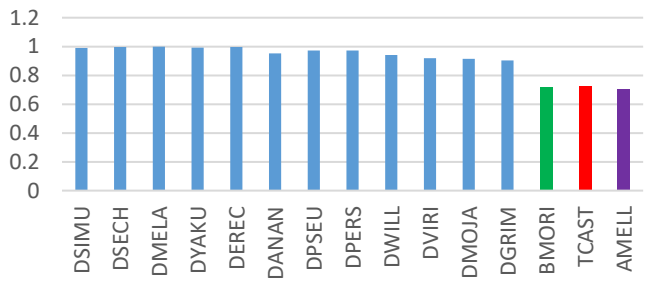
Ubx: bit score/minimal length



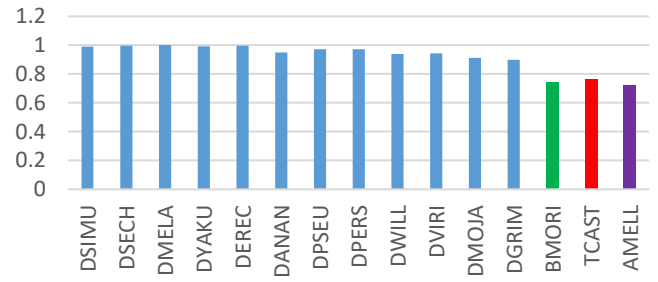
Ubx : bit score/aligned length



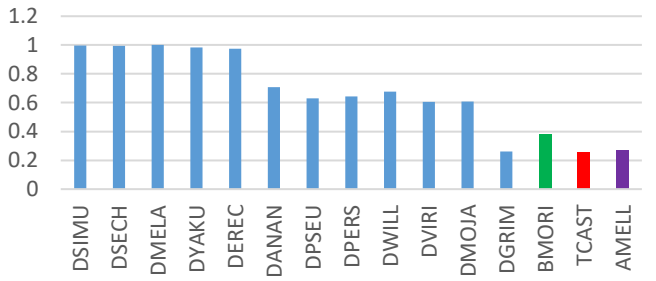
Cpr450: bit score/ minimal length



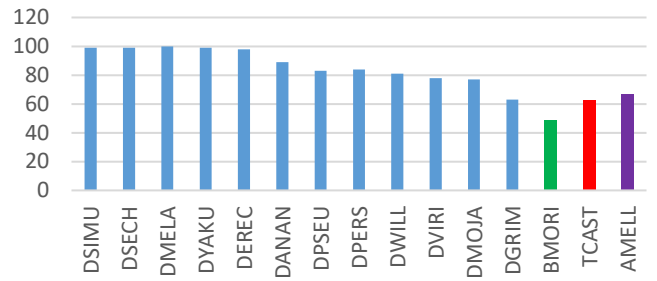
Cpr450: bit score/aligned length



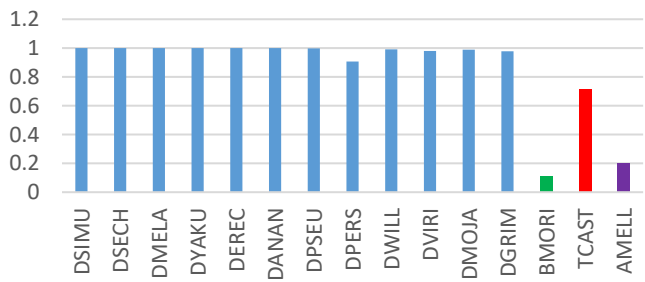
Hairy: bit score/ minimal length



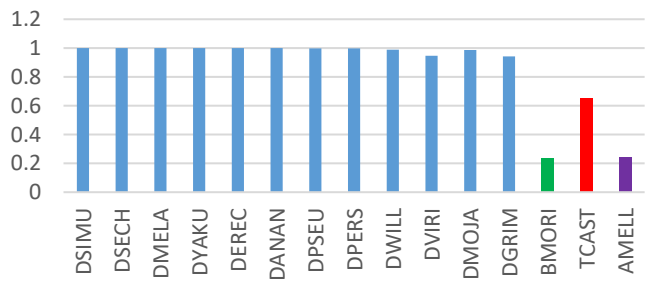
Hairy: Percentge Positives



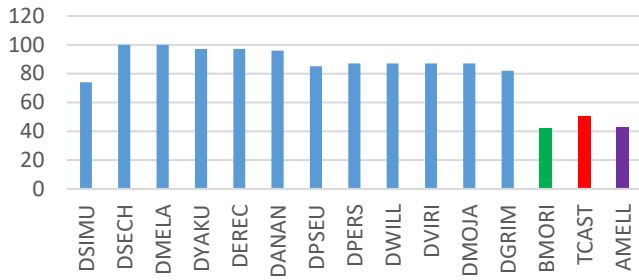
Hth:bit score/ minimal length



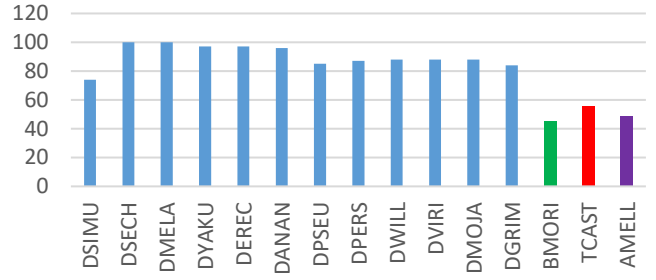
Hth: bit score/aligned length



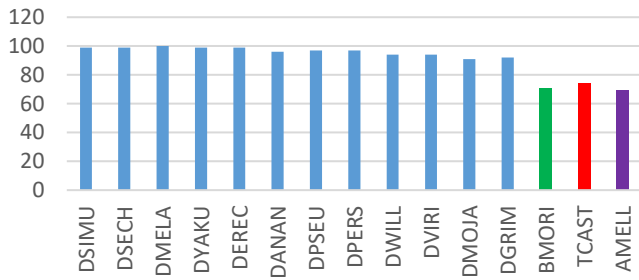
Ubx: Percentage Identities



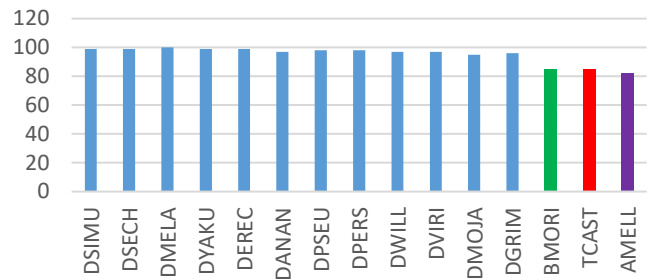
Ubx: Percentage Positives



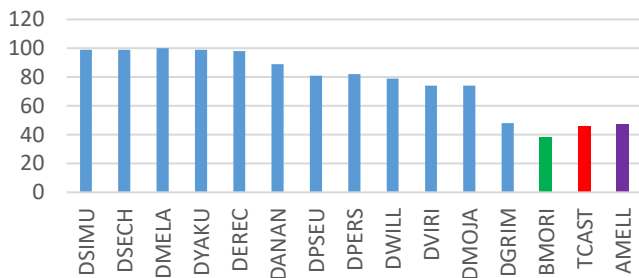
Cpr450: Percentage Identities



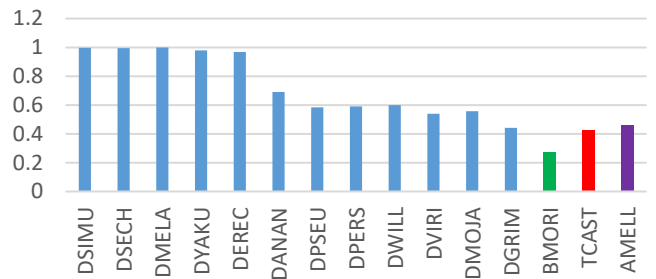
Cpr450: Percentage Positives



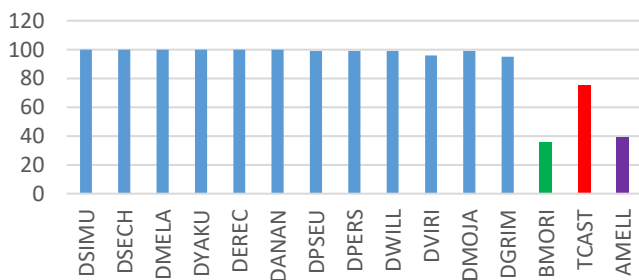
Hairy: Percentage Identities



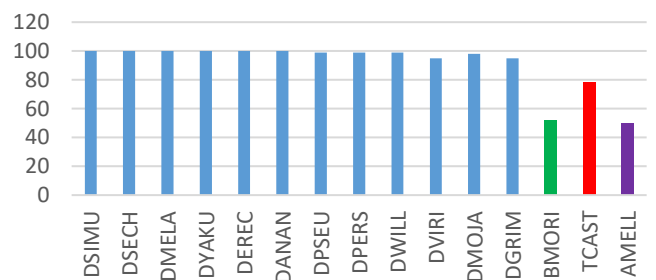
Hairy: bit score/aligned length



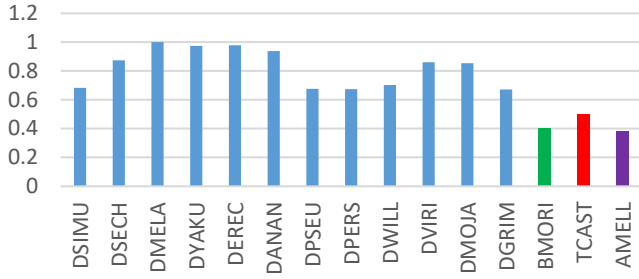
Hth: Percentage Identities



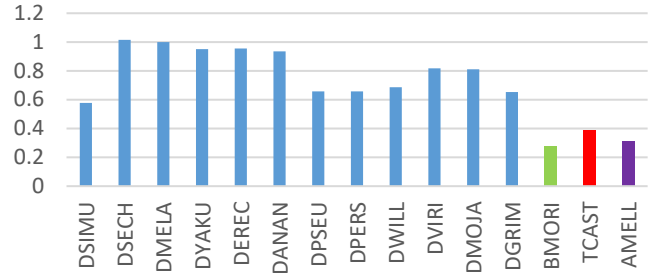
Hth: Percentage Positives



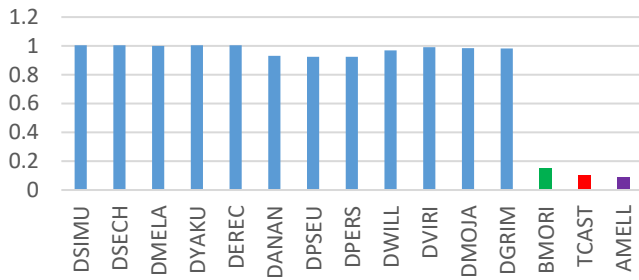
Ubx: bit score/minimal length



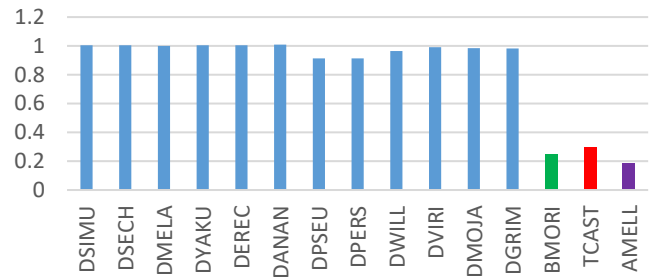
Ubx : bit score/aligned length



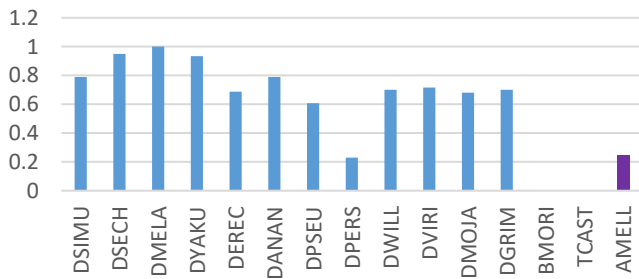
Adf-1:bit score/ minimal length



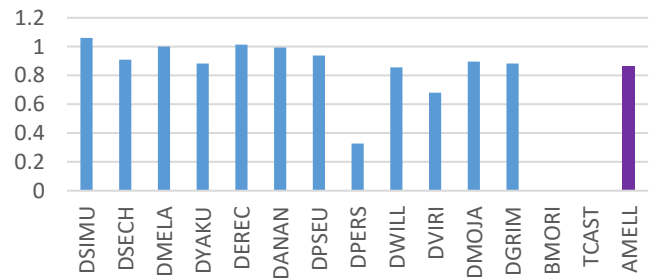
Adf-1:bit score/aligned length

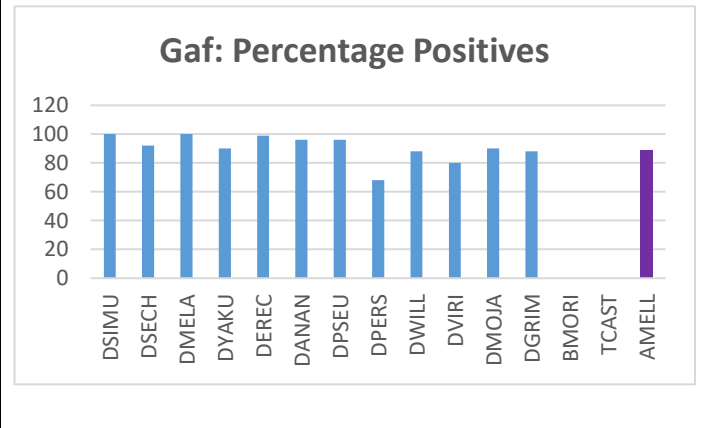
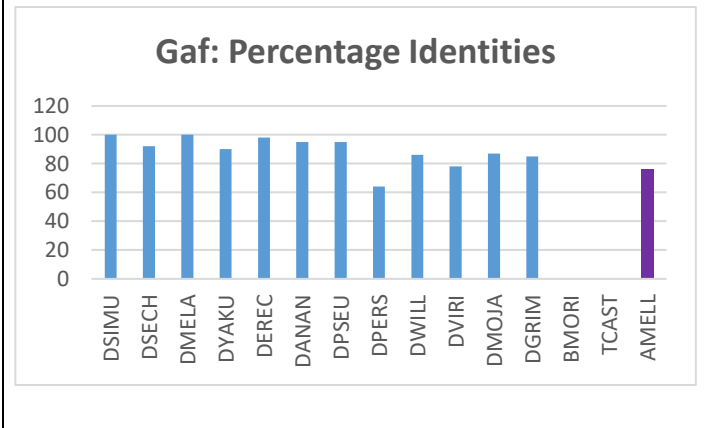
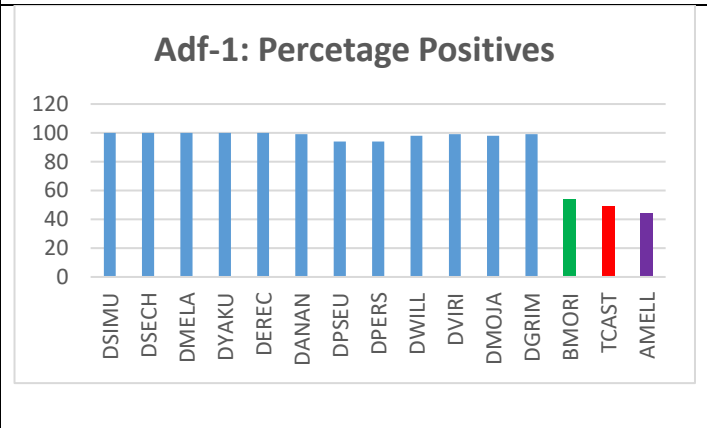
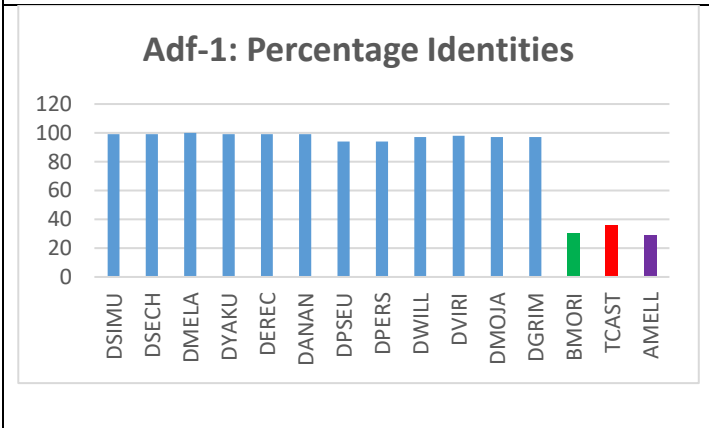
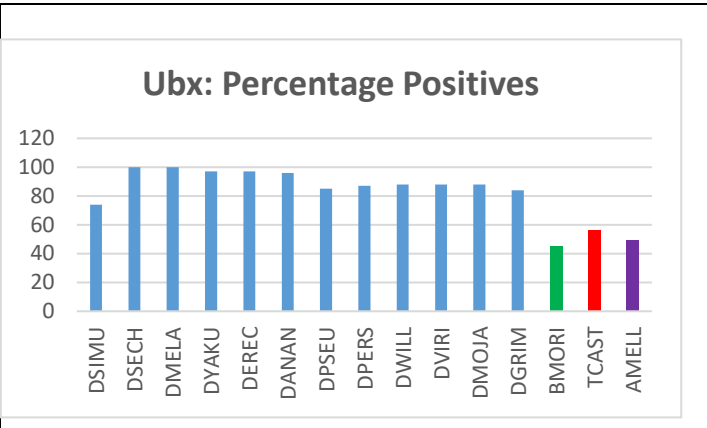
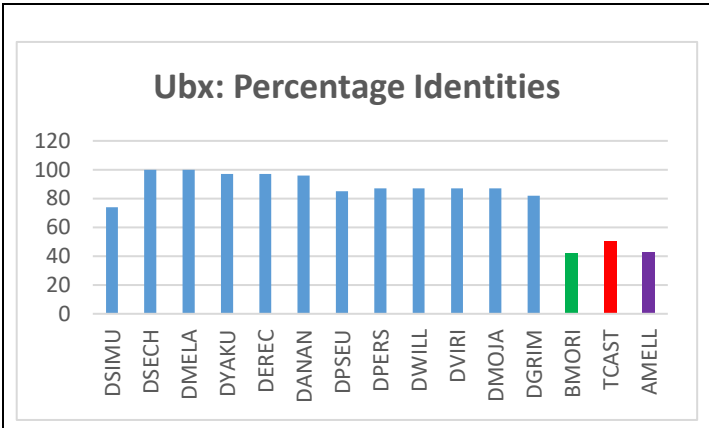


Gaf: bit score/ minimal length



Gaf:bit score/ aligned length





Comparison of Ubx and other cofactor proteins based on four different homology parameters across twelve *Drosophila* species and *Bombyxmori*, *Apis mellifera* and *Tribolium castaneum*

6.2 Sequence alignments-

2) Sequence alignments

Name	Length			
	DMELA	AMELL	TCAST	BMORI
Ubx	389	330	314	254
E2f1	805	450	457	456
Mad	455	468	468	422
Hth	487	485	456	388
Exd	376	418	259	409
En	552	349	284	132
Elf-1	1333	843	628	363
Myc	717	445	397	376
Gaf	581	474	a	a
Adf-1	262	515	390	210
Snail	390	525	364	272
Hairy	337	446	371	217
Dorsal	999	436	556	375
pan	751	860	220	130
slp-1	322	427	311	233

Local alignments				
	Percentage Identities			
	DMELA	AMELL	TCAST	BMORI
Ubx	100	43	50	42
e2f1	100	46	48	46
Mad	100	83	80	83
hth	100	39	75	36
exd	100	83	88	75
en	100	79	66	80
elf1	100	61	67	68
myc	100	46	44	36
GAF	100	76	a	a
Adf-1	100	29	23	30
snail	100	53	54	52
hairy	100	47	46	38
Dorsal	100	67	73	61
pan	100	50	50	72
Slp-1	100	51	75	50

Global Alignment			
Percentage Identities			
DMELA	AMELL	TCAST	BMORI
100	45	49	46
100	20	21	20
100	83	79	83
100	31	75	30
100	73	63	76
100	29	29	17
100	30	31	20
100	11	-	-
100	25	-	-
100	-	18	-
100	20	22	28
100	28	30	33
100	24	29	17
100	15	13	13
100	32	37	35

2) Cofactor alignment

	Exd	En	MAD	E2f1	Gaf	Myc
Exd			209-410	14-201		
			3-288	23-223		
En	237-301		135-238	438-745		
	453-514		251-404	62-416		
MAD				539-791		
				2-274		
E2f1						
Gaf	3-222	10-138	138-210	267-637		14-362
	281-474	418-556	333-403	60-565		145-497
Myc	98-374	21-497	141-370	283-754		
	61-306	290-713	212-431	37-550		

			Aligned Length			
	E2f1	En	Exd	Gaf	MAD	Myc
E2f1						
En	407		65		154	
Exd	218				298	
Gaf	527	150	245		75	405
MAD	291					
Myc	598	560	311		274	

			Identity			
	E2f1	En	Exd	Gaf	MAD	Myc
E2f1						
En	20.1		32.3		23.4	
Exd	21.1				18.8	
Gaf	19.4	31.3	20		29.3	19.5
MAD	22.3					
Myc	19.7	20.7	19.6		21.5	

			Similarity			
	E2f1	En	Exd	Gaf	MAD	Myc
E2f1						
En	32.7		52.3		39	
Exd	34.9				30.9	
Gaf	31.7	45.3	36.3		41.3	33.1
MAD	35.1					
Myc	31.6	32.1	34.1		33.6	

			Score			
	E2f1	En	Exd	Gaf	MAD	Myc
E2f1						
En	157.5		95		89	
Exd	72.5				59	
Gaf	122.5	148.5	70		46	113
MAD	88.5					
Myc	154.5	117	73		72	

			Gaps			
	E2f1	En	Exd	Gaf	MAD	Myc
E2f1						
En	35.4		4.6		32.5	
Exd	21.6				36.2	
Gaf	33.6	21.3	31		8	26.7
MAD	56					
Myc	35.1	39.1	31.8		35.8	