

Predicting *tbx22* Zebrafish Protein Structure Using Multi-Level Prediction Tools and Demonstration of Conserved Structural Domains in Relation to Orthologous *tbx22* Proteins in Humans

Vijay P. Boominathan, Tracie Ferreira

Department of Bioengineering, University of Massachusetts Dartmouth, North Dartmouth, MA, USA
Email: vboominathan@umassd.edu, tferreira@umassd.edu

Received 16 February 2016; accepted 23 March 2016; published 30 March 2016

Abstract

Biological functions of proteins play a key role in the development of any organism. The gene *tbx22* is a member of a phylogenetically conserved family of genes, which share a common DNA binding domain: T box. This study examines the similarity in the developmental pattern influenced by the transcription factor TBX22 and *tbx22* in *H. sapiens* and *D. rerio* respectively. Secondary and tertiary structures of the proteins are predicted using standard structure prediction software's like Phyre 2, Predict Protein, SWISSMODEL, PSIPRED and the homology of the proteins were compared to each other. Protein homology prediction shows more than 65% between the 2 organisms. Superimposing the predicted protein structures reveals conserved domains between the human and zebrafish proteins. Additional supporting data from Genomatix MATBASE, MATINSPECTOR show higher matrix family scores for BRAC (Brachury gene mesoderm developmental factor) in Human and Zebrafish. Transcription factor and promoter element analysis with Transcriptome Viewer, Gene 2 Promoter and Genomeinspector reveal a high degree of homology between the 2 organisms. Bioinformatic-Proteomics and protein structural analysis approaches shown here explain in detail the relationship between the Human and Zebrafish *tbx22* Gene-Protein-Transcription factor. These studies also support zebrafish as a predictive model for numerous developmental patterning events in higher vertebrates.

Keywords

T-box, *tbx22*, Transcription Factors, Genomatix, Protein Prediction, Superimposed Models

1. Introduction

This gene is a member of a phylogenetically conserved family of genes, which share a common DNA-binding domain T-box. The T-box genes are an ancient family of transcription factors that are well conserved throughout

How to cite this paper: Boominathan, V.P. and Ferreira, T. (2016) Predicting *tbx22* Zebrafish Protein Structure Using Multi-Level Prediction Tools and Demonstration of Conserved Structural Domains in Relation to Orthologous *tbx22* Proteins in Humans. *Journal of Biosciences and Medicines*, 4, 79-92. <http://dx.doi.org/10.4236/jbm.2016.43014>

all metazoans [1] [2] Genes associated with T-box are significantly involved in the regulation of the developmental processes of early mesoderm induction, specification, patterning and somite formation in many organisms [3]. All the T-box genes share a conserved DNA domain which is homologous, encoding the T-box [4]. Mutations in human TBX22 cause X-linked cleft palate with anklyoglossia syndrome affecting palatogenesis [5]. Characterization of several Tbx genes in humans suggests a role for these genes in craniofacial development [6]. The human form of TBX22 has 5 transcripts with 2 involved in protein coding. TBX22 is a part of the TBX1 subfamily which also consists the genes TBX1, 10, 15 18 and 20. While TBX1 has been shown to regulate palatogenesis, a similar complex defect is found in DiGeorge syndrome. The DiGeorge pattern of defects has been observed in the van gogh/tbx1 zebrafish mutant [7]. The two zebrafish *tbx22* splice isoforms, *tbx22a* and *tbx22b*, encode proteins of 444 and 400 amino acids, respectively. Zebrafish *tbx22* mRNA expression mirrors mammalian TBX22 expression and is consistent with early patterning of the vertebrate face [8].

Protein homology models can be computed by the SWISSMODEL server homology modelling pipeline [9]-[11]. This fully automated protein structure server is accessible via the ExPasy web server to match models online with multiple templates and established database. Three different types of modelling requests (*automated mode*, *alignment mode*, *project mode*) are provided, which differ in the amount of user intervention. Modelling template utilizes top ranking templates using PROMOD-II and MODELLER. Models are built based on the target-template alignment using Promod-II. Coordinates which are conserved between the target and the template are copied from the template to the model. Insertions and deletions are remodeled using a fragment library. Side chains are then rebuilt. Finally, the geometry of the resulting model is regularized by using a force field. The global and per-residue model quality has been assessed using the QMEAN scoring function [12]. For improved performance, weights of the individual QMEAN terms have been trained specifically for SWISS-MODEL [11].

Phyre 2 is a suite of tools available on the web to predict and analyze protein structure, function and mutations. This protocol will guide users from submitting a protein sequence to interpreting the secondary and tertiary structure of their models, domain composition and model quality. A range of additionally available tools are described to find a protein structure in a genome, to submit large number of sequences at once and to automatically run weekly searches for proteins that are difficult to model. The focus of Phyre 2 is to provide biologists with a simple and intuitive interface to state-of-the-art protein bioinformatics tools [14]. Phyre 2 encourages model building by 4 main stages: Gathering homologous sequences, Fold library scanning, Loop with multiple template modelling and Side chain placement. Developed Protein Data bank (PDB) files are then super imposed on top of one another to identify the common sharing areas. The SuperPose is a rapidly and robustly calculating web server, which calculates both pairwise and multiple protein structure superpositions using a modified quaternion eigenvalue approach. SuperPose is composed of two parts, a front-end web interface (written in Perl and HTML) and a back-end for alignment, superposition, RMSD calculation and rendering (written in Perl and C). The front-end accepts two kinds of input, PDB text files or PDB accession numbers or any combination of both [15].

2. Materials and Methods

2.1. Sequence Analysis and Model Development

Aminoacid sequences of Human TBX22 and Zebrafish *tbx22* were obtained from the UniProt database collection of proteins [16], Supplemental Data S1. The accession code for them is Q9Y458 for human and C7FDJ3-1, C7FDJ4-1 for Zebrafish spliceforms. Bioinformatic work has been performed by using CLC Main Workbench 7.6.4. Neighbor joining (NJ) method was employed to generate the phylogenetic relationship of TBX22 between different organisms of interest and comparison. Neighbor Joining algorithm, K-mer based approach, where $k = 15$ and Distance measure = Mahalanobis was used to develop the phylogram.

2.2. Transcription Factor Binding Factor Sites Prediction

TBX22 comes under the V\$BRAC-Brachury gene family of genes, which is a mesoderm developmental factor. TBX22 gene for humans against MATTINSPECTOR tool from Genomatix has shown the location of the gene on the chromosome X and ALTS_NT_187635 of human GRCh 38 (GeneID: 50945 GXL_16356/50945/GXL_1808093) respectively. The *tbx22* for zebrafish is located in the GRCz10 chromosome (GeneID: 556143

GXL_891394) with 1 locus, 3 identified transcripts and 1 unique promoter sequence. Transcription Factor (TF) binding site analysis was compared with other TF binding site prediction within Genomatix systems like Transriptome viewer and Gene2Promoter tools [17].

2.3. Structural Modelling

Sequence alignments were also used to identify the match/gaps/mismatch between sequences. There are no established Protein Data Bank (PDB) files for the HumanTBX22 and Zebrafish tbx22. Generation of PDB structures were based on Prediction softwares like Phyre 2, SWISSMODEL. Swissmodel employs building models based on the target-template alignment using Promod-II. Coordinates which are conserved between the target and the template are copied from the template to the model. Insertions and deletions are remodeled using a fragment library. Side chains are then rebuilt. Finally, the geometry of the resulting model is regularized by using a force field. In case loop modelling with ProMod-II does not give satisfactory results, an alternative model is built with MODELLER [18]. Superimpose alignment of 2 sequences is achieved by SuperPose, SuperPose is a protein superposition server. SuperPose calculates protein superpositions using a modified quaternion approach. From a superposition of two or more structures, SuperPose generates sequence alignments, structure alignments, PDB coordinates, RMSD statistics, Difference Distance Plots, and interactive images of the superimposed structures. Molecular graphics and analyses were performed with the UCSF Chimera package [19].

2.4. Mammalian Cell Transfection and Expression of tbx22 Proteins

HEK-293, human embryonic kidney cells 293 (epithelial morphology) were used in this study. The cells were cultured in the base medium, Dulbeccos's Modified Eagle's medium (Thermo Fisher, Agawam, MA, USA), 10% Heat inactivated Fetal Bovine Serum (HI FBS, Qualified, Thermo Fisher), and 1% Penicillin-Streptomycin (Thermo Fisher) in a 37°C incubator with 5% (v/v) CO₂ injected. The doubling times of the cells were calculated by identifying the cell growth over time. The assay was performed by seeding 0.3×10^6 cells in a 6-well plate and they were allowed to adhere and grow. Cells were prepared for density calculations by treatment with trypsin and dilution in media, followed by manual cell counts using a hemocytometer. Cell density was taken into consideration and cells were seeded at a standard density for initiation of the assay, reaching confluency at 1.2×10^6 cells at the end of the time point. Cells were then transfected with custom-made tbx22 gene constructs using Lipofectamine 3000 (Life Technologies, L3000008). A DNA concentration of 1 µg was transfected under standard conditions. Transfected cells were then allowed to grow for 3 - 4 days. The media was then changed to Neomycin enhanced media to screen for cells. Cells were then scaled up to standard 25 cm² flasks for further protein production, followed by testing the supernatant for presence of specific proteins.

3. Results

3.1. Sequence Identification

For each identified template, the template's quality has been predicted from features of the target-template alignment. The templates with the highest quality have then been selected for model building. Model building results are based on the primary amino acid sequence. Identified sequences were used for Alignment and tree construction (**Figure 1**). Sequences from UNIPROT and NCBI were crosschecked before the development of relatedness.

3.2. Transcription Factor Binding Site Prediction

Based on the EIDorado database, 2 loci were identified for humans with 20 transcripts and 8 unique promoters. 1 transcript, GXT_26252783 from the GeneID: 50945 GXL_16356 and 9 transcripts from the maternally expressed X chromosome, GXT_22770400, GXT_24460915, GXT_26251083, GXT_26251084, GXT_26251085, GXT_26251086, GXT_26255662, GXT_26251089 and GXT_22770402. Whereas the identified zebrafish transcripts are GXT_23655886, GXT_24578594 and GXT_26537083. Matinspector tool from Genomatix identified the Individual binding sites in a promoter. Functional assessment of binding sites is further confirmed with other tools, e.g. Model Inspector, Comparative Genomics, Frame Worker, Genomatix Pathway System Identification of distance phylogenetically is significant for the analysis of TBX22 between different species and also between

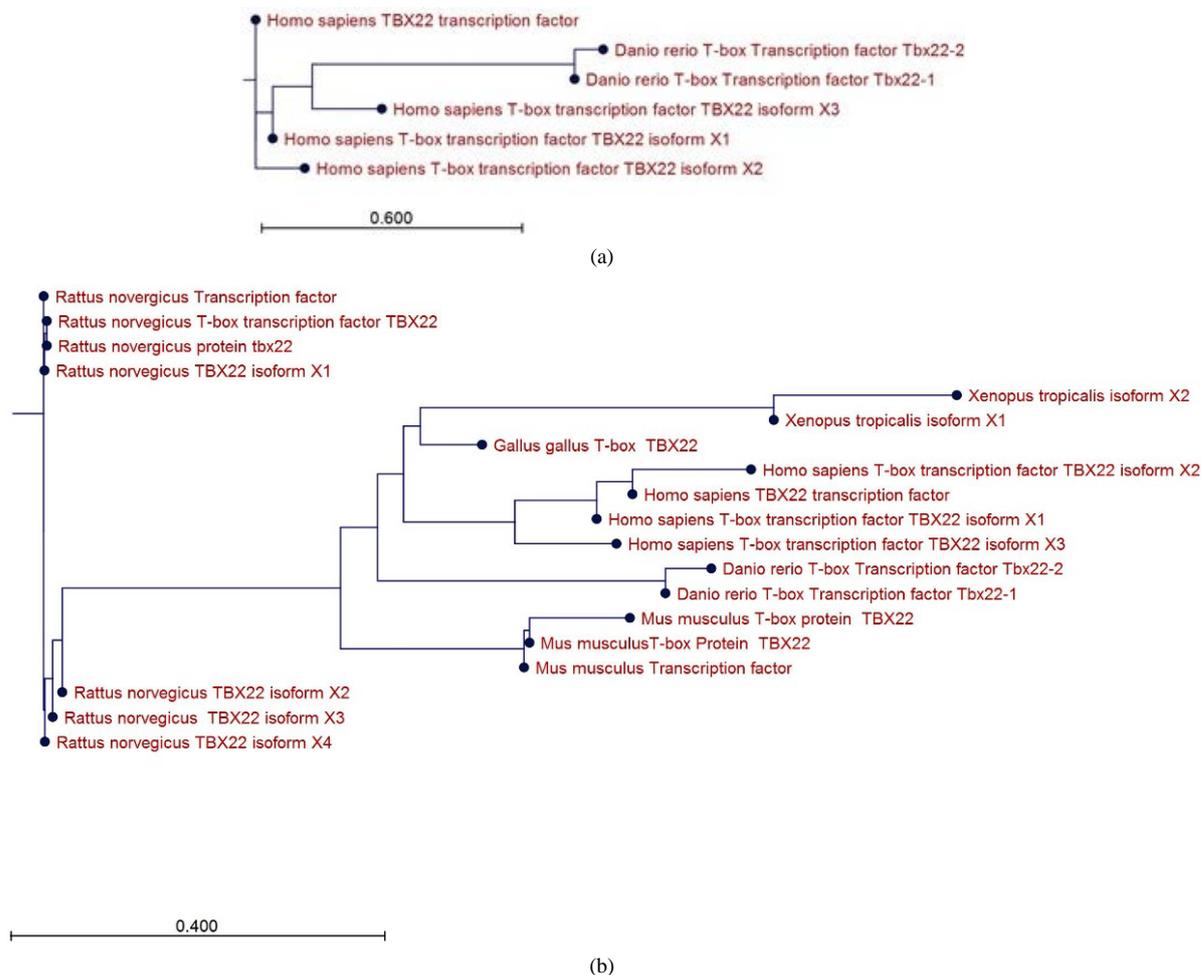


Figure 1. (a) Detailed Analysis of Phylogram using NJ showing the Distance of TBX22 protein and Transcription factor among orthologous sequences between different organisms and (b) Comparison and broad understanding of the Phylogram developed between the TBX22 transcription factors between *Homo sapiens* and *Danio rerio* isoforms.

Humans and Zebrafish to bridge the gap of closely relatedness (**Figure 1(a)**). Phylogram relationship between other organisms and Humans-Zebrafish has been shown individually and in detail (**Figure 1(b)**).

Both transcription factors and isoforms have been used for the development to make sure no significant details are missed during the tree building. The human transcription factor is related to its three different isoforms X1, X2, X3, which are further closely related to the zebrafish isoforms 1 and 2. (**Figure 1(b)**). Comparative statistics of the protein information developed with the repeating aminoacids show their structural similarity with closely equal numbers (**Figure 2**).

3.3. T-Box Genes and Brachury Gene Matrix Family

Table 1 depicts the V\$BRAC matrix family, explaining the common set of genes for the BRACHURY gene family in both humans and Zebrafish. Certain genes exist in both organisms as splice variant forms, while few inclusions are restricted to their own such as Human TBX23P and Zebrafish tbx16, 19. TBX22 and tbx22 are sequentially similar to that of the organism with little mismatches. All the isoforms of the TBX22 gene were examined for the presence of the BRAC matrix family genes (**Figure 3**). The presence of the matrix family on the gene sequence explains that the T-box domain is contained within and its location on the gene. Gene identification with all matrix families and V\$BRAC family has been shown with their location (**Figures 3(a)-(d)**). Sequence alignment of AA sequences of PDB structures are aligned together and sequence homology between them was identified before superimposing structures. Sequence identity/similarity scores between tbx22-1 and

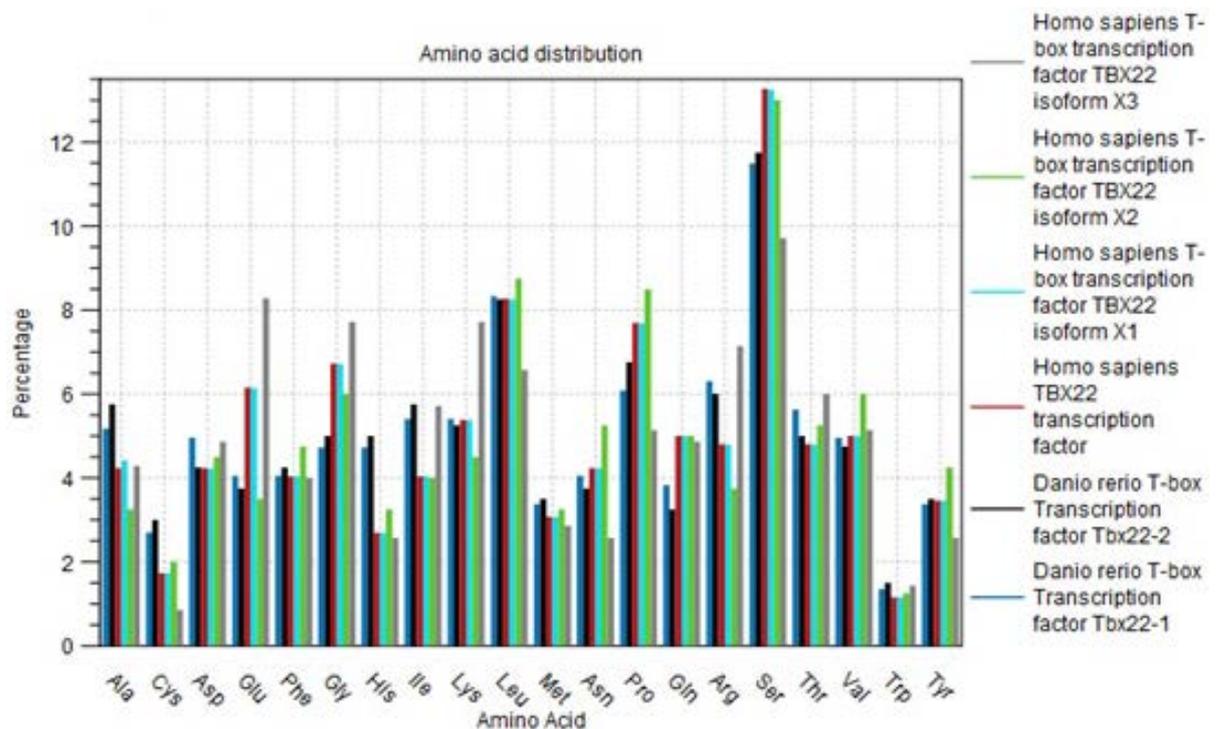


Figure 2. Comparative statistics layout in the form of histogram, showing amino acids distribution between Human TF, isoforms and Zebrafish TF isoform.

Table 1. Brachury gene Family and associated homologous tbx genes in Humans and Zebrafish.

Humans: EOMES, MGA, T, TBR1, TBX1, TBX10, TBX15, TBX18, TBX19, TBX2, TBX20, TBX21, TBX22, TBX23P, TBX3, TBX4, TBX5, TBX6	Zebrafish: eomesa, mgaa, ta, tbr1a, tbr1b, tbx1, tbx15, tbx16, tbx18, tbx19, tbx20, tbx21, tbx22, tbx2a, tbx2b, tbx3a, tbx4, tbx5a, tbx5b, tbx6
--	---

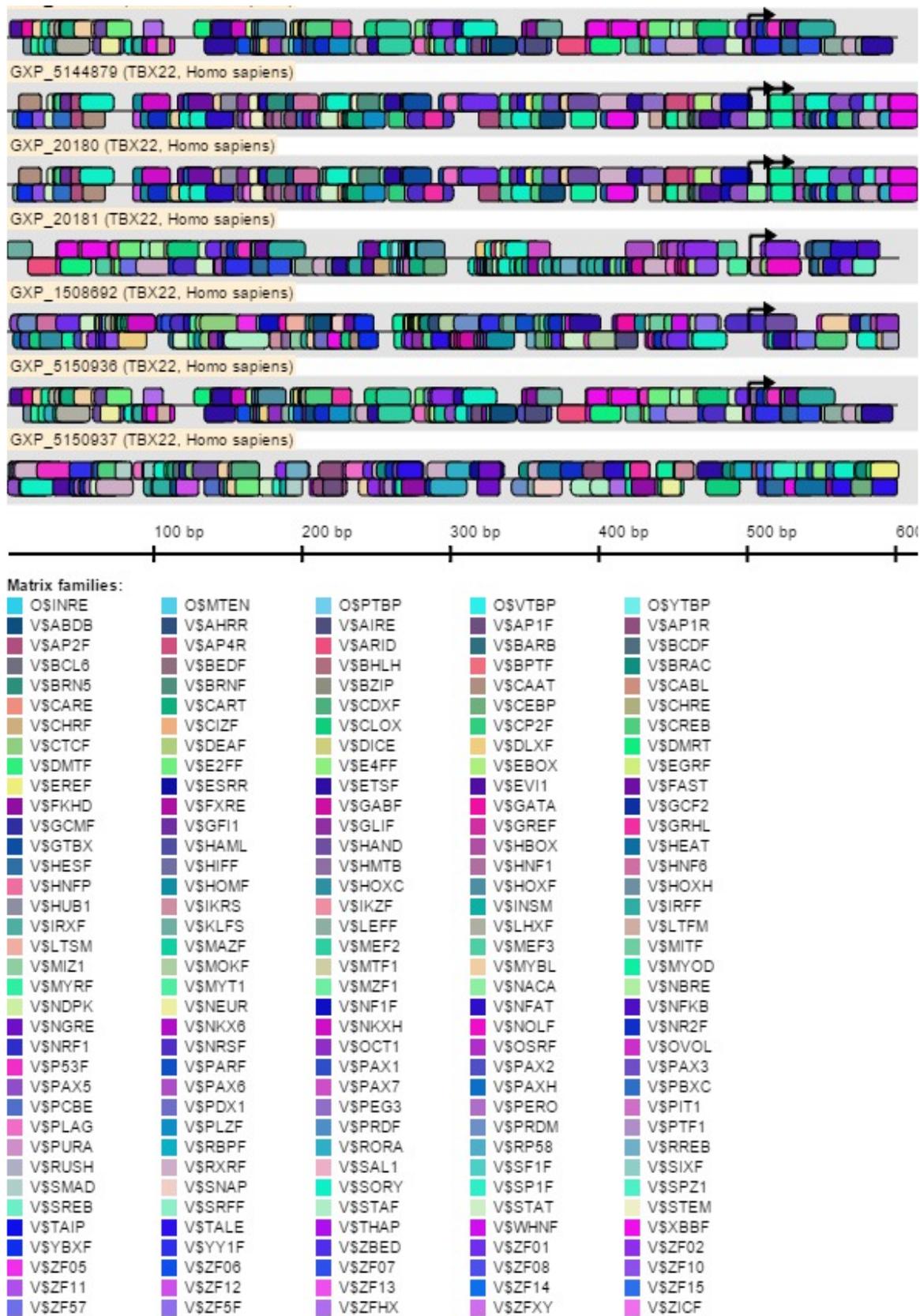
22-2 spliceforms were 89.2%/89.4%, while tbx22 vs TBX22 has a score of 40.3%/53.3% (**Table 2** and **Table 3**).

3.4. Protein Homology Modeling

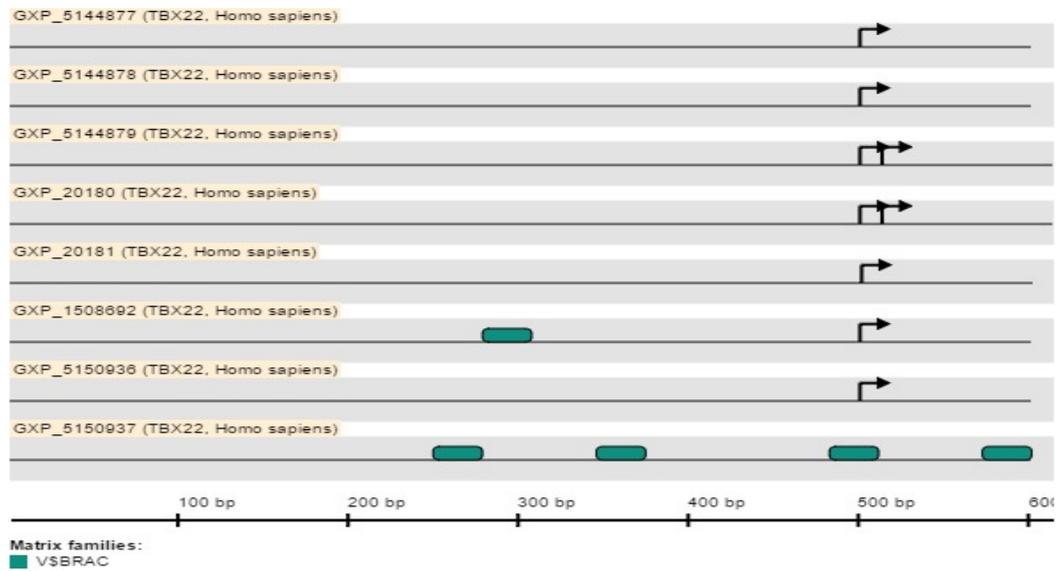
The SWISS-MODEL template library (SMTL version 2016-01-20, PDB release 2016-01-15) was searched with Blast and HHblits [12] [13] for ZF tbx22-1 and tbx22-2 evolutionary related structures matching the target sequence for Transcription factor 5 and 1 as the closest PDB template. Close to 30 templates were built for the sequence in Table S1 The newly built homology models were then linked to one another and an ideal model for the amino acid sequences was constructed for every TF of interest. SuperPose uses structural alignment to guide the superposition of two or more structures. The SuperPose web server supports the submission of either PDB-formatted files or PDB accession numbers. Structures generated here are PDB formatted files using Phyre 2. Structures of Tbx TF 22-1 and 22-2 of Zebrafish are visualized along with their secondary structure (**Figure 4(a)**, **Figure 4(b)**). Human TBX22 TF is also shown separately (**Figure 4(c)**). Superimposing one protein over another is performed by Structure Comparison-MatchMaker tool. The resulting sequence of tbx22-1 and tbx22-2 of zebrafish is shown with similar moieties (**Figure 5**). Superimposing structural alignment of third PDB is generated by laying out Human TBX22 TF over the result of 22-1 and 22-2 (**Figure 6**). Structural differences and overhanging sequence can be picturized by the generated model superimposed using UCSF Chimera 1.10.2 [17].

3.5. Tbx22 Protein Expression in Mammalian Cell System

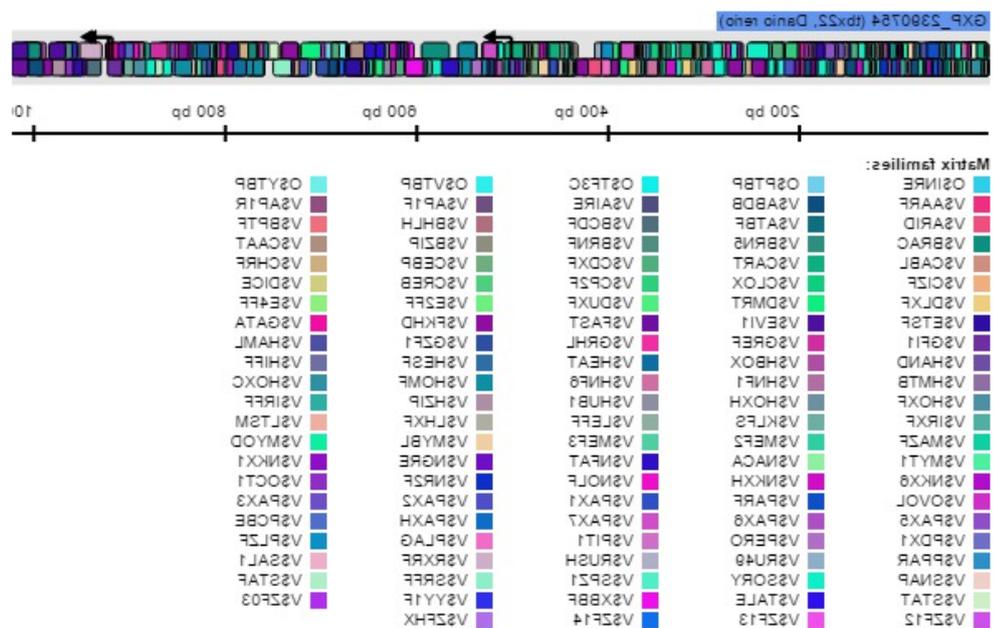
Cells transfected and grown in standard 25cm² flasks were transferred to grow in standard media which has neomycin. Presence of Neomycin resistance gene in the custom plasmid construct makes only the selected-transfected cells grow. The antibiotic resistance gene acts as a screening tool to eliminate the non-transfected



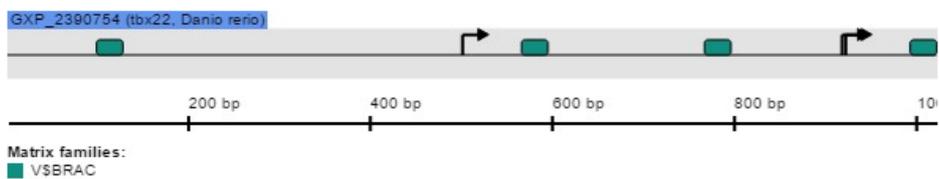
(a)



(b)



(c)



(d)

Figure 3. Promoter sequences of (a) TBX22 gene of human genome and (c) tbx22 of Zebrafish genome analyzed against EIDorado database showing all different matrix families. Presence and location of the brachury gene-V\$BRAC matrix family is illustrated on the gene sequence for Human TBX22 in (b) and Zebrafish tbx22 (d) respectively.

Table 2. Alignment of tbx22-1 and tbx22-2 of zebrafish.

```

# Program: needle
# Align_format: srspair
# Aligned_sequences: 2
# 1: PDBA_model_default_chain_default
# 2: PDBB_model_default_chain_default
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 444
# Identity:      396/444 (89.2%)
# Similarity:   397/444 (89.4%)
# Gaps:         44/444 ( 9.9%)
# Score: 2127.0

PDBA_model_de      1 MQSITRQVDHTSGNRCKMKKVENTENESLQDVLEYDSDTLRRTLFDQAQDL      50
                                     . . . : | |
PDBB_model_de      1 -----MACEDL      6

PDBA_model_de      51 RKSAKERKPAGAEAVRKVRVDLQGS DLWKR FHEIGTEMIITKVGRRMFPS      100
    |||
PDBB_model_de      7 RKSAKERKPAGAEAVRKVRVDLQGS DLWKR FHEIGTEMIITKVGRRMFPS      56

PDBA_model_de     101 IRVKVHNLDP LQQYSIAMDIMPMD SKKYRYVYHSSQWVIAGNTDHSCIPP      150
    |||
PDBB_model_de     57 IRVKVHNLDP LQQYSIAMDIMPMD SKKYRYVYHSSQWVIAGNTDHSCIPP      106

PDBA_model_de     151 HLYVHPDSPCSGENWMRQVISFDRVKLTNNELDDRGHIILKSMHKYRPRI      200
    |||
PDBB_model_de     107 HLYVHPDSPCSGENWMRQVISFDRVKLTNNELDDRGHIILKSMHKYRPRI      156

PDBA_model_de     201 HVILHCPPECLSKRLLSLPADGVFTFSFPETQFTT V TAYQNQQITK LKID      250
    |||
PDBB_model_de     157 HVILHCPPECLSKRLLSLPADGVFTFSFPETQFTT V TAYQNQQITK LKID      206

PDBA_model_de     251 RNPFAKGFRE RNGAVLDGILESYSWHSPFNGFKSLAMELQGR CFGSSASG      300
    |||
PDBB_model_de     207 RNPFAKGFRE RNGAVLDGILESYSWHSPFNGFKSLAMELQGR CFGSSASG      256

PDBA_model_de     301 ITPSVSSAQAVHPTAFSSPQCCKILPSSCFLAYRAYCSICLN NYTGLRPM      350
    |||
PDBB_model_de     257 ITPSVSSAQAVHPTAFSSPQCCKILPSSCFLAYRAYCSICLN NYTGLRPM      306

PDBA_model_de     351 LDLPLLTSLPVKKGDRCRSHWLYNSAGTSMHPIHNTLSGWASENENTASA      400
    |||
PDBB_model_de     307 LDLPLLTSLPVKKGDRCRSHWLYNSAGTSMHPIHNTLSGWASENENTASA      356

PDBA_model_de     401 FMSPHSYSFPI SHRFSTDT HHLKLAHSESTVLQMPMISSIGYSP      444
    |||
PDBB_model_de     357 FMSPHSYSFPI SHRFSTDT HHLKLAHSESTVLQMPMISSIGYSP      400

```

Table 3. Alignment of tbx22 of zebrafish and TBX22 human.

Program: needle
 # Align_format: srspair
 # Aligned_sequences: 2
 # 1: SUPO_model_tbx22
 # 2: PDBB_model_Human TBX22
 # Matrix: EBLOSUM62
 # Gap_penalty: 10.0
 # Extend_penalty: 0.5
 # Length: 538
 # Identity: 217/538 (40.3%)
 # Similarity: 287/538 (53.3%)
 # Gaps: 112/538 (20.8%)
 # Score: 911.0

SUPO_model_1_	1	MQSITRQVDHTSGNRCKKMKVE---NTENESLQDVLEYDSDLRRTL--	44
PDBB_model_de	1	-----MALSSRARAFSVEALVGRPSKRKLQDPIQAEQPELREKKGG	41
SUPO_model_1_	45	FDAQDLRKSAK-----ERKP-----AGAEA-----VRK	67
PDBB_model_de	42	EEEEERRSSAAGKSEPLEKQPKTEPSTASSSGCGSDSGYGNSSSESLEEKD	91
SUPO_model_1_	68	VRVDLQGS DLWKR FHEIGTEMIITKVGRRMFPSIRVKVHNL DPLQQYSIA	117
PDBB_model_de	92	IQMELQGS SELWKR FHDIGTEMIITKAGRRMFPSVRVKVGLDPGKQYHVA	141
SUPO_model_1_	118	MDIMPMDSKKYRYVYHSSQWVIAGNTDHSCIPPHLYVHPDSPCSGENWMR	167
PDBB_model_de	142	IDVVPVDSKRYRYVYHSSQWMVAGNTDHL CIIPRFYVHPDSPCSGETWMR	191
SUPO_model_1_	168	QVISFDRVKLTNNELDDRGHII LKSMHKYRPRIHVILHCPPECLSKRLLS	217
PDBB_model_de	192	QIISFDRM KLTNNEMDDKGHIILQSMHKYKPRVHVIEQGSVDLS-QIQS	240
SUPO_model_1_	218	LPADGVFTFSFPETQFTTTV TAYQNQQITK LKIDRNPFAKGFRE--RNGAV	265
PDBB_model_de	241	LPTEGVKTF SFKETEFTTTV TAYQNQQITK LKIERNPFAKGFRTGRNRGV	290
SUPO_model_1_	266	LDGILESYSWHSPFN-GFKSLAMELQGR CFGSSASGITPSVSSAQAVHP-	313
PDBB_model_de	291	LDGLELETYPWRP SFTLDFKTFGADTQ-----SGSSGSSPVTSSGGAPSPL	335
SUPO_model_1_	314	TAFSSPQCK---ILPSS-----CFLAY-----RAYC-SICLNNTYGLRP	349
PDBB_model_de	336	NSLLSPLCFSPMFHLPTSS LGMPCEAYL PNVNLP LICYICPTNFWQQQP	385
SUPO_model_1_	350	-----MLDLPLL TSLPV---KKGDRCRSHWLYNS	375
PDBB_model_de	386	LVL PAPERLASSNSQSLAPLMMEVPMLS SLGVTNSKSGSSEDSSDQYLQ	435
SUPO_model_1_	376	AGTSMHP IHNTLSGWAS-----ENENTASAFMSP-----HSYSFPIS	412
PDBB_model_de	436	APNSTNQ M---LYGLQSPGNIFLPNSITPEALSCSFHPSYDFYRYNFSMP	482
SUPO_model_1_	413	HRFSTDTHHLKLAHSESTV LQMPMISSIGYSP-----	444
PDBB_model_de	483	SRLISGSNHLKVND DSQVSFGEGKCNH VHWYPAINHYL	520

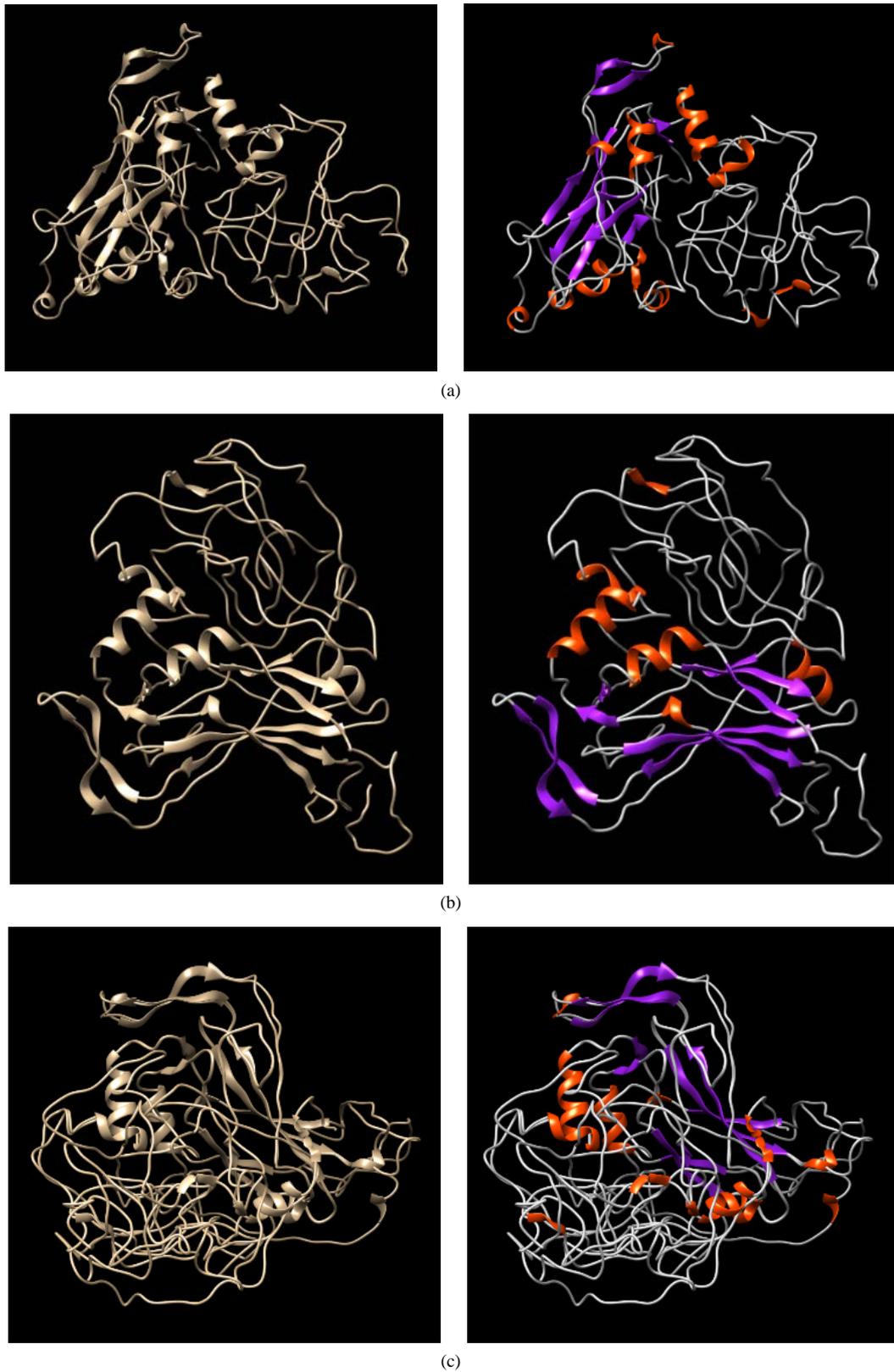


Figure 4. a, Generated PDB view and secondary structure of (a) Zebrafish tbx22 22-1 TF, (b) Zebrafish tbx22-2TF and (c) Human TBX22 TF.

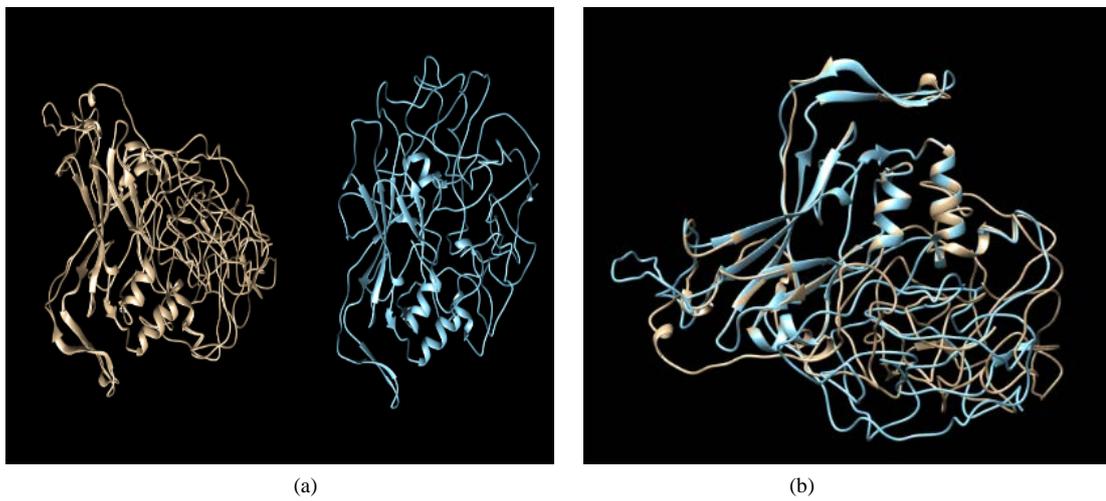


Figure 5. Zebrafish TF tbx22-1 and tbx22-2 tiled up for superimpose and b) superimposed TF 22-1 and 22-2 on each other.

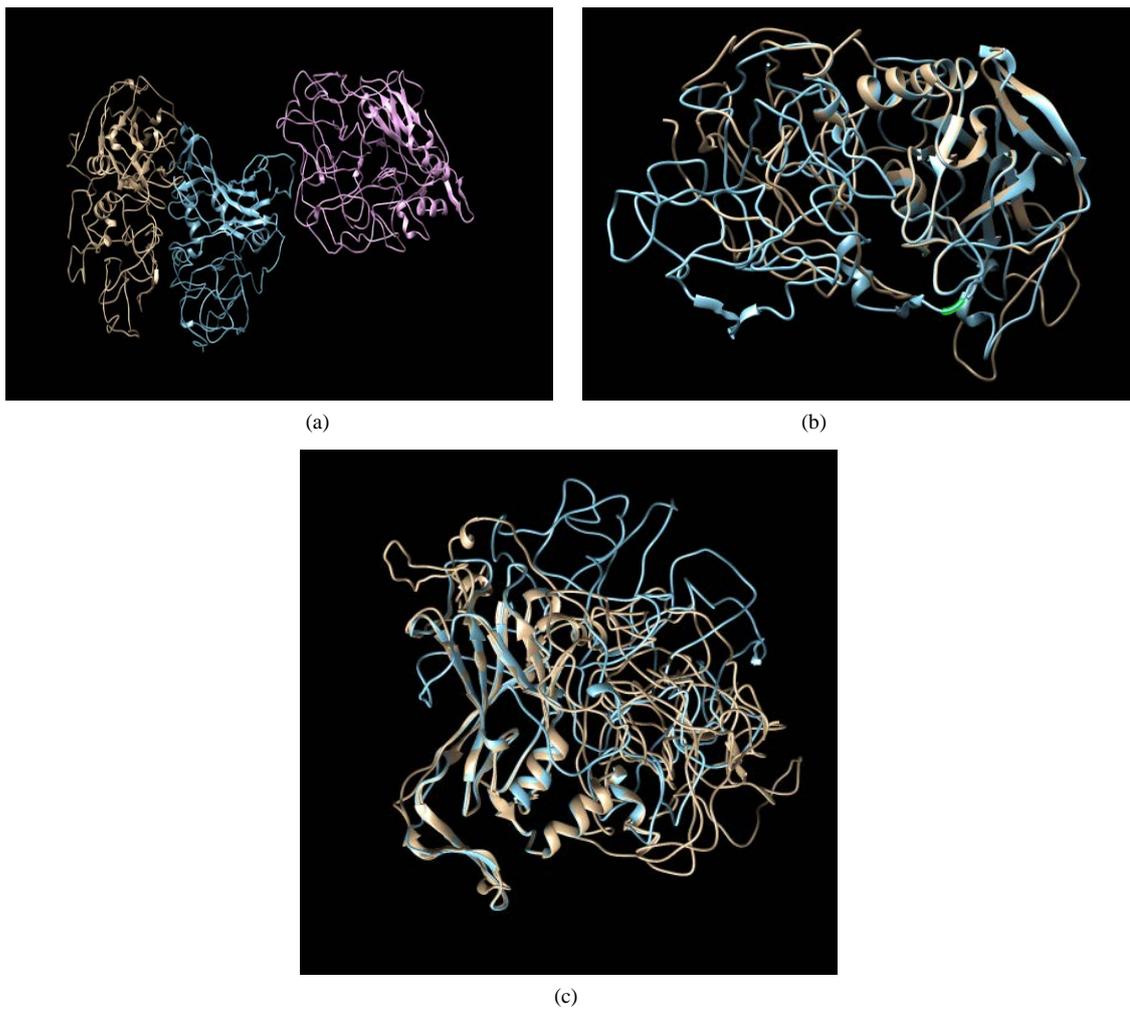


Figure 6. (a) PDB generated structures tiled up sidewise tbx22-1, tbx22-1 and TBX22. (b) Superimposed structure of tbx22-1 and tbx22-2. (c) Final PDB superimposed structural comparison of TF tbx22-1, tbx22-2 and TBX22 using Match Maker tool from UCSF CHIMERA.

cells. Protein production in the cells is confirmed by isolating the supernatant after 4 - 7 days, purifying them against Magnetic bead separation, followed by SDS-PAGE analysis. Confirmation of the size was examined by performing a Silver stain analysis to confirm the size of the protein (**Figure 7**). Expressed protein size has been estimated to have a molecular weight of 53 kDa which is the same as the *tbx22* protein.

4. Discussion

The aim of this study was to build PDB files for Human TBX22 and ZF *tbx22* transcription factors and determine the best model fit to superimpose the protein structures on top of each other. In order to achieve that, a combination of Bioinformatics applications were utilized starting with development of a NJ phylogram, identifying the distant relatedness followed by Structure prediction using Phyre 2 and SWISSMODEL. The final model was built by SuperPose and Chimera tools.

Phylogenetic analysis of relatedness shows the ZF and human transcription factors are closely related based on the nodal distance between them. Human TBX22 shows that it is more closely related to *Gallus gallus* and *Xenopus tropicalis* than to ZF *tbx22*, and is more distantly related to *Rattus norvegicus* and *Mus musculus* making the zebrafish an ideal model for comparative studies. When only Human and Zebrafish TF's are compared in a phylogram the distancescores reveal that the two are close to each other. Based on the NJ tree scores, TF binding site prediction has been analyzed for Humans versus ZF and it shows the common binding sites and common matrix families which share binding motifs. Analysis of the identified genes has shown the location of the T-Box domain to be the same as the V\$BRAC sequence position. The T-Box binding domain is approximately 180 AA in length and is highly conserved among T homologous proteins from different species. The Brachury TF is bound as a dimer, interacting with the major and minor grooves of the DNA. Human TBX22 TF PDB model mimics the T-box transcription factor 1,5 and 3 with a sequence identity of 56.08%, 53.26% and 51.85% respectively. While the Zebrafish *tbx22* TF PDB model also mimics T-Box Transcription factors 1, 5 and 3 with a sequence identity of 53.72%, 52.17% and 51.10%.

The predicted models of human TBX22 and ZF *tbx22* have similar binding areas with the established T-Box TF PDB structures showing the location of common sequence areas with the T-Box domain of TF's. Assembly of the the 3 different protein models remained a major bottleneck in developing and interpreting the material

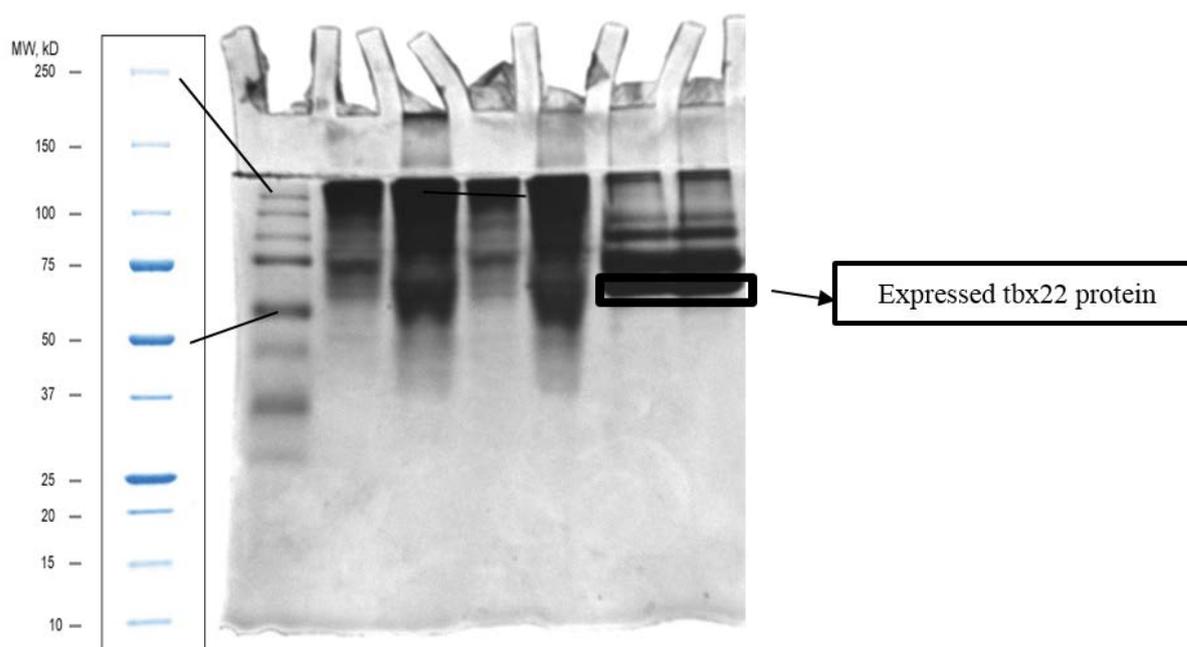


Figure 7. *tbx22* protein expression in mammalian HEK cells. Lane 1 MW marker, Biorad Preciion Plus. Lane 2 Control with normal media, Lane 3 Unpurified *tbx22*-1, Lane 4 Control with selective media, Lane 5 Unpurified *tbx22*-2 supernatant, Lane 6 Purified protein *tbx22*-1, Lane 7 Purified protein *tbx22*-2. Protein size confirmation of purified protein around 53kDa visualized against silver staining protocol.

before and after simulation. Structural modelling of TBX22 shows the predicted template model to be built with more than 90% confidence level for 58% of the sequence. ZF tbx22 predicted template model to established PDB model had 53% and 48% sequence identity with more than 90% confidence level. Predicted structures of the models have been superimposed on one another and shown to share the major binding areas. Superimposed structures of independent proteins one on each other has shown to have a more reliable and common areas that have domains in the stereo view of the protein. Though the predicted models show maximum evidence to the structure of TBX22 and tbx22 to be a part of the T-box protein family, built models are not suggested to be taken if it has low identity percentage. While all the proteins over the study have been verified to have at least 50% sequence identity to the simulated templates against the established PDB models and more than 60% sequence identity among themselves. Having completed this simulation the predicted models show promise in solving this proteins structure. Supporting evidence has been also shown by expressing the protein in mammalian cells. Protein purification followed by coomassie/silver staining technique has shown the expressed protein of an expected size of 53 kDa. Were a ZF specific antibody for tbx22 protein available, we could confirm this easily by Western blotting. Based on the data in hand it is also evident that Zebrafish would be an ideal predictive model to identify numerous developmental patterning events translatable to human development.

5. Conclusion

Protein prediction methods have allowed biologists to identify protein structure, functionality, orientation and other important general information. Available knowledge of existing proteins can guide us with a template to approach a solution to unknown structures and function. Uncharacterized proteins can now be defined with the advantage of multiple public databases, with increased set of established data sources combined with bioinformatics tools. As more tools are accessed computational biologists can unravel mysteries of unidentified/unpredicted biological molecules to solve known problems leading to advancements and new ideas. The goal of this paper was to show the sequence similarity between T-box22 of 2 different species and show how they are sequentially and structurally similar to each other. Supporting documentation available from databases and the usage of the right tools has shown enough evidence to prove that the ZF tbx22 TF shares functional homology to the Human TBX22 TF.

References

- [1] Papaioannou, V.E. and Silver, L.M. (1998) The T-Box Gene Family. *BioEssays*, **20**, 9-19. [http://dx.doi.org/10.1002/\(SICI\)1521-1878\(199801\)20:1<9::AID-BIES4>3.0.CO;2-Q](http://dx.doi.org/10.1002/(SICI)1521-1878(199801)20:1<9::AID-BIES4>3.0.CO;2-Q)
- [2] Wardle, F.C. and Papaioannou, V.E. (2008) Teasing out T-Box Targets in Early Mesoderm. *Genes & Development*, **18**, 418-425. <http://dx.doi.org/10.1016/j.gde.2008.07.017>
- [3] Kiefer, J.C. (2004) The Tbx Files: The Truth Is out There. *Developmental Dynamics*, **231**, 232-236. <http://dx.doi.org/10.1002/dvdy.20122>
- [4] Foppiano, S., Hu, D. and Marcucio, R.S. (2007) Signaling by Bone Morphogenetic Protein Directs Formation of an Ectodermal Signaling Center That Regulates Craniofacial Development. *Developmental Biology*, **312**, 103-114. <http://dx.doi.org/10.1016/j.ydbio.2007.09.016>
- [5] Braybrook, C., Doudney, K., Marciano, A.C.B., Arnason, A., Bjornsson, A., Patton, N.A., et al. (2001) The T-Box Transcription Factor Gene TBX22 Is Mutated in X-Linked Cleft Palate and Ankyloglossia. *Nature Genetics*, **29**, 179-183. <http://dx.doi.org/10.1038/ng730>
- [6] Jezewski, P.A., Fang, P.K., Payne-Ferreira, T.L. and Yelick, P.C. (2009) Alternative Splicing, Phylogenetic Analysis, and Craniofacial Expression of Zebrafish tbx22. *Developmental Dynamics*, **238**, 1605-1612. <http://dx.doi.org/10.1002/dvdy.21962>
- [7] Piotrowski, T., Ahn, D., Schilling, T.F., Nair, S., Ruvinsky, I., Geisler, R., Rauch, G., Haffter, P., Zon, L.I., Zhou, Y., Foott, H., Dawid, I.B. and Ho, R.K. (2003) The Zebrafish Van Gogh Mutation Disrupts tbx1, Which Is Involved in the DiGeorge Deletion Syndrome in Humans. *Development*, **130**, 5043-5052. <http://dx.doi.org/10.1242/dev.00704>
- [8] Silva, T.L., Cox, A.A., Boominathan, V.P., Jezewski, P.A. and Ferreira, T.L. (2012) Early Expression of the tbx22 Gene in Zebrafish Influences Positioning of Pharyngeal Arch Cartilages. *American Journal of Molecular Biology*, **2**, 318-331. <http://dx.doi.org/10.4236/ajmb.2012.24033>
- [9] Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L. and Schwede, T. (2014) SWISS-MODEL: Modelling Protein Tertiary and Quaternary Structure Using Evolutionary Information. *Nucleic Acids Research*, **42**. <http://dx.doi.org/10.1093/nar/gku340>

- [10] Benkert, P., Biasini, M. and Schwede, T. (2011) Bioinformatics. *Structural Bioinformatics*, **27**, 343-350. <http://dx.doi.org/10.1093/bioinformatics/btq662>
- [11] Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J.E. (2015) The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nature Protocols*, **10**, 845-858. <http://dx.doi.org/10.1038/nprot.2015.053>
- [12] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res*, **25**, 3389-3402. <http://dx.doi.org/10.1093/nar/25.17.3389>
- [13] Remmert, M., Biegert, A., Hauser, A. and Söding, J. (2011) HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat Methods*, **25**, 173-175.
- [14] Kelley, L.A. and Sternberg, M.J.E. (2009) Protein Structure Prediction on the Web: A Case Study Using the Phyre Server. *Nat. Protoc.*, **4**, 363-371. <http://dx.doi.org/10.1038/nprot.2009.2>
- [15] Maiti, R., Van Domselaar, G.H., Zhang, H.Y. and Wishart, D.S. (2004) SuperPose: A Simple Server for Sophisticated Structural Superposition. *Nucleic Acids Res.*, **32**.
- [16] Sequences: <http://www.uniprot.org/uniprot/?query=tbx22&sort=score>
- [17] Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. and Werner, T. (2005) MatInspector and Beyond: Promoter Analysis Based on Transcription Factor Binding Sites. *Bioinformatics*, **21**, 2933-2942.
- [18] Sali, A. and Blundell, T. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.*, **234**, 779-815. <http://dx.doi.org/10.1006/jmbi.1993.1626>
- [19] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J Comput Chem*, **25**, 1605-1612. <http://dx.doi.org/10.1002/jcc.20084>

Supplemental Data S1

Amino Acid sequences of T-box 22 Transcription factor

Homo sapiens TBX22 Transcription factor sequence

MALSSRARAFSVEALVGRPSKRKLQDPIQAEQPELREKKGEEEEEERRSSAAGKSEPLEKQPKTEPSTASSGCGSDSGYGNSSSE
 LEEKDIQMELQGSSELWKRFDHIGTEMIITKAGRRMFPSVRVKVKGDPGKQYHVAIDVVPVDSKRYRYVYHSSQWMVAGNTD
 HLCIIPRFYVHPDSPCSGETWMRQIISFDRMKLTTNEMDDKGHIILQSMHKYKPRVHVIEQGSSVDLSQIQSLPTEGVKTFSEKETE
 FTTVTAYQNQQITKLIKIRNPFKAGFRDTRGRNRGVLGLETYPWRPSFTLDFKTFGADTQSGSSGSSPVTSSTGGAPSNLLSPL
 CFSMPFHLPSTSSLGMPCEAYLPNVNPLCYKICPTNFWQQPLVLPAPERLASSNSSQSLAPLMMMEVPMSSSLGVTNSKSGSSED
 SSDQYLQAPNSTNQMLYGLQSPGNIFLPSNITPEALSCSFHPSYDFYRYNFSMP SRLISGSNHLKVNDDSQVSFGEKCNHVHWY
 PAINHYL

Danio rerio Transcription factor Tbx22-1

MQSITRQVDHTSGNRCKMKVENTENESLQDVLEYDSDTLRRTLFDAQDLRKS AKERKPAGAEAVRKVRVDLQGS DLWKR FH
 EIGTEMIITKVGRRMFPSIRVKVHNLDPLQQYSIAMDIMPMDSKKYRYVYHSSQVIAGNTDHS CIPPHLYVHPDSPCSGENW MRQ
 VISFDRVKLTNNELDDRGHILKSMHKYRPRIHVILHCPPECLSKRLLSLPADGVFTFSFPETQFTVTAYQNQQITKLIKIRNPFK
 GFRERNGAVLDGILESYSWHSPFNGFKSLAMELQGRFCGSSASGITPVSSAQAVHPTAFSSPQCKILPSSCFLAYRAYCSICLNN
 YTGRLPMLDPLLTSLPVKKGDRCSRHWLYNSAGTSMHPIHNTLSGWASENENTASAFMSPHSYSPISHRFSTDTTHHLKLAHSE
 STVLQMPMISSIGYSP

Danio rerio Transcription factor Tbx22-2

MACEDLRKS AKERKPAGAEAVRKVRVDLQGS DLWKR FEIGTEMIITKVGRRMFPSIRVKVHNLDPLQQYSIAMDIMPMDSKK
 YRYVYHSSQVIAGNTDHS CIPPHLYVHPDSPCSGENW MRQVISFDRVKLTNNELDDRGHILKSMHKYRPRIHVILHCPPECLS
 KRLLSLPADGVFTFSFPETQFTVTAYQNQQITKLIKIRNPFKAGFRERNGAVLDGILESYSWHSPFNGFKSLAMELQGRFCGSSA
 SGITPVSSAQAVHPTAFSSPQCKILPSSCFLAYRAYCSICLNNYTGRLPMLDPLLTSLPVKKGDRCSRHWLYNSAGTSMHPIH
 NTLSGWASENENTASAFMSPHSYSPISHRFSTDTTHHLKLAHSESTVLQMPMISSIGYSP