Dynamics and recognition of homeodomain containing protein-DNA complex of IRX4

Adil Malik[1], Neha S. Gandhi[1], and Jyotsna Batra[1]

[1]Queensland University of Technology

January 20, 2022

Abstract

Iroquois Homeobox 4 (IRX4) belongs to a family of homeobox TFs having roles in embryogenesis, cell specification and organ development. Recently, Large scale Genome-Wide Association studies and epigenetic studies have highlighted the role of *IRX4* and its associated variants in prostate cancer. No studies have investigated and characterized the structural aspect of the IRX4 homeodomain and its potential to bind to DNA. The current study uses sequence analysis, homology modelling and molecular dynamics simulations to explore IRX4 homeodomain-DNA recognition mechanisms and the role of somatic mutations affecting these interactions. Using publicly available databases, gene expression of *IRX4* was found in different tissues, including prostate, heart, skin, vagina, and the protein expression was found in cancer cell lines (HCT166, HEK293), B cells, ascitic fluid and brain. Sequence conservation of the homeodomain shed light on the importance of N- and C-terminal residues involved in DNA binding. The specificity of IRX4 homodimer bound to consensus human DNA sequence was confirmed by molecular dynamics simulations, representing the role of conserved amino acids including R145, A194, N195, S190, R198 and R199 in binding to DNA. Additional N-terminal residues like T144 and G143 were also found to have specific interactions highlighting the importance of N-terminus of the homeodomain in DNA recognition. Additionally, the effects of somatic mutations, including the conserved Arginine (R145, R198 and R199) residues on DNA binding elucidated the importance of these residues in stabilizing the protein-DNA complex. Secondary structure and hydrogen bonding analysis showed the roles of specific residues (R145, T191, A194, N195, R198 and R199) in maintaining the homogeneity of the structure and its interaction with DNA. The differences in relative binding free energies of all the mutants shed light on the structural modularity of this protein and the dynamics behind protein-DNA interaction. We also have predicted that the C-terminal sequence of the IRX4 homeodomain could act as a potential cell-penetrating peptide, emphasizing the role these small peptides could play in targeting homeobox TFs.

1. Introduction

Homeobox genes belong to a family of homeodomain-containing TFs (TFs), have been vastly studied for their roles in development, physiology and tissue homeostasis [1]. Even though some members of the homeodomain family, comprising of HOXs, Hepatocyte nuclear factors (HNFs) and NANOGs (NKX genes), are well characterized for their role in various cancers, the mechanistic function of Iroquois (IRX) proteins in tumorigenesis and their DNA binding sequence is still not fully explored[2-4]. New studies on some HOX genes have identified their roles in various cancers, but their functional mechanisms are still to be explored[5-7]. For instance,*HOXA9* has been found to be a tumour suppressor/oncogene in breast cancer and leukemia [1]. Another HOX gene, *HOXB13,* has been well studied in prostate development and tumorigenesis, with inherited mutations having a genetic contribution to prostate cancer[1]. These classes of proteins usually function as complexes (homo or hetero dimers) to exert their regulatory function, altering their binding preference[8]. Limited diversity in eukaryotes has been observed in the recognition and binding of homeodomains to DNA [9]. This could be due to a specific constraint in the specific amino acids associated with the homeodomain architecture and its preference for specific DNA recognition sequences[10,11].

The *IRXs* are one of the newly added members of homeodomain TF family that have been found to play an

important role in developmental processes [12]. IRX proteins contain the unique Iro-box motif, a conserved motif of 13 amino acid residues in the carboxyl-terminal region. They also have an atypical homeodomain with three extra amino acids between the first and second alpha helices, which groups them in the 3-amino-acid-loop-extension (TALE) family of TF[13]. These homeobox TFs play important roles in embryogenesis, cell specification and differentiation and organ development. The human *IRX* complex is composed of six genes, found in two clusters of three genes, each in chromosome 5 (*IRX1, 2* and *4*) and 16 (*IRX3* , *5*and *6*) [14-17].

Recently IRX TFs have been studied in different cancers, suggesting aberrant expression of these proteins in contributing to tumorigenesis.*IRX5* has been reported to be regulated by vitamin D3 in prostate cancer involved in regulating cell cycle and apoptosis[18]. Knockdown of*IRX5* was observed to reduce the cell viability of androgen-sensitive LNCaP cells. *IRX2* protein expression has been correlated with breast tumour size, indicating its oncogenic function in breast cancer [19]. Genome-Wide Association Studies (GWAS) identified *IRX4* as a causative gene in prostate cancer susceptibility[20]. Additionally, alternate splicing of *IRX4* has also been recently studied in prostate cancer, highlighting differential regulation in prostate tumorigenesis and progression[21]. Epigenetic studies in pancreatic cancer found the *IRX4* promotor region to be hypermethylated, influencing increased cell growth[22]. *IRX4* has also been described as a tumour suppressor in prostate cancer via vitamin D interactions[23]. Other studies have also suggested the potential oncogenic roles of *IRX4* in breast cancer and non-small cell lung cancer (NSCLC)[24,25]. Other differential roles of *IRXs* have been reported linking it to multiple mechanisms associated with tumour progression[26-28].

Although *IRX* gene clusters are now being identified as novel therapeutic targets in carcinogenesis[27], their protein structure, which may help to understand their functions, has not been biophysically characterized using techniques like Nuclear Magnetic Resonance (NMR) and X-radiation crystallography (X-ray). Various studies have used homology modelling and molecular dynamics (MD) simulations to understand the molecular mechanisms of TF binding to DNA. A recent study on HOXB13 used computationally modelled protein structures to predict the effect of single nucleotide polymorphisms (SNPs) on the non-homeobox region [29]. Additionally, this approach was also used to model HOXB13 protein and predict the functional role of SNPs in prostate cancer, demonstrating genotype-phenotype effects and paving the way for further clinical studies highlighting its theranostic applications[29]. Furthermore, a study on transcription regulator SoxR (Sulphur Oxidation) predicted DNA binding residues of these proteins using homology modelled structures[30]. Structural construction using homology modelling of E2F1 TF revealed dimerization partner domains and the efficiency to bind to DNA[31]. The structure of a protein is linked to its stability, function and its interaction. Although the Protein Data bank (PDB) has a good number of crystallographic structures, not enough information is available regarding the human proteome. The use of computationally modelled structures to understand the physical and chemical properties of TFs has great benefits. It is well established that missense mutations play an important role in diseases affecting the core tertiary structure of a protein[32,33]. One of the key benefits of these approaches is analyzing the effect of mutations on the protein structure and its binding capacity. Interpretation of mutants and their association to diseases can be significantly influenced using this technique[34]. A recent modelling study in the zyxin family of proteins LIM1-3 domains has indicated new insights into protein-protein interactions and potential nucleic acid binding platforms of these proteins, highlighting opportunities for therapeutic development[35].

We studied the sequence conservation of the amino acids present in the homeodomain in this work. We have built a homology model of IRX4 homeodomain and used MD- simulations and free energy calculations to provide insight into the mechanistic of protein-DNA binding. We also checked the mutations on the DNA binding domain and its effect on homeodomain stability. A classical modelling approach has been used in this work over Alphafold[36] as the prediction of protein-DNA interactions using the Alphafold approach is still in its infancy.

2. Computational Methods

2.1 Gene and Protein Expression analysis

Expression of the *IRX4* gene has been well established and identified in different tissues at the RNA level by RNA seq data. There is yet a gap in determining the expression of IRX4 at the protein level. To fill this hole, we initially searched different databases to identify the presence of IRX4 protein in healthy tissues and disease states. Proteomics database (ProteomicsDB) contains quantitative mass spectrometry data for most human proteins[37]. Next, we used PaxDb (Protein Abundances across organisms), a database to identify protein abundance at the proteome-wide level[38]. Protein quantification was reported as parts per million (ppm), describing the protein quantification with reference to the entire proteome.

2.2 Sequence analysis

The protein sequence of Human Iroquois family members like IRX1 (P78414), IRX2 (Q9BZI1), IRX3 (P78415), IRX4 (P78413), IRX5 (P78411) and IRX6 (P78412) were retrieved from UniProt (accession numbers in brackets). IRX4 (P78413) is a 519 amino acid long protein with DNA binding region, in positions 143-204. The homeodomain consists of 62 amino acids in the IRX4 protein. Multiple sequence alignment and phylogenetic tree analysis of IRX family members was performed using Clustal Omega [39]. The 62 amino acid homeodomain sequence was checked for mutations in the region using the Catalogue of Somatic Mutations in Cancer (COSMIC) database [40].

2.3 Secondary structure and conservation prediction

Secondary structure prediction of full-length IRX4 protein was performed using the PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/)[41]. ConSurf server was used to estimate and visualize the evolutionary conservation of residues in the homeobox region of IRX4 (https://consurfdb.tau.ac.il/)[42]. The maximum number of homologs for conservation analysis was kept at 150, and the E-value cut-off of 0.0001. The phylogenetic tree was constructed using neighbour-joining algorithm.

2.4 Homology modelling of IRX4 homeodomain and DNA

PSI-BLAST was used to search specific templates for IRX4 homeobox. The crystal structure of TALE homeodomain TF TGIF1 (PDB code: 6FQP), an X-RAY Diffraction structure with a resolution of 2.42 Å, came up as the top hit in the BLAST search and therefore was selected as a template to model IRX4. Pairwise sequence alignment of the residues highlighted more than 60% of residues matching with the template protein. Homology modelling of the IRX4 homeodomain was carried out using SWISS-MODEL (https://swissmodel.expasy.org/)[43]. The template structure is present as dimer bound to DNA in PDB. Initially, using the protein structure template, the homeodomain of IRX4 was modelled. Next, two copies of the protein structure were superimposed to the template (6FQP) using UCSF Chimera (http://www.rbvi.ucsf.edu/chimera). The WT IRX4 homeodomain protein was bound to DNA as a homodimer. It has been previously established IRX TFs form homodimers and heterodimers, contributing to gene regulation[44]. Individual mutagenesis was carried out in both protein chains using the Rotamer tool in Chimera. The reliability of the models was verified using the Ramachandran plot using PROCHECK server (https://servicesn.mbi.ucla.edu/PROCHECK/). The interactions between protein-DNA and secondary structures involved were analyzed using DNAproDB [45] and PDBsum [46].

2.5 Physio-Chemical characterization

ExPASy's ProtParam web server tool was used to characterize the physio-chemical properties of the WT protein and the mutants[47]. The server provides an *in silico* approach to determine the physical properties of proteins based on their sequences.

2.6 Effects of mutations on IRX4 protein

I-mutant server was used to calculate the stability of the WT protein and mutants [48]. Default parameters of the server such as pH value of 7 and temperature 25 were used for this analysis. The stability of the WT and mutant protein structures were calculated based on amino acid properties, relative solvent accessibility surface, evolution and structural information of the protein [49].

2.7 Molecular dynamics (MD) simulations of homeodomain and its interaction

The native and mutant protein structures were subjected to MD simulations using the PMEMD program in AMBER16 package[50]. All the systems were solvated using an octahedral water box at an edge distance of 12 Å using a 3-point water model (TIP3P)[51]. Besides an ionic strength of 150mM, an appropriate number of sodium and chloride counterions were added to neutralize the charge of each system. In this study, AMBER ff14SB force field[52] for protein and BSC0 for DNA [53] were used to perform MD simulations. To remove high energy contacts and steric clashes between atoms, each system was minimized by the steepest descent minimization of 2500 steps followed by a conjugate gradient minimization of 4000 steps. Then, each system was gradually heated from 0 to 298.15 K for 50 ps in NVT ensemble and equilibrated for another 600 ps with the protein-DNA position restrained using a force constant of 10 kcal/mol·Å^2 and 2 kcal/mol·Å^2, respectively. Simulations employed periodic boundary conditions, and long-range electrostatic interactions were estimated through Particle Mesh Ewald (PME) approaches. SHAKE was applied to bonds involving hydrogen, and the cutoff value for direct-space nonbonded interactions was set to 12.0 Å [54]. The time step was set to 2 fs. The Langevin thermostat with a collision frequency of 2.0 ps^{-1} was adopted to regulate the temperature of the systems. The equilibration for 20 ns was carried out 298.15 K at 1 bar pressure without any restrictions and followed by a production run for 200 ns at the same conditions were used to record conformations. Root mean square fluctuation (RMSF) and hydrogen bonds were analyzed using the CPPTRAJ module of AMBER Tools[55]. Hydrogen bond profiles of all the simulated structures were calculated to find the number of hydrogen bonds associated with each structure. The schematic diagrams of the protein-DNA complex were deduced using DNAproDB. UCSF Chimera 1.14 [56] was used to visualize protein structures and MD trajectories and prepare images. Fifteen systems including wild type-DNA and mutated IRX4-DNA systems were simulated and the interactions analysed.

2.8 Calculation of free energy of Protein-DNA complex

The molecular mechanics generalized Born surface area (MM-GBSA) method has been applied to compute the relative binding free energies for the protein-DNA complexes[57,58].

$\Delta G_{bind} = \Delta E_{bonded} + \Delta E_{vdW} + \Delta E_{ele} - T\Delta S + \Delta G_{egb} + \Delta G_{esurf}$ (Equation 1)

in which the first three terms represent binding free energy components in gas phase and the last two terms denote solvation free energies. ΔE_{bonded} includes energetics for bond, angle, and dihedral terms. ΔE_{vdW} and ΔE_{ele} are van der Waals, and electrostatic interactions between protein and DNA and these two components were calculated using molecular mechanics force fields. ΔG_{egb} is polar solvation energy and can be calculated with the GB model developed by Onufriev et al. (igb=5)[59]. The last term ΔG_{esurf} is non-polar solvation free energy, which is determined using the following equation.

$\Delta G_{esurf} = \gamma \times \Delta SASA + \beta$ (Equation 2)

where the parameters γ and ΔSASA respectively denote the surface tension and the difference in the solvent accessible surface areas caused by ligand associations. For our current work, γ and β were assigned as 0.0072 kcal·mol·Å^{-2} and 0 kcal*mol^{-1}, respectively[59]. Due to the high cost in the calculation of TΔS term, we did not perform entropy calculations. Further, it is usually approximated that in comparing the relative stabilities of different complexes, the entropic contributions are not significant, and they were not included in this study. The other free energy components were calculated based on the snapshots extracted from the last 150ns of the production MD trajectories. An ionic strength of 0.15M was also used during the GB calculations.

The relative free energies calculated by MM-GBSA were further decomposed into residue contributions. It is important to note that the interaction energies calculated in this manner are not directly comparable to experimental results, nor will they sum to the total binding energy but can be compared relatively to the mutants. This type of decomposition analysis is useful for identifying residues that have the most considerable effects on the binding energy.

2.9 Cell-penetrating peptide identification

The homeodomain sequence of IRX4 was subjected to the CPPsite 2.0, a database of cell-penetrating

peptides[60]. CPP site 2.0 contains experimentally validated cell-penetrating peptides and this server was used to determine the efficacy of IRX4 sequence to be considered in that category. SkipCPP-Pred was used to predict the sequence having the highest cell-penetrating potential[61]. SkipCPP-Pred (http://server.malab.cn/SkipCPP-Pred/Index.html) is a sequence-based computational predictor of cell-penetrating peptides that uses an algorithm k-skip n-gram combined with a random forest classifier.

3. Results and Discussion

3.1 IRX4 expression from publicly available datasets

We initially checked the expression of *IRX4* gene from the Genotype-Tissue Expression (GTEX) portal[62] and found expression in multiple tissues, including breast, oesophagus, heart, salivary gland, prostate, skin and vagina (Figure S1). We found an increased expression of this gene in the prostate, skin, vagina and oesophagus compared to other tissues. Furthermore, we checked if the expression at the transcript level correlated with any cancer-specific RNA seq data. We used GeneVestigator to identify the top hits for higher expression of*IRX4* gene [63]. Amongst the initial hits, the expression of *IRX4* was found to be higher in malignant prostate cancer cells and oesophagus squamous cell carcinoma (Figure S2).

The transcriptomic data also indicated the expression of this gene in breast, skin and brain tumours highlighting the importance of this gene in cancer study. Next, we went ahead to check the expression of IRX4 protein using multiple databases and datasets. This protein was found to be expressed only in the MCF7 breast cancer cell line according to the ProtemicsDB, which uses Mass Spectrometry (MS) data to quantify protein (Figure S3). To further identify if the expression of IRX4 did correlate with any MS specific data, we used PaxDB to identify proteins and proteome-wide levels (Figure 1). Interestingly, we found evidence of the protein in multiple datasets, which are represented as ppm. H727 cells, which are epithelial lung bronchus cells, showed higher protein expression than its total proteome. Low expression of the gene was observed in the lung tissue. The Global Proteome Machine (GPM) is a well-curated protein expression database consisting of peptides from multiple species [64]. According to this database, the expression of IRX4 was found in cell-lines HCT116, epithelial colorectal carcinoma cells, HEK293 and embryonic human kidney cells. B cells and Ascitic fluid too showed expression of the protein, highlighting the vast range of expression patterns for this protein.

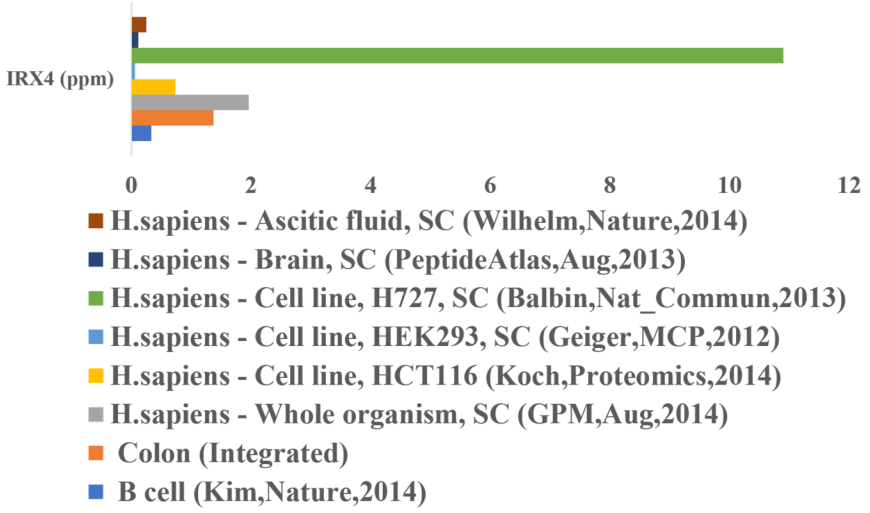

Figure 1: Protein expression analysis using PaxDB database. The expression of IRX4 was checked with multiple mass spectrometry datasets. The highest expression was observed in H727 cells representing the lung bronchus. Substantial expression was found in other tissues and cells, notably HCt116, Brian, HEK293 and B cells.

3.2 Sequence analysis

The full-length IRX protein sequences were aligned using Clustal omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) to identify sequence similarities between the family members (Figure 2). Interestingly, when the homeodomain region of the IRXs were aligned, there was a high sequence similarity except for the N-terminal. The DNA binding domain of IRX4 is comprised of 62 highly conserved residues.

The sequence similarity in the homeodomain of all IRX family members as depicted in Figure 3 shows the conservation of specific residues. Phylogenetic analysis between IRX members indicates a divergence in the N-terminal region of the homeodomain compared to the C-terminus, which has highly conserved amino acids. This shows the evolution of IRX family homeodomain and the functional specificity associated with the N and C terminal amino acid residues. The sequence has a relatively high hydrophobic residue count in the N-terminus and central region compared to the C-terminus, which has a stretch of basic residues. Various studies on homeodomain have highlighted the importance of the N-terminus, which has an essential role in DNA recognition and binding as well as for its transcriptional regulation[65]. The N-terminus in HOX proteins has also been observed to confer stable protein-protein interactions and also activate transcription and affect DNA binding activities in HOX proteins[66]. This could explain the specificity of IRX proteins in binding to different sequences and thereby regulating a wide variety of specific genes.

```
IRX4      MSYPQFGYPYSSAPQFLMATNSLSTCCESGGRTLADSGPAASAQAPVYCPVYESRLLATA  60
IRX6      MSFPHFGHPYRGASQFLASASSSTTCCESTQRSVSDVASGSTPAPALCCAPYDSRLLGSA  60
IRX2      MSYPQ-GYLYQAP-GSLALY-----SCPAYG-------------ASALAAPRSE---ELA  37
IRX5      MSYPQ-GYLYQPS-ASLALY-----SCPAYS-------------TSVISGPRTD---ELG  37
IRX1      MSFPQLGYPQYLSAAGPGAY-----GGERPGVL---AAAAAAAAAASSGRPGAA---ELG  49
IRX3      MSFPQLGYQYIRP-----LY-----PSERPGAA---GGSGGSAGARGGLGAGAS---ELN  44
          **:*: *:

IRX4      RHELNSAAALGVYGGPYGGS--------QGYGNYVTYGSEAS-----AFYSLN--SFDSK  105
IRX6      RPELG--AALGIYGAPYAAAA-----AAQSYPGYLPYSPEPP-----SLYGALNPQYEFK  108
IRX2      RSASGSAF------SPYPGSAAFTAQAATGFGSPLQYSADAA-AAAAGFPSYMGAPYDAH  90
IRX5      RSSSGSAF------SPYAGSTAFTAP-SPGYNSHLQYGADPAAAAAAAFSSYVGSPYD-H  89
IRX1      GGA-GAAAVTSVL-GMYAAAG--PYAGAPNYSAFLPYAADLS------LFSQMGSQYELK  99
IRX3      ASGSLSNVLSSVYGAPYAAAA--AAAAAQGYGAFLPYAAELP------IFPQLGAQYELK  96
                    . * .:         .:   : *. :        :       :: :

IRX4      DGSGSAHGGLAPAAAAYYPYEPALGQYPYDRYGTMDS--GTRRKNATRETTSTLKAWLQE  163
IRX6      EAAGSFTSSL-AQPGAYYPYERTLGQYQYERYGAVELSGAGRRKNATRETTSTLKAWLNE  167
IRX2      T-TG------MTGAISYHPYGSAAYPYQLND---------PAYRKNATRDATATLKAWLNE  135
IRX5      T-PG------MAGSLGYHPYAAPLGSYPYGD---------PAYRKNATRDATATLKAWLNE  134
IRX1      DNPGVHPATF---AAHTAPAYYPYGQFQYGD---------PGRPKNATRESTSTLKAWLNE  148
IRX3      DSPGVQHPAAAAAFPHPHPAFYPYGQYQFGD---------PSRPKNATRESTSTLKAWLNE  148
             *            *      :                 *****::*:******:*

IRX4      HRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKENKMTWPPRNKCADEKRPYAE  223
IRX6      HRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKENKMTWAPKNKGGEERKAE--  225
IRX2      HRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKENKMTWAPRNKSEDEDEDEGD  195
IRX5      HRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKENKMTWTPRNRSEDEEEEENI  194
IRX1      HRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKENKVTWGARSKDQEDGALFGS  208
IRX3      HRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKENKMTWAPRSRTDEEGNAYGS  208
          ****************************************:**  :.:  ::

IRX4      GEEEEGGEEEAREEPLKSSKNAEPVGKEEKELELSDLDDFDPLEAEPPACELKPPF----  279
IRX6      -----GGEEDSLG-CLTADTKEVTASQEARGLRLSDLEDLEEEEEEEEEAEDEEVVATAG  279
IRX2      ATRSKDES-PDKAQE--GT----ETSAEDEGISLHVD------SLTDHSCSAESD-----  237
IRX5      DLEKNDEDEPQKPED-KGD----PEGPEAGGAEQK----------AASGCER--------  231
IRX1      DTEGDPEKAEDD---EEID----LESIDIDKIDEHDGDQ----SNEDDEDKAEAP-----  252
IRX3      EREEEDEEEDEEDGKRELE----LEEEELGGEEEDTGGE----GLADDDEDEEID-----  255
                .                  :                  .

IRX4      ---HSLDGGLERVPAAPDGPVKEASGALRMSLAAGGGAALD------EDLERARSCLRSA  330
IRX6      DRLTEFRKGAQSLP----GPC--------------AAARE------GRLER-RECGLAA  313
IRX2      ---------GEKLPCRAGDPLCE-------SGSECKDKYDDLEDDEDDDEEGER-GL-AP  279
IRX5      ------------L----QGPPTP-------AGKETEGSLSDSDFKEPP-SEGRLDALQGP  267
IRX1      ----------------HAPAAP--SALARDQGS---PLAAADVL---K-------PQDS  280
IRX3      ---------LENLDGAATEPELSLAGAARRDGDLGLGPISDSKNS---DSEDSSEGLEDR  303
                           *

IRX4      ----------------------AA------GPEPLPGA------EGGPQVCEAKLGFVPAG  357
IRX6      ----------------------PR------FSFNDPSG------SEEADFLSAETGSPRLT  340
IRX2      PKPVTSSPLTGLEA----------PLLSPPPEAAPRGGRKTP-----------QGSRTSP  318
IRX5      PRTGGPSPAGPAAARLAEDPAPHYPAGAPAPGPHPAAGEVPPGPGGPSV----IHSPPPP  323
IRX1      PL----------------------GLAKEAPEPGSTRLL-SPG------------AAAGG  305
IRX3      PL----------------------PVLSLAPAPPPVAVA-SPSLPSPPV--SLDPCAPAP  338


IRX4      ASAGLEAKPRIWSLAHTATAAAAAATSL--SQTEFPSCMLK---------RQGPAAPAAV  406
IRX6      MHYPCLEKPRIWSLAHTATASAVEGAPPARPRPRSPECRMI---------PGQPPASA--  389
IRX2      GAPPPASKPKLWSLAEIATSDLKQPSLGP--G-----CGP----------PGLPAAAAP-  360
IRX5      PPPAVLAKPKLWSLAEIATSSDKVKDGGG--GNEGSPCPPCPGPIAGQALGGSRASPAP-  380
IRX1      LQGAPHGKPKIWSLAETATSPDGAPKASP--PPPAGHPGAHG-PSA--------------  348
IRX3      APASALQKPKIWSLAETATSPDNPRRSPP--GAGGSPPGAAVAPSA--------------  382
                ***::****. **:

IRX4      SSAPATSPSVALPHSGALDRHQD--SPV-TSLRNWVDGVFHDPIL--RH--STLN----Q  455
IRX6      -------RRLSVPRDSACDESS------------CIPKAFGNPKF--AL--QGLP-----  421
IRX2      ASTGAPPGGSPYPASPLLGRPLYYTSPFYGNYTNYGNLNAALQGQGLLR-----------  409
IRX5      APSRSPSAQCPFPGGTVLSRPLYYTAPFYPGYTNYGSFGHLHGHPGPGPGPTTGPG---S  437
IRX1      --------GAPLQHPAFLPSHGLYTCHI-GKFSNWTNSAFLAQG--SLLNMRSFLGVGAP  397
IRX3      --------LQLSPAAAAAAAHRLVSAPL-GKFPAWTNRPFPGPPPGPRLHPLSLLGSAPP  433


IRX4      AWATAKGALLDPGPLGRSLGAGANVLT----------------APLAR----------AFP  490
IRX6      ---------LN---------------C----------------APCPR----------RSE  432
IRX2      ----YNSAAAA---PGEALHTAPKAASDAG--KAGAHPLESHYRSPG---------GGYE  451
IRX5      HFNGLNQTVLN---RADALAKDPKMLRSQS--Q-----LDL---CKD---------SPYE  475
IRX1      HAAPHGPHLPAPPPPQPPVAIAPGALNGDKASVRSSPTLPERDLVPRPDSPAQQLKSPFQ  457
IRX3      HLLGLPGAAGH--PAAAAAFARPAEPEGGT-----------DRCSALEVEKKLLKTAFQ  479


IRX4      PAV----PQDAPAAGAARELLALPKAGGKPFCA      519
IRX6      PAV----QGQYPSGAEAG----------------      446
```

Figure 2: Sequence alignment of full-length IRX proteins using Clustal omega. The highlighted sequence is the homeodomain in the IRX proteins. Hydrophobic amino acids (A, V, F, P, M, I, L and W) are coloured red. Acidic amino acids (D and E) are coloured blue. Basic amino acids (R, H and K)-magenta, Others (S, T, Y, H, C, N, G and Q) are green.

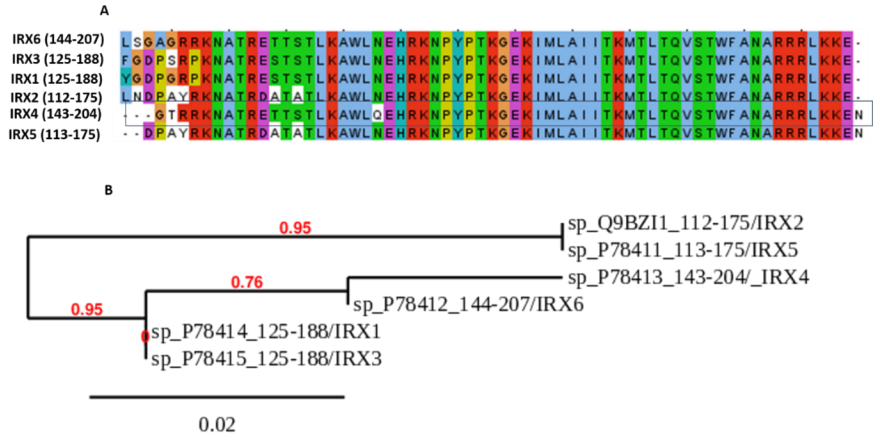

Figure 3: IRX homeodomain protein sequences and their conservation . A) Multiple sequence alignment shows high conservation in the DNA binding domain of IRX protein sequences. The difference at the N-terminal can explain the recognition specificity of these proteins to target DNA sequence B) Phylogenetic tree analysis highlighting highly conserved sequences. The tree shows evolution of the IRX homeodomain and also the functional specificity of the members of the family. The tree was constructed using the maximum likelihood method and derived from multiple sequence alignments of all IRX homeodomain region. The bootstrap values at the nodes of the branches indicate the maximum-likelihood of the sequence conservation between IRX proteins.

3.3 Modelling of homeodomain using SWISS-MODEL

The three-dimensional structure of the homeodomain of human IRX4 is not yet characterized using any biophysical methods like X-Ray or NMR. Secondary structure prediction using PSIPRED was carried out on a full-length sequence to check the ratio of secondary structures in the sequence (Figure 4A). Homeodomain from residues 143-204 showed the existence of 3 helices combined by the coiled region in between the helices. The PSIPRED results matched with the modelling result from SWISS-MODEL of the homeodomain, which correlated with the existence of 3 helices. The generated homology models were further subjected to an overall model quality check. The models were subjected to backbone dihedral angles (phi and psi) of the amino acid residues in the protein structure. Ramachandran plot (Figure S5) generated by PROCHECK showed 98.2% of the residues in the most favoured regions, highlighting the suitability and accuracy of the generated model[67]. The homology model shared similar structural folds with the template, i.e., 4 α-helices, 3 helix-helix interactions, 2 β-turns, and 1 γ-turn (Figure S6). We further analyzed the conservation profile of individual amino acids in the homeodomain structure. As expected, the most conserved amino acids were exposed to the surface with direct contact with DNA (Figure 4B). These residues were classified as conserved (N148, P168, W160, Y169, P170, K175, Q188, W192, F193, N195, R197, R198, R199) and semi-conserved (R146, K147, A149, L157, K172, A194). Apart from A194, all the other residues were found to be exposed on the surface, potentially contacting DNA, while all non-conserved residues are present on the distal surface. Interestingly, there are no somatic mutations reported for this residue.

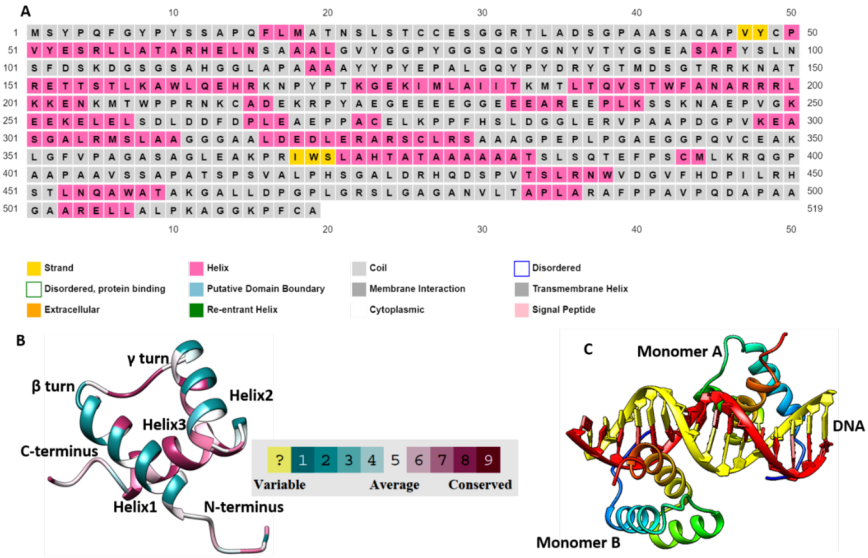

Figure 4: Secondary structure prediction and homology model of IRX4 homeodomain A) Secondary structure of IRX4 as predicted by PSIPRED. B) 3D structure of IRX4 homeodomain as modelled by SWISS-MODEL (figure generated using UCSF Chimera). The three conserved helices of the homeodomain as highlighted in pink. The conservation map of the IRX4-DNA Binding Domain was calculated using the ConSurf server. Ribbons are displayed based on conservation scores. C) The model of IRX4 dimer bound to DNA. The ribbons displayed in rainbow colours are IRX4-DNA Binding Domain dimers bound to DNA (coloured from blue to red representing N- to C-terminus of protein).

3.4 Identification of amino acid residues bound to DNA

We next modelled our homeodomain structure as a dimer to DNA based on the template 6FQP and identified specific amino acid residues interacting with DNA. As previously reported, homeodomain proteins in eukaryotes have similar DNA sequence binding preferences[68], we went ahead with the binding of IRX4 to DNA consensus binding site (ATTGACAGCTGTCAAT)[69]. DNAproDB was used for analysing interactions of protein-DNA complexes[70]. Before modelling the structure of IRX4 homodimer onto the DNA, TGIF1 PDB structure was analysed for its binding amino acids. As depicted in Figure S6, Arginine, Isoleucine and Asparagine on chains A and B of the dimer constitute the majority of interactions between DNA and protein. However, upon analyzing the binding of the IRX4 protein to DNA, specific amino acid residues, notably R145, T191, A194, N196, R198, and R199, were found to be interacting with DNA molecule (Figure 5). These interactions correlated with the high conservation of the residues bound to DNA based on Consurf. The evolutionarily conserved amino acid residues have been found to have well-defined structures involved in bound and unbound states[71]. Previous studies have also found conserved positions of amino acids to be structurally aligned and similar, providing key insights into the sequence-structure relationship [72]. MD simulations were carried out on this protein-DNA complex on the Wild Type (WT) and the mutant protein complexes to check the effect of mutations on the binding. The free energies of binding were calculated for the systems using Molecular Mechanics/Generalized Born Surface Area (MM/GBSA).

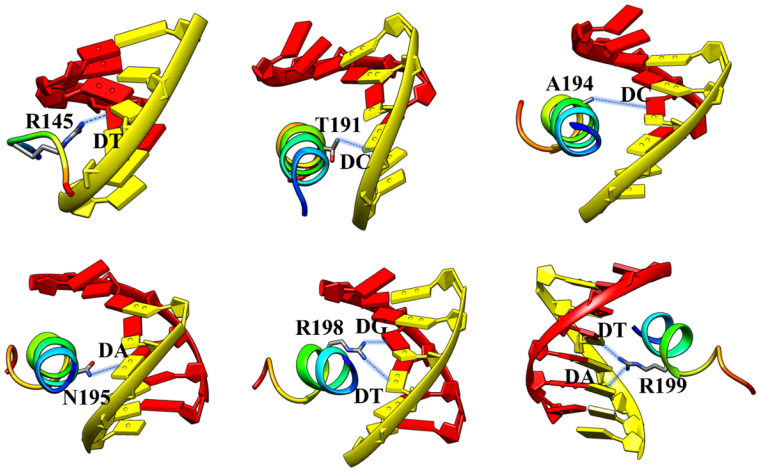

Figure 5: Representation of the energy minimized models of IRX4 homeodomain bound to DNA generated using UCSF Chimera . Arginine on the N-terminal of the homeodomain on both the monomers binds to the minor grove of DNA, highlighting the importance of arginines in this complex. Additional residues in the helix, including T191, A194, N196, R198, R199 bound to DNA. Red and yellow are the two strands of DNA. The dotted lines represent the interactions between amino acids and DNA bases. (coloured from blue to red represents N- to C-terminus of protein)

3.5 MD simulations of the WT protein-DNA complex

The protein-DNA system was subjected to 200 ns long MD simulations to examine the stability and interactions within the complex. The WT protein-DNA complex attained showed strong interactions with DNA at both N- and C-terminal of the homeodomain. Additional interactions of residues G143, T144, N148 and S190 were observed resulting in a more compact and stable protein-DNA complex (Figure 6). The interactions from the pre-MD model represents a snapshot from the MD-simulations after the minimisation step whereas the post-MD model snapshots represents an image of the interactions after 200ns. When comparing it to the template 6FQP, a vast array of amino acids have been found to be interacting on the both chains of the protein to DNA. This highlights the importance of strong and stable molecular interactions arginine, asparagine, threonine and alanine in stabilizing the IRX4 protein-DNA complex.

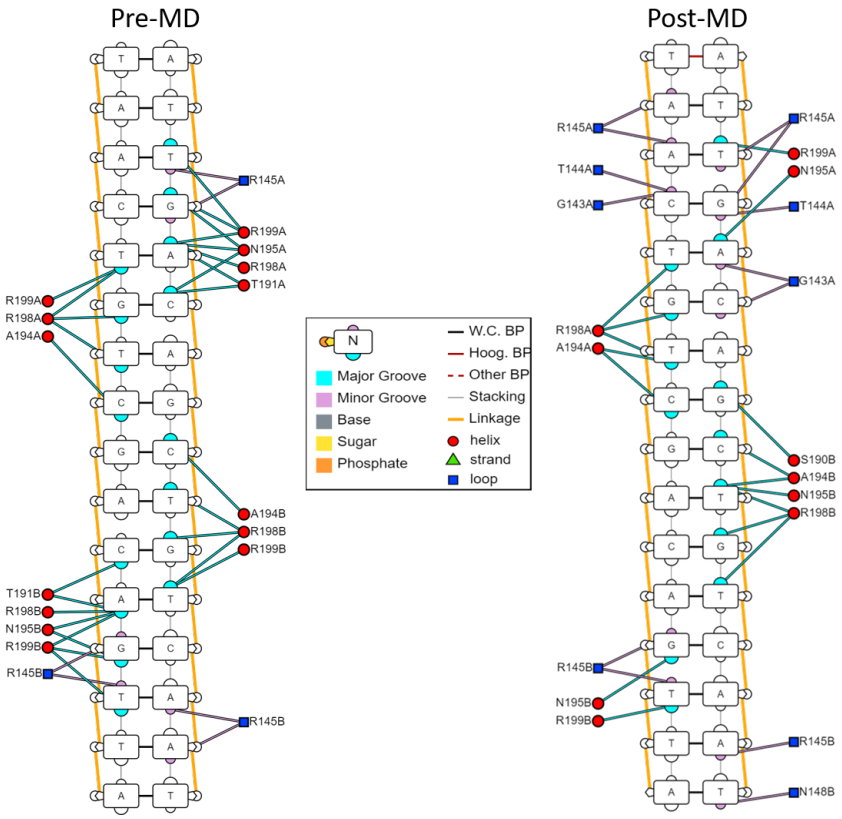

Figure 6: The interaction profile of WT IRX4 homeodomain-DNA complex before and after MD simulation . The amino acid residues from the αhelix of the protein is shown to have interactions in the pre-MD model of the complex. In contrast, the amino acids in the loop region of the protein structure seem to play an important role in stabilizing the complex post simulations. The pre-MD snapshot were taken after minimisation and post-MD snapshot after 200ns of simulations.

3.6 Retrieval of IRX4 mutations (Dataset)

The human IRX4 protein contained 253 somatic mutations retrieved from COSMIC database. The database consists of 163 missense, 11 nonsense, 72 coding silent, 3 each insertions and deletion mutations (Figure 7) wherein the DNA binding domain consists of 33 missense mutations, 1 nonsense and 15 coding silent mutations. High number of missense mutations are present in genes that are relevant to a disease phenotype. These mutations generate protein variants with a single amino acid change and are of particular interest in biomedicine[73]. Given that these single amino acid substitutions have an adverse effect on protein stability and may cause structural changes leading to aberrant binding surfaces and impairing protein function, their functional importance need to be studied using computational approaches before addressing their functional effects. The missense mutations in the homeodomain regions were selected for further predictive analysis. To examine the effect of substitutions on these residues, we performed an extensive comparative analysis of physicochemical properties such as theoretical Isoelectric point (PI), Instability Index (II), Aliphatic Index (AI) for the WT as well as mutant variants using Protparam as detailed in Table 1. The properties primarily PI, molecular weight, AI and extinction coefficient, showed minimal or no change due to the point mutations. However, the Grand Average of Hydropathicity (GRAVY) calculated based on the hydropathy values of each amino acid to the full length of the sequence, indicated a change in hydrophobicity[74]. Increasing positive scores for some of the mutants showed a greater hydrophobicity of those mutants compared to the native protein. Additionally, II of the protein predicts the stability of the protein in a solution state in a test tube.

Out of all the mutants, 15 showed an II of less than 40. The more hydrophobic protein mutants would be difficult to solubilize and are less likely to be resolve on a 2D gel electrophoresis[75].

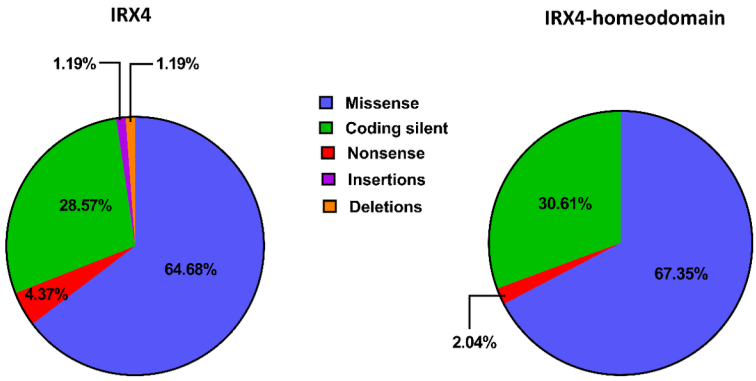

Figure 7: A statistical representation of the distribution of somatic mutations in the IRX4 full-length protein and the homeobox region . The profile indicates a higher percentage of missense mutations in both the full-length protein and the homeobox region.

Table 1: Physicochemical parameters computed using ProtParam tool

Protein	CDS Mutation	COSMIC ID	MWt	pI	EC	II	AI	GRAVY
WT-homeodomain			7301.51	11.38	12490	40.09	63.06	-1.037
G143S	c.427_428delinsTC	COSM387516	7331.54	11.38	12490	41.46	63.06	-1.044
T144R	c.431C>G	COSM8703091	7356.59	11.59	12490	50.7	63.06	-1.098
R145W	c.433C>T	COSM8591278	7331.54	11.1	17990	28.43	63.06	-0.979
R145L	c.434G>T	COSM4802014	7258.48	11.1	12490	33.96	69.35	-0.903
R146H	c.437G>A	COSM8557474	7282.46	11.1	12490	37.78	63.06	-1.016
A149V	c.446C>T	COSM9103754	7329.56	11.38	12490	38.72	66.13	-0.998
R151C	c.451C>T	COSM3381062	7248.46	10.77	12490	40.09	63.06	-0.924
R151H	c.452G>A	COSM245062	7282.46	11.1	12490	40.09	63.06	-1.016
E152K	c.454G>A	COSM3013209	7300.57	11.66	12490	40.09	63.06	-1.044
T156M	c.467C>T	COSM1254958	7331.6	11.38	12490	40.09	63.06	-0.995
E163K	c.487G>A	COSM5868355	7300.57	11.66	12490	38.2	63.06	-1.044
H164R	c.491A>G	COSM1066962	7320.56	11.59	12490	50.55	63.06	-1.058
Y169F	c.506A>T	COSM4141745	7285.51	11.69	11000	44.32	63.06	-0.971
T171I	c.512C>T	COSM3615067	7313.57	11.38	12490	38.72	69.35	-0.953
K172N	c.516G>T	COSM1254959	7287.44	11.37	12490	36.61	63.06	-1.031
I176V	c.526A>G	COSM8374433	7287.48	11.38	12490	40.09	61.45	-1.042
I180T	c.539T>C	COSM7350603	7289.46	11.38	12490	40.09	56.77	-1.121
I181V	c.541A>G	COSM8210484	7287.48	11.38	12490	37.35	61.45	-1.042
M184I	c.552G>A	COSM8467851	7283.48	11.38	12490	33.93	69.35	-0.995
T185S	c.554C>G	COSM4788359	7287.48	11.38	12490	47.65	63.06	-1.039
Q188E	c.562C>G	COSM5952047	7302.5	11.06	12490	45.63	63.06	-1.037
N195S	c.584A>G	COSM1210929	7274.49	11.38	12490	40.09	63.06	-0.994
A196T	c.586G>A	COSM6687514	7331.54	11.38	12490	38.72	61.45	-1.077
R197S	c.589C>A	COSM354476	7232.4	11.1	12490	33.96	63.06	-0.977

Protein	CDS Mutation	COSMIC ID	MWt	pI	EC	II	AI	GRAVY
R197C	c.589C>T	COSM3013206	7248.46	10.77	12490	37.94	63.06	-0.924
R197H	c.590G>A	COSM1066958	7282.46	11.1	12490	29.49	63.06	-1.016
R198W	c.592C>T	COSM4838791	7331.54	11.1	17990	30.85	63.06	-0.979
R199C	c.595C>T	COSM6451720	7248.46	10.77	12490	33.96	63.06	-0.924
R199H	c.596G>A	COSM8480424	7282.46	11.1	12490	33.96	63.06	-1.016
K201M	c.602A>T	COSM3941284	7304.53	11.37	12490	41.46	63.06	-0.944
K202M	c.605A>T	COSM4159835	7304.53	11.37	12490	45.35	63.06	-0.944
E203K	c.607G>A	COSM1695357	7300.57	11.66	12490	40.09	63.06	-1.044
N204S	c.611A>G	COSM3827909	7274.49	11.38	12490	43.2	63.06	-0.944

EC- Extinction coefficients, II- Instability index, AI-Aliphatic index, GRAVY- Grand average of hydropathicity

3.7 Screening for destabilizing mutants by I-Mutant Server and MD simulations

As reported in COSMIC, the mutations were mapped onto the 3D IRX4 homeodomain structure. These mutants were further checked for a change in protein stability using I-mutant server. ΔΔG value predicted by I-mutant server is prediction of the protein stability changes upon single point mutations in the protein. In contrast, the Reliability Index (RI) is a neural network predictor to check the overall accuracy of the function to the mutations to increase or decrease protein stability. Among the 33 mutations submitted, I-mutant predicted an increase in stability of 8 mutant proteins, namely R145L, A149V, R151C, Y169F, I180T, Q188E, K201M and E203K. All the conserved residues barring Y169F and A149V showed a decrease in protein stability, further highlighting the role of these mutants in DNA binding efficiency. Consurf (Figure S4) predicted most of the conserved residues and associated mutants showed a higher RI, which showed a decrease in stability of the IRX4 protein (Table 2).

Table 2: Predicting the stability of mutant proteins by I-Mutant Server

Protein	**RI**	**ΔΔΓ αλυε (Κςαλ/μολ)**	**I mutant score**
WT	Nil	Nil	Nil
G143S	9	-1.12	Decrease
T144R	6	-0.28	Decrease
R145W	1	-0.41	Decrease
R145L	1	0.29	Increase
R146H	2	0.1	Decrease
A149V	1	0.06	Increase
R151C	3	-0.44	Increase
R151H	3	0.17	Decrease
E152K	9	-1.37	Decrease
T156M	4	-0.52	Decrease
E163K	8	-1.44	Decrease
H164R	5	-0.78	Decrease
Y169F	3	0.52	Increase
T171I	6	-1.17	Decrease
K172N	7	-1.07	Decrease
I176V	7	-0.04	Decrease
I180T	1	1.17	Increase
I181V	6	0.5	Decrease
M184I	3	0.04	Decrease
T185S	4	0.03	Decrease

Protein	RI	ΔΔΓ αλυε (Κςαλ/μολ)	I mutant score
Q188E	4	0.3	Increase
N195S	6	-0.24	Decrease
A196T	8	-1.26	Decrease
R197S	9	-2.36	Decrease
R197C	4	-0.73	Decrease
R197H	9	-1.32	Decrease
R198W	7	-0.96	Decrease
R199C	3	-0.83	Decrease
R199H	8	-0.89	Decrease
K201M	3	0.06	Increase
K202M	2	0.19	Increase
E203K	9	-1.14	Decrease
N204S	2	0.04	Decrease

Moreover, the energy minimised structures of the mutant protein-DNA complex showed weak interactions throughout the complex. (Figure S7). For example, when R145 was replaced with W and L amino acids, weak interactions between these amino acids and DNA resulted in compensation by other amino acids of the protein to stabilize the protein. Furthermore, to highlight the effects of the mutations, energy, stability, hydrogen bond and secondary structure analysis were carried out.

3.8 Binding Mechanisms

3.8.1 MM/GBSA free energy calculations

To understand the thermodynamics behind the protein-DNA complexes, MM-GBSA was calculated to determine the binding mechanisms. The calculations were performed to calculate the binding energies of WT protein comparing it to mutant protein in the DNA-protein complex. Based on the 200ns MD trajectories of WT and variants the energy analysis and its corresponding component as calculated and reported in Table 3. The results indicated high binding energy (-216.33 Kcal/mol) compared to most other variants. The point mutations decreased the positive polar term resulting in overall increase of the negative term promoting DNA-protein complex formation. The energy components included here are Van der Waals, electrostatic interaction, non-polar and polar energies contributing to the total binding energy and favouring complex formation. Primarily, the stability of the complex was majorly due to the electrostatic interaction energy whereas other energy terms contributed very less to the total energy of the complex. The spontaneity of this interactions is highlighted by the high negative value of the electrostatic interactions, whilst polar solvation energies hinders the interactions between DNA and protein. Interestingly, some of the residues showed higher binding stability when mutated resulting in higher negative values in T144R, A149V and Q188E. N-terminal R (arginine) showed a loss in interaction energy (-188.74 Kcal/mol) when mutated to lysine. However, when mutated to tryptophan, it showed considerable increase in binding energy (-224.41) of the mutant. Studies have highlighted the role of tryptophan in DNA binding domain to be important, which might be the case in this mutant also[76]. The same effect was observed on the C-terminal R when mutated to tryptophan (R198W= -224.43 Kcal/mol). But when mutated to cysteine (-186.36 Kcal/mol) and histidine (-185.16 Kcal/mol) there was a loss of interaction energy of the mutant. These residues were also found to be highly conserved in the Consurf analysis. Furthermore, this also implies about the role of these conserved residues to predominantly be the fundamental residues driving a compact interaction with DNA.

Table 3: Decomposition of calculated binding energies using MM/GBSA

System	Energy Components (Kcal/mol)	Energy Components (Kcal/mol)	Energy Components (Kcal/m
	ΔδΩ	ΔΕΕΛ	$\Delta\Gamma_{\gamma\alpha\varsigma}$ (δΩ +ΕΕΛ)

System	Energy Components (Kcal/mol)	Energy Components (Kcal/mol)	Energy Components (Kcal/m
WT	-179.58	-12741.14	-12920.73
T144R	-200.83	-13973.14	-14173.97
R145L	-154.25	-11330.80	-11485.05
R145W	-176.69	-11768.18	-11944.88
A149V	-199.55	-12685.84	-12885.39
Y169F	-191.86	-12220.75	-12412.62
K172N	-177.34	-11831.65	-12008.99
Q188E	-191.04	-11901.53	-12092.57
N195S	-175.27	-12576.62	-12751.90
R197C	-175.22	-10911.09	-11086.32
R197H	-183.60	-11180.89	-11364.50
R197S	-197.59	-11790.22	-11987.81
R198W	-193.84	-11505.21	-11699.05
R199C	-164.24	-11674.88	-11839.13
R199H	-172.79	-11847.76	-12020.56

3.8.2 Stability of IRX4 homeodomain-DNA complex

The protein-DNA complex was subjected to long MD simulations of 200ns time period. We calculated root mean square fluctutation (RMSF) for homeodomain-DNA complex. From the RMSF plot (Figure 8), it was observed that the residues directly bound to DNA experienced slightly high fluctuations in all the protein structures. The residues on the C-terminal of the homeodomain were found to have slightly higher RMSF values indicating flexibility of this region of the homedomain. The residues 155-165 had lower RMSF values reflecting these residues to have a good potential to be bound to DNA. Important residues, primarily the one which were found to have binding to DNA showed less flexibility in their conformational state. The mutants K172N and Y169F showed higher fluctuations when the original amino acid were mutated compared to the other mutants. Subsequent energy analysis of these residues and mutants were done to confirm the effect of the mutations on binding to DNA.

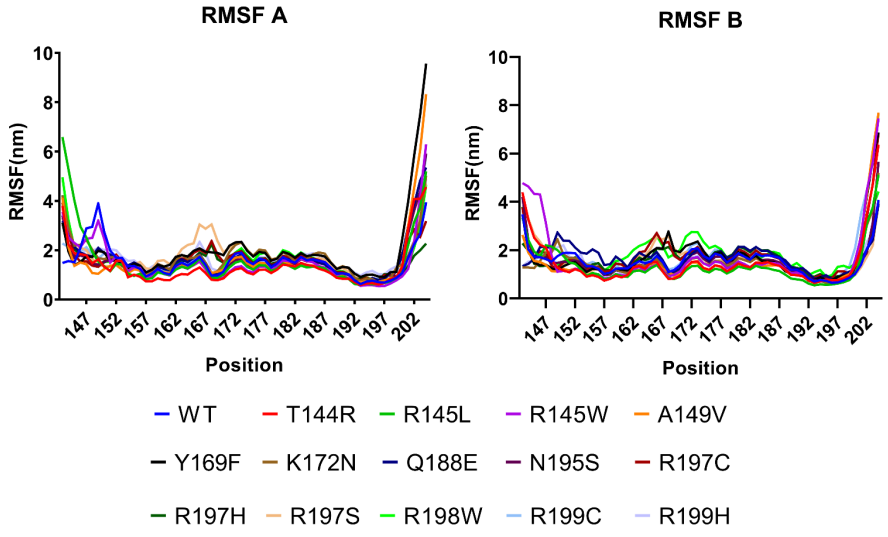

Figure 8: The representative RMSF values of native IRX4 homeodomain protein structure and the mutants. RMSF-A and RMSF-B represents the fluctuations in amino acid residues in Monomer A and B.

3.8.3 Hydrogen bond analysis of protein-DNA complex

Hydrogen bonds were analyzed between the IRX4 homeodomain-DNA complex to find important residues involved in hydrogen bonding. The key criteria for this analysis were as previously reported[77]. A total of 7 hydrogen bonds were observed at the protein-DNA interface for the WT IRX4 protein as reported in Table 4. The hydrogen bonds are illustrated as donor-acceptor bonds between DNA and protein. The atoms represent the side chain and backbone of the DNA and protein. The residues K145, A149, Y169, S190, R197 and R199 were involved in forming multiple hydrogen bonds and stabilizing the interactions. Hydrogen bonds having fractions of more than 0.5 were considered as having interaction between DNA and amino acids. There was a loss of hydrogen bonds in the arginine mutant variants which highlighted the conservation of these residues in DNA recognition. There was loss of hydrogen bonds in all the mutants. Weak hydrogen bonds were formed between residues compared to WT protein which showed change in flexibility of the residues binding to DNA. The Arginines on the C-terminal side of the homedomain when mutated affected hydrogen bonding shown in Table 5. The N-terminal Arginines showed hydrogen compensation when mutated resulting in formation of weak hydrogen bonds to maintain the stability of protein-DNA complex.

Table 4: The hydrogen bonds between WT IRX4 homeodomain and DNA

Donor	Acceptor
R-145-NH1	DT3-N3
A-149-N	DT3-OP2
Y-169-OH	DC9-OP1
ARG-197-NH2	DC9-OP2
R-199-NH2	DT19-OP2
R-197-NE	DC9-OP2
S-190-OG	DG8-OP2

Table 5: Hydrogen bonds in the homeodomain mutants

Mutants	Hydrogen Bonds
T144R	R197-DC9 S190-DG8 Y169-DC9 T154-DG20 R199-DG20 W192-DG20
R145L	Y169-DC9 R197-DC9 T154-DG20 S190-DG8 R199-DG20 Q188-DA21 R198-DG11 R197-DC9
R145W	Y169-DC9 R197-DC9 T154-DG20 N148-DT3 S190-DG8 R151-DT3
A149V	R145-DT19 R199-DG20 Y169-DC9 R146-DA21 R197-DC9 R198-DG11
Y169F	R199-DG20 R198-DG11 R197-DC9 K147-DT3 N195-DA21
K172N	Y169-DC9 R197-DC9 R198-DG11 T144-DA15 S190-DG8
Q188E	R197-DC9 Y169-DC9 R199-DT19 R145-DA14 S190-DG8
R197C	R199-DG20 R145-DT19 R146-DA21 N195-DA21 R198-DG11
R197H	R199-DG20 A149-DG20 R145-DT19 N195-DA21 R146-DA21
R197S	T144-DG20 R199-DG20 R198-DG11 Q188-DA21 R151-DT18

Mutants	Hydrogen Bonds
R198W	Y169-DC9 R197-DC9 T144-DG20 S190-DG8 R146-DC22
R199C	Y169-DC9 R197-DC9 S190-DG8 R198-DG11
R199H	Y169-DC9 R197-DC9 R145-DT19

3.8.4 Secondary structure changes during simulation

The average secondary structure transitions in each of the simulations were monitored using Define Secondary Structure of Proteins (DSSP) program [78]. Simulations started from residues in contact with the interface (DNA); these residues include both the N- and C-termini of both the monomers of IRX4. A scan be seen in Fig S8 and S9, the N-terminal region of the homeodomain have attained quite a distinct secondary structure compare to the mutants. Secondary structure analysis was used to identify the extent of changes which helps in promoting the structural integrity of the binding. Compared to the native structure, A149V after initial 50 ns showed a robust helical propensity compared to other mutants. The helical conformation amongst the other mutants were consistently altered throughout the 200ns simulations involving turn structures. Interestingly the smallest helix consisting of 9 amino acid (174-182) didn't show any major differences along the different mutants. Examination of individual residue energy difference highlights the fact there is reorientation of the protein from its initial orientation in the mutants, highlighting the effect of the variants on DNA binding. The final orientation and secondary structure in all the variants is different, exhibiting loss of loss or gain in bending of the helix.

3.9 Residue-wise energy decomposition analysis

3.9.1 Energy landscape changes in nucleotides

The simulated complex of the IRX4-DNA was subjected to energy calculation difference amongst the different nucleotide residues in the DNA (Figure 11). When compared to the native protein, all the variants showed a decrease in energy profiles amongst the nucleotide. This highlights the fact that there was a decrease in binding potential of the residues to DNA nucleotides. The maximum fluctuations were seen at the DNA binding residues pronouncing the role of these residues in binding. Consequently, all the variants were rigid in nature with changes in the amino acid sequence resulting in decrease in the interaction of the molecules. Importantly the residue R197, R198 and R199 showed the most fluctuations in the energy landscape of the DNA molecule, showing the importance of the C-terminal Arginines in stabilizing the DNA fluctuations. Previous results on arginine mutants in the DNA binding domain of STAT3 have shown changes in intracellular shuttling and phosphorylation[79]. This also confirms that point mutations at the DNA binding residues causes a destabilizing effect leading to protein compactness and change in binding potential. Additionally, these mutations can also generate neomorphic functions resulting in disease phenotype.

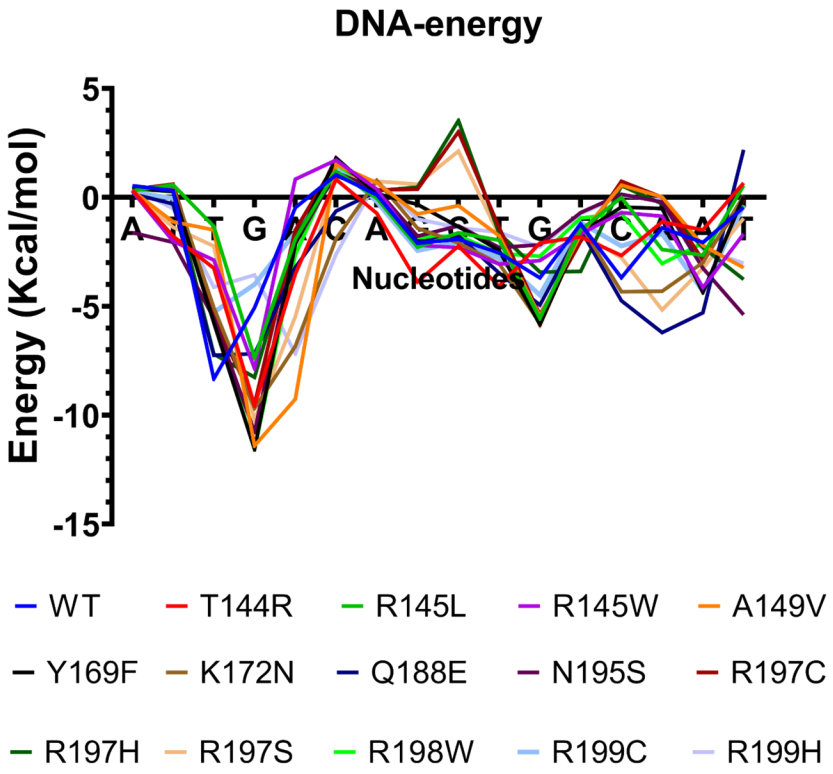

Figure 9: Average free energy of binding per nucleotide in the DNA for all variants. Graph showing energy difference amongst the native and mutant proteins with respect the DNA nucleotide sequence.

3.9.2 Energy landscape per amino acid in the homeodomain

Multi-dimensional energy plots, attributed to each amino acid in the homeodomain were generated to analyze the change in energy landscape across the sequence. As depicted in Fig S10 and S11, energy plots for both monomers show a variation in the energy dynamics of the protein structure after simulations. The changes in the C-terminal domain of the protein can be seen in all the variants compared to the WT protein. Interestingly, the amino acid residues interacting with DNA showed a decreased in binding potential in all the mutants. The N- terminal residues K147, N148, A149 and Y169 showed a decrease in binding energy in the variants. The highlight were the C-terminal residue mutants which changed the overall binding potential of the protein, affecting the flexibility of the helix and decreasing the potential of IRX4 to bind DNA.

3.10 Cell-penetrating peptide sequence in IRX4 homeodomain

TFs were considered as undruggable targets. Several approaches to target these TFs, such as inhibiting the protein-protein interactions and TF-DNA binding have been demonstrated in recent years[80]. DNA has been the target for various antiviral, anti-cancer and antimicrobial drugs wherein ligands bind to the major groove for inhibiting DNA-protein interactions [81]. Interactions of short peptides with DNA and their effectiveness as anti-tumour, antibacterial or anti-inflammatory have recently been gaining attention. These peptides despite being an attractive entity, need work for improving their stability and their ability to penetrate the cells [82]. Cell penetrating peptides (CPPs) have shown to be a potential therapeutic delivery agent entering cells via endocytosis. These CPPs have been found to be positively charged due to presence of several Arginine/lysine residues[83]. To check the efficacy of the homeodomain penetrating the cell membrane, we used different tools and databases including CPP site2.0[84], a database of cell-penetrating peptides and SkipCPP-Pred[61]. Using the CPP site2.0, we found more than 50% similarities in the C-terminal sequence,

residues from 184-204 (MTLTQVSTWFANARRRLKKEN) of IRX4. The presence of positively charged amino acids in the C-terminal region highlights the potential of this region to be effective in cell penetration.

Additionally, this sequence from C-terminal homeodomain was subjected to SkipCPP-Pred to predict the effectiveness of this sequence (Table 6). The full-length 62 amino acid homeodomain had a low prediction confidence of 0.57 compared to the C-terminal residues, highlighting their potential to be in the class of cell penetrating peptides. The C-terminal sequence of IRX4 shares similarity with other IRXs. However, N-terminal of the homeodomain in all the IRX proteins offer specificity to these peptides.

Table 6: Prediction of the cell-penetrating ability of homeodomain sequence

Residue numbers	Sequence	Prediction Confidence
143-204	IRX4-homeodomain	0.57
184-204	MTLTQVSTWFANARRRLKKEN	0.59
185-204	TLTQVSTWFANARRRLKKEN	0.71
186-204	LTQVSTWFANARRRLKKEN	0.67
187-204	TQVSTWFANARRRLKKEN	0.62
188-204	QVSTWFANARRRLKKEN	0.61
189-204	VSTWFANARRRLKKEN	0.68
190-204	STWFANARRRLKKEN	0.73
191-204	TWFANARRRLKKEN	0.75
192-204	WFANARRRLKKEN	0.75
193-204	FANARRRLKKEN	0.80
194-204	ANARRRLKKEN	0.84
195-204	NARRRLKKEN	0.84

Additionally, we also checked the ability of the mutant protein to be a cell-penetrating peptide sequence. As depicted in Table 7, the mutation of Arginines on the C- and N-terminal of the homeodomain leads to a decrease in confidence of the peptide to cell penetrating. As previously mentioned, cellular uptake peptides is increased due to multiple arginines by attaching fatty acid to N-terminal of the peptide[85]. It has also been mentioned that the presence of more than six arginine residues is critical for efficient cell-penetrating functions[85]. The ability of these peptides for convenient cellular uptake needs to be studied in detail to find their efficacy as therapeutic agents targeting IRX homeodomain family in different cancers.

Table 7: Prediction of the cell-penetrating ability of mutant homeodomain

Mutants	Prediction confidence
T144R	0.60
R145L	0.52
R145W	0.51
A149V	0.55
Y169F	0.59
K172N	0.50
Q188E	0.52
N195S	0.59
R197C	0.54
R197H	0.53
R197S	0.56
R198W	0.53
R199C	0.54
R199H	0.53

4. Conclusion

The three-dimensional structure of the DNA binding domain of IRX4 is central to the protein being bound to DNA culminating in context-specific transcriptional programs. To better understand the structural basis of DNA binding and the effect of mutations, we integrated homology modelling and MD simulations. Our results suggest that the amino acid residues that are in contact with DNA be highly conserved across protein families. These residues provide a platform for stable DNA-protein interactions. Upon analyzing the IRX4 homeodomain sequence, we tried to investigate if the amino acid residues bound to DNA have high levels of conservation. Unhighly specific proteins like the Iroquois family, base contacting residues are highly conserved, allowing member proteins to recognize the same target sequence. Here, we found strong interactions of R145, T191, A194, N195, R198 and R199 to the DNA molecule, which also showed higher confidence in the conservation scale. Post-MD simulations additional residues including N- terminal residues were found to interact to DNA nucleotides.The mutations on residues interacting with DNA may disable to protein to recognize the target sequences and bind to DNA.

Hydrogen bonds and hydrophobic interactions play a significant role in stabilizing protein: DNA interaction. MD simulations of 200ns were used to check the behaviour of the interaction profile. The residues that were found to interact with DNA bases G143, T144, R145, N148 formed part of the protein loop region. Additionally, amino acids in the helix, S190, A194, N195, R198 and R199 formed strong interaction with the DNA post MD simulations. Mutations affecting the binding amino acids were also screened for affecting the interaction. RMSF showed greater fluctuations in the mutants which were directly interacting with the DNA. Alternatively, the interaction energy profile showed similar trends as the total energy of the interaction complex decreased compared to WT. The mutants at R145 (R145L), Y169 (Y169F). R197 (R197C and R197H) and R199 (R199C and R199H) showed a decrease in total energy and stability of the complex. Protein-DNA recognition is a critical component of gene regulation and several amino acid residues play important roles in this process. The Arginine at the N-terminal of the homeodomain has been found to serve as core element in recognizing DNA and mutations at this positions have markedly reduced DNA binding activity as per previous reports. Additional, Arginine at the C-terminal region of homeodomains is essential for conformational stability of the recognition helix for optimal DNA recognitions. Our data correlates with previous findings wherein mutations at important residues have resulted in a decrease in protein stability as predicted by I-mutant as well as reduced DNA binding. These hotspots seem to be very important in the IRX4 homeodomain region which might cause severe change in the phenotype of diseases.

Interestingly, the Arginine at the C-terminal sequence is part of the peptide that is highly confident of being cell-penetrating. The C-terminal arginine-rich sequence provides an interesting side to the use of these peptides to knockdown representative binding of oncogenic homeodomain TFs. Taken together, we examined the distinct role of IRX4 homeodomain, accentuating the mechanism of DNA recognition and the stability of the complex. Our outcome delivers fundamental insights into the structural and thermodynamic stability of IRX4-DNA binding which could have implications in various cancers. These mutations if validated experimentally could have a significant effect in the regulation of downstream genes affected by IRX4. This work offer insight into the role of these mutations in thermodynamic genesis during the development of tumorigenesis and having specific phenotypic effects.

Author Contributions

JB designed the research study, NSG and AM performed the computational research. NSG and JB supervised the study and gave intellectual input. AM wrote the manuscript. All authors contributed to finalizing the manuscript. All authors listed have made a substantial, direct and intellectual contribution to the work and approved it for publication.

Funding

JB and NSG were supported by Advance Queensland Industry Research Fellowship. JB acknowledges support from NHMRC Career Development Fellowship. AM acknowledges support by QUTPRA scholarship.

Conflict of Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Acknowledgements

Computational (and/or data visualisation) resources and services used in this work were provided by the eResearch Office, Queensland University of Technology, Brisbane, Australia. A big thanks to Ms. Dulari K Jayarathna and Ms. Shubhra Chandra for their contribution and help in preparing images and moral support.

References

1. Li B, Huang Q, Wei G-H. The Role of HOX Transcription Factors in Cancer Predisposition and Progression.*Cancers (Basel).* 2019;11(4):528.

2. Abate-Shen C. Deregulated homeobox gene expression in cancer: cause or consequence? *Nature reviews Cancer.* 2002;2(10):777-785.

3. Samuel S, Naora H. Homeobox gene expression in cancer: insights from developmental regulation and deregulation. *European journal of cancer (Oxford, England : 1990).* 2005;41(16):2428-2437.

4. Shah N, Sukumar S. The Hox genes and their roles in oncogenesis. *Nature reviews Cancer.*2010;10(5):361-371.

5. Lu J, Song G, Tang Q, et al. IRX1 hypomethylation promotes osteosarcoma metastasis via induction of CXCL14/NF-kappaB signaling. *The Journal of clinical investigation.* 2015;125(5):1839-1856.

6. Megias-Vericat JE, Montesinos P, Herrero MJ, et al. Impact of novel polymorphisms related to cytotoxicity of cytarabine in the induction treatment of acute myeloid leukemia.*Pharmacogenetics and genomics.* 2017;27(7):270-274.

7. Wu Y, Davison J, Qu X, et al. Methylation profiling identified novel differentially methylated markers including OPCML and FLRT2 in prostate cancer. *Epigenetics.*2016;11(4):247-258.

8. Mann RS, Lelli KM, Joshi R. Chapter 3 Hox Specificity: Unique Roles for Cofactors and Collaborators. In:*Current Topics in Developmental Biology.* Vol 88. Academic Press; 2009:63-101.

9. Berger MF, Badis G, Gehrke AR, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell.* 2008;133(7):1266-1276.

10. Pomerantz JL, Sharp PA. Homeodomain Determinants of Major Groove Recognition.*Biochemistry.* 1994;33(36):10851-10858.

11. Connolly JP, Augustine JG, Francklyn C. Mutational analysis of the engrailed homeodomain recognition helix by phage display. *Nucleic Acids Res.*1999;27(4):1182-1189.

12. Bürglin TR. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. *Nucleic Acids Research.* 1997;25(21):4173-4180.

13. Cavodeassi F, Modolell J, Gomez-Skarmeta JL. The Iroquois family of genes: from body building to neural patterning. *Development (Cambridge, England).*2001;128(15):2847-2855.

14. Cheng CW, Chow RL, Lebel M, et al. The Iroquois homeobox gene, Irx5, is required for retinal cone bipolar cell development. *Developmental biology.*2005;287(1):48-60.

15. Feijoo CG, Manzanares M, de la Calle-Mustienes E, Gomez-Skarmeta JL, Allende ML. The Irx gene family in zebrafish: genomic structure, evolution and initial characterization of irx5b. *Development genes and evolution.* 2004;214(6):277-284.

16. Gomez-Skarmeta JL, Modolell J. Iroquois genes: genomic organization and function in vertebrate neural development. *Current opinion in genetics & development.*2002;12(4):403-408.

17. Kerner P, Ikmi A, Coen D, Vervoort M. Evolutionary history of the iroquois/Irx genes in metazoans.*BMC Evolutionary Biology.* 2009;9:74-74.

18. Myrthue A, Rademacher BL, Pittsenbarger J, et al. The iroquois homeobox gene 5 is regulated by 1,25-dihydroxyvitamin D3 in human prostate cancer and regulates apoptosis and the cell cycle in LNCaP prostate cancer cells.*Clinical cancer research : an official journal of the American Association for Cancer Research.* 2008;14(11):3562-3570.

19. Werner S, Stamm H, Pandjaitan M, et al. Iroquois homeobox 2 suppresses cellular motility and chemokine expression in breast cancer cells. *BMC Cancer.* 2015;15(1):896.

20. Xu X, Hussain WM, Vijai J, et al. Variants at IRX4 as prostate cancer expression quantitative trait loci.*European journal of human genetics : EJHG.* 2014;22(4):558-563.

21. Fernando A, Liyanage C, Moradi A, Janaththani P, Batra J. Identification and Characterization of Alternatively Spliced Transcript Isoforms of IRX4 in Prostate Cancer.*Genes.* 2021;12(5):615.

22. Chakma K, Gu Z, Motoi F, Unno M, Horii A, Fukushige S. Abstract 821: DNA hypermethylation of IRX4 is a frequent event that may confer growth advantage to pancreatic cancer cells. *Cancer Research.* 2019;79(13 Supplement):821-821.

23. Ha Nguyen H, Takata R, Akamatsu S, et al. IRX4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin D receptor, conferring prostate cancer susceptibility. *Human Molecular Genetics.* 2012;21(9):2076-2085.

24. Zhang D-L, Qu L-W, Ma L, et al. Genome-wide identification of transcription factors that are critical to non-small cell lung cancer. *Cancer Letters.* 2018;434:132-143.

25. Corrêa S, Panis C, Binato R, Herrera AC, Pizzatti L, Abdelhay E. Identifying potential markers in Breast Cancer subtypes using plasma label-free proteomics. *Journal of Proteomics.* 2017;151:33-42.

26. Morey SR, Smiraglia DJ, James SR, et al. DNA Methylation Pathway Alterations in an Autochthonous Murine Model of Prostate Cancer. *Cancer Research.*2006;66(24):11659-11667.

27. Wang P, Zhuang C, Huang D, Xu K. Downregulation of miR-377 contributes to IRX3 deregulation in hepatocellular carcinoma. *Oncol Rep.* 2016;36(1):247-252.

28. Holmquist Mengelbier L, Lindell-Munther S, Yasui H, et al. The Iroquois homeobox proteins IRX3 and IRX5 have distinct roles in Wilms tumour development and human nephrogenesis. *The Journal of Pathology.* 2019;247(1):86-98.

29. Chandrasekaran G, Hwang EC, Kang TW, et al. Computational Modeling of complete HOXB13 protein for predicting the functional effect of SNPs and the associated role in hereditary prostate cancer. *Sci Rep.* 2017;7:43830-43830.

30. Bagchi A, Roy D, Roy P. Homology modeling of a transcriptional regulator SoxR of the Lithotrophic sulfur oxidation (Sox) operon in alpha-proteobacteria. *J Biomol Struct Dyn.* 2005;22(5):571-577.

31. Nayan MY, Jusoh SA, Mutalip SSM, Mohamed R. Homology modeling of the DNA binding and dimerization partner domains of E2F1 transcription factor protein in homo sapiens. Paper presented at: 2012 IEEE Symposium on Business, Engineering and Industrial Applications; 23-26 Sept. 2012, 2012.

32. Gao M, Zhou H, Skolnick J. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure.* 2015;23(7):1362-1369.

33. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol.* 2005;353(2):459-473.

34. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol.* 2019;431(11):2197-2212.

35. Siddiqui MQ, Badmalia MD, Patel TR. Bioinformatic Analysis of Structure and Function of LIM Domains of Human Zyxin Family Proteins. *Int J Mol Sci.* 2021;22(5):2647.

36. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold.*Nature.* 2021;596(7873):583-589.

37. Samaras P, Schmidt T, Frejno M, et al. ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Research.*2019;48(D1):D1153-D1163.

38. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines.*Proteomics.* 2015;15(18):3163-3168.

39. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7:539.

40. Forbes SA, Bhamra G, Bamford S, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet.* 2008;Chapter 10:Unit-10.11.

41. Buchan DWA, Jones DT. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research.*2019;47(W1):W402-W407.

42. Ashkenazy H, Abadi S, Martz E, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research.* 2016;44(W1):W344-W350.

43. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 2018;46(W1):W296-w303.

44. Bilioni A, Craig G, Hill C, McNeill H. Iroquois transcription factors recognize a unique motif to mediate transcriptional repression <em>in vivo</em&gt.*Proceedings of the National Academy of Sciences of the United States of America.* 2005;102(41):14671.

45. Sagendorf JM, Berman HM, Rohs R. DNAproDB: an interactive tool for structural analysis of DNA-protein complexes. *Nucleic acids research.* 2017;45(W1):W89-W97.

46. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. *Protein Science : A Publication of the Protein Society.*2018;27(1):129-134.

47. Wilkins MR, Gasteiger E, Bairoch A, et al. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol.* 1999;112:531-552.

48. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research.*2005;33(Web Server issue):W306-W310.

49. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature methods.* 2010;7(4):248-249.

50. Case D, Betz R, Cerutti DS, et al. *Amber 16, University of California, San Francisco.* 2016.

51. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics.*1983;79(2):926-935.

52. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation.* 2015;11(8):3696-3713.

53. Pérez A, Marchán I, Svozil D, et al. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J.*2007;92(11):3817-3829.

54. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.*Journal of Computational Physics.* 1977;23(3):327-341.

55. Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation.*2013;9(7):3084-3095.

56. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem.* 2004;25(13):1605-1612.

57. Kollman PA, Massova I, Reyes C, et al. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research.* 2000;33(12):889-897.

58. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery.* 2015;10(5):449-461.

59. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics.* 2004;55(2):383-394.

60. Agrawal P, Bhalla S, Usmani SS, et al. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Res.*2016;44(D1):D1098-1103.

61. Wei L, Tang J, Zou Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics.*2017;18(7):742.

62. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.*2013;45(6):580-585.

63. Hruz T, Laule O, Szabo G, et al. Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes. *Advances in Bioinformatics.* 2008;2008:420747.

64. Craig R, Cortens JP, Beavis RC. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *Journal of Proteome Research.*2004;3(6):1234-1242.

65. Simard A, Di Giorgio L, Amen M, Westwood A, Amendt BA, Ryan AK. The Pitx2c N-terminal domain is a critical interaction domain required for asymmetric morphogenesis.*Dev Dyn.* 2009;238(10):2459-2470.

66. Di Rocco G, Mavilio F, Zappavigna V. Functional dissection of a transcriptionally active, target-specific Hox–Pbx complex. 1997;16(12):3644-3654.

67. Hollingsworth SA, Karplus PA. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts.* 2010;1(3-4):271-283.

68. Bobola N, Merabet S. Homeodomain proteins in action: similar DNA binding preferences, highly variable connectivity. *Current opinion in genetics & development.*2017;43:1-8.

69. Guca E, Suñol D, Ruiz L, et al. TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF-β signaling. *Nucleic Acids Res.* 2018;46(17):9220-9235.

70. Sagendorf JM, Markarian N, Berman HM, Rohs R. DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Research.* 2019;48(D1):D277-D287.

71. Saldaño TE, Monzon AM, Parisi G, Fernandez-Alberti S. Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLOS Computational Biology.*2016;12(3):e1004775.

72. Friedberg I, Margalit H. Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function.*Protein science : a publication of the Protein Society.*2002;11(2):350-360.

73. Petrosino M, Novak L, Pasquo A, et al. Analysis and Interpretation of the Impact of Missense Variants in Cancer. *Int J Mol Sci.* 2021;22(11):5416.

74. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 1982;157(1):105-132.

75. Chang KY, Yang J-R. Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS One.* 2013;8(8):e70166-e70166.

76. Kanei-Ishii C, Sarai A, Sawazaki T, et al. The tryptophan cluster: a hypothetical structure of the DNA-binding domain of the myb protooncogene product. *The Journal of biological chemistry.* 1990;265(32):19990-19995.

77. Yuan L, Wu H, Zhao Y, Qin X, Li Y. Molecular simulation of the interaction mechanism between CodY protein and DNA in Lactococcus lactis. *Frontiers of Chemical Science and Engineering.* 2019;13(1):133-139.

78. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 1983;22(12):2577-2637.

79. Ginter T, Fahrer J, Kröhnert U, et al. Arginine residues within the DNA binding domain of STAT3 promote intracellular shuttling and phosphorylation of STAT3. *Cellular signalling.* 2014;26(8):1698-1706.

80. Bushweller JH. Targeting transcription factors in cancer — from undruggable to reality.*Nature Reviews Cancer.* 2019;19(11):611-624.

81. Kolchina N, Khavinson V, Linkova N, et al. Systematic search for structural motifs of peptide binding to double-stranded DNA. *Nucleic Acids Research.*2019;47(20):10553-10563.

82. Sorolla A, Wang E, Golden E, et al. Precision medicine by designer interference peptides: applications in oncology and molecular therapeutics. *Oncogene.*2020;39(6):1167-1184.

83. Patel SG, Sayers EJ, He L, et al. Cell-penetrating peptide sequence and modification dependent uptake and subcellular distribution of green florescent protein in different cell lines. *Sci Rep.* 2019;9(1):6298.

84. Gautam A, Singh H, Tyagi A, et al. CPPsite: a curated database of cell penetrating peptides.*Database (Oxford).* 2012;2012:bas015.

85. Oh D, Nasrolahi Shirazi A, Northup K, et al. Enhanced cellular uptake of short polyarginine peptides through fatty acylation and cyclization. *Mol Pharm.*2014;11(8):2845-2854.

List of figures