STRUCTURAL BIOINFORMATICS ANALYSIS OF THE CANDIDATE TUMOR SUPPRESSOR PROTEIN CTCF

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
ENGINEERING AND NATURAL SCIENCES
OF ISTANBUL MEDIPOL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING AND BIOINFORMATICS

By

Sinem Darwısh

September, 2021

STRUCTURAL BIOINFORMATICS ANALYSIS OF THE CANDIDATE TUMOR SUPPRESSOR PROTEIN CTCF

By Sinem Darwısh

22 September, 2021

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

_____

Assist. Prof. Kıvanç Kök (Advisor)

_____

Assist. Prof. Özge Şensoy

_____

Assist. Prof. Mehmet Hikmet Üçışık

Approved by the Graduate School of Engineering and Natural Sciences:

_____

Prof. Dr. Yasemin Yüksel Durmaz

Director of the Graduate School of Engineering and Natural Sciences

I hereby declare that all information in this document has been obtained and presented in accordance with the academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Signature :

Name, Surname: SİNEM DARWISH

ACKNOWLEDGEMENT

I would like to express my gratitude and deep appreciation to my advisor Assist. Prof. Kıvanç Kök for his frequent supervision on my Master's thesis research, help, and support despite his heavy schedule. Without his mentorship, this research wouldn't have been brought to light. I want to express my gratitude towards his supervision on my thesis, which we have finally at hand. In particular for keep the research valid and ensured it meets the standards of the University and the academic discipline. I have to admit that his evaluation of each piece of this study fueled my energy to discover and research the next sections and so on. So, finally, many thanks for his guidance and help from the beginning of my Master's thesis project until today.

I would first like to express my gratitude and deep appreciation to Assistant Professor Özge Şensoy, whose expertise was valuable in formulating the research methodology. I also thank her for continuous guidance, help, and support. Without her support, I would not have gotten my Master's degree project finished. Honestly, she made it possible to get through several obstacles and move forward with my Master's thesis. I would like to mention that acquiring this level of knowledge wouldn't have happened without her as well. Many thanks for her guidance and help since I started my thesis. Her generosity in answering my questions allowed me to see from different perspectives and it does mirror her profound knowledge of the subject. Therefore, her contribution is highly appreciated. I would also like to thank Sensoy lab members for teaching me about MD analysis and clarifying my questions during my Master's project research.

I would like to thank Asmaa Samy for her instructions related to molecular dynamics techniques. I have learned a lot from her. Special thanks for her suggestions related to the DCCM analysis method (regarding the related R codes). I thank Metehan İlter for his instructions related to PCA analysis (regarding the related Python script).

Besides, I'm immensely blessed to have my family and my husband Ahmed Darwısh standing by me. My sincere thanks to each family member who supported me. Feeling their material and moral support in my life was and remains essential.

Sinem Darwısh

September, 2021

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Δ : Delta
Å : Angstrom
π : Pi
φ : Phi
ψ : Psi

ABBREVIATIONS

AA	: Amino acid
AAN	: Amino acid network
BORIS	: Brother of Regulator of Imprinted Sites
C2H2-ZF	: Cys2-His2 zinc finger
CBS	: CTCF binding site
CHARMM-GUI	: Chemistry at HARvard Molecular Mechanics - Graphical User Interface
COSMIC	: Catalogue of Somatic Mutations in Cancer
CTCF	: CCCTC-binding factor
CTS	: CTCF target site
CUPSAT	: Cologne University Protein Stability Analysis Tool
D(2)P(2)	: Database of Disordered Protein Prediction
dbSNP	: Single Nucleotide Polymorphism database
DCCM	: Dynamic cross correlation map
DHA	: Donor-hydrogen-acceptor
DNA	: Deoxyribonucleic acid
ECR	: Evolutionary constrained region
GROMACS	: GROningen MAchine for Chemical Simulations
Hi-C	: High-throughput chromosome conformation capture (High-throughput 3C)
HOPE	: Have (y)Our Protein Explained
MD	: Molecular dynamics
PCA	: Principal component analysis
PDB	: Protein Data Bank
ProSNEx	: Protein Structure Network Explorer
PROVEAN	: Protein Variation Effect Analyzer
PTM	: Post translational modification
RIN	: Residue interaction network
RING	: Residue Interaction Network Generator
RMSD	: Root-mean-square deviation
RMSF	: Root-mean-square fluctuation
RNA	: Ribonucleic acid
SIFT	: Sorting Intolerant From Tolerant
TAD	: Topologically associating domain
VMD	: Visual Molecular Dynamics
WT	: Wild type
ZF	: Zinc finger
Zn	: Zinc

ADAY TÜMÖR BASKILAYICI PROTEİN CTCF'İN YAPISAL BİYOENFORMATİK ANALİZİ

ÖZET

Sinem Darwısh

Biyomedikal Mühendisliği ve Biyoenformatik, Yüksek Lisans

Tez Danışmanı: Dr.Öğr.Üye. Kıvanç Kök

Eylül, 2021

Genomun ‘’baş dokumacısı’’ olarak yaygın bir şekilde tanınan CCCTC-bağlayıcı faktör (CTCF) çok iyi korunmuş bir transkripsiyon faktörü ve mimari proteindir. CTCF üzerine yapılan konvansiyonel araştırmalar esasen karşılaştırmalı genomik, CTCF hedef bölgeleri ve 3D genom organizasyonu perspektifinden fonksiyonel özelliklerine odaklanmıştır. Bununla ilgili olarak, bu düzenleyici proteinin kanserle ilişkili yönleri büyük ölçüde bilinmemektedir. CTCF'in işlevsel çok yönlülüğü ve ilgili deneysel bulguların eksikliğinden nedeniyle, şu anki literatür CTCF'i kanserle ancak zayıf bir şekilde, aday tümör baskılayıcı protein olarak ilişkilendiriyor. Son zamanlardaki çabalar Çinko Parmak (ÇP) 3'te oluşan, kanserle ilgili üç tek amino asit CTCF varyasyonunu (isimlendirecek olursak R339W, K344E, H345R'yi) vurguladı. Bu mutasyonlar üzerine CTCF'in DNA hedef bölgelerine bağlanmasının azaldığı araştırılan CTCF hedef bölgelerin çoğunda gösterildi. Yine de, gözlemlenen farklı bağlanma modellerinin mekanik yönleri halen büyük ölçüde anlaşılmamıştır. Buna paralel olarak, yapısal biyoinformatik ıslak laboratuvar gözlemlerinin tamamlamada ve aday tümör baskılayıcı protein hipotezini geliştirmede yararlı olabilir. Bu tez, yukarıda bahsedilen üç varyasyona odaklanarak moleküler dinamik (MD) simülasyonları ile birlikte, birincil dizi ve 3D yapı tabanlı hesaplama tahminlerini kullanarak bu konudaki araştırma boşluğunu doldurmayı amaçlamıştır. İlk olarak, mevcut protein yapısal verinin koruma, düzensizlik ve B-faktörü bilgisi ile entegrasyonu CTCF proteini içindeki korunmuş ve düzensiz bölgeleri betimledi. Bu adım ayrıca, araştırılan üç amino asidin hayvanlarda yüksek oranda korunduğunu belirlemiştir. Bu da ilgili mutasyonlar için geniş bir fonksiyonel sonuç öne sürer. İkinci olarak, seçilen varyasyonların protein yapısal kararlılığı üzerindeki patojenitesi ve etkisi, bilinen CTCF dizisi ve yapısına dayalı olarak biyoinformatik veri tabanları ve hesaplamalı biyoloji araçları kullanılarak tahmin edildi. Sonuçlar araştırılan sübstitüsyonların hepsinin hasar verici etkisinin bulunabileceğini ve aralarında R339W varyasyonunun en patojenik olduğunu öne sürüyor. Son olarak, vahşi tip ve mutant CTCF proteinlerinin karşılaştırmalı moleküler dinamik (MD) analizi, incelenen dizi-yapı ilişkisi hakkında derinlemesine bir bakış kazandırdı. MD yörünge analizleri, global düzeyde önemli değişiklikler tespit etmedi, fakat bunun yerine analiz edilen sistemler arasında, özellikle ZF3'te, mutasyonların yerel etkisini gösteren bazı bölgesel farklılıkları ortaya çıkardı. Şaşırtıcı bir şekilde, gerçekleştirilen yapısal biyoinformatik analizleri, bir yandan en şiddetli yapısal etkiye sahip mutasyon olarak R339W varyasyonunu yeniden vurgularken, diğer yandan da H345R mutasyonu üzerine Zn iyonunun kaybını ortaya çıkardı. Sonuç olarak, elde edilen bulgular, bu iki mutasyonun fenotipte ortaya çıkma

olasılığının K344E'den daha yüksek olduğunu gösterir. Genel olarak, bu çalışma, yapısal özelliklerini aydınlatarak ve seçilen kanserle ilgili mutasyonların onkojenisitesini aydınlatarak varsayılan tümör baskılayıcı CTCF'in işlevi hakkında yeni bakışlar sağlar. CTCF ile ilgili gelecekteki çalışmaların kanserin daha derin anlaşılmasına katkıda bulunması bekleniyor. Bu tür potansiyel ilerleme sonuçta yeni translasyonel uygulamaların yolunu açabilir.

Anahtar sözcükler: CCCTC-bağlayıcı factor proteini; aday tümör baskılayıcı protein; sekans-yapı ilişkisi; hesaplamalı tahminler; moleküler dinamik analizi.

STRUCTURAL BIOINFORMATICS ANALYSIS OF THE CANDIDATE TUMOR SUPPRESSOR PROTEIN CTCF

ABSTRACT

Sinem Darwısh

M.S. in Biomedical Engineering and Bioinformatics

Advisor: Asst. Prof. Kıvanç Kök

September, 2021

CCCTC-binding factor (CTCF) is an extremely well-conserved transcription factor and architectural protein, which is widely recognized as a ‘’master weaver’’ of the genome. Conventional research on CTCF has mainly focused on its functional properties from the perspective of comparative genomics, CTCF target sites, and 3D genome organization. Relatedly, cancer-associated aspects of this regulatory protein remain largely unknown. Due to the functional versatility of CTCF and lack of related experimental findings, current literature only loosely implicates this protein in cancer as a candidate tumor suppressor protein. Recent efforts have highlighted three cancer-related CTCF single amino acid variations, namely R339W, K344E, H345R, which occur in the Zinc Finger (ZF) 3. Decreased CTCF binding to its DNA target sites upon these mutations was demonstrated in the majority of the investigated CTCF target sites. Still, mechanistic aspects of the observed differential binding patterns have remained mostly elusive. Concordantly, structural bioinformatics can be instrumental in complementing wet-lab observations and refining the candidate tumor suppressor protein hypothesis. This thesis aimed to fill this research gap using primary sequence and 3D structure-based computational predictions along with molecular dynamics (MD) simulations by focusing on three aforementioned variations. Firstly, integration of the available protein structural data together with the conservation, disorder, and B-factor information delineated conserved and disordered regions within CTCF protein. This step also determined that the three investigated amino acids are highly conserved in animals, which suggests a broad functional consequence for the respective mutations. Secondly, the pathogenicity and impact of the selected variations on the protein structural stability were predicted using bioinformatics databases and computational biology tools based on the known CTCF sequence and structure. The results suggest that all investigated substitutions have a damaging effect and that the R339W variation is the most pathogenic one among them. Finally, comparative molecular dynamics (MD) analysis of wild-type and the mutant CTCF proteins yielded an in-depth view of the studied sequence-structure relationship. MD trajectory analyses did not detect substantial changes on the global level but rather revealed some regional differences between the analyzed systems, especially in the ZF3, hinting at the local effect of the mutations. Intriguingly, the performed structural bioinformatics analysis on the one hand re-emphasized the R339W variation as the mutation with the most drastic structural impact and on the other hand uncovered loss of Zn ion lost upon H345R mutation. All in all, the obtained finding indicates that these two mutations but are likely to be manifested in the phenotype than K344E. Overall, this work

provides novel insights into the putative tumor suppressor function of CTCF by unraveling its structural features and elucidating the oncogenicity of the selected cancer-relevant mutations. Future work on CTCF is expected to contribute to a deeper understanding of cancer. Such potential progress can ultimately pave the way for new innovative translational applications.

Keywords: CCCTC-binding factor protein; candidate tumor suppressor protein; sequence-structure relationship; computational predictions; molecular dynamics analysis.

CHAPTER 1

1. INTRODUCTION

1.1. Regulatory protein CTCF

CCCTC-binding factor (CTCF) is an extremely well conserved, multifunctional, 11-zinc finger protein. This transcription factor is a key protein for the regulation of chromatin loops formation and 3D genome organization, along with transcription activation and repression, RNA splicing and enhancer/promoter insulation [1]. As a transcriptional repressor CTCF, was discovered firstly within the CT-rich promoter regulatory region of the chicken c-myc gene [2]. While the presence of CTCF has been validated until now in vertebrates, flies, and some nematodes, it is absent in yeast, plants and *Caenorhabditis elegans* [1]. Importantly, CTCF participates in many cellular processes, such as growth, proliferation, differentiation, apoptosis, cell cycle and X-chromosome inactivation [3]. Relatedly, the study on pre-implantation development in mice with a complete loss-of-function Ctcf knockout allele resulted into early embryonic lethality [4]. Consequently, CTCF is regarded as an essential gene for embryonic development [5].

1.2. CTCF as a master weaver of the genome

One of the principal regulatory roles of the transcription factor CTCF is forming loops with distant elements of the genome in order to orchestrate the nuclear organization [6]. Due to this extraordinary role in establishing the 3D chromatin architecture, this multifunctional architectural protein is widely recognized as the “master weaver of the genome” [7], [8]. Genomes of eukaryotes are packaged into several levels of organization, from nucleosomes to chromatin fibers and then to large-scale chromosomal domains within each chromosome [9]. Nucleosomal fibers are formed by DNA, which is wrapped around histones. The resulting structure is then folded and looped into higher-order structures [10].

The complexity of genomic interactions in mammalian chromatin networks assists in mediating intrachromosomal loops and interchromosomal contacts [10]. The chromosome conformation capture-related (C3-related) approaches have revealed that CTCF mediates both inter-chromosomal and intra-chromosomal interactions between among sites in the genome [9]. By mediating internuclear contacts, CTCF controls enhancer-promoter interactions throughout the genome and supports organization of Topologically Associating Domains (TADs) [11]. Application of high-throughput chromosome conformation capture, shortly, high-throughput 3C (Hi-C) methods showed that genome could be subdivided into circa 2000 TADs, which have strong contacts within each TAD but totally weak contacts with other TADs [12]. CTCF demarcates the single TAD boundaries, as result of its ability of blocking interactions across its binding sites [12].

The cohesin protein is another important component of the chromatin loops. In the mammalian genome, CTCF and cohesin are involved differentially in establishment or maintenance of the topological domain architecture, in the regions enriched with their binding sites [13]. Remarkably, CTCF and cohesin cooperate in mediating long-range interactions in the genome [14]. With respect to this observation, CTCF and the cohesin complex colocalize largely throughout mammalian genomes [15]. Concordantly, half of the DNA binding sites of cohesin in post-mitotic cells are overlapping with those of CTCF [16]. Interestingly, CTCF and cohesin have high interaction and gene expression levels during cell differentiation, which indicates their roles in mediating impacts of chromatin structure on gene regulation [15]. The depletion of CTCF on the one hand reduces the intradomain contacts and on the other hand, in contrast, causes increased interdomain interactions between neighboring chromatin domains. Relatedly, disruption of the cohesin complex causes loss of shorter range chromatin interactions without disturbing topological domain organization [15]. In the same study, in the light of the obtained findings, it was proposed that in the absence of CTCF, cohesin does not localize properly and consequently cannot reach beyond boundaries with nonspecific interaction [15].

1.3. Highly conserved transcription factor CTCF

CTCF is a multifunctional protein, which is ubiquitously expressed and extremely well conserved from Drosophila to human [17]. Importantly, CTCF has been intensively investigated from the perspective of comparative genomics. For example, a pioneering

study, published by Heger et al. in 2012, suggests that the CTCF protein and a genome-wide abundance of CTCF-specific binding motifs are unique to bilaterian phyla. In contrast, this transcription and the mentioned motifs are absent in other eukaryotes. In the same study, based on the obtained findings, the authors proposed the presence of a Hox-CTCF kernel as major organizer of bilaterian body plans [18]. This work re-emphasized the fundamental role of CTCF as developmentally important protein. As mentioned above, concerning the chromatin architecture, CTCF binding to DNA assists in the formation of a higher-order genome structure by demarcating the boundaries of large-scale TADs [19]. Clustered CTCF binding has been suggested to maintain topologically associating domains. In another words, dynamic conservation of CTCF site clusters appears to be a critical property underlying functional stability of a higher-order chromatin structure [19]. In agreement with this proposed paradigm, the TADs demarcated by CTCF are remarkably well conserved between species, which previously motivated another study on conserved spatial chromatin organization [20]. In summary, the progress in understanding conservation of the CTCF protein and its DNA binding motifs considerably contributes to our understanding of chromatin special organization, s conservation and dynamics.

1.4. Structure of CTCF

The CTCF gene codes for the gene product, which is comprised of 727 amino acids and 11 zinc finger (ZF) domains. CTCF is characterized by the central region of the protein, which includes a cluster (array) of eleven domains of C2H2-type zinc fingers. Five of ZFs bind to a long DNA sequence conserved in most animals in a specific manner [21]. With regard to ZF types, CTCF is classified as the Cys2-His2 zinc finger (C2H2-ZF) protein. As such, CTCF is a typical example for DNA-binding transcription factors, which contains clusters of C2H2-type zinc fingers [21]. While the first 10 ZFs correspond to the classical Cys2-His2 types, the 11th ZF possesses an unusual Cys2-His-Cys sequence [22]. Proteins containing Cys2His2 ZFs are frequently present in eukaryotic organisms. While ZF domains are well-known for serving protein-DNA interactions, they also actually facilitate some protein-RNA and protein-protein interactions.

In the course of specific binding of CTCF to DNA, each of ZFs 3-6 recognizes 2-3 bases. In contrast, each of ZFs 9-11 only reads 1-2 bases. Interestingly, ZF7 rotates 35° more than ZFs 3-6 and contacts 6 bases. Remarkably, ZF8 contacts the spacer between CBS modules 1 and 2 in a non-specific manner and does not provide extra binding to the DNA

[23]. Finally, the terminal zinc fingers ZF1, ZF10 and ZF11 contribute to interactions of CTCF with other proteins or to binding of CTCF to RNA, rather than specific DNA binding [24].

Interestingly, DNA regions that bind by CTCF are not bent. Still, DNA might be bent from outside with other protein collaboration [23]. Therewith, CTCF is not a typical DNA bending protein [25].

1.5. Zinc Ion Coordination in Cys2His2 Zinc Fingers

Zinc-finger domains provide stable structures for coordination between zinc ions and a group of histidine and cysteine residues. The most common ZF is named as ”classical zinc fingers” or ”Cys2His2 zinc fingers”. It uses two histidines and two cysteines for interaction with zinc ion (Zn^{+2}). Herein, Cys2His2 ZFs coordinate a zinc ion tetrahedrally, to form a stable structure including a β-sheet and an α-helix (abbreviated as “ββα fold”). β-sheet includes two cysteines at the N-terminus of protein while a-helix has two histidines on the turns at the C-terminus [26]. Hereby, this type of zinc fingers uses zinc ion (Zn^{+2}) coordination to stabilize protein folding and as a consequence the Cys2His2 ZF domain folds in the existence of Zn^{+2} to form a compact ββα domain. More specifically, each zinc finger incorporates a single Zn^{+2} in sandwich form between the two-stranded antiparallel β-sheets and the α-helix. Herewith, Zn^{+2} is tetrahedrally coordinated between two cysteines at one end of the β-sheet and two histidines in the C-terminal portion of the α-helix [27]. With regard to the ZF3, the zinc ion is coordinated by the residues Cys324, Cys327, Hsd340, Hsd345 as shown in **Figure 1.1**, which is adopted from the publication by Filippova et al., 2002 [28].

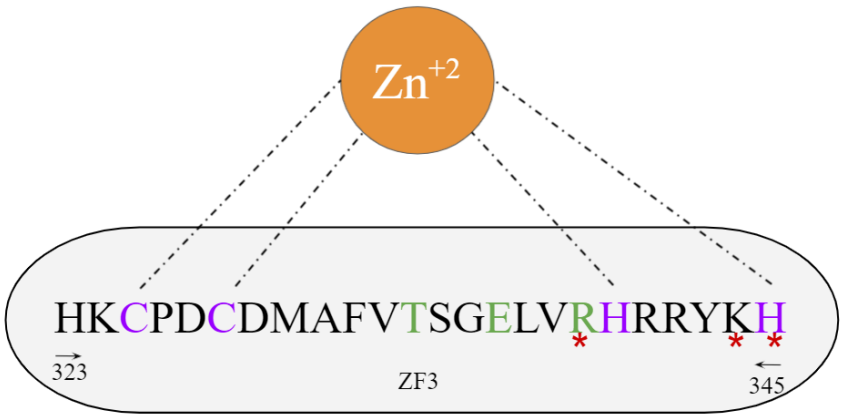


Figure 1.1: CTCF ZF3 amino acid sequence and its interaction with zinc ion (Zn^{+2}). Amino acid, interacting with Zn^{+2}, are colored in violet. Mutation sites, which were the

focus of this study, are labeled with red star. DNA-binding amino acids are colored green. Zn^{+2} is depicted as orange circle. Adopted from the publication by Filippova et al., 2002 [28].

1.6. CTCF Binding Sites

CTCF occupies almost 80,000 sites on mammalian chromosomes [24]. To expose the CTCF binding site (CBS) interactions and the CTCF structure, the CTCF-DNA complex was solved in two different studies. This was achieved with CORE sequence [24] and CTCF-binding site sequence of the HS5-1a enhancer of the *Pcdh* cluster, respectively [23]. C2H2 ZF proteins interact with the triplet element of DNA, which comprises three adjacent base pairs. While ZFs recognize one strand of DNA from 3′ to 5′ linearly, with their protein sequence proceeding from N to C termini, the side chains bind to DNA with N termini side of each helix and contact the major groove [24]. Not all ZFs but only the ZFs 3-7 were found to be mediating CTCF binding to the 15bp CORE sequence motif in CTCF binding site (CBS). Each of the ZFs contacts accordingly 3 bps (triplet) using three amino acids. In the 15bp CORE sequence, the actual DNA sequence generally differs to some extent from the known consensus 15bp CORE sequence. This is also the case in the CTCF crystal structure which was published and made available by Hashimoto et al in 2017. This structure includes majority of CTCF ZFs and the corresponding bound DNA. Regarding the ZF3 in the specified structure, the amino acids T333, E336 and R339) constitute the actual DNA contact site which binds to a specific, respective base in the investigate CBS. The performed crystallization provided, among others, details of these inter-molecular interactions for the first bp of triplet in the form of inter-atomic distances (in angstroms) [24]. The interaction sites of CTCF are as following: ZF6 with GCA, ZF5 with GGG, ZF4 with GGC and ZF3 with GCT in the 3′ sequence. Furthermore, ZF7 interacts with CCA in 5′ sequence [24].

CTCF ZFs with CBSs in *Pcdh* enhancers and promoters, recognize different regions of DNA [23]. More specifically, in this regard, 4 modules of CBS motifs interact with CTCF ZFs. Although ZF9, ZF10, ZF11 interact with module 1,most CBSs lack module 1 [23]. Module 2 is commonly recognized by ZF6 and ZF7. Relatedly, the important specific recognition of the central asymmetrical A24 by Gln418 is crucial for distinguishing the CBS orientation. Module 3 is primarily recognized by ZF4 and ZF5. Interestingly, the base-specific interactions of ZFs 4-7 are with the three most highly conserved nucleotides (C20, G25 and G28). These nucleotides form two hydrogen bonds with side chains of

Asp451, Arg396 and Arg368, respectively [23]. This reveals why CBS is to a very great degree conserved within modules 2-3. In the ZFs2-8-HS5-1aE complex crystal structure, ZF2 is disordered, however the α-helix of ZF3 is localized in the major groove of module 4. Furthermore, C30 and G31 nucleotides form hydrogen bonds with side-chain of Arg339 in ZF3. Plus, C30 and G31 are located in the "CGCT" box, which is the core sequence within conserved sequence elements in the promoter region of members of mammalian *Pcdh* gene family [23]. Module 4 is an asymmetric sequence. Hence, the binding of ZF3 to module 4 designates the directionality of CTCF binding to many oriented CBSs. This is crucial for correct chromatin looping during 3-dimensional genome organization.

1.7. Candidate tumor suppressor protein CTCF

Traditional approaches to deciphering the biological role of CTCF have mainly concentrated on its functional properties from the viewpoint of comparative genomics, DNA target site occupancy and 3D genome organization. In the earliest report, published by Filippova et al. in 1998, the authors mentioned that CTCF is a new candidate tumor suppressor gene for the first time. Cytogenetic studies demonstrated that chromosomal deletions at 16q22.1 locus occur in multiple different types of human malignancies. Thus, the region was predicted to harbor tumor suppressor genes. CTCF is one of the genes located at 16q22.1 that overlaps with the most frequently occurred chromosomal deletions found in breast and prostate cancer. Thus, with regard to the cancer-associated chromosome region 16q22.1 CTCF was proposed as a novel candidate tumor suppressor [29]. Besides, the in vivo study using a bioluminescent orthotopic breast cancer model in mice showed that BORIS, a paralog of CTCF, and CTCF expression suppresses breast cancer growth, providing additional evidence for CTCF's tumor suppressor role [30]. Furthermore, another study implicates disruption or inactivation of CTCF-DNA complex in cancer development [31].

CTCF mutations mostly occur in the DNA binding or zinc ion coordinating amino acids in the zinc finger domains [30]. CTCF is among the 127 genes, which were identified as significantly mutated genes in endometrial cancer. CTCF mutations have also been reported in different cancer types, including Wilms' tumor, leukaemia, breast and prostate cancer [32]. In another study, 15 transients abnormal myelopoiesis and 14 Down syndrome-related acute megakaryoblastic leukaemia samples were investigated using

whole genome/exome sequencing. As result of this study, CTCF was found to be mutated or deleted in 10 Down syndrome-related acute megakaryoblastic leukaemia cases, in 1 Transient abnormal myelopoiesis case and in 4 non-Down syndrome-related acute megakaryoblastic leukaemia cases [33]. Conceivably, mutations occurring in zinc finger domains can disrupt tumor suppressor function of CTCF, thereby contributing to tumor growth and development.

Publications on association of CTCF with-cancer related were searched in PubMed using "CTCF cancer mutation" and "CTCF tumor suppressor" search terms. Number of publications per year and cumulative number of publications are shown in **Figure 1.2**. Detailed information is shown in **Table A.1**.

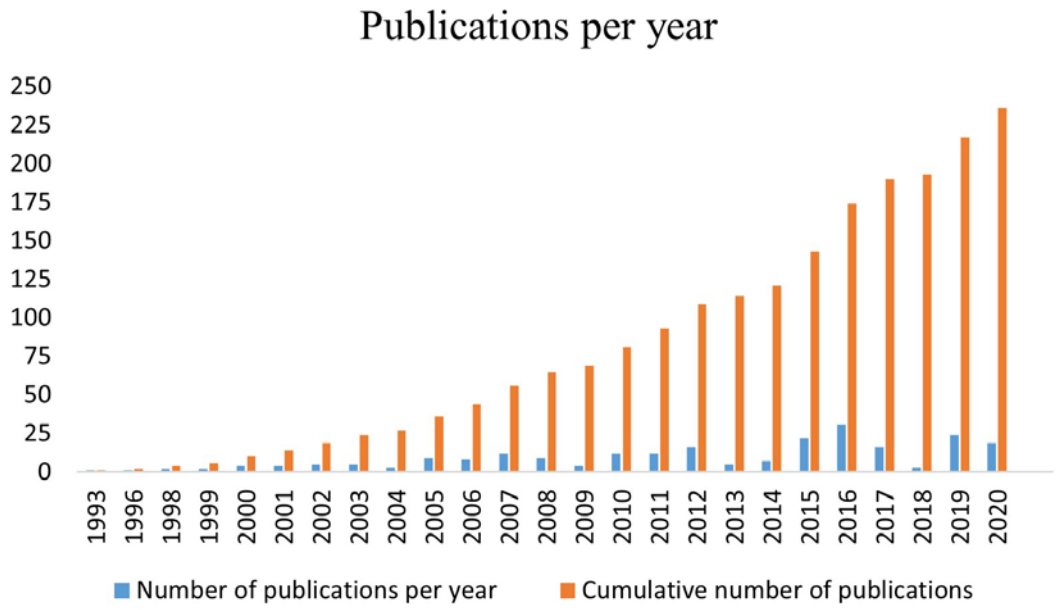


Figure 1.2: Number of publications per year and cumulative number of publications for the "CTCF cancer mutation" and "CTCF tumor suppressor" search terms in PubMed.

One of the related studies highlighted R339W and K344E mutations from the list, related to Wilms and Breast Tumor, respectively [28]. The mutations are located on 3rd zinc-finger. Another mutation on 3rd zinc-finger, namely H345R, which is related to Prostate tumor, was also reported in the study. The three mutations are critical for ZF formation or DNA base recognition.

The binding of CTCF, depending on the sequence of the CTSs (CTCF Target Sites), were assessed using in vitro study for three variants of the ZF3 domain The result of this study, among others, show that tumor-specific variants in CTCF ZF3 completely changed the range of CTSs recognized by the full-length CTCF (**Table 1.1**) [34].

Table 1.1: Binding affinity of CTCF and CTS for WT, R339W, H345R, K344E systems. Adopted from results in Figure 2C in [3].

CTSs	WT	R339W	K344E	H345R
Myc A	Strong binding	No binding	Partial loss of binding	No binding
Myc B	Strong binding	No binding	No binding	No binding
Ch Myc V	Strong binding	No binding	Partial loss of binding	No binding
PLK	Strong binding	No binding	No binding	No binding
PIM	Strong binding	No binding	No binding	No binding
P19ARF	Strong binding	No binding	No binding	No binding
H19 DMD 4	Strong binding	No binding	No binding	No binding
H19 DMD 7	Strong binding	No binding	No binding	No binding
R 144	Strong binding	Partial loss of binding	Partial loss of binding	No binding
Amyloid PP	Strong binding	Strong binding	Strong binding	Strong binding
Globin F II	Strong binding	Strong binding	Strong binding	Strong binding
Lys F1	Strong binding	Strong binding	Strong binding	Strong binding

1.8. Aims

Conventional research on CTCF has mainly concentrated on its functional properties from the perspective of comparative genomics, DNA binding sites and 3D genome organization. When compared to these topics, cancer-related aspects of CTCF are understudied and the candidate tumor suppressor protein questions remain open. This study aimed to complement experimental indications in this area by addressing the issue of CTCF's cancer candidate tumor suppressor protein candidacy by focusing on structure-function relationship using a structural bioinformatics-based approach. Many studies, based both on genome-level and gene-level experiments, have been reported on the various regulatory roles of CTCF. However, the published literature on the tumor suppressor role of this multi-functional protein remains largely inconclusive. To this end, CTCF is currently regarded as a candidate tumor suppressor based on *in vitro* studies. The related findings in this field are still inconclusive due to the functional versatility of

CTCF and lack of related experimental findings. Notably, recent efforts have highlighted three cancer-related CTCF single amino acid variations, namely R339W, K344E, H345R, which occur in the Zinc Finger (ZF) 3. The purpose of this *in silico* study was to complement such experimental efforts in this direction by unraveling molecular dynamics (MD) properties of the wild-type protein and elucidating structural aspects of pathogenicity of 3 specific cancer-associated mutations, namely R339W, K344E, H345R. As the starting point of this thesis, it was hypothesized that the specified mutations lead to the observed phenotype (partial or complete loss of interaction with DNA) by disrupting CTCF structural stability and/or functional integrity. Consequently, the structural bioinformatics focusing on the functional consequences of these mutations was performed to address this issue. More specifically, impact of the selected mutations on the protein function was predicted using set of bioinformatics tools and databases, which was followed by the MD analysis of the corresponding variants. In this endeavor, the mutants were compared to the wild type (WT) using multiple bioinformatics tools and databases Overall, the goal of accomplished structural bioinformatics analysis was to provide novel insights into the structural features of the candidate tumor protein CTCF in the context of cancer by focusing on the three selected mutations.

CHAPTER 2

2. THEORETICAL PART

2.1. Conservation, Disorder and B-factor analysis

2.1.1. Conservation analysis

Changes in protein sequence during evolution occur inversely with power of specific constraints at all sites. If the site is in constrained region, it is considered that they have role in protein functions. The evolutionarily constrained regions (ECRs) can help to point protein functions and effects or mutations. therefore, the defining ECRs will help to discover domains of proteins and important residues. The importance of the residues in the protein, will provide information about pathogenicity of the variation in the residues. Because if the residue is in the ECRs region, mutation might affect function of protein. In light of this, to identify of ECRs in interested proteins Aminode is developed. Firstly, orthologs and paralogs of protein are collected and performed multisequence alignment to calculate relative rate of amino acid substitution. Secondly, number of amino acid substitution is calculated based on phylogenetic relationships between examined proteins. The result is used to compute relative rate of amino acid substitution and the rates are plotted in their positions along multi-sequence alignment of proteins. ECRs are shown as colored horizontal blocks at the top of the plot [35].

COBALT is a tool to align multiple protein sequences automatically with high speed and accuracy. COBALT collects pairwise constraints obtained from databases and finds sequence similarity to input. After, combining the pairwise constraints, multiple sequence alignment is generated [36].

2.1.2. Disorder analysis

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) achieve their functions without depending on a single known conformation. Such highly

flexible regions are characterized as an ensemble of conformations without stable 3D structure. IDPs can serve as linkers within ordered regions or recognize other specific molecules for protein-protein interactions. Disordered regions can adapt a more ordered and structured conformation, also exist as unfolded and folded states [37]. The IDPs or molecular recognition features (MoRFs) occur mostly in regulation or signaling proteins. IUPred2A server is used for prediction of protein disorder that combines prediction method of general disorder IUPred and disordered binding region ANCHOR [38].

2.1.3. B-factor analysis

Protein folding states have major role for dynamics, thermostability and protein function. The shape of protein depends on amino acid sequence, their physicochemical properties, and chemical interactions. Understanding the role of each residue in protein shape is possible with network-based interactions of amino acids within 3D structures. ProSNEx (Protein Structure Network Explorer) builds protein structure networks, besides, analyzes amino acid conservation, flexibility, and annotation. PSNs generates by adding nodes which represent residues and interactions between with residues are edges. Edges are specified as carbon-alpha, carbon beta or atom-pair contact networks [39]. B-factor is a temperature factor that measures uncertainty or flexibility of atoms in a 3D structure. B-factor demonstrates disorder of proteins and deeply buried regions located in core of protein has mostly low B-factor [40].

2.2. Databases and Tools for Prediction of Mutations Effects

2.2.1. Databases for Pathogenicity of Variants

Next-generation sequencing technology has been developed in last twenty years and generated big genome data. Bioinformatics tools are also developed to afford the computational needs to analyze the genome data. One of the bioinformatics approaches to NGS analysis is in silico prediction tools. The in silico tools predict effects of variations for disease. Many types of variants might cause diseases, such as substitutions, deletion, insertion, nonsense, and frameshift mutations.

SIFT prediction is based on conservation of amino acid in protein family. The assumption of method is mutations in the conserved positions are prone to be deleterious. In light of this, if the position includes hydrophobic amino acids along protein family, while hydrophobic amino acids substitution will be tolerated, hydrophilic or other charged

amino acids will not be tolerated. In the input amino acid sequence, SIFT detects other related proteins and obtains a multi-sequence alignment then, each amino acid in alignment and their probability being tolerated is calculated. Predicted value between 0 and 0.05 means substitution has potential to be deleterious [41].

PROVEAN aligns distantly related and homologous sequences which collect from NCBI NR protein databases. The most similar sequences are clustered and generated a supporting set for input sequence. For supporting sequences, BLOSUM62 is computed and obtained PROVEAN score. As a result, PROVEAN predicts effect of variation based on similarity between input protein sequence and known homologous protein and calculated alignment score. PROVEAN score threshold (−2.282) is calculated to be best threshold between the deleterious and neutral variants, the variants below threshold are deleterious [42].

MutationTaster uses available databases such as 1000G, ClinVar, HGMD, HapMap for the prediction. If a variant exists more than 4 times as homozygous in 1000G or HapMap databases, variant is reported as neutral. If the variant is reported as pathogenic with a known phenotype in ClinVar, it is predicted as disease causing.

2.2.2. 3D Structure-based Prediction of Mutation Impacts

2.2.1.1. CUPSAT

Cologne University Protein Stability Analysis Tool (CUPSAT) is a webserver for prediction of how the stability of a protein structure changes upon amino acid point mutations. CUPSAT algorithms is assessing mutated amino acid environment by distribution of torsion angle and amino acid-atom potentials. Amino acid environment can be distinguished using solvent accessibility of amino acids and specificity of secondary structures. CUPSAT builds the prediction from atom and torsion angle potentials. Atom potentials and torsion angle potentials were obtained from PISCES webserver including 4024 proteins. Atomic potential of amino acids based on radial pair distribution is calculated by Boltzmann's principle:

$$\Delta G_{\alpha} = -kT \ln\left(\frac{g_{ij}(r_d)}{g(r_d)}\right) \quad \textbf{(2.1)}$$

where *gij(rd)* places as radial pair distribution function of the pair *i* and *j* which has distance *rd.g(rd)* between. Then, calculated Boltzmann energy values are used for mean

force potentials of all amino acids to assign favorable values in the neighboring amino acids orientations by observed φ – ψ combinations. It shows how adapting altered orientations after mutation. Unfolding Gibbs free energy change (ΔG), and $\Delta\Delta G$ *[*ΔG (mutant – wild-type)] also calculated. $\Delta\Delta G$ scores range from negative to positive scores that means negative scores are destabilizing, positive scores are stabilizing. The results are comparative two different methods which are thermal and chemical denaturation. The prediction model of CUPSAT was mostly optimized for mutation with thermal $\Delta\Delta G$ [43].

2.2.1.2. HOPE Analysis

HOPE is developed for fully automatic analysis of structural and functional effects of amino acid mutations. Program collects data from other sources using WHAT IF Web services. The tools HOPE uses are sequence annotations by UniProt, predictions by DAS services, homology models by YASARA etc. These data are used for decision scheme to determine effect of mutation. The decision scheme contains contacts with other molecules and inter-protein interactions, structural domains, existing modifications, known variants and polymorphisms, conservation of amino acids, amino acid properties such as size, charge, and hydrophobicity. As a result, HOPE generates a report with explaining effects and changes upon mutations and also plots figures and animations.

2.2.1.3. RIN Analysis

The amino acids in secondary structure of protein, interact each other and protein folds into 3D structures. Recently, representing 3D structure of protein as residue interaction networks (RINs) has become popular practice. These networks are also known as amino acid networks (ANNs). In these network models, single amino acid is a node and edges demonstrate their interactions. AANs allow to focus on single amino acid and its interactions. By doing so, AANs provide additional knowledge into protein structure, folding and function and wholistic view of protein interactions [44]. Residue Interaction Network Generator (RING) 2.0 generates interaction networks that are accurate and reliable through an experimental re-parameterization method of distance thresholds performed on the whole PDB database. Hydrogen bonds are computed by using DHA angle $\leq 63^{\circ}$ limitation, denoting valid donor or acceptor atoms. C–C and C–S pairs are assented valid VDW interactions and their atom radius are not involved in distance. N and O atoms in sidechain of asparagine and glutamine are also considered as VDW interaction. Distance thresholds are for strict model; 3.5 Å, 4.0 Å, 2.5 Å, 0.5 Å, 6.5 Å for

hydrogen bond, ionic, disulphide, VDW, π-π stacking, respectively. Distance thresholds are for relaxed model; 5.5 Å, 5.0 Å, 3.0 Å, 0.8 Å, 7.0 Å for hydrogen bond, ionic, disulphide, VDW, π-π stacking respectively. Distribution of hydrogen bonds was observed as 5.01-5.04 Å in helices separated by a turn. FRST and TAP scores are determined for conformational energy. RING 2.0 is a fast software to generate intra-chain or inter-chain interactions including ligand and solvent atoms. Calculations of atomic interactions are according to geometrical criteria, without using force fields, so obtains reliable interaction network rapidly [45].

2.3. MD Analysis

The molecular dynamics study of proteins demonstrates structural or conformational changes and their interatomic movements by time based on a general physics model. By doing Molecular Dynamics (MD) simulations, conformational changes, protein folding mechanisms, ligand binding etc. studies can be elucidated at atomic level. The basic of MD is calculation of applied force of each atom to all other atoms in a system. By using Newton's laws of motion, force on each atom in a specific time and updated its velocity and position by the force is calculated. As a result, a trajectory is obtained which is a 3D film shows configuration in atomic level of systems during simulation. The technique can explain motions and positions of each atom in a specific time interval with known and controllable conditions such as temperature, mutations, PTMs, ligands which is difficult in experimental studies. The forces are computed by a known molecular mechanics model which is a force field. The force field are consist of the exact experimental measurements and mechanical calculations of quantum [46]. CHARMM36 is an accurate representation of experimental bilayer properties by pressure, NPT ensemble and constant particle number [47]. CHARMM36m is developed to generate ensembles of polypeptide backbone conformation for IDPs [48].

2.3.1. RMSD

Root-mean-square fluctuation (RMSD) is average displacement of atoms in structure at specific times comparative to a reference structure which is crystallographic structure or initial frame of trajectory. Time-dependent motions of protein are analyzed to figure out if the structure is stable or diverging from initial coordinates. Usually, if diverging initial coordinates are observed, it might show that non-equilibrated structure during simulation [49]. RMSD calculates the deviation between equivalent atoms in two structures by the

following formula:

$$rmsd = \sqrt{\frac{\sum_i d_i^2}{n}} \quad \textbf{(2.2)}$$

where *d* is the distance between each of the *n* pairs of equivalent atoms in the two structures [50].

2.3.2. RMSF

Root-mean-square deviation (RMSF) is one of the common measurements of structural fluctuation. Displacement of a specific atom or set of atoms comparative to a reference structure is calculated by RMSF [49]. RMSF is calculated according to following formula:

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\langle [r_i'(t) - r_i(t_{\mathrm{ref}})]^2 \right\rangle} \quad \textbf{(2.3)}$$

where *T* is trajectory time, *tref* is time of reference frame, *ri* is the position of *i*th residue and *ri* is the position of atoms that occurred of *ith* residue in a frame after superimposing with reference frame [51].

2.3.3. PCA

Principal Component Analysis (PCA) is covariance matrix diagonalization which calculates eigenvectors and eigenvalues of atomic fluctuations. Eigenvectors are the directions in 3N-dimensional (N is number of atoms) configurational space. Eigenvectors describe collective motions of atoms. Eigenvalues define total mean square motion of systems in the course of eigenvectors. The eigenvalues define the mean square fluctuation of the motion along the eigenvectors. The overall global motion of the protein can be expressed by first eigenvector which is largest eigenvalue [52]. The covariance matrix is computed by the following formula:

$$\mathrm{C_{ij}} = \langle \mathrm{M_{ij}} \Delta \mathrm{r_i} \Delta \mathrm{r_j} \rangle \quad \textbf{(2.4)}$$

where C ij is covariance matrix, $\mathbf{M_{ij}}\,\Delta \mathbf{r_i} \Delta \mathbf{r_j}$ is a positional change from average time structure for every coordinate of Ca atoms i and j [53].

2.3.4. DCCM

Dynamic cross-correlation matrix (DCCM) is a matrix of time-wise displacement correlation of amino acids between each other along protein. The significance of displacement of pair-wise atoms is examined with cross-correlation coefficients. The cross-correlation coefficient, *Cij* defines correlative motion of two atoms or two residues and can be calculated by the following formula:

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\left(\langle \Delta r_i^2 \rangle \cdot \left\langle \Delta r_j^2 \right\rangle \right)^{1/2}} \quad \textbf{(2.5)}$$

where *i* and *j* are *ith* and *jth* atom or residue, *Dri* and *Drj* mean the displacement vector corresponding with *ith* and *jth* residue or atom, and 〈...〉 denotes the ensemble average [54]. In the case of C*ij* equal to 1, fluctuations of considered atoms are totally correlated and -1 means anticorrelated. If C*ij* is equal to 0 there is no correlation between atoms [55].

CHAPTER 3

3. EXPERIMENTAL PART

3.1. Workflow of this Study

First, conservation and disorder analysis were performed evaluating the structural properties of the CTCF protein in the context of amino acid sequence variation and structural flexibility. In this step, primary sequence or available PDB structure was used as input. As a subsequent step, three mutations associated with the CTCF were selected by literature screening. Following this, the study was investigated by in silico analyses. The applied method steps and the used programs are represented as workflow in **Figure 3.1.**

Analyses were done using different input data types of CTCF, such as amino acid sequence, 3D structure and MD simulation. Most programs require CTCF ID and obtain the amino acid sequence or 3D structure. Types of data, together with the corresponding analyses programs are summarized in **Table 3.1**.

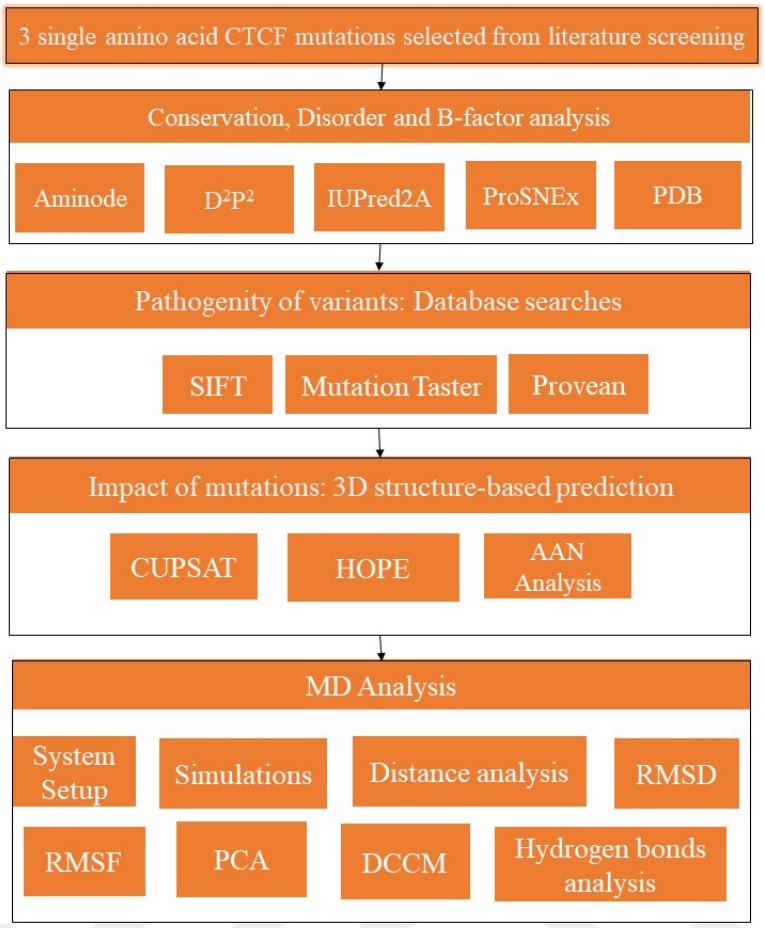


Figure 3.1: Workflow and used tools of the study.

Table 3.1: Technical overview of for the performed structural bioinformatics analysis.

Input	Program	Analysis method
CTCF ID	Aminode- Webserver	Conservation analysis
CTCF ID	PDB (DataBank) Webserver	Protein Data Bank
CTCF ID	IUPred2A- Webserver	Disorder analysis
CTCF ID	D2P2 – Webserver	Disorder analysis
CTCF ID	ProSNEx- Webserver	B-factor analysis
CTCF ID	SIFT- Webserver	Prediction tool
CTCF ID	Provean-Webserver	Prediction tool
CTCF ID	MutationTaster- Webserver	Prediction tool
CTCF ID	HOPE- Webserver	Mutation effect
Amino acid sequence	NCBI-COBALT-Webserver	Multiple sequence alignment

Amino acid sequence	EMBOSS Needle-Webserver	Multiple sequence alignment
3D structure	PyMOL	Structure alignment
3D structure	CUPSAT- Webserver	Stability
3D structure	RING 2.0- Webserver	Amino acid network (AAN) analysis
MD simulation	GROMACS	RMSD
MD simulation	GROMACS	RMSF
MD simulation	GROMACS	Hydrogen bonds analysis
MD simulation	GROMACS, VMD	Distance Analysis
MD simulation	GROMACS	PCA
MD simulation	R-Bio3D package	DCCM
MD simulation	VMD	Zinc atom localization

3.2. Conservation, Disorder and B-factor analysis

3.2.1. Conservation Analysis

Protein conservation information related to a set of representative animals was first retrieved and visualized using Aminode (http://www.aminode.org) tool [35]. This tool provided pre-computed results regarding Relative Substitute Scores (which correspond to the relative rate of amino acid substitutions) and Evolutionary Constrained Regions (ECRs) based on the alignment of the respective protein sequences. This provided overview of variable and conserved regions in the protein.

In addition to conservation of the whole CTCF protein sequence, we also check the conservation of the ZF domain by performing multiple sequence alignment of all 11 ZF domains of CTCF. Herein, NCBI tool COBALT was used to compute the multiple amino acid [63]. BLOSUM80 was used as coloring method in the visualization of the generated alignment.

3.2.2. Disorder Analysis

Disorder analysis was performed using first IUPred2A tool (http://iupred2a.elte.hu) [38]. Furthermore, additional disorder information was also obtained from the Database of

Disordered Protein Prediction (D(2)P(2)) (https://d2p2.pro/) [37]. Finally, the Protein Data Bank (PDB) (www.rcsb.org) [56] was used as an additional source of disorder information, which also integrated information about available, related PDB structures and their coverage.

3.2.3. B-factor Analysis

In order to complement conservation and disorder analysis, B-factor analysis was performed using the Protein Structure Network Explorer (ProSNEx) tool (http://prosnex-tool.com) [39]. This solution provided information about B-factor values not only across the primary sequence (covering the whole sequence), but also for a selected, available PDB structure, namely 5T0U (covering only included portion of the protein).

3.3. Obtaining CTCF Protein-DNA Complex

For the purpose of the study, CTCF protein structure is needed to perform next in silico analyses. There are in total 18 structures associated with CTCF in Protein Data Bank and 12 of them are CTCF-DNA complexes. These available structures are the result of two studies published in 2017: one study by Yin et al. and another by Hashimoto et al. The CTCF-DNA complexes do not include all zinc finger domains in one structure so each of them has different numbers of zinc fingers. To obtain the almost complete structure, different PDB files including different zinc fingers were decided to combine. For this, the combination, of PDBID:5T0U and PDBID:5UND was used. These structures were generated in a previous study [24]. The included zinc fingers are as in **Table 3.2**. The combination is the best option to have the longest CTCF-DNA complex. X-ray diffraction structures of CTCF-DNA complexes are at 3.20 Å and 2.55 Å resolutions respectively PDBID:5T0U and PDBID:5UND.

Table 3.2: PDB structures and length information of used 3D structures and the combined CTCF-DNA complex.

Alignment	PDB IDs	Zinc Fingers	AA sequence span	AA sequence length	DNA bp (length)
Input	5T0U	2-7 ZFs	293-462	170	23
Input	5UND	4-10 ZFs	348-518	171	26
Overlapping region	-	4-7 ZFs	348-462	115	14
Output	Combined structure	2-9 ZFs	293-518	226	35

EMBOSS Needle [57] performed sequence alignments of protein (**Figure B.1**) and DNA (**Figure B.2** and **Figure B.3**) between 5T0U and 5UND fasta files. To combine two structures, structure alignment was performed in pyMOL (Version 1.8.4.0 Schrödinger, LLC) for overlapping regions of two PDB files. By doing so, same coordinate for both structures was saved. As result, the combined structure was obtained. Details of the combination are presented in **Table 3.2**.

3.4. Predicting pathogenicity of variants

3.4.1. Results from the Databases Focusing on Pathogenicity of Variants

SIFT, Provean, MutationTaster have selected for investigation the pathogenicity of amino acid changes. Sorting Intolerant from Tolerant (SIFT) is an prediction algorithm that shows how coding variation affects on function of proteins [41]. Protein Variation Effect Analyzer (PROVEAN) is another prediction tool, which provides a general approach prediction of functional effects of amino acid substitutions or indels [42]. MutationTaster is a more recent tool than SIFT and Provean, which estimates the pathogenicity of DNA sequence alterations including amino acid substitutions, insertion/deletion, intronic and synonymous variants [58]. The protein sequence by CTCF ID was submitted to tools.

3.4.2. 3D Structure-based Prediction of Mutation Impacts

CUPSAT is used for how the stability of a protein changes after mutations. 3D structure of CTCF was uploaded and interested mutation site was chosen. The results are information about protein stability and torsion angles of the mutation sites for 19 possible amino acid substitutions.

Have (y)Our Protein Explained (HOPE) is a fully automatic webserver for analyzing structural and functional changes of single mutations. CTCF UniProt ID was given to start analysis. The output of HOPE is collections from other sources like WHAT IF Web services, predictions by DAS services, UniProt database, homology models by YASARA [59].

Amino acid network (AAN) analysis of residue interactions as performed by Residue Interaction Network Generator (RING). PDB files of 4 CTCF-DNA complex systems were submitted as input files. Parameters are applied as the closest network policy, showed only one the most important type of interaction, distance thresholds were relaxed,

and water molecules were skipped. The generated networks were visualized in Cytoscape [60].

3.5. MD Analysis

3.5.1. Preparing MD Simulation Files

For the molecular dynamics simulations, charmm-gui is used to prepare inputs files [61]. Chemistry at HARvard Molecular Mechanics- Graphical User Interface (Charmm-gui) is a web-based tool to generate effective simulation inputs. Solution Builder is an easy way to generate a series of input files for molecular dynamics simulation of the molecule in aqueous solvent environments [47], [62]. For the wild-type and mutant systems, 15.0 Å edge distance rectangular type water box was generated. Solution was ionized by 0.15 M KCl at 310 K temperature. As ion placing method, Monte-Carlo was used. (Access: 9 December 2019)

3.5.2. MD Simulation

MD simulations of wild type and mutant structures were performed with GROningen MAchine for Chemical Simulations (GROMACS) 5.1.4 using the CHARMM36m force field [48]. GROMACS is one of the engines for performing molecular dynamics simulations and energy minimization [63].

GROMACS run in mpi parallel system with 4 node, 44 core per node on Load Sharing Facility (LSF). MD simulations aimed at 500.000-ps (500-ns) MD runs. For all systems, two replicates were started with a different initial velocity and run 500-ns.

3.5.3. Analyses of MD Simulations

3.5.3.1. RMSD

RMSD is calculated by gmx rms module. RMSD was calculated by C-alpha atoms of structures. Starting structures of the simulations were used as reference structures. Moving average of RMSD is calculated with 50 range.

3.5.3.2. RMSF

Root mean square fluctuation (RMSF) is calculated by gmx rmsf module for C-alpha atoms of residues fluctuation during the trajectory.

3.5.3.3. PCA

PCA analysis was performed to show importance of collective motion of the studied systems. The gmx covar module of GROMACS was used to compute the covariance matrix of Cα atoms of MD simulation trajectories. The eigenvalues and their eigenvectors were obtained by diagonalizing the covariance matrix by gmx covar module of GROMACS. The first two eigenvectors were analyzed with gmx anaeig module. The resulting PCA plot helped to visualize the span of confirmations for each system.

3.5.3.4. DCCM

DCCM analysis was performed using the Bio3D package in R that plots the atom-wise cross-correlations. DCCM is obtained via calculation of the covariance matrix of atomic fluctuations during MD simulation trajectory. DCCM was calculated by C-alpha atoms of structures.

3.5.3.5. Hydrogen bond analysis

Number of hydrogen bonds is calculated by gmx hbond for last 400ns. The hydrogen bonds analysis was done for protein itself and between protein and DNA, pairs within 0.35 nm.

3.5.3.6. Distance Analysis

Distance between DNA and CTCF backbone was obtained from Visual Molecular Dynamics (VMD) and graphs generated by xmgrace. gmx analyze module computed the average distances with error estimates. gmx mindist computed the number of contacts between DNA and CTCF.

Distance between zinc atom and zinc-coordinating residues during 500 ns trajectory was determined by VMD. The zinc ion localization graphs in ZF3 were obtained from VMD [64].

CHAPTER 4

4. RESULTS AND DISCUSSION

4.1. Conservation, disorder, and B-factor aspects of the CTCF protein

4.1.1. Conservation of the primary sequence

To approach the CTCF tumor candidacy issue from a broader perspective, the first step in the analysis was to examine the conservation-related aspects on the protein level using Aminode. Results of the Amino Acid Evolutionary Constrained Analysis were obtained for CTCF using this bioinformatics source. The obtained conservation-oriented overview of CTCF demonstrated differences in the conservation along the sequence, highlighting variable areas towards the flanking regions and conserved areas in the central region **(Figure 4.1**). Such conserved areas are predicted as Evolutionary Conserved Regions (ECRs) and are thought to be of functional significance.

Intriguingly, the main ECR, which occupies the center of the protein, also included the position of the three studied mutations, namely. This prompted the further, in-depth analysis of this conservation pattern.

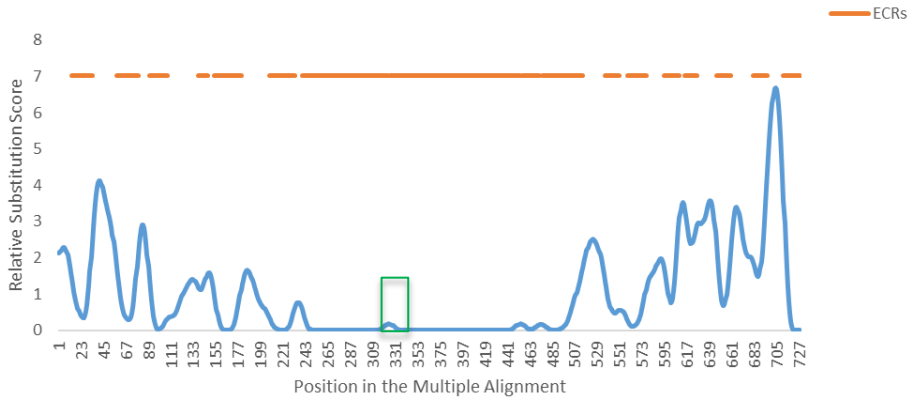


Figure 4.1: Overview of the conservation in the CTCF based on the analysis results provided by the database Aminode. Relative Substitution Score is shown as curve and Evolutionary Conserved Regions (ECRs) are shown as yellow lines. The position of ZF3 is highlighted with green box.

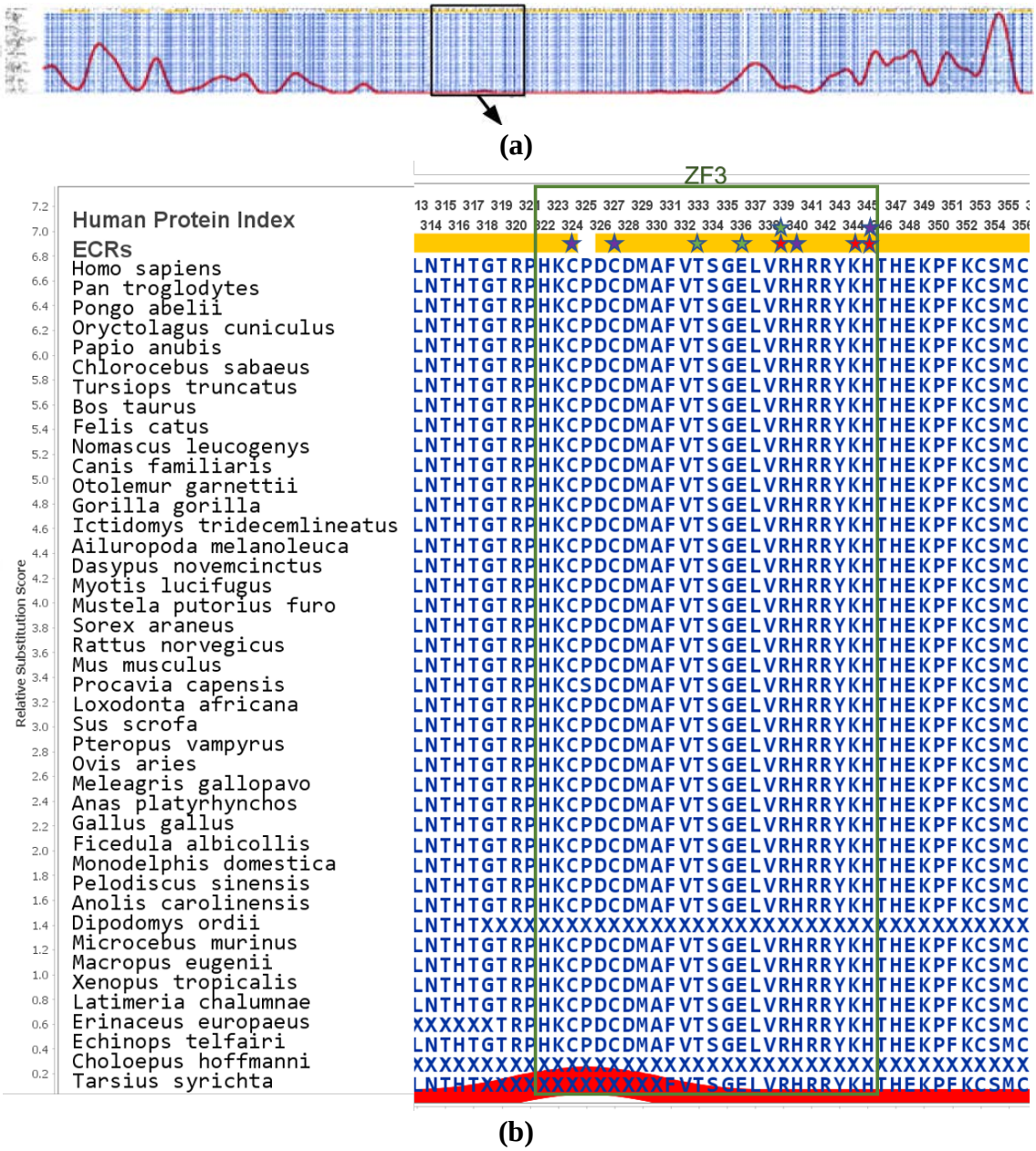


Figure 4.2: Detailed visualization of CTCF of protein conservation pattern CTCF plotted on the top of the representative animal protein sequences. **a)** Regions with low Relative Substitution Score were predicted as the Evolutionary Constrained Regions. Evolutionary Conserved Regions (ECRs), which are the areas with low Relative Substitution Score, are highlighted in yellow**. b)** Conservation in the area which contains ZF3. ZF3 is demarcated with box using a green line. Positions in the human CTCF protein index corresponding to the studied single amino acid mutations, DNA-contacting amino acids of ZF3 and ZF-binding amino acids are labeled with red, green and, violet stars, respectively.

CTCF is known to contain the conserved ZF domain, belonging to the “C2H2” (classical) type. To check this in detail, multiple alignments of 11 zinc fingers of CTCF was performed. The obtained result highlights extreme conservation of the ZF domain sequences. **Figure 4.3** shows the multiple alignments of ZF 1-11. Only ZF11 is different zinc finger class. The DNA-contacting residues C, C, H and H were shown to be

conserved in the positions 3, 6, 19 and 24 (according to the numbering shown at the top of the figure) of the alignment (except for the ZF11, in which the H at the end of the alignment is absent). In general, these residues were found to be more conserved than other residues in the aligned sequences. Such extreme conservation hints for essential functional relevance. Hence, mutation in one of them, is expected to cause more change in protein when compared to other (less conserved) residues. In this regard, the H345R substitution, which is one of the 3 substitutions selected for investigation in this study, could be resulting in a substantial change of protein structure and function. This conserved residue occupies position 24 in the obtained alignment (according to the numbering shown at the top of the alignment).

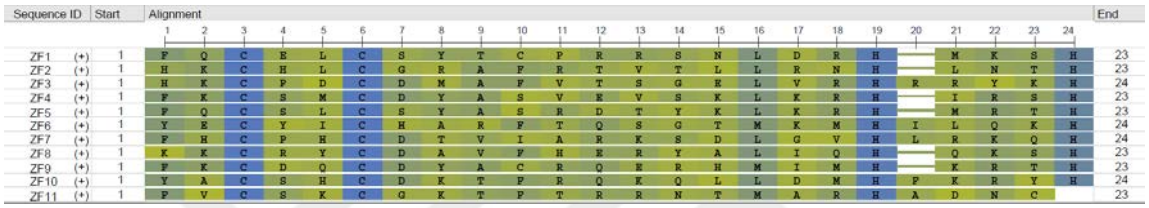


Figure 4.3: Alignment of 11 ZFs amino acid sequence. Coloring applied based on BLOSUM62 selection. Gaps represented as white, The four DNA binding residues are shown in blue color.

4.1.2. Disorder tendency in the structure

Disorder of the CTCF protein was investigated using 3 consecutive steps. In each step, a separate software was applied. First, prediction of disorder tendency in the CTCF structure was performed using the IUPred2 tool based on two methods (ANCHOR2 and IUPred2). The obtained results highlight higher disorder in the flanking regions when compared to the central region. In another word, in the central region, which harbors the 11 ZFs, the disorder tendency was found to be the lowest (in general, below the score 0.5) (**Figure 4.4**). The highest disorder was predicted for the right flanking region. Both of the implemented methods revealed comparably similar results. In sum, this analysis provided a detailed information about disorder levels across the amino acid sequence of investigated protein.

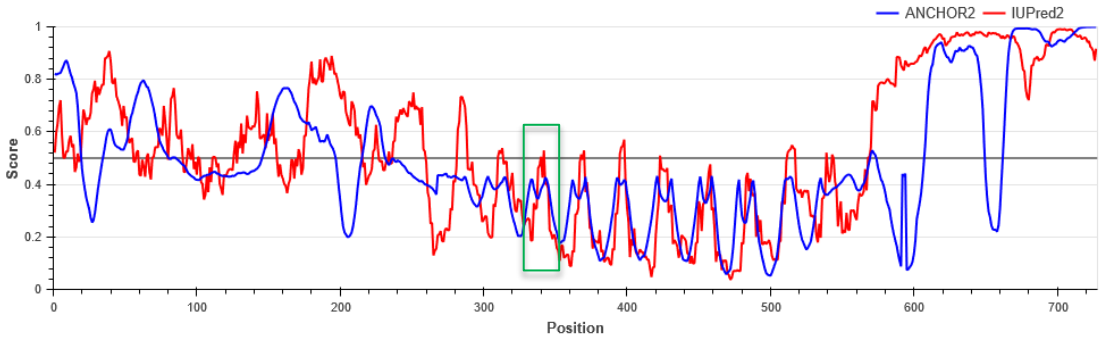


Figure 4.4: Prediction of disorder tendency in the CTCF structure based on two methods (ANCHOR2 and IUPred2) highlights higher disorder in the flanking regions when compared to the central region. The position of ZF3 is highlighted with green box.

Second, in order to obtain a more comprehensive and integrative view into CTCF disorder, the d2p2 Pro tool was applied. This software provides results for 9 separate methods, focusing on predicting disorder, along with the related consensus result in the form of prediction agreement. In addition to this comprehensive summary, the program also integrates and presents some relevant structural properties, namely predicted domains, MoRF Regions and PTM Sites. All in all, this analysis provides a broader view into CTCF structure when compared to the IUPred2 analysis. In concordance with the previous step, more disorder was observed in the flanking regions, when compared to the center of the protein. Interestingly, the areas of low or absent disorder corresponded to the locations of the predicted ZF domains.

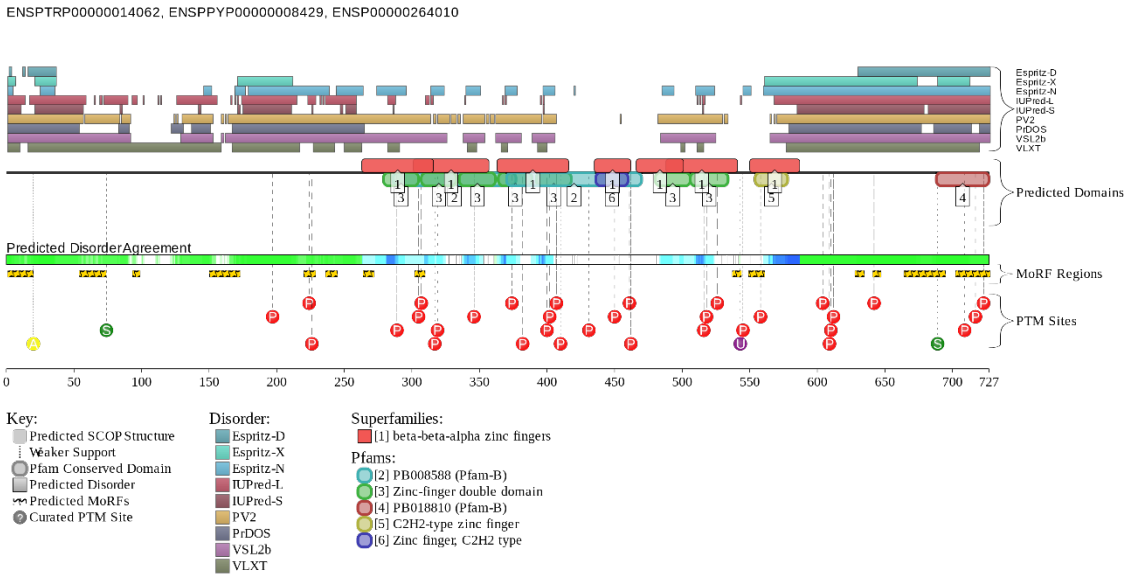


Figure 4.5: Predicted disorder agreement (based on 8 methods), along other structural features, such as predicted domains and PTMs, across the CTCF primary sequence highlights weak or absent agreement in the area corresponding to the predicted ZF domains. The green, blue and white color indicates strong, weak, and absent disorder agreement, respectively.

Finally, the CTCF-related record (in the form of a dedicated webpage) in the PDB website was used to retrieve and visualize disorder-related data in combination with other structurally-relevant information. Collectively, this visualization provided a unified view into known structural aspects of CTCF. Again (like in the previous two steps), the flanking regions were found to contain more disordered areas, when compared to the center of the protein. Importantly, the obtained visualization also highlights the position of 11 ZF motifs, along with the span of the available PDB structures. Remarkably, these 11 PDB structures mostly correspond to the region of low disorder. Importantly, each PDB structure covers only limited number of the ZFs (**Figure 4.6**). Among the available PDB structures, the combination of 5T0U and 5UND structures was used in the later steps of this study in order to incorporate a longer sequence of ZFs, namely ZFs 2-9 (see section 3.2).

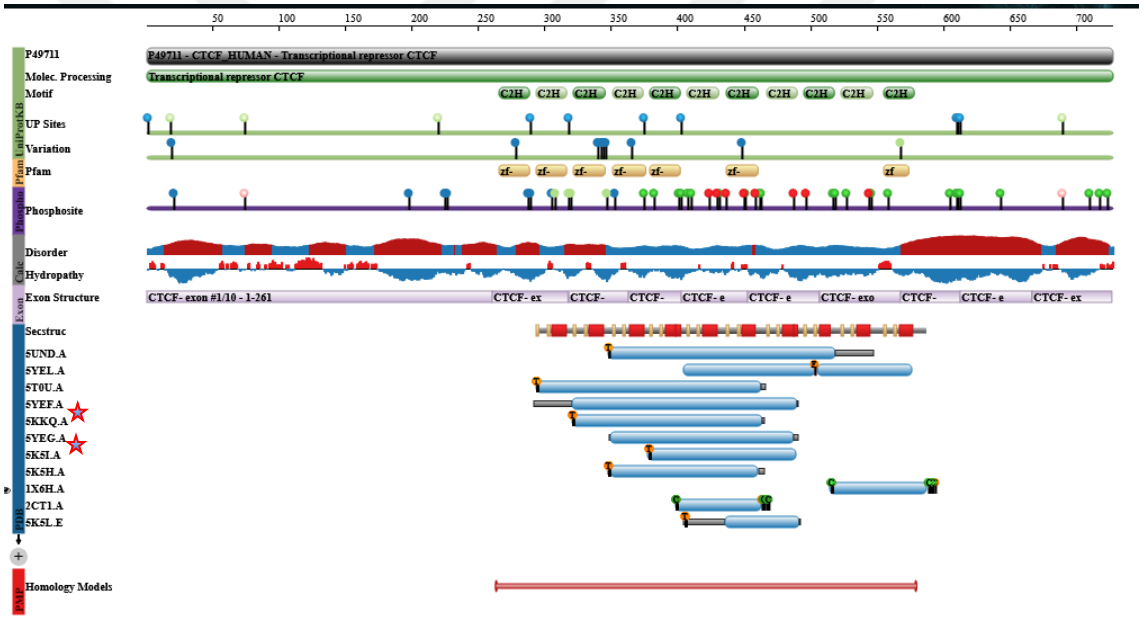


Figure 4.6: Overview of disorder in the CTCF protein along with other structural aspects, such as the position of the 11 C2H2 ZF motifs and the coverage of the available PDB structures. The PDB structures used in this study a labeled with red stars. Source: PDB website.

Could disorder be manifested in the temperature of the protein and hint for flexibility? As a complementary approach to disorder, in order to gain more insights into flexibility of the protein, B-factor analysis was performed using ProSNEx tool. As input, the PDB structure 5T0U was used. B-factor analysis based on an available PDB structure highlights differential values along the CTCF protein.

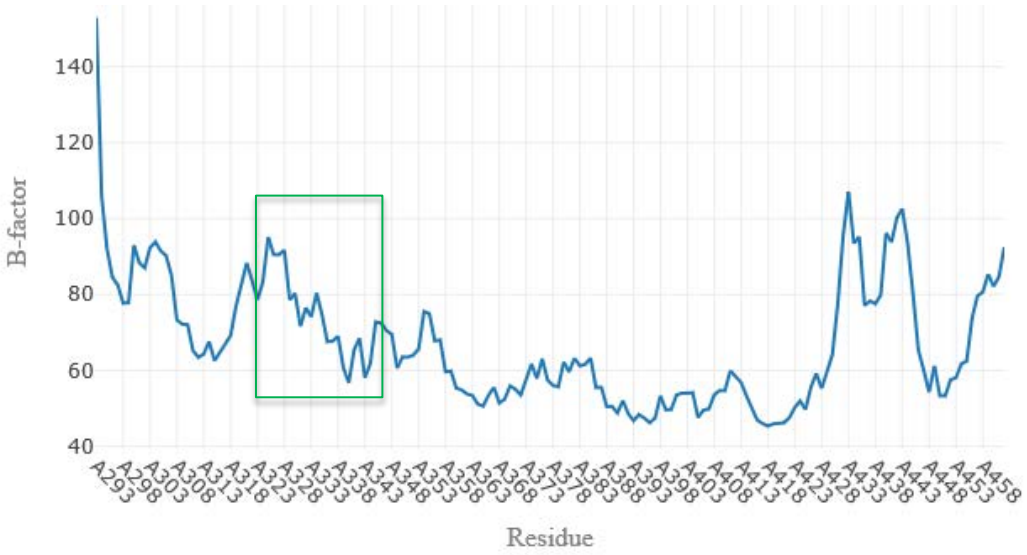


Figure 4.7: B-factor values across the primary sequence.ZF3 region is shown green box.

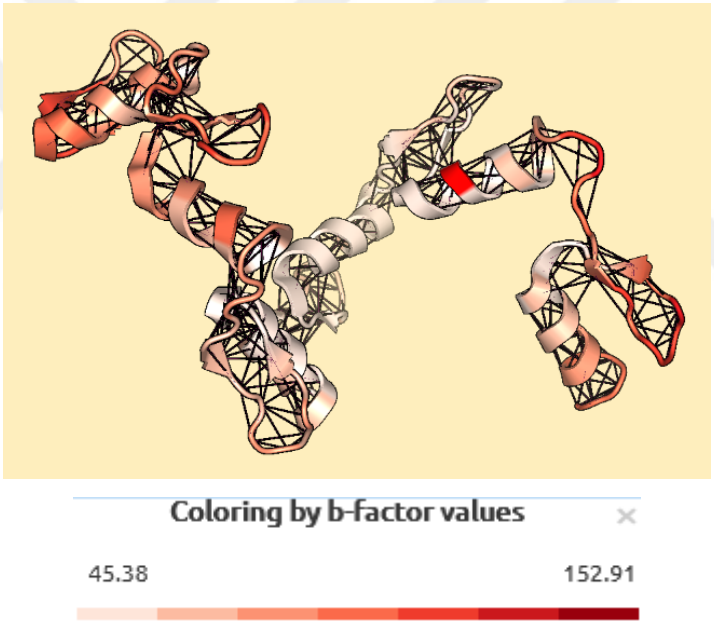


Figure 4.8: Coloring of 3D structure according to B-factor values.

4.2. Structure of combined CTCF-DNA complex

In **Figure 4.6** starred PDB IDs are used to have completed CTCF. As result of combination, the resulting structure consisted of 293rd- 432nd AA of 5T0U and 433rd-518th AA of 5UND. The resulting CTCF protein-DNA complex has contains “293rd AA to 518th AA” CTCF protein structure, which is 226 AA long (with regard to the protein), and also 35 bp long (with regard to the DNA). **Figure 4.9** shows the overlapping regions. Appendix has alignment of fasta files.

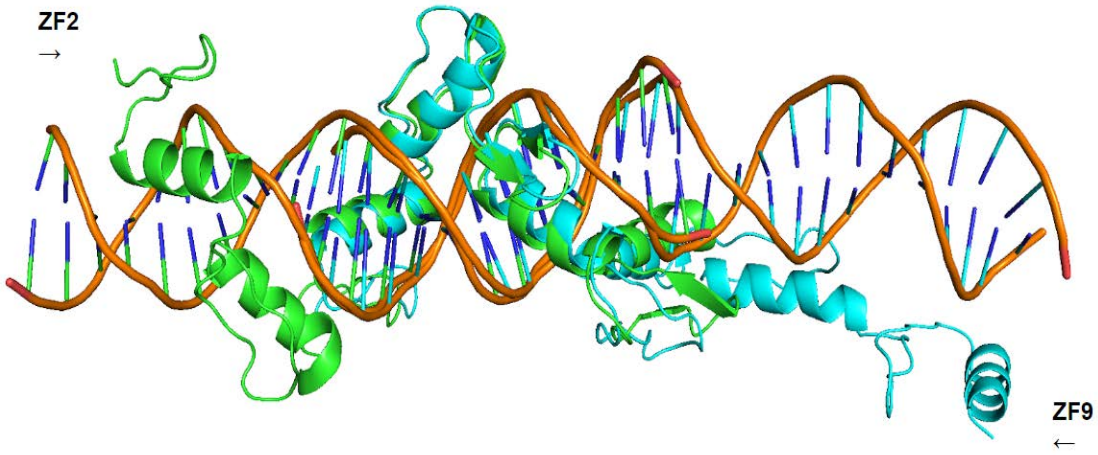


Figure 4.9: Alignment of PDB:5T0U (green) and PDB:5UND (cyan) shows overlapped regions of CTCF-DNA complex. Structural alignment and visualization by pyMOL.

Since the 10th Zinc Finger could not be observed in X-ray crystallography of the 5UND PDB structure, it is missing. Therefore, the final obtained structure is comprised of only 2-9 ZFs.

4.3. Molecular Visualization of Mutation Sites

As a result of combining two different CTCF-DNA complexes, the final CTCF protein-DNA structure was successfully obtained. Structure is shown in **Figure 4.10** visualized from UCSF Chimera [65] program with DNA and 2-9 zinc fingers of CTCF.

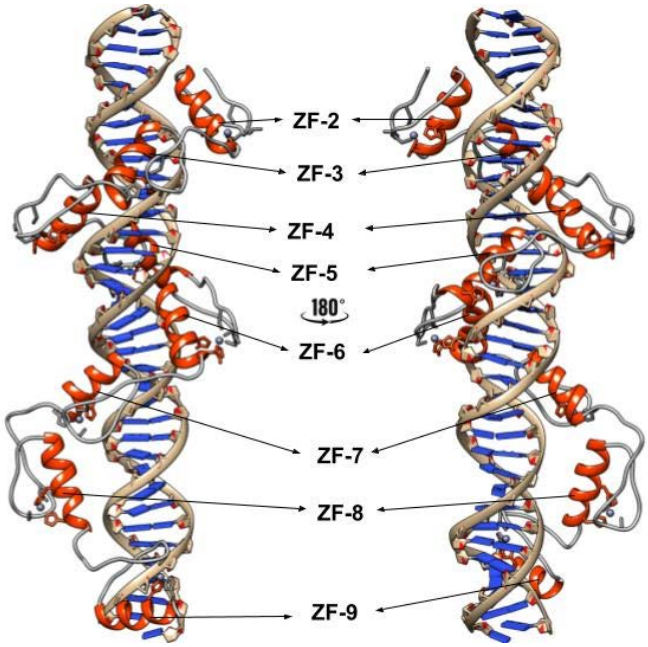


Figure 4.10: Final CTCF-DNA complex ZF2-ZF9 visualization by using Chimera. The figure is adopted from Figure 4 in [23].

In literature review, 17 missense CTCF variants pointed that are reported as tumor-specific mutations that alteration of zinc finger domains effects its binding to DNA. Detailed information about 17 missense variants is on **Table A.2**. The table includes information about variants which are dbSNP and COSMIC IDs, which cancer type and occurred regions. COSMIC is the largest database to explore the impact of expert manually curated somatic mutation in human cancers [66]. The Single Nucleotide Polymorphism database (dbSNP) is an open-source genetic polymorphisms archive that contains many kind of human variation, population frequency and informative genomic mappings for polymorphism and clinical mutations. dbSNP accepts submissions from public laboratories and private organizations [67]. By literature and database screening, three point mutations, namely H345R, K344E, R339W were prioritized and selected for in-depth investigation in this study. The three studied point mutations are located in 3rd Zinc Finger domain and it can be seen in **Figure 4.11** marked in orange. **Figure 4.11** also shows wild type amino acids of the studied mutation sites as purple licorine.

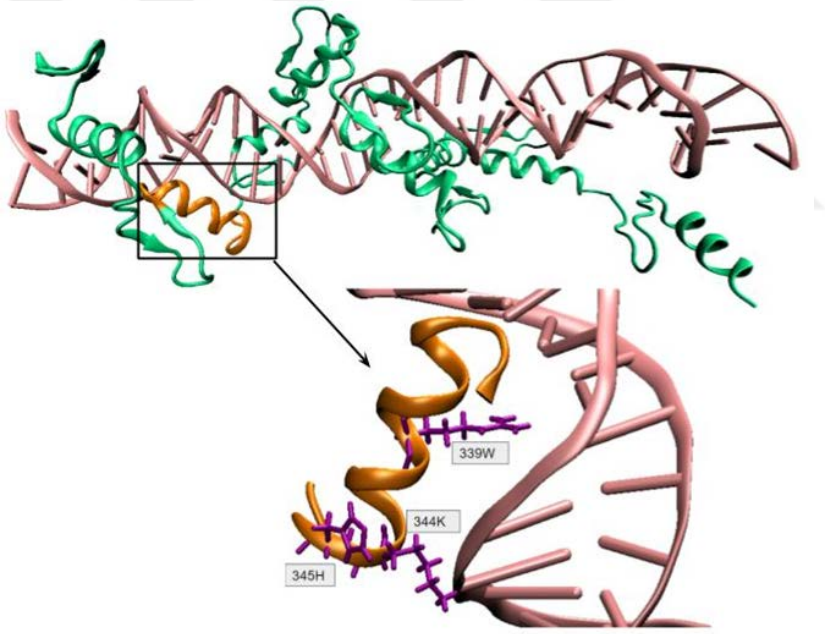


Figure 4.11: Mutation sites of interest, namely 339W,344K and 345H visualized by using VMD. 3rd zinc finger colored as orange; rest of the protein is lime. DNA colored as pink. Studied mutation sites visualized as purple licorine.

R339W mutation that positively charged arginine mutated to aromatic group tryptophan. K344E mutation in which lysine is positively charged amino acid mutated to negatively charged glutamate. H345R mutation location is end of 3rd zinc finger and histidine is neutral amino acid mutated to positively charged arginine.

In addition to the neutral state (which is a more standard case), histidine can also be positively charged in some conditions. However, in the used structure it is neutral, as shown in **Figure C.1**.

HOPE server is used for both visualization and prediction of effects on mutation sites. The following figures include the chemical structures of the wild-type and the mutant amino acids. The backbone of amino acid is common for both, the side chain, unique for every amino acid. Common backbone colored as red and color of changed residue is black. And their structural representations are shown by pyMOL above the chemical structures. WT amino acids are marked in green licorine, mutant amino acids are marked in red licorine. The final picture of figures is superimposition result from HOPE server. 339th amino acid arginine turns to tryptophan (**Figure 4.12**), 344th amino acid lysine turns to glutamate (**Figure 4.13**), 345th amino acid histidine turns to arginine (**Figure 4.14**).

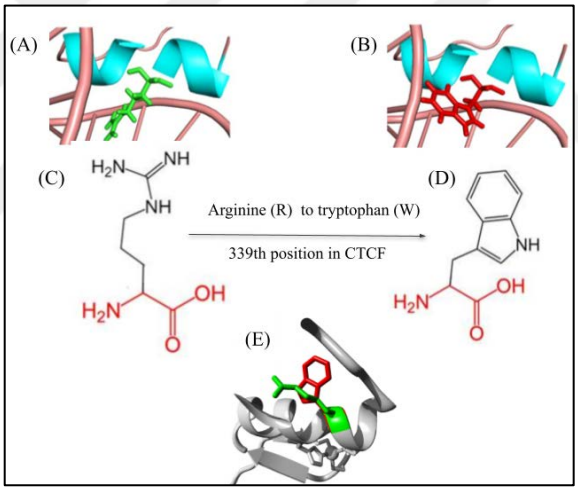


Figure 4.12: Amino acid changes of R339W. **a)** Structure of R339 and **b)** W339. Chemical structure of **c)** arginine and **d)** tryptophan**. e)** Superimposition of R339 to W339 on CTCF.

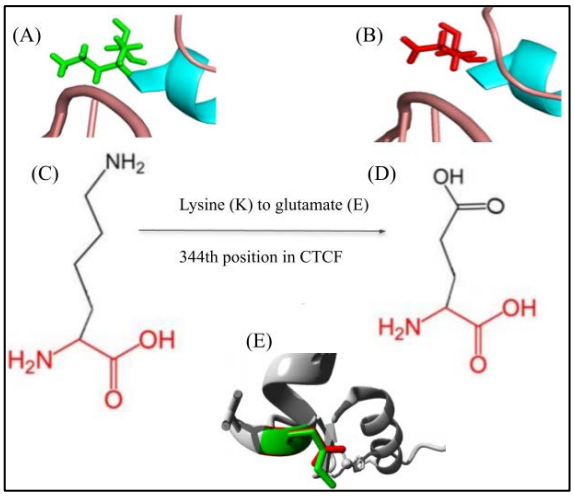


Figure 4.13: Amino acid changes of K344E. **a)** Structure of K344 and **b)** E344. Chemical structure of **c)** lysine and **d)** glutamate**. e)** Superimposition of K344 to E344 on CTCF.

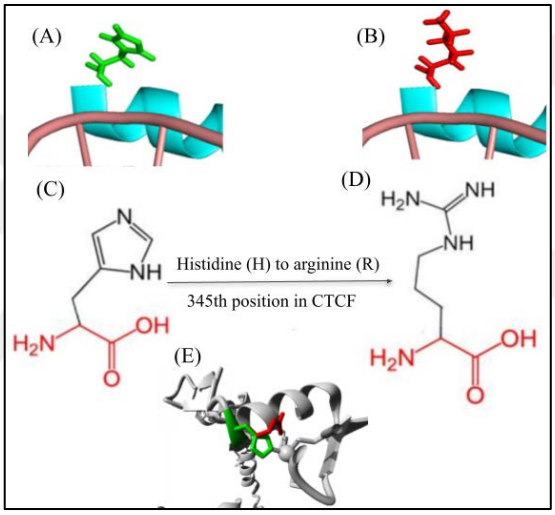


Figure 4.14: Amino acid changes of H345R. **a)** Structure of H345 and **b)** R345. Chemical structure of **c)** histidine and **d)** arginine**. e)** Superimposition of H345 to R345 on CTCF.

4.4. Effect of Mutations

The relation of CTCF with cancer has been mentioned many times in the literature. However, the last status of the studies emphasizes that CTCF is still a candidate tumor suppressor protein. The aim of this study is to provide in silico approaches to literature, in addition to these studies and to illuminate the CTCF protein structure with MD simulation study. For these, 3 mutations on the third zinc finger mentioned in the literature were selected and their effect on CTCF protein was demonstrated by many analysis methods. The R339W, H345R, K344E mutations were experimentally demonstrated that CTCF does not bind to CTSs on DNA, or partial binding occurred (**Figure 4.15**). These results of 3 mutations have been shown to be associated with cancer. Considering these,

mutation prediction tools were used at the beginning, to have foresight for the next analyses and to validate the experiments.

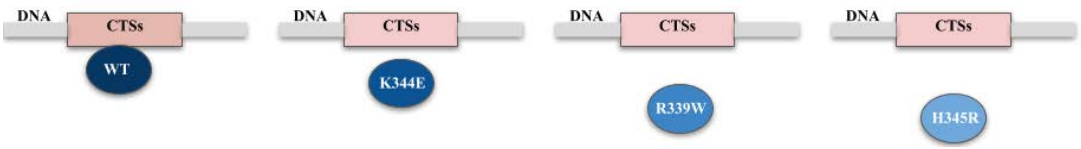


Figure 4.15: Graphical representation of overall binding affinity between CTCF and CTSs for WT and the three investigated mutants. Derived from **Table 1.1**, which is adopted from [34].

Every amino acid has different properties as charge, hydrophobicity, and size. WT and mutant amino acids diverge in the properties. Summary of the changes get from HOPE server analysis in these properties are in **Table 4.1.**

Table 4.1: Changes in size, charge, and hydrophobicity results from HOPE server.

Mutations	Size	Charge	Hydrophobicity
R339W	Increased	positive to neutral	more
K344E	Decreased	positive to negative	no change
H345R	Increased	neutral to positive	no change

The changes in R339W mutations, W339 is bigger than R339 and more hydrophobic. Differences in size and hydrophobicity might impact forming hydrogen bonds. After mutation arginine to tryptophan, the size differs and it might result in not to form the same hydrogen bond as the WT residue, because of wrong position. While WT residue arginine is positively charged, mutated residue tryptophan is a neutral residue, so charge is lost after mutation.

In K344E mutation, the mutant residue has smaller size than the WT residue. The difference in size between WT and mutant amino acid, might disturb hydrogen bonds. While WT residue charge was positive, mutant residue charge is negative. K344E mutation introduces the opposite charge to the position and it is possible disturbing contacts with other molecules.

In H345R mutation, the mutant residue has bigger size than the WT residue and size difference between WT and mutant residue can result arginine not being in the correct position to form hydrogen bonds. Histidine is neutral but the mutant residue has positive charge. The mutation introduces positive charge to the position that might cause repulsion between the mutant residue and neighboring residues.

R339 residue is buried in the center of a ZF3 domain. The changes between the WT and mutant residue might result disturbing the core structure of ZF3. Also, the mutant residue is bigger and possibly will not fit in the central region of ZF3. R339W might cause losing hydrogen bonds in the core of the protein and might result misfolding.

H345 and K344 residues are located on the surface of the ZF3 but function is unknown. The residues were not in contact with other ZF domains of which their function is known within the used 3D structure. However, interaction with other molecules or other parts of the protein is still possible and might be affected by H345R and K344E mutations. According to HOPE server, the differences in amino acid properties by mutations might disturb zinc finger domain which is known to bind DNA which might be disturbed by this mutation.

Prediction of the effects of mutations using *in silico* tools is frequently used approach. Most preferable tools as SIFT, Provean, MutationTaster predict effects of the mutations from primary structure of CTCF are in **Table 4.2**. Mutations' ID in COSMIC and dbSNP databases are in **Table 4.3.**

Table 4.2: *In silico* prediction results and scores of studied mutations.

AA Change	Cancer type	SIFT	Provean	MutationTaster
R339W	Wilms tumor	Score:0 Damaging	Score: -7.24 Deleterious	Disease causing
K344E	Breast tumor	Score: 0.006 Damaging	Score: -3.26 Deleterious	Disease causing
H345R	Prostate tumor	Score: 0.001 Damaging	Score: -7.24 Deleterious	Disease causing

Table 4.3: Studied variants and their dbSNP and COSMIC IDs.

Variants	dbSNP	COSMIC
R339W	rs1243010179	COSM6908613
K344E	rs1215280530	COSM6908612
H345R	-	COSM280199

In silico results show that the studied mutations have pathogenic effects on CTCF Prediction tools are a cheap and time-saving methods for verifying experimental results and before experimental studies. For this reason, the tools are developed to have high accuracy, specificity, and sensitivity. The three studied mutations associated with CTCF show essentially high pathogenicity. These mutations can be considered as disease causing mutations. R339W shows higher pathogenicity in Provean and SIFT scores. The K344E may be less damaging compared to the other two mutations.

4.5. Mutations affect stability of CTCF

Point mutations may affect stability of a protein and its folded state by altering the interactions intra-molecule such as hydrogen bonds, hydrophobic and electrostatic interactions. CUPSAT webserver was used to calculate the protein structural stability to consider its effect and predict the level of destabilization. The tables below show predicted protein stability changes on studied point mutations. Result table indicates experimental thermal denaturation methods. Result in **Table 4.4** includes changes in the overall stability calculated by using atom and torsion angle potentials. Secondly, it includes the adaptation of the torsion angle (phi, psi) are as favorable or unfavorable. Last, the predicted ΔΔG positive value means stabilize, negative value destabilize effect.

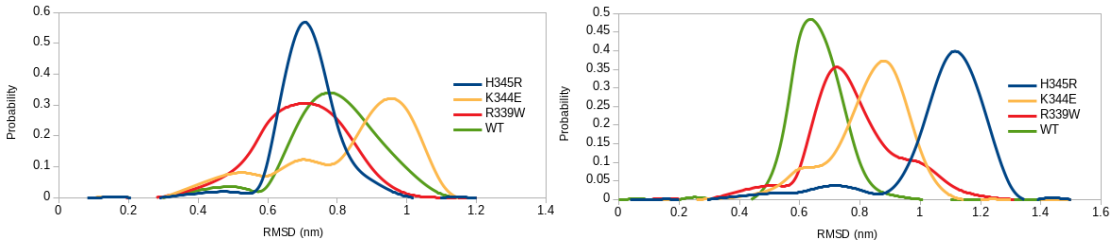
Table 4.4: Structure-based stability result of studied mutations from CUPSAT webserver [68].

Amino acid changes	Overall Stability	Torsion	Predicted ΔΔG (kcal/mol)
R339W	Destabilising	Favourable	-1.78
K344E	Destabilising	Favourable	-0.13
H345R	Destabilising	Unfavorable	-0.97

Based on CUPSAT calculation results, mutations are destabilized. R339W mutation decreases the stability of CTCF presenting with the most negative free energy change (ΔΔG) value. The destabilizing of R339W mutation might be result of losing the binding affinity to DNA. K344E mutation is the one closer to wild type status with negative free energy change (ΔΔG) value and favorable torsion angles. While H345R decreases stability, torsion angles are unfavorable after mutation. The unfavorable torsion angles of H345R could be result of hardly adapting to new environment after destabilized protein. The results are predicted based on 3D structures and to know how the mutations destabilize CTCF, more reliable results are obtained after MD studies.

RMSD of C-alpha atoms were calculated for wild-type and mutant proteins with respect to their initial structures during 500 ns MD. Probability distribution of the RMSD of the 4 systems both replicates are shown in **Figure 4.16**. The graph of RMSD over time (ns) is in **Figure D.1** and the moving average of RMSD is in **Figure D.2**. For 1st replicate result, K344E has higher fluctuation than wild-type otherwise H345R and R339W have lower fluctuation. For 2nd replicate, wild-type has lowest fluctuation than other systems.

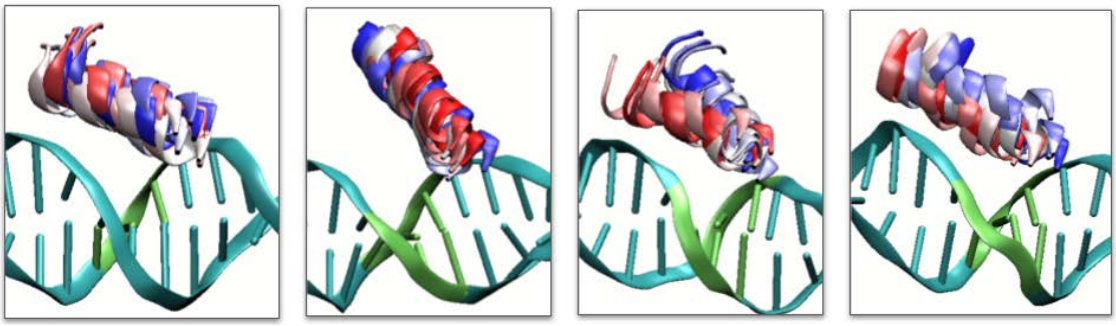
In 2nd replicate, H345R system presented similar RMSD to 1st replicate in first 50 ns but after 50 ns some reasons kept H345R with higher RMSD during trajectory. Reason is fluctuation of ZF2 is very high in 2nd replicate and affects total fluctuation of system. without including ZF2 for RMSD calculation, H345R system represents same fluctuation to 1st replicate as 0.68 nm average. As a result, RMSD increased after K344E mutation and protein destabilized. However, R339W and H345R mutations represent similar RMSD to wild-type means that systems are still stable after mutation.



(a) 1st replicate **(b)** 2nd replicate

Figure 4.16: Probability distribution of RMSD (nm) for **a)** 1st and **b)** 2nd replicates.

Besides RMSD results, considering changes and fluctuation of protein by time will give different aspects to dynamics of protein. **Figure 4.17** shows only ZF3 fluctuation, complete protein timestep fluctuation in **Figure D.4**. No change in stability does not indicate that the orientation of the protein remains the same. R339W and WT showed similar RMSD, while the ZF3 changed direction and the orientation drastically (**Figure 4.17b**). Simulation of W339 started with different orientation than wild-type. The reason is modeling program did not place W339 in same volume as R339. Such orientational changes in structure and still being stable may have different results. As expectation of MD, in K344E, stability decreased while ZF3 dislocated (**Figure 4.17c**). On the other hand, H345R was able to maintain its stability and conformation after mutation (**Figure 4.17d**).



(a) WT **(b)** R339W **(c)** K344E **(d)** H345R

Figure 4.17: Timestep analysis of 1st replicate trajectory. Coloring of helix; from initiation position to final position is red to blue. Lime color represents triplet nucleic acids of ZF3 binding site. Rest of the DNA is cyan.

RMSF calculates the fluctuation of atoms at positions in the trajectory. RMSF of C-alpha atoms is calculated to measure overall flexibility of the wild-type and mutant complexes (**Figure 4.18**). ZF3 and ZF4 regions become more flexible after mutations. Here we see the local effect of mutation due to ZF3 is the mutation site and ZF4 is the neighboring zinc finger. Rest of the protein is less flexible after R339W and K344E mutations. K344E mutation increased the flexibility in ZF2 and B-sheet region of ZF8. H345R mutation does not affect flexibility of CTCF generally. ZF2 highly fluctuated in H345R and its effects RMSD of CTCF. Also, ZF8 is less fluctuated after H345R. Our structure starts with ZF2 so amino acid is numbered based on first amino acid in structure which is 293.

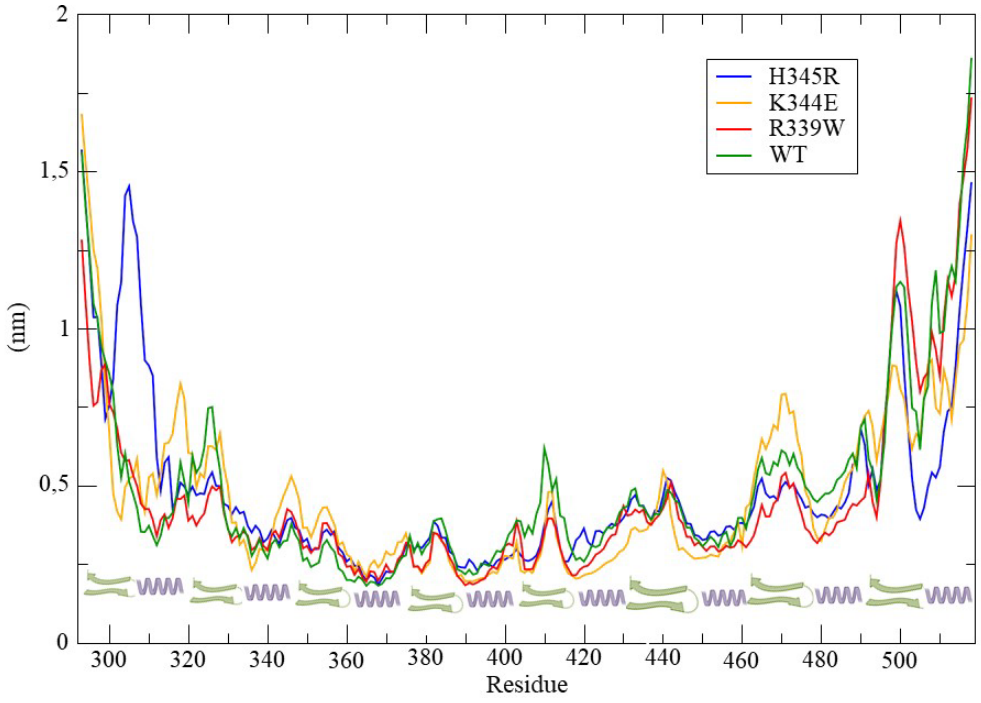
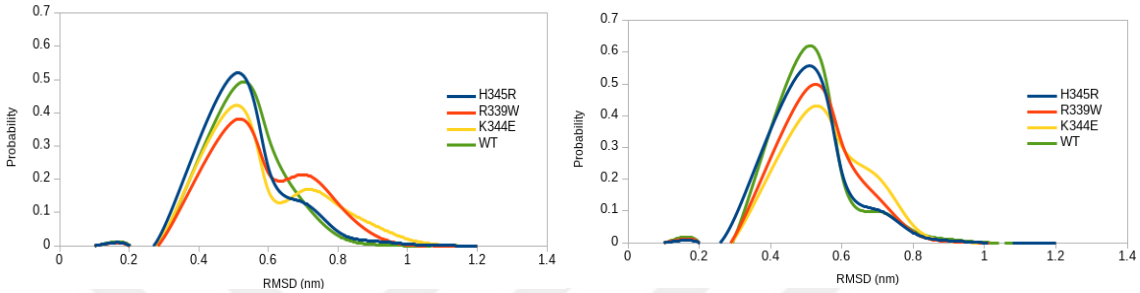


Figure 4.18: RMSF graphs of C-alpha atoms of wild-type and mutant systems obtained from combined replicates. Helix and B-sheet regions are shown along CTCF ZF2-9.

RMSD of P atoms in DNA backbone for 5’ and 3’ chains were calculated for both replicates. RMSF of P atoms in DNA backbone for 5’ and 3’ chains were calculated for combined replicates (**Figure D.3**). Combined probability distribution graphs present that mutant systems’ DNA has less probability to have same stability as wild-type. Specially K344E and R339W systems have the possibility for more fluctuation (~0.2 nm). Any small changes in fluctuation of DNA might disturb the binding because of specific bindings.



(a) RMSD of 3’ strand of DNA **(b)** RMSD of 5’ strand of DNA

Figure 4.19: RMSD graphs of phosphate backbone of DNA strands.

4.6. Conformational differences between studied four systems

PCA is used to figure out the significance of the global motion of protein. In this regard, these visualizations help to compare the span of confirmations between each system in this study. The gmx covar module was used to calculates and diagonalizes covariance matrix of Cα atoms of wild-type and mutant systems trajectories. Eigenvectors were analyzed using gmx anaeig module. As expected, result, first two eigenvectors contribute almost 50% of the global motion of protein. Calculated cumulative fluctuation of eigenvalues by python script is shown in **Table E.1**. Almost all systems have 50% contribution in first two eigenvectors. Based on these results, two eigenvectors were considered for rms fluctuations and 2D projections calculations.

RMSF of first two eigenvectors C-alpha atoms. The figure shows most dominant motion of protein contributes to RMSF. It gives overall perspective for some regions flexibility. Details for mutations systems comparing to WT are in **Figure E.1, Figure E.2, Figure E.3.**

The 2D projection of the motion over the first two eigenvectors is generated by using Principal Component Analysis. The displacement of C-alpha atoms along the eigenvector 1 and 2 is that projection of the most global motions for wild-type and mutant systems in

phase space along the PC1 and PC2. As shown in **Figure 4.20**, wild-type covers larger region than mutant systems. Specially 1st replicate results indicate that mutants are in narrow spaces. It might show that protein is more rigid conformation after mutation. Wild-type has more flexible conformations to occur. As conclusion, while wild-type covers heterogeny regions, mutants are more in homogeny regions.

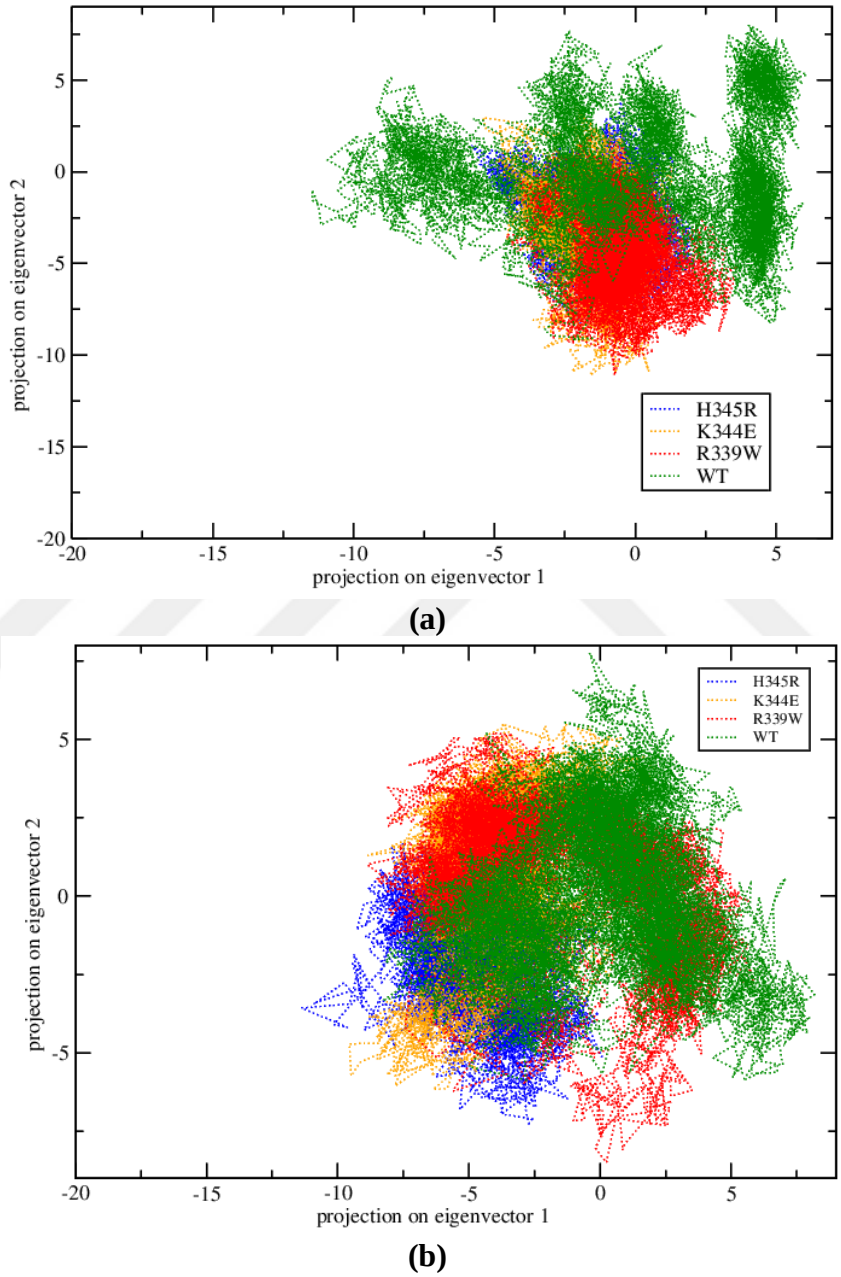


Figure 4.20: PCA-based 2D projection of the trajectories based on eigenvectors 1 and 2 **a)** 1st replicate and **b)** 2nd replicate.

After seeing how global motion changes, to ensure the inter-protein differences change after mutations, dynamical cross-correlation map (DCCM) analysis was performed using the Bio3D package in R that plots the cross-correlations in atom-wise. The DCCM plot

is a graphical representation from a matrix of all cross-correlations in atom-wise. Negative correlation indicates the atoms move in opposite directions, while positive correlation indicates the atoms move in the same direction [69]. **Figure 4.21** shows ΔDCCM plots of systems. The results are average, raw results of WT, R339W, K344E, H345R are in **Figure F.1, Figure F.2, Figure F.3** and **Figure F.4** respectively. Residues are numbered starts from 1 different other than other graphs.

According to results, positive correlation is increased for ZF3 domain for R339W and H34R. ZF3 in K344E has both positive and negative correlations. As shown in **Figure 4.21**, K344E moves different directories during trajectory. The effect of mutation also changes the correlation for neighbor domains ZF4 and ZF5. H345R increased positive correlation for ZF4. R339W changed direction of ZF8 according to ZF5-7.

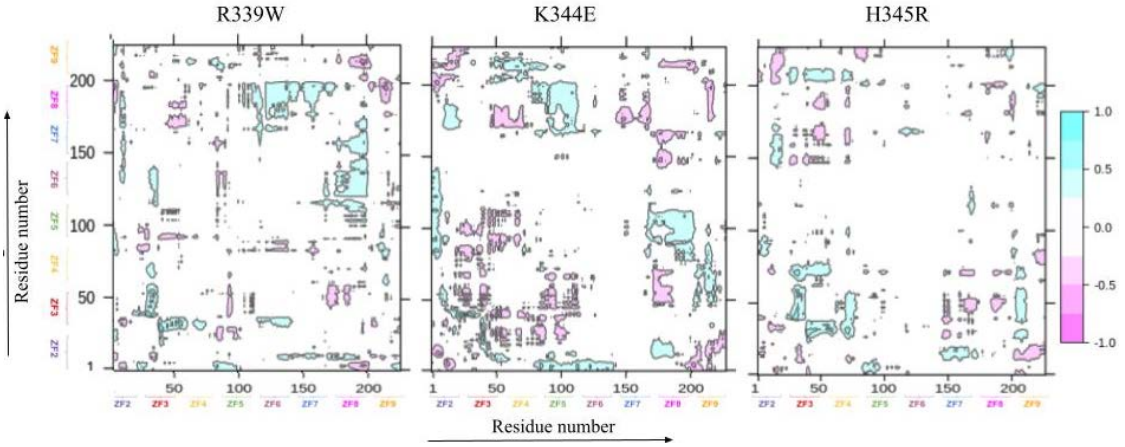


Figure 4.21: ΔDCCM plots of mutant systems. Positive correlation to negative correlation is gradient from cyan to pink as color bar. White denotes no correlation.

4.7. Intra-molecular and inter-molecular interactions after CTCF mutations.

Amino acid network (AAN) was done to represent residue interactions of the protein for analyzing local effect of mutations. Each residue is described as a node, and the physicochemical interactions between two nodes are described as edges [45]. These interactions include hydrogen bond, hydrophobic interactions, π-π stacking, electrostatic interactions and van der Waals forces. Besides individual interactions of mutation sites, the acquired knowledge about how mutations affect each residue along protein gives another perspective on the effect of mutations. **Figure 4.22** shows how degree of residues changes after mutation comparing to wild-type along ZF3 domain. The summary of interested residue interactions is as in **Table 4.5**. The visualization of R339W, K344E,

H345R systems of **Table 4.5** is in **Figure G.3, Figure G.4** and **Figure G.5** respectively. The network of residues in **Figure G.1** is to show the interaction map and degree distribution along wild-type CTCF calculated with RING 2.0, visualized with Cystoscape.

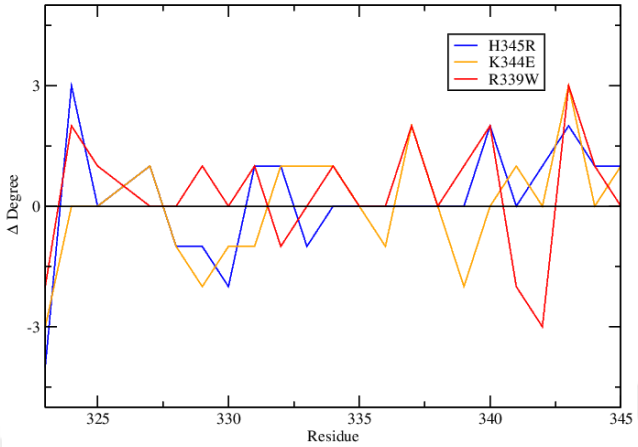


Figure 4.22: Δ Degree of residues along ZF3 domain after mutations.

Table 4.5: Interaction network of mutation sites compared to wild-type with relaxed and strict option. * indicates interactions between the residue, which were investigated in this thesis.

Mutation sites	Residues	Interaction	In mutation (relaxed)	In mutation (strict)
R339W	334:SER	HBOND	kept	-
R339W	335:GLY	HBOND	kept	kept
R339W	336:GLU	HBOND	kept	-
R339W	342:ARG	HBOND	kept	kept
R339W	343:TYR	HBOND	kept	kept
R339W	**344:LYS***	HBOND	kept	-
R339W	361:VAL	VDW	turn to HBOND	-
R339W	362:GLU	VDW	lost	-
R339W	350:PRO	VDW	gain	-
R339W	351:PHE	PIPISTACK	gain	gain
K344E	**339:ARG***	HBOND	kept	-
K344E	340:HIS	HBOND	turn to IONIC	kept
K344E	341:ARG	HBOND	lost	-
K344E	347:HIS	HBOND	gain	-
H345R	340:HIS	PIPISTACK	turn to HBOND	lost
H345R	341:ARG	HBOND	kept	kept
H345R	342:ARG	HBOND	kept	-
H345R	324:CYS	VDW	gain	-

Interactions between neighboring aromatic rings are called aromatic stacking or π-π stacking which refer to a specific type of intermolecular interaction distinct from the non-covalent interactions unique to aromatic molecules [70]. There is π-π stacking between two zinc ions coordinating residues H340 and H345. When the π-π stack interaction between H340 and H345 is lost by H345R mutation, results in loss of Zn ion from ZF3. Interestingly, after H345R substitution, R345 has van der Waals interaction between C324 zinc ion coordinating residue (See section 3.8). Another important π-π stacking occurs after R339W mutation between W339 and F351 residues.

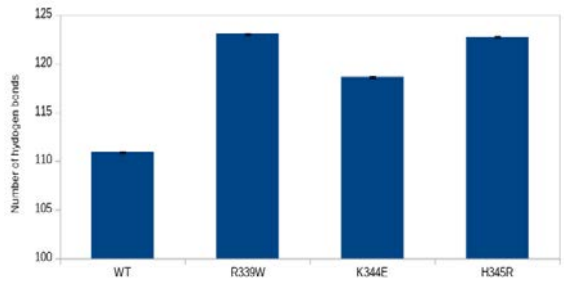
R339 interacts with ZF4 with V361 and E362 while W339 has the interaction with closer residues which are P350, F351. System lost one VDW interaction with E362 and gain with closer residue P350. On the other hand, V361 VDW interaction turns to hydrogen bond which is stronger. These results show that ZF3 became more rigid than wild-type by R339W mutation.

K344E mutation causes ionic interaction which is salt bridges at H340 instead of hydrogen bond and gains a hydrogen bond on loop between ZF3 and ZF4 H347 position. These new interactions might affect DNA binding. Moreover, both mutation sites K344 and R339 have hydrogen bonds that still keep after R399W and K344E substitutions.

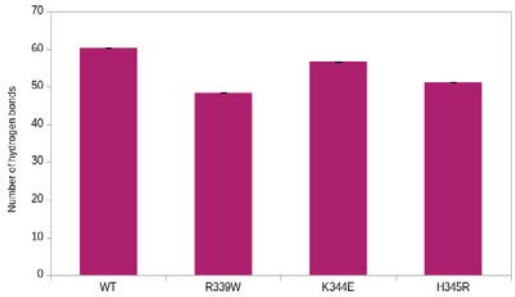
The interactions mentioned in **Table 4.5**, can be seen in the appendix figures generated by Cytoscape. Residues were colored by their degree which means number of interactions of each. Edges show any type of interaction of only interested residue in each figure. Colors represent type of interaction, hydrogen bond, π-π stacking, van der Waals or ionic interaction (salt bridges).

The network analysis shows what might happen between residues based on distances and energy calculations from static structure of systems. After MD dynamics, we get more reliable results for interactions. The number of hydrogen bonds is calculated by gmx hbond for last 400ns. Formation of hydrogen bonds inter-protein or between DNA is a major force for stability of protein structure. The hydrogen bonds analysis done for protein itself and between protein and DNA. Results are represented as average of

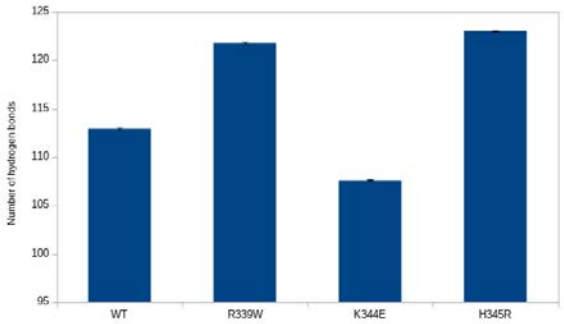
hydrogen bonds occurred. The mean and median values are in **Table H.1** and **Table H.2**.



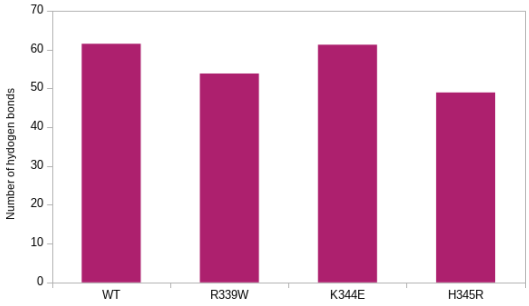
(a) Protein-protein, 1st replicate



(b) Protein-DNA 1st replicate



(c) Protein-protein, 2nd replicate



(d) Protein-DNA 2nd replicate

Figure 4.23: Number of hydrogen bonds (pairs within 0.35 nm). Between **a)** protein-protein in 1st replicate, **b)** protein-DNA in 1st replicate. **c)** protein-protein in 2ndreplicate, **d)** protein-DNA in 2nd replicate.

H-bonds analysis in the course of trajectory shows that while h-bonds between protein-DNA break, h-bonds in protein occur. The changes during trajectory is in **Figure H.1** and **Figure H.2.** This will result into less interaction between protein-DNA, while protein becomes more rigid. These results show that the H345R and R339W mutant systems are more stable and rigid than wild-type. K344E system is do not indicate the result due to differences in two replicates.

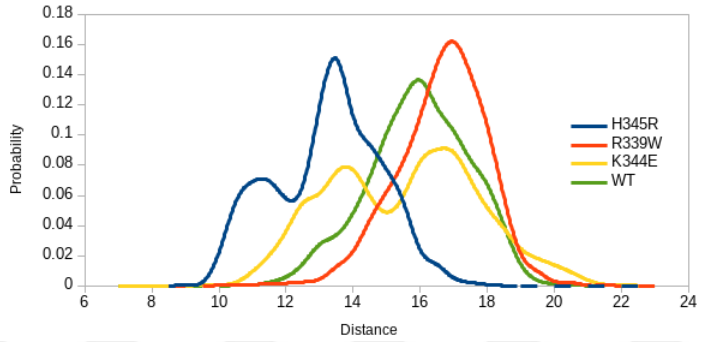
Previous results indicate the effects of mutations on CTCF structurally and interaction with residues. After all, DNA-CTCF interactions were observed. Distance analysis is done to determine protein-DNA binding affinity of all systems. Distance changes during 500 ns trajectory were analyzed by VMD for both replicates. In this analysis, we examined the distance of the third zinc finger, which is the mutation site, from DNA to demonstrate the local effect of the mutation.

Distance analysis were done between 3 amino acids that bind to the DNA binding site specified in the study, which expose CTCF-DNA protein [24], used for the thesis. By calculating the distances over the backbone, C-alpha atom of amino acids and phosphate atom of nucleic acids, it is aimed to ignore fluctuations as much as possible. The

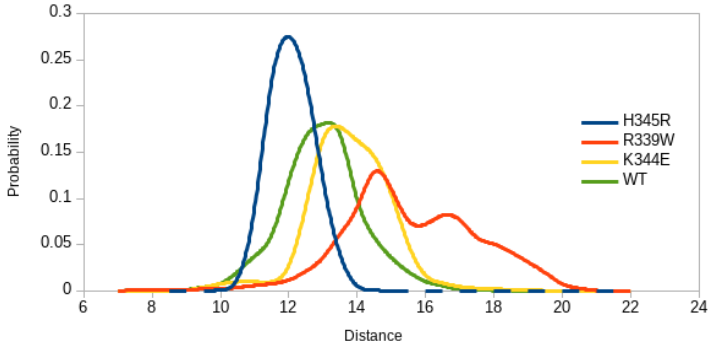
calculations were done between two-pairs atoms as following:

- THR333:CA and THY28:P
- GLU336:CA and CYT27:P
- ARG339:CA and GUA26:P

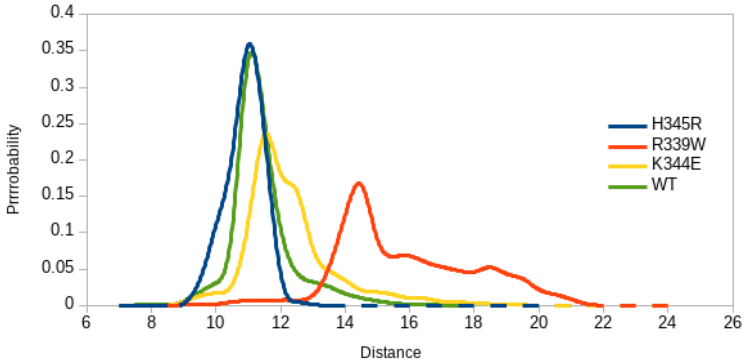
The two-pairs atom distances of crystal structure of wild-type are respectively 10.23, 10.64, 10.64 Å. These distances were used as reference distance for time-wise distance graphs are T333 (**Figure I.1**), E336 (**Figure I.2**) and R339 (**Figure I.3**) binding sites. The next graphs show probability distribution of distance for total of 1 microsecond. The average bar plots of distances are in **Figure I.4, Figure I.5** and **Figure I.6** for T333, E336 and R339 respectively.



(a) T333 DNA binding site



(b) E336 DNA binding site



(c) R339 DNA binding site

Figure 4.24: Probability distribution of distance in binding sites respectively **a)** T333, **b)** E336, and **c)** R339 DNA binding sites.

According to the conformational changes in **Figure 4.17,** mutation alters the interaction with DNA in ZF3. The distance results show that three mutated CTCF have different affinities for DNA. R339W mutation on **Figure 4.17b** moves away from DNA obviously. Also, in distance analysis we see that after the binding site R339 mutated, it moves away from DNA. After calculation number of contacts, R339 and neighbor residues 338 and 340 has no contact with DNA (cut off < 3 Å). The zero contact indicates that W339 has no affinity for binding DNA. T333 and E336 amino acids have also less attraction to DNA in comparison to other systems. On the other hand, K344E mutation has similar binding behavior with wild-type. Distance changes in R339 and E336 binding site of K344E is similar to wild-type. In T333 binding site, it is complex due to different results of 1st and 2nd replicates. Nonetheless, specific interaction of E344 decreased shown with number of contact analysis results (cut off < 3 Å). H345R mutation has makes CTCF have more affinity to DNA than wild-type. In all graphs we see H345R is closer to DNA than wild-type.

4.8. Zinc atom localization

HOPE server analysis results mentioned zinc ion localization and how affects the stability. According to HOPE server result, R339 residue does not contact with a metal which is a zinc ion, but neighboring residues make a metal-contact. It is found to be involved in a metal ion contact in the 3D structure of wild-type K344. In the 3D structure can be seen that the wild-type H345 is involved in a metal ion contact. The representation of zinc ion localization in interested residues is shown in **Figure 1.1**.

(a) R339W **(b)** K344E **(c)** H345R

Figure 4.25: Superimposition of wild-type and mutant residues generated from HOPE server. Color of protein is grey; side chains of wild-type and the mutant residue are colored as green and red respectively. Zinc ion is shown as grey sphere.

According to HOPE server analysis, zinc ion interaction is affected after mutations. Zinc ion might be affected by the R339W in its vicinity due to neighboring-metal-contact

residue. After K344E mutation, the size differences between the lysine and glutamate might disturb the interaction with the zinc ion and also, the difference in charge might disturb the interaction with zinc ion which may lead to destabilization of the domain. Also, H345R differs in size between wild-type and mutant residue so, might disturb the interaction with the zinc ion. Because a positive charge is introduced very close to the positively charged zinc ion there might be repulsion and destabilization of the domain.

The results of HOPE server are a kind of prediction from 3D structure. After MD analysis we checked the zinc ion coordination of ZF3 domain. MD analysis confirmed that H345R mutation loses zinc ion of ZF3 domain. **Figure 4.26** represents zinc ion localization of wild-type and three mutant systems during 500 ns trajectory. Results are same for both replicates and figures are from second replicate trajectory. While wild-type and R339W, K344E systems keep interaction with zinc ion, H345R loses zinc ion from the beginning of trajectory. While H345 residue which coordinates anion zinc ion is neutral, R345 residue is positively charged. So H345R mutation caused zinc ion loss because of repulsion of positive ions.

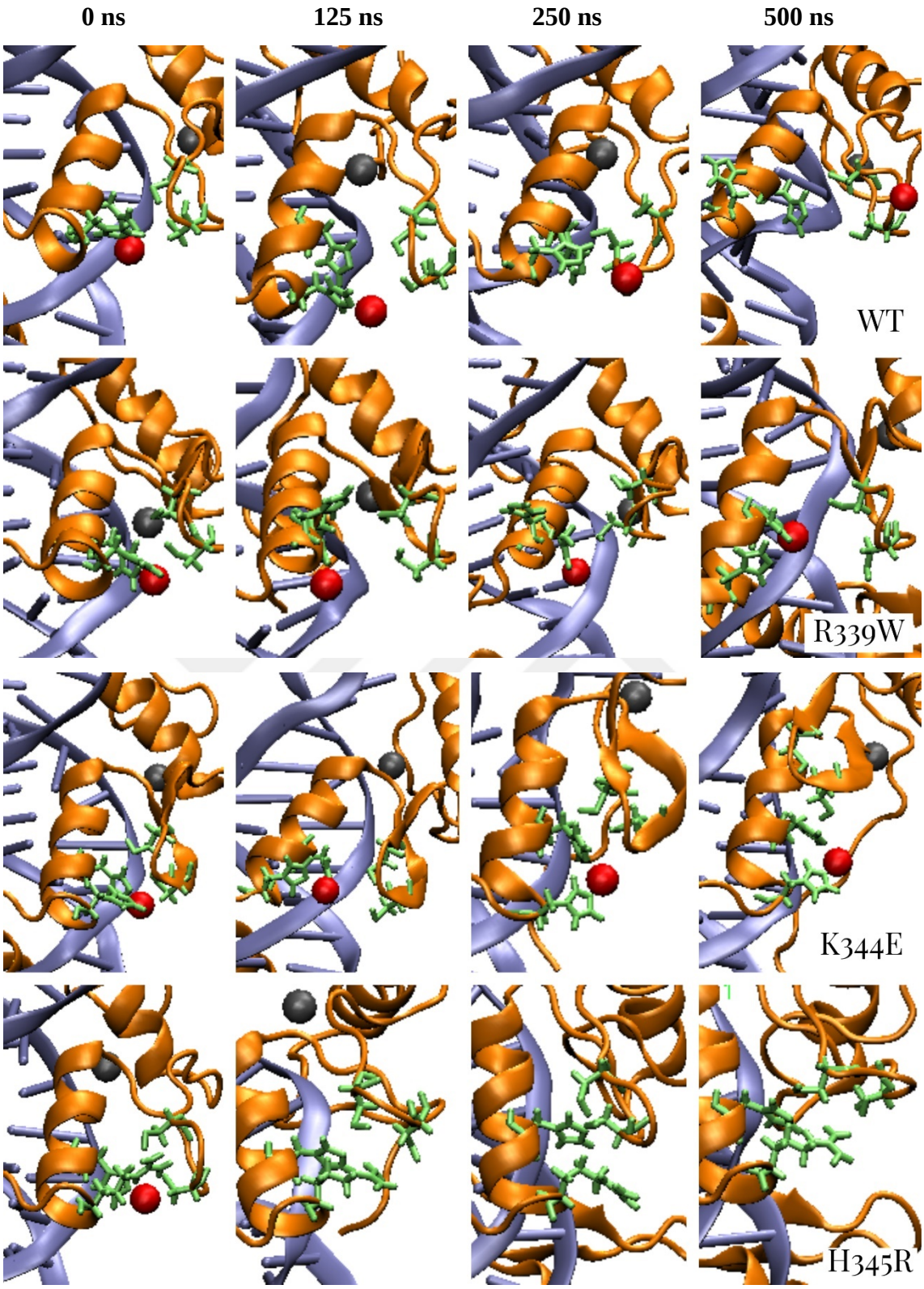


Figure 4.26: ZF3 loses the Zn ion upon the H345R mutation while the WT protein, and the R339W, K344E mutant proteins maintain it. DNA is colored ice blue; protein is colored orange; C324, C327, H340, H345 are highlighted as licorine lime. Zinc ion of ZF3 showed as red sphere while rest of zinc ions are gray spheres. Images were generated using VMD.

4.9. Discussion

4.9.1. The studied mutations occur in the conserved, ordered region of CTCF

CTCF is known as highly conserved protein. Our analysis shows that DNA binding regions which are zinc finger domains are the conserved areas. Additionally, these areas are Evolutionary Conserved Regions (ECRs). While the regions out of the of 11 ZFs are found highly disordered, the central region of CTCF which harbors 11 ZFs, the disorder was found to be the lowest. In another sight, disordered regions have higher temperatures as shown in **Figure 4.7.** The findings of conservation analysis show that conserved areas are functionally important so, mutations on these areas will be affected.

4.9.2. Prediction of mutation effect

Amino acids differ in charge, size or hydrophobicity and these differences can result affecting function of protein. According to location of amino acids on helix, R339W faces to DNA and its size is bigger anymore. The bulky residue will not fit the contact area and will lose connection with DNA. Histidine is neutral in this structure. By mutation, 345th amino acid becomes positively charged and it might result repulsion between other residues but more affinity to DNA. K344E mutation results extra negative charge in outer region of ZF3 and it might disturb contacts with other molecules. Also, E344 is smaller than K344 probably new residue will not fit correctly.

The most used 3 tools SIFT, Provean and MutationTaster are developed to have high accuracy, specificity, and sensitivity. The results of SIFT, Provean and MutationTaster show that 3 mutations are reported as damaging/disease causing. The mutation R339W, which is thought to have the strongest pathogenic effect, has the highest score. The K344E mutation was reported less pathogenic score than the other two mutations as experimentally demonstrated. The H345R mutation kept its pathogenic status. The parallelism between prediction tools and the experimental results formed a foreknowledge for the next studies.

4.9.3. Mutations affect stability of CTCF

3D structure of protein provides valuable information about the changes upon mutation with altered stability and flexibility of the protein. To understand how stability changes by mutations, MD simulation was done. Differences in RMSD and RMSF analyses of wild-type and mutant trajectories explain the change in stability and impact of amino acid

substitution in the protein. Before MD simulation, CUPSAT calculated protein stability upon 3D structure and it shows us CTCF was destabilizing after mutations. It was one of the motivations to start MD analysis.

RMSD of R339W was found to be in consistent with the wild-type in both replicates. K344E fluctuated more than wild-type. H345R systems represented fluctuation more than other systems, the reason ZF2 loses inter-residues communication and fluctuates highly. Without considering ZF2 fluctuation rate, H345R has similar fluctuation to 1st replicate. As result, H345R fluctuation is like wild-type. On the other hand, the results do not support CUPSAT calculations. We expected K344E to be more similar to wild-type, but it was the one more destabilized. This shows that MD analysis is more accurate on the protein structure.

ZF2 and ZF9 as free ends, have high fluctuations. On the other hand, ZF9 is mentioned for specific binding to DNA. Reason of ZF9 high fluctuation is free end but on the other hand it might be because of ZF9 does not specifically bind to DNA. Also, the flanking regions of DNA are highly fluctuated (**Figure D.3**).

In the DNA side, calculated RMSD, RMSF shows how the backbone of DNA has fluctuated. The calculation of possibility distribution of DNA backbone RMSD shows that DNA is stable with mutant systems, but according to stable condition of DNA of wild-type structure, K344E and R339W mutations have less possibility to be in same flexibility. If specific interactions are less between protein and DNA it might result more flexible DNA. The results do not promise effects of mutations but give motivation to next results.

4.9.4. Conformational differences between studied four systems

After RMSD and RMSF analysis, we wanted to see global differences between four systems. PCA analysis represents that CTCF behavior is different after mutation. It might result decreasing interaction with other molecules. It is clear that mutations affected the global dynamic of CTCF. To see what might change along CTCF, DCCM analyses were done. To comparing, DCCM plots were generated with differences from wild-type (ΔDCCM plots, **Figure 4.21**). In three systems, communication of zinc finger domains is changed. Specially ZF3 in K344E disturb ZF4 and ZF5 communication. Interestingly, ZF8 and ZF9 correlation changed between other domains. H345R mutation might, have

local affect as shown positive correlation inter-domain in ZF3 and ZF4. R339W is affected locally as ZF3 correlation is high but ZF8 correlation between ZF6 and ZF7 is also increased. The global effects are important, these results giving insight to focus on local effects.

4.9.5. Intra-molecular and inter-molecular interactions after CTCF mutations.

To have idea what might be changed in mutation sites, we used RING 2.0 algorithm which generates residue-residue network by atomic interactions based on geometrical criteria and optimized distance threshold. The results are from static structures and our result does not rely on this analysis. The important result from RING 2.0 is π–π stacking interaction between studied residues. The noncovalent interactions that aromatic rings get included such as π–π, anion–π and cation–π have constitutive roles for intermolecular relationships which are important for protein folding, drug-receptor interactions and crystal engineering [71]. After the gain of the aromatic ring with R339W mutation, new π-stacking interaction will occur. Also, π-π interaction between two zinc-coordinating histidine turns to stronger bond.

Hydrogen bonds indicate specific interaction along protein. It is important aspect for the dynamic of protein. Hydrogen bonds were calculated after MD analysis between protein-protein and protein-DNA. Results showed that the mutations increased the number of hydrogen bonds in the CTCF. This increase might result in more stable. On the other hand, the new hydrogen bonds that make undesirable interactions may cause structural changes in protein. Hydrogen bond numbers are decreasing between DNA-protein in mutant proteins. By losing the specific interactions between protein-DNA might result in more flexible DNA to move. According to this result, it is observed that mutations negatively affect the interaction between CTCF and CTSs. The interactions which do not exist in wild-type are likely to affect protein folding and interaction with other proteins.

By distance analysis, we got for the interaction of mutated residues. According to the results, we see that in R339W mutation, the ZF3 moves away from the DNA, and it does not bind to the DNA specifically, as shown in the number of contact analyses. While it is an important DNA binding site for ZF3, it can be said that the loss of interaction of this point disturbs CTCF function.

H345R moves closer to DNA. The reason might be positively charge of R345 get

influenced by negatively charged DNA. Binding to DNA with high affinity may cause it not to break away when necessary. While lysine is positively charged amino acids and friendly with DNA, turns to negatively charged glutamate amino acids. It changes stability of K344E system, and we see it moves away from DNA in some regions also keeps itself in same distance as wild-type.

4.9.6. Zinc ion coordination

Zinc ion is a factor for function and stability on ZF domains in CTCF. Losing a metal ion after mutation might disturb the function of zinc finger. According to HOPE server's prediction analysis based on 3D structure, all three mutations might result loss zinc ion. R339W site is a neighbor of H340 zinc-ion-coordinating residue, but the mutation did not affect zinc ion unlike HOPE server result. Also HOPE server mentioned K344 as a zinc ion coordinating residue but it is not, so we do not see zinc ion is affected by K344E mutation or cause of being neighbor of H345 zinc-ion-coordinating residue.

After the prediction of HOPE, MD confirmed that zinc ion is lost in ZF3 of H345R system. Zn ion is important for recognition the DNA and losing the metal ion from zinc finger will disturb the interaction with DNA.

CHAPTER 5

5. CONCLUSIONS AND FUTURE WORK

A common notion in structural bioinformatics is that mutation of conserved amino acids is expected to have a higher effect on protein structure and stability due to keeping known. According to this notion, the sequence conservation is due to a greater functional relevance of wild-type amino acids for the function of protein. Departing from this paradigm, CTCF (in general) and the three studied mutations (specifically) were assessed from the primary sequence conservation perspective. If the amino acid is not conserved but its vicinity (the region) was conserved, the functional consequence might be also negatively. We showed that the central region in the protein, which also covers the three selected mutation sites, is highly conserved. Furthermore, the central region of CTCF is ordered but the flanking regions (towards the C and N termini) are disordered and correspond to higher B-factors. Concordantly, we observed similar overall pattern in RMSF. While helixes in central region of the investigated protein structure are less flexible, flanking regions and loop regions have higher fluctuation. Intriguingly, all three residue sites, corresponding to the studied mutations, were found to be conserved, hinting for the possible pathogenicity of single amino acid substitution at that position.

According to literature, ZF3 provides the orientation of reading DNA sequence for CTCF [23]. The changes in ZF3 and its binding for CTSs might disturb other zinc finger bindings as well.

Based on the findings obtained from this study, R339W mutations appear to disturb the DNA binding because this mutation is the DNA contact site on ZF3. Relatedly, the aromatic ring of W339 might be stabilized with DNA sugar backbone. This could be due to the stabilization of protein structure upon distancing of R339W from DNA and loss of its DNA specific interaction, CTCF is still stable. The new residue with aromatic ring is bulkier than the wild-type residue and could stabilize itself with non-covalent interaction.

While R339W loses interaction with DNA, specific interactions occur within CTCF that might facilitate the stabler form.

K344E mutation introduces negative charge in ZF3. The effect of negative charge was shown to destabilized protein and negatively affect special correlation within ZF3 and in the neighboring ZF4 (based on DCCM analysis). In addition, due to the fact that the mutation site is on the outer side of the helix, it may affect interactions with other proteins or within protein rather than altering its interaction with DNA. Because when its interaction with DNA was examined, it does not move away from DNA and the hydrogen bonds between it and DNA do not seem to substantially decrease. In this case, its effect might be related with other proteins, namely CTCF interaction partners.

The obtained results suggest that H345R mutation can disturb function of CTCF. The positive charge forming after H345R mutation, could be the affinity of CTCF to DNA. In contrast, the new charge of the studied R339 residue most likely caused repulsion of zinc ion from ZF3.

The study by Yin et al. [23] mentioned specific binding of ZF9, but another structure study [24] did not report specific interaction for this domain with DNA. According to the first mentioned study in [24], of thesis finding, RMSF graph shows that ZF9, in the end part of protein structure, shows high fluctuation It seems more likely that there is no specific interaction on this zinc finger. It has also been reported that ZF2 does not interact with DNA specifically. High fluctuation in RMSF can be seen because it is both free end in structure and does not interact with any DNA, RNA, or other protein during trajectory. On the other hand, ZF2 and ZF9 are disordered regions in CTCF.

CTCF should bind to DNA with high affinity. Cancer cells can activate oncogenes if the interaction between CTCF and DNA decreases and DNA becomes more accessible.
The overall results of the thesis are summarized in **Table 5.1** and **Table 5.2**.

Table 5.1: General perspective to obtained result.

	Experimental Result	**Conservation and order**	**AA Size**	**Charge**	**Prediction information from databases**
R339W	No binding	Conserved and ordered	Increased	Positive to neutral	Damaging
K344E	Partially binding	Conserved and ordered	Decreased	Positive to negative	Damaging

H345R	No binding	Conserved and ordered	Increased	Neutral to positive	Damaging

Table 5.2: General perspective to obtained result (continues).

	RMSD	**PCA**	**DCCM**	**Distance to DNA**	**Zinc ion loss**
R339W	Stable	Narrower span	Observed local effect	Increased	Kept
K344E	Destable	Narrower span	Observed local effect	Same	Kept
H345R	Stable	Narrower span	Observed local effect	Decreased	Loss

Currently, the CTCF protein is widely regarded as a candidate tumor suppressor protein. For instance, the related record in the UniProt database, which is a primary source for protein research, states accordingly that CTCF “Seems to act as tumor suppressor.” (https://www.uniprot.org/uniprot/P49711). This thesis research aimed at providing novel insides into this tumor suppressor candidacy using by means of bioinformatics analysis of CTCF structure-function relationship. The results of this work protein provide new insights into the structural biology of CTCF in the context of cancer. By highlighting the mechanisms of mutations in the binding site and the zinc-finger region of protein shows us what changes after mutations how might lead to cancer states. Phosphorylation plays an important role in the regulation of CTCF modulation. In future studies, it is possible to search WT and phosphorylated-WT by MD simulations. Also, methylation can be considered for how distance would be change in WT ZF3 and methylated CTSs. How interactions would be effected can be searched by MD simulations.

Future research on CTCF focusing on its implication in cancer is expected to result in deeper understanding of tumorigenesis and can eventually help in combating cancer development. The related translational approaches can, for example, facilitate design of novel therapeutics. Thus, more efforts in this research direction are required.

BIBLIOGRAPHY

[1] L. Braccioli and E. De Wit, “CTCF: A Swiss-army knife for genome organization and transcription regulation,” *Essays Biochem.*, vol. 63, no. 1, pp. 157–165, 2019, doi: 10.1042/EBC20180069.

[2] V. V. Lobanenkov *et al.*, “A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5’-flanking sequence of the chicken c-myc gene,” *Oncogene*, vol. 5, no. 12, pp. 1743–1753, Dec. 1990, Accessed: Apr. 30, 2021. [Online]. Available: https://europepmc.org/article/med/2284094.

[3] R. Ohlsson, V. Lobanenkov, and E. Klenova, “Does CTCF mediate between nuclear organization and gene expression?,” *BioEssays*, vol. 32, no. 1. Bioessays, pp. 37–50, 2010, doi: 10.1002/bies.200900118.

[4] J. M. Moore *et al.*, “Loss of maternal CTCF is associated with peri-implantation lethality of CtCf null embryos,” *PLoS One*, vol. 7, no. 4, 2012, doi: 10.1371/journal.pone.0034915.

[5] C. G. Bailey *et al.*, “CTCF expression is essential for somatic cell viability and protection against cancer,” *Int. J. Mol. Sci.*, vol. 19, no. 12, 2018, doi: 10.3390/ijms19123832.

[6] V. V. A. Steven M. Singer, Marc Y. Fink, “A CTCF Code for 3D Genome Architecture,” *Physiol. Behav.*, vol. 176, no. 3, pp. 139–148, 2019, doi: 10.1016/j.cell.2015.07.053.A.

[7] J. E. Phillips and V. G. Corces, “CTCF: Master Weaver of the Genome,” *Cell*, vol. 137, no. 7. pp. 1194–1211, 2009, doi: 10.1016/j.cell.2009.06.001.

[8] C. Bailey *et al.*, “CTCF Expression is Essential for Somatic Cell Viability and Protection Against Cancer,” *Int. J. Mol. Sci.*, vol. 19, no. 12, p. 3832, Nov. 2018, doi: 10.3390/ijms19123832.

[9] C. T. Ong and V. G. Corces, “CTCF: An architectural protein bridging genome topology and function,” *Nat. Rev. Genet.*, vol. 15, no. 4, pp. 234–246, 2014, doi: 10.1038/nrg3663.

[10] V. C. JE Phillips, “CTCF: Master Weaver of the Genome,” *Cell*, vol. 23, no. 1, pp. 1–7, 2009, doi: 10.1038/jid.2014.371.

[11] E. P. Nora *et al.*, “Spatial partitioning of the regulatory landscape of the X-inactivation centre,” *Nature*, vol. 485, no. 7398, pp. 381–385, 2012, doi: 10.1038/nature11049.

[12] R. Ghirlando and G. Felsenfeld, “CTCF: making the right connections,” *Genes Dev.*, vol. 30, no. 8, pp. 881–891, Apr. 2016, doi: 10.1101/gad.277863.116.

[13] J. R. Dixon *et al.*, “Topological domains in mammalian genomes identified by analysis of chromatin interactions,” *Nature*, vol. 485, no. 7398, pp. 376–380, 2012, doi: 10.1038/nature11082.

[14] K. S. Wendt *et al.*, “Cohesin mediates transcriptional insulation by CCCTC-binding factor,” *Nature*, vol. 451, no. 7180, pp. 796–801, 2008, doi: 10.1038/nature06634.

[15] J. Zuin *et al.*, “Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 3, pp. 996–1001, 2014, doi: 10.1073/pnas.1317788111.

[16] S. J. Bastiaan Holwerda and W. de Laat, “CTCF: The protein, the binding partners, the binding sites and their chromatin loops,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1620. 2013, doi: 10.1098/rstb.2012.0369.

[17] K. Jabbari, P. Heger, R. Sharma, and T. Wiehe, “The diverging routes of BORIS and CTCF: An interactomic and phylogenomic analysis,” *Life*, vol. 8, no. 1, 2018, doi: 10.3390/life8010004.

[18] P. Heger, B. Marin, M. Bartkuhn, E. Schierenberg, and T. Wiehe, “The chromatin insulator CTCF and the emergence of metazoan diversity,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 43, pp. 17507–17512, Oct. 2012, doi: 10.1073/pnas.1111941109.

[19] E. Kentepozidou *et al.*, “Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains,” *Genome Biol.*, vol. 21, no. 1, pp. 1–19, Jan. 2020, doi: 10.1186/s13059-019-1894-x.

[20] L. J. T. Kaaij, F. Mohn, R. H. van der Weide, E. de Wit, and M. Bühler, “The ChAHP Complex Counteracts Chromatin Looping at CTCF Sites that Emerged from SINE Expansions in Mouse,” *Cell*, vol. 178, no. 6, pp. 1437-1451.e14, Sep. 2019, doi: 10.1016/j.cell.2019.08.007.

[21] O. G. Maksimenko, D. V. Fursenko, E. V. Belova, and P. G. Georgiev, “CTCF As an Example of DNA-Binding Transcription Factors Containing Clusters of C2H2-Type Zinc Fingers,” *Acta Naturae*, vol. 13, no. 1, pp. 31–46, 2021, doi: 10.32607/actanaturae.11206.

[22] M. Renda *et al.*, “Critical DNA binding interactions of the insulator protein CTCF: A small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci,” *J. Biol. Chem.*, vol. 282, no. 46, pp. 33336–33345, 2007, doi: 10.1074/jbc.M706213200.

[23] M. Yin *et al.*, “Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites,” *Cell Res.*, vol. 27, no. 11, pp. 1365–1377, 2017, doi: 10.1038/cr.2017.131.

[24] H. Hashimoto, D. Wang, J. R. Horton, X. Zhang, V. G. Corces, and X. Cheng, “Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA,” *Mol. Cell*, vol. 66, no. 5, pp. 711-720.e3, 2017, doi: 10.1016/j.molcel.2017.05.004.

[25] M. J. MacPherson and P. D. Sadowski, “The CTCF insulator protein forms an unusual DNA structure,” *BMC Mol. Biol.*, vol. 11, 2010, doi: 10.1186/1471-2199-11-101.

[26] G. Malgieri, M. Palmieri, L. Russo, R. Fattorusso, P. V. Pedone, and C. Isernia, “The prokaryotic zinc-finger: Structure, function and comparison with the eukaryotic counterpart,” *FEBS Journal*, vol. 282, no. 23. Blackwell Publishing Ltd, pp. 4480–4496, Dec. 01, 2015, doi: 10.1111/febs.13503.

[27] A. V. Persikov, R. Osada, and M. Singh, “Predicting DNA recognition by Cys2His2 zinc finger proteins,” *Bioinformatics*, vol. 25, no. 1, pp. 22–29, 2009, doi: 10.1093/bioinformatics/btn580.

[28] G. N. Filippova *et al.*, “Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity,” *Cancer Res.*, vol. 62, no. 1, pp. 48–52, 2002, Accessed: Aug. 19, 2021. [Online]. Available: https://cancerres.aacrjournals.org/content/62/1/48.short.

[29] G. N. Filippova *et al.*, “A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22. 1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers.,” *Genes, Chromosom. Cancer*, vol. 36, no. April 1997, pp. 26–36, 1998.

[30] J. C. Tiffen *et al.*, “The cancer-testis antigen BORIS phenocopies the tumor suppressor CTCF in normal and neoplastic cells,” *Int. J. Cancer*, vol. 133, no. 7, pp. 1603–1613, 2013, doi: 10.1002/ijc.28184.

[31] A. D. Marshall, C. G. Bailey, and J. E. J. Rasko, “CTCF and BORIS in genome regulation and cancer,” *Curr. Opin. Genet. Dev.*, vol. 24, no. 1, pp. 8–15, 2014, doi: 10.1016/j.gde.2013.10.011.

[32] A. D. Marshal *et al.*, “CTCF genetic alterations in endometrial carcinoma are pro-tumorigenic,” *Oncogene*, vol. 36, no. 29, pp. 4100–4110, 2017, doi: 10.1038/onc.2017.25.

[33] K. Yoshida *et al.*, “The landscape of somatic mutations in Down syndrome-related myeloid disorders,” *Nat. Genet.*, vol. 45, no. 11, pp. 1293–1301, 2013, doi: 10.1038/ng.2759.

[34] E. K. Rolf Ohlsson, Victor Lobanenkov, “Does CTCF mediate between nuclear organization and gene expression?,” *Physiol. Behav.*, vol. 176, no. 1, pp. 139–148, 2016, doi: 10.1002/bies.200900118.

[35] K. T. Chang, J. Guo, A. Di Ronza, and M. Sardiello, “Aminode: Identification of Evolutionary Constraints in the Human Proteome,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–8, Jan. 2018, doi: 10.1038/s41598-018-19744-w.

[36] J. S. Papadopoulos and R. Agarwala, “COBALT: Constraint-based alignment tool for multiple protein sequences,” *Bioinformatics*, vol. 23, no. 9, pp. 1073–1079, May 2007, doi: 10.1093/bioinformatics/btm076.

[37] M. E. Oates *et al.*, “D2P2: Database of disordered protein predictions,” *Nucleic Acids Res.*, vol. 41, no. D1, 2013, doi: 10.1093/nar/gks1226.

[38] B. Mészáros, G. Erdös, and Z. Dosztányi, “IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding,” *Nucleic Acids Res.*, vol. 46, no. W1, pp. W329–W337, 2018, doi: 10.1093/nar/gky384.

[39] R. M. Aydlnkal, O. Serçinoğlu, and P. Ozbek, “ProSNEx: a web-based application for exploration and analysis of protein structures using network formalism,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W471–W476, 2019, doi:

10.1093/nar/gkz390.

[40] Q. Liu, Z. Li, and J. Li, “Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts,” *BMC Bioinformatics*, vol. 15, no. 16, pp. 1–11, Dec. 2014, doi: 10.1186/1471-2105-15-S16-S3.

[41] P. C. Ng and S. Henikoff, “SIFT: predicting amino acid changes that affect protein function,” doi: 10.1093/nar/gkg509.

[42] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, “Predicting the Functional Effect of Amino Acid Substitutions and Indels,” *PLoS One*, vol. 7, no. 10, Oct. 2012, doi: 10.1371/journal.pone.0046688.

[43] P. V, G. MM, and S. D, “CUPSAT: prediction of protein stability upon point mutations,” *Nucleic Acids Res.*, vol. 34, no. Web Server issue, Jul. 2006, doi: 10.1093/NAR/GKL190.

[44] W. Yan, J. Zhou, M. Sun, J. Chen, G. Hu, and B. Shen, “The construction of an amino acid network for understanding protein structure and function,” *Amino Acids*, vol. 46, no. 6. pp. 1419–1439, 2014, doi: 10.1007/s00726-014-1710-6.

[45] D. Piovesan, G. Minervini, and S. C. E. Tosatto, “The RING 2.0 web server for high quality residue interaction networks,” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W367–W374, 2016, doi: 10.1093/nar/gkw315.

[46] S. A. Hollingsworth and R. O. Dror, “Molecular Dynamics Simulation for All,” *Neuron*, vol. 99, no. 6. pp. 1129–1143, 2018, doi: 10.1016/j.neuron.2018.08.011.

[47] J. Lee *et al.*, “CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field,” *J. Chem. Theory Comput.*, vol. 12, no. 1, pp. 405–413, Jan. 2016, doi: 10.1021/acs.jctc.5b00935.

[48] J. Huang *et al.*, “CHARMM36m: An improved force field for folded and intrinsically disordered proteins,” *Nat. Methods*, vol. 14, no. 1, pp. 71–73, 2016, doi: 10.1038/nmeth.4067.

[49] L. Martínez, “Automatic Identification of Mobile and Rigid Substructures in Molecular Dynamics Simulations and Fractional Structural Fluctuation Analysis,” 2015, doi: 10.1371/journal.pone.0119264.

[50] O. Carugo and S. Pongor, “A normalized root-mean-spuare distance for comparing protein three-dimensional structures,” *Protein Sci.*, vol. 10, no. 7, pp. 1470–1473, Jul. 2001, doi: 10.1110/PS.690101.

[51] X. Zhang, Y. Sun, Z. Zhang, H. Wang, J. Wang, and D. Zhao, “Intermolecular insights into allosteric inhibition of histone lysine-specific demethylase 1,” *Biochim. Biophys. Acta - Gen. Subj.*, vol. 1865, no. 11, p. 129990, Nov. 2021, doi: 10.1016/j.bbagen.2021.129990.

[52] D. M. F. Van Aalten, D. A. Conn, B. L. De Groot, H. J. C. Berendsen, J. B. C. Findlay, and A. Amadei, “Protein dynamics derived from clusters of crystal structures,” *Biophys. J.*, vol. 73, no. 6, pp. 2891–2896, 1997, doi: 10.1016/S0006-3495(97)78317-6.

[53] S. S. Narang, S. Shuaib, D. Goyal, B. Goyal,] Simranjeet, and S. Narang, "Assessing the effect of D59P mutation in the DE loop region in amyloid aggregation propensity of β2-microglobulin: A molecular dynamics simulation study † Running head: Effect of D59P mutation in β2m aggregation," doi: 10.1002/jcb.26241.

[54] H. Wan, J. P. Hu, X. H. Tian, and S. Chang, "Molecular dynamics simulations of wild type and mutants of human complement receptor 2 complexed with C3d," *Phys. Chem. Chem. Phys.*, vol. 15, no. 4, pp. 1241–1251, 2013, doi: 10.1039/c2cp41388d.

[55] G. E. Arnold and R. L. Ornstein, "Molecular dynamics study of time-correlated protein domain motions and molecular flexibility: Cytochrome P450BM-3," *Biophys. J.*, vol. 73, no. 3, pp. 1147–1159, 1997, doi: 10.1016/S0006-3495(97)78147-5.

[56] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1. Nucleic Acids Res, pp. 235–242, Jan. 01, 2000, doi: 10.1093/nar/28.1.235.

[57] N. SB and W. CD, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, Mar. 1970, doi: 10.1016/0022-2836(70)90057-4.

[58] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow, "Mutationtaster2: Mutation prediction for the deep-sequencing age," *Nature Methods*, vol. 11, no. 4. pp. 361–362, 2014, doi: 10.1038/nmeth.2890.

[59] H. Venselaar, T. A. te Beek, R. K. Kuipers, M. L. Hekkelman, and G. Vriend, "Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces," *BMC Bioinforma. 2010 111*, vol. 11, no. 1, pp. 1–10, Nov. 2010, doi: 10.1186/1471-2105-11-548.

[60] P. Shannon *et al.*, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, 2003, doi: 10.1101/gr.1239303.

[61] S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM," *J. Comput. Chem.*, vol. 29, no. 11, pp. 1859–1865, Aug. 2008, doi: 10.1002/jcc.20945.

[62] B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[63] M. J. Abraham *et al.*, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1–2, pp. 19–25, 2015, doi: 10.1016/j.softx.2015.06.001.

[64] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–38, 1996, doi: 10.1016/0263-7855(96)00018-5.

[65] E. F. Pettersen *et al.*, "UCSF Chimera - A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct.

2004, doi: 10.1002/jcc.20084.

[66] J. G. Tate *et al.*, “COSMIC: the Catalogue Of Somatic Mutations In Cancer,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D941–D947, Jan. 2019, doi: 10.1093/NAR/GKY1015.

[67] S. T. Sherry *et al.*, “dbSNP: the NCBI database of genetic variation,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001, doi: 10.1093/NAR/29.1.308.

[68] M. M. Gromiha, “Prediction of protein stability upon point mutations,” in *Biochemical Society Transactions*, 2007, vol. 35, no. 6, pp. 1569–1573, doi: 10.1042/BST0351569.

[69] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. D. Caves, “Bio3d: An R package for the comparative analysis of protein structures,” *Bioinformatics*, vol. 22, no. 21, pp. 2695–2696, 2006, doi: 10.1093/bioinformatics/btl461.

[70] C. R. Martinez and B. L. Iverson, “Rethinking the term ‘pi-stacking,’” *Chemical Science*, vol. 3, no. 7. pp. 2191–2201, 2012, doi: 10.1039/c2sc20045g.

[71] X. Lucas, A. Bauzá, A. Frontera, and D. Quiñonero, “A thorough anion-π interaction study in biomolecules: On the importance of cooperativity effects,” *Chem. Sci.*, vol. 7, no. 2, pp. 1038–1050, 2016, doi: 10.1039/c5sc01386k.

APPENDIX A

Summary of the Literature on the Candidate Tumor Suppressor Gene CTCF

Table A.1: Quantitative Summary of the Literature on the Candidate Tumor Suppressor Gene CTCF.

Years	Number of publications per year	Cumulative number of publications
1993	1	1
1996	1	2
1998	2	4
1999	2	6
2000	4	10
2001	4	14
2002	5	19
2003	5	24
2004	3	27
2005	9	36
2006	8	44
2007	12	56
2008	9	65
2009	4	69
2010	12	81
2011	12	93
2012	16	109
2013	5	114
2014	7	121
2015	22	143
2016	31	174
2017	16	190
2018	3	193
2019	24	217

2020	19	236

Table A.2: Information about cancer-related CTCF variants obtained using literature and database searches. The mutations investigated in this study are highlighted with red stars.

Variants	COSMIC	dbSNP	Cancer type	Zinc-Finger
R29W	COSM6845191	rs771567232	Endometrioid carcinoma	-
R278C	COSM2995136	rs1266478000	Colon Cancer	ZF1
R339W*	COSM6908613	rs1243010179	Leukaemia, Wilms tumor	ZF3
R342C	COSM194431	rs1131691283	Leukaemia	ZF3
R342H	COSM1716423	-	Endometrioid carcinoma	ZF3
K344E*	COSM6908612	rs1215280530	Breast cancer	ZF3
H345R*	COSM280199	-	Prostate, Colon Cancer	ZF3
K352N	COSM1716419	-	Endometrioid carcinoma	ZF4
K365T	COSM972486	-	Endometrioid carcinoma	ZF4
H373L	COSM1716409	-	Endometrioid carcinoma	ZF4
R377C	COSM972487	-	Endometrioid endometrial carcinomas	on loop between ZF4-ZF5
R377H	COSM303879	rs968244943	Endometrioid endometrial carcinomas	on loop between ZF4-ZF5
P378L	COSM972488	rs1167223790	Endometrioid carcinoma	ZF5
F416L	COSM1716410	-	Endometrioid carcinoma	ZF6
R448Q	COSM6908614	-	Acute lymphoblastic leukaemia	on loop between ZF7-ZF8
A694V	COSM1716416	-	Endometrioid carcinoma	-
L722I	COSM1716417	-	Endometrioid carcinoma	-

APPENDIX B

Alignment of used CTCF-related PDB files

```
5UND              1 ------------------------------------------------      0
5T0U              1 PLGSTHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCPDCDMAFVTSGELV     50
5UND              1 -------GSEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCSYAS     43
                          ..||||||||||||||||||||||||||||||||||||||||||
5T0U             51 RHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCSYAS    100
5UND             44 RDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFH     93
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
5T0U            101 RDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFH    150
5UND             94 CPHCDTVIARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKS    143
                    |||||||||||||||||||||||||||
5T0U            151 CPHCDTVIARKSDLGVHLRKQHSYIEQ----------------------    177
5UND            144 HKNEKRFKCDQCDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLL    193
5T0U            178 ------------------------------------------------    177
5UND            194 DMHFKRYHD    202
5T0U            178 ---------    177
```

Figure B.1: Alignment of amino acid sequences of the 5UND and 5T0U PDB structures. Blue line and red line denote the first AA the last AA in each PDB file, respectively. The green box demarcates the span of ZF3.

DNA-chain B

```
5UND              1 ---------CGCCCCCTGCTGGCCTCTGTGGGCACTG     28
                            |||||||||||||||
5T0U              1 CCTCACTAGCGCCCCCTGCTGGC--------------     23
```

Figure B.2: Alignment of nucleic acid sequences of the chain B of the 5UND and 5T0U PDB structures. Blue line and red line denote the first base and the last base in each PDB file, respectively.

DNA-chain C

```
5und              1 CAGTGCCCACAGAGGCCAGCAGGGGGCG---------     28
                                ||||||||||||||
5t0u              1 -------------GCCAGCAGGGGGCGCTAGTGAGG     23
```

Figure B.3: Alignment of nucleic acid sequences of the chain C of the 5UND and 5T0U PDB structures. Blue line and red line denote the first base the last base in each PDB files, respectively.

APPENDIX C

Charge of residues in the obtained combined structure

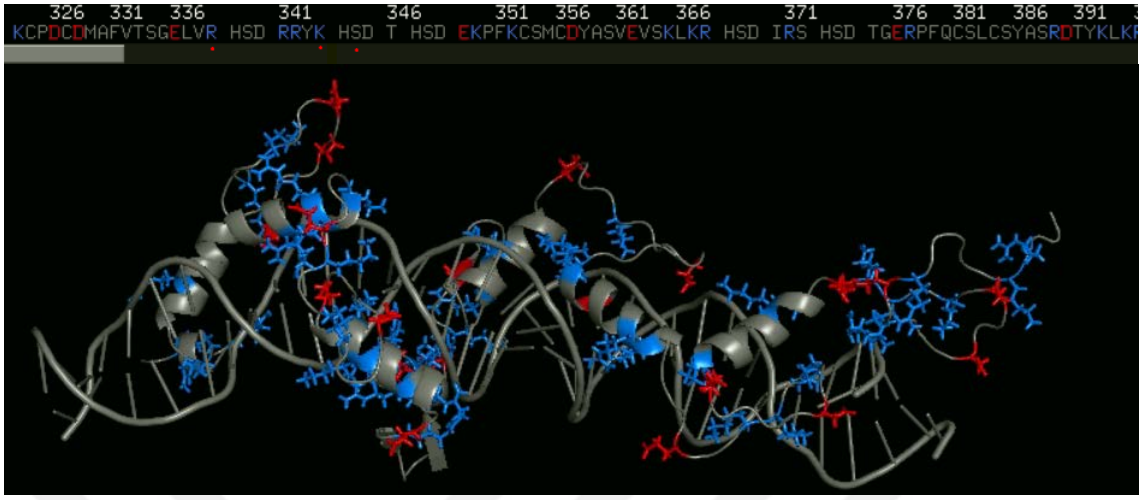
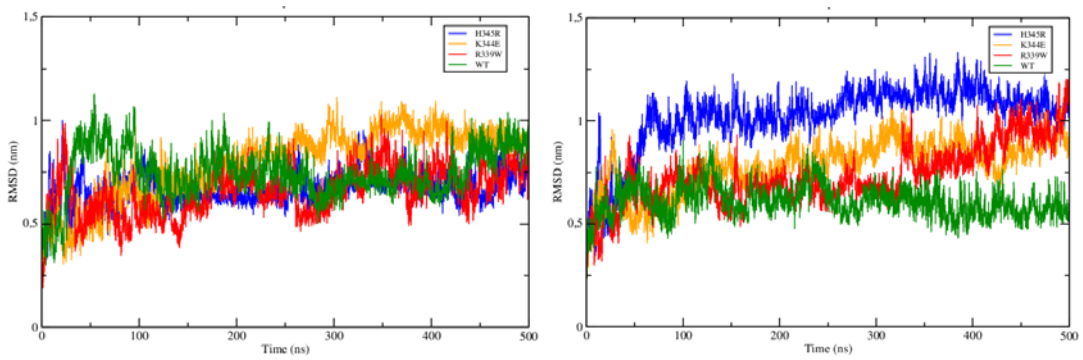


Figure C.1: Visualization of charged residues on the combined WT structure, obtained after aligning 5UND and 5T0U PDB files. The amino acid sequence is shown above and the structure (which also contains the corresponding DNA part) is shown below. Figure was generated using pyMOL. Positively charged residues and negatively charged residues are shown using blue licorice and red licorice representation, respectively, whereas the neutral residues are shown in grey. The same color applies for the amino acid sequence. Thereat, the three mutation positions, which were the focus of this study, are denoted with red star: The arginine (R) and the lysine (K) are positively charged, whereas the histidine, which is herein encoded as HSD, is its neutral form.

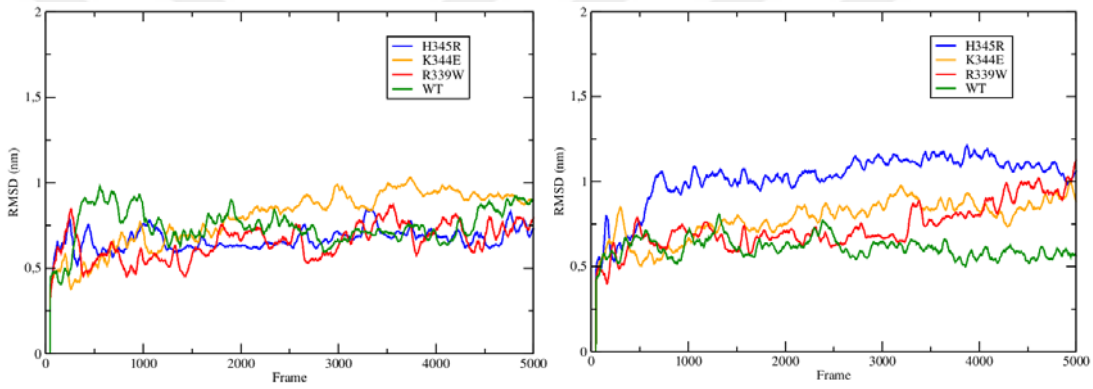
APPENDIX D

Trajectory-based analysis of CTCF dynamics



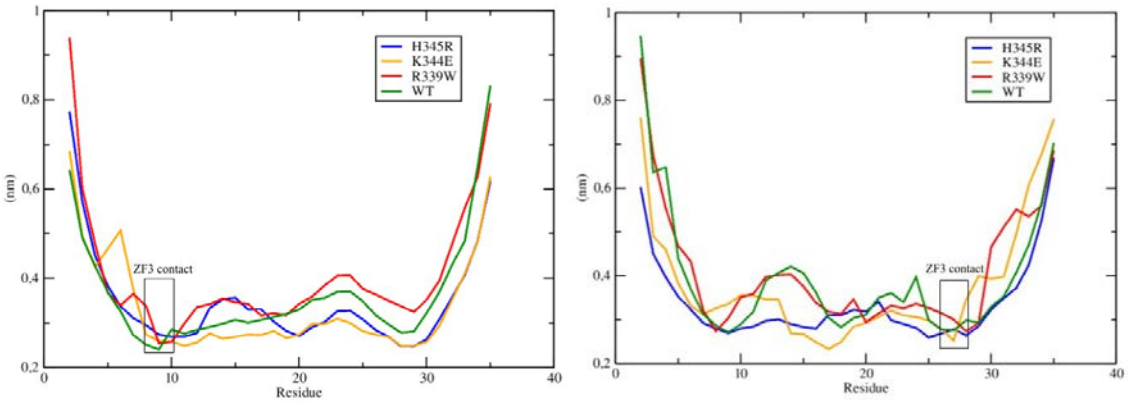
(a) **(b)**

Figure D.1: RMSD of WT and the three mutants during 500 ns for the 1st replicate **a)** and the 2nd **b)** replicate.



(a) **(b)**

Figure D.2: Moving average of RMSD of WT and the three mutants during 500 ns for the 1st replicate **a)** and the 2nd **b)** replicate. The window of 50 frames was applied.



(a) **(b)**

Figure D.3: RMSF of P atoms of DNA of the 5' strand **a)** and the 3' strand **b)**. RMSF is shown for the WT and the three mutant structures. The ZF3 contact region is highlighted using a box, which has black border.

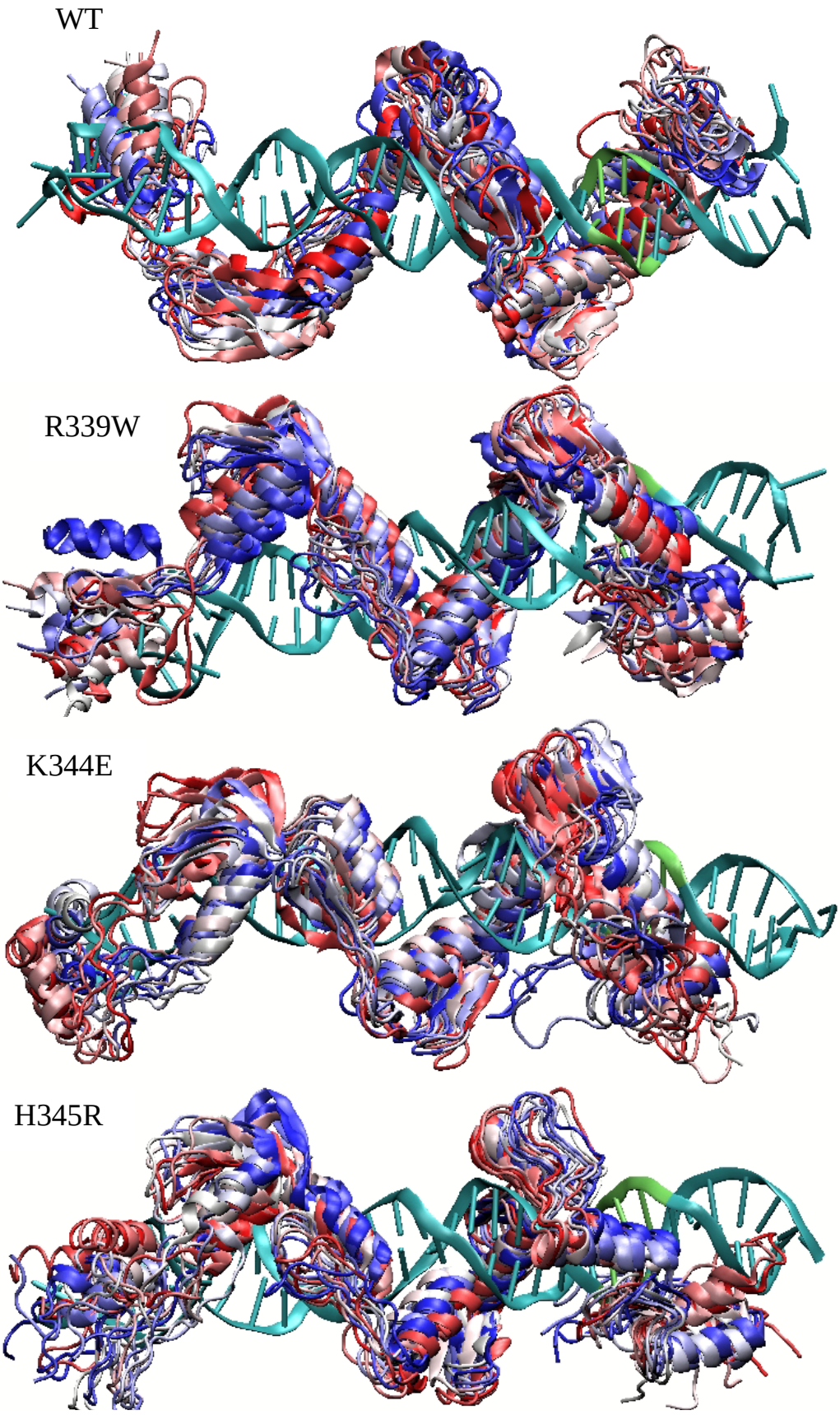


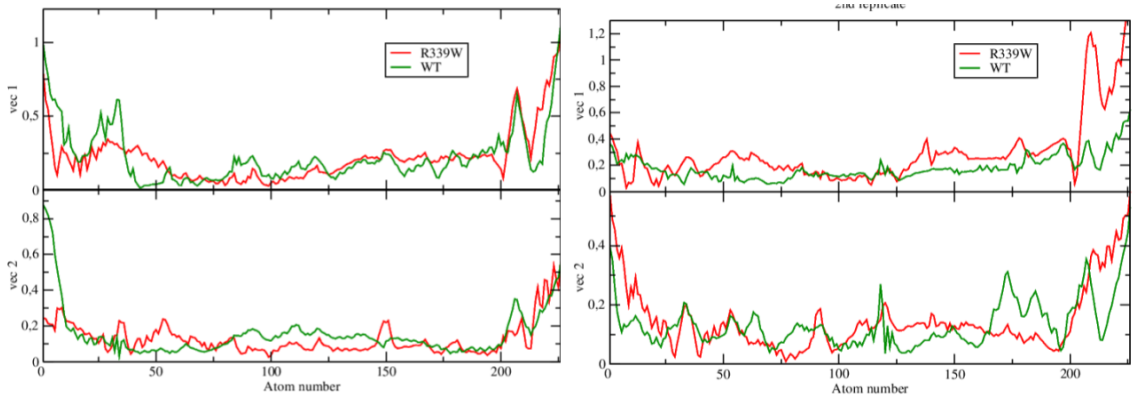
Figure D.4: Time-wise fluctuation of WT and the three mutant structures during 500 ns of the 1st replicate. The initial confirmation is shown in red color and the subsequent fluctuations are illustrated in blue color. For all structures, the helix structure of ZF9, corresponding to its initial (WT) state, could not be shown (as its actual conformation) due to a possible technical issue in VMD.

APPENDIX E

PCA Analysis

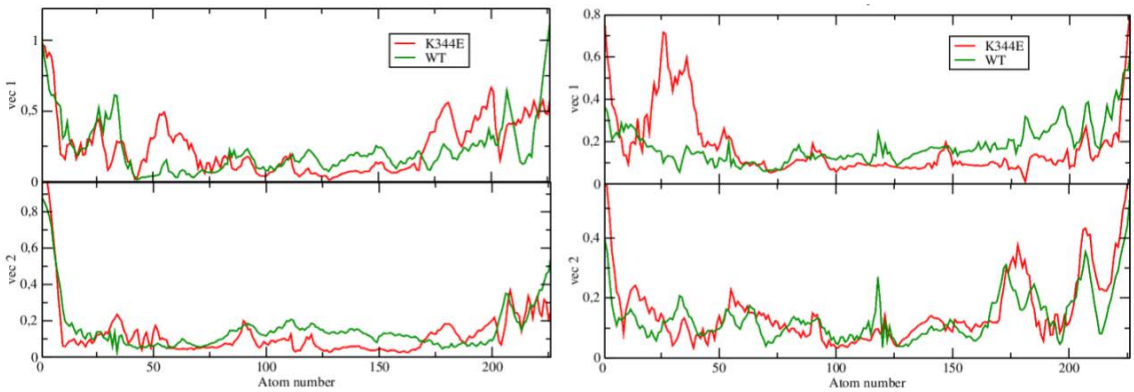
Table E.1: Cumulative variation explained by eigenvectors of PCA.

Eigenvectors	1st replicate	2nd replicate
WT		
1	35.658	26.672
2	53.076	42.544
3	66.030	56.452
4	72.235	64.779
5	77.067	69.936
R339W		
1	40.751	54.181
2	53.396	66.649
3	63.798	75.326
4	70.638	81.149
5	75.620	84.850
K344E		
1	40.127	31.295
2	56.473	52.691
3	69.009	62.775
4	75.031	68.725
5	79.236	73.624
H345R		
1	27.553	39.659
2	48.392	52.875
3	62.453	63.194
4	69.426	72.135
5	74.702	76.889



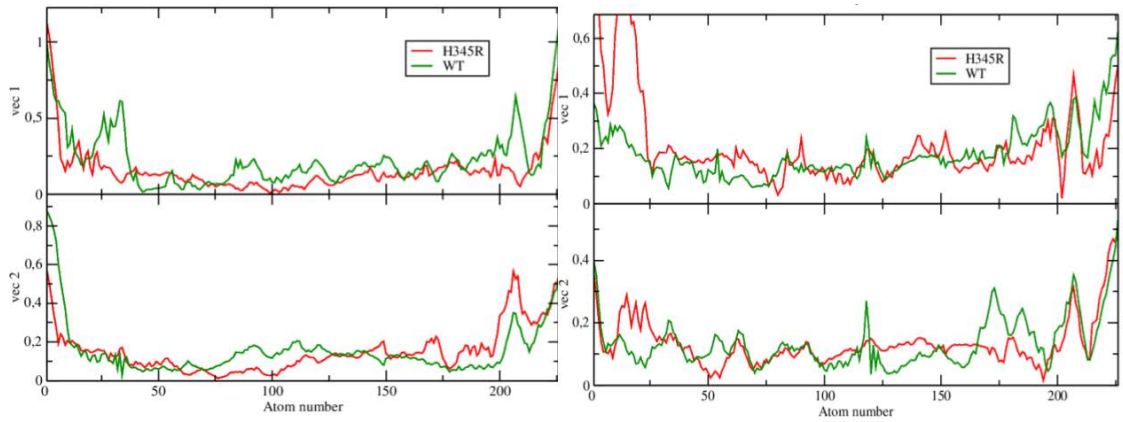
(a) 1st replicate **(b)** 2nd replicate

Figure E.1: RMSF (nm) of R339W and WT, based on the first and the second eigenvectors of PCA.



(a) 1st replicate **(b)** 2nd replicate

Figure E.2: RMSF (nm) K344E and WT, based on the first and the second eigenvectors of PCA.

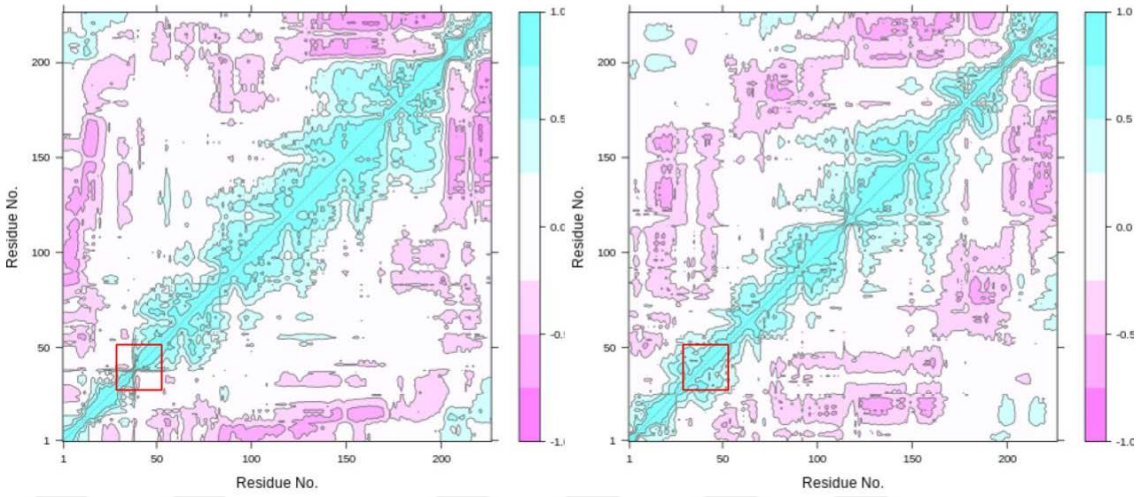


(a) 1st replicate **(b)** 2nd replicate

Figure E.3: RMSF (nm) of H345R and WT, based on the first and the second eigenvectors of PCA.

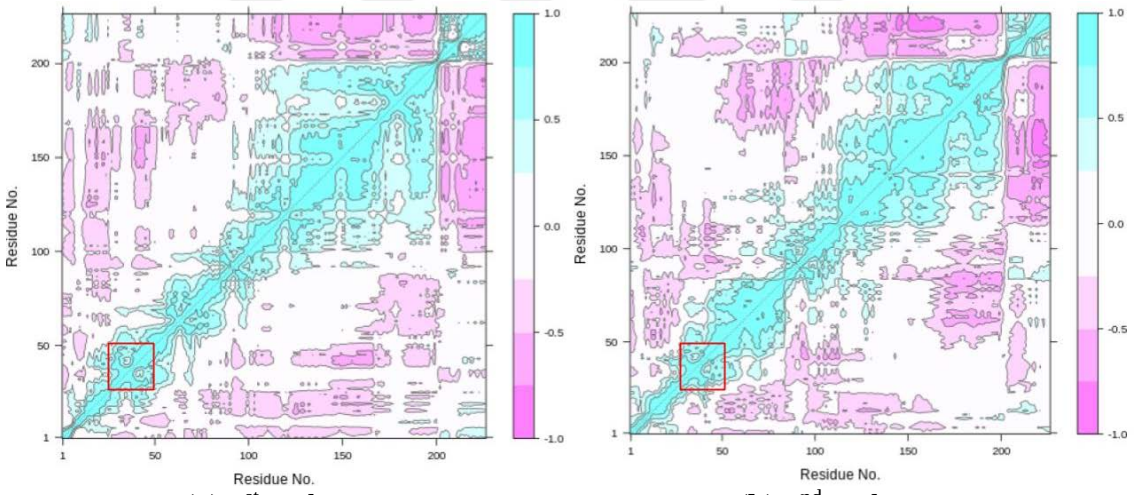
APPENDIX F

DCCM Analysis



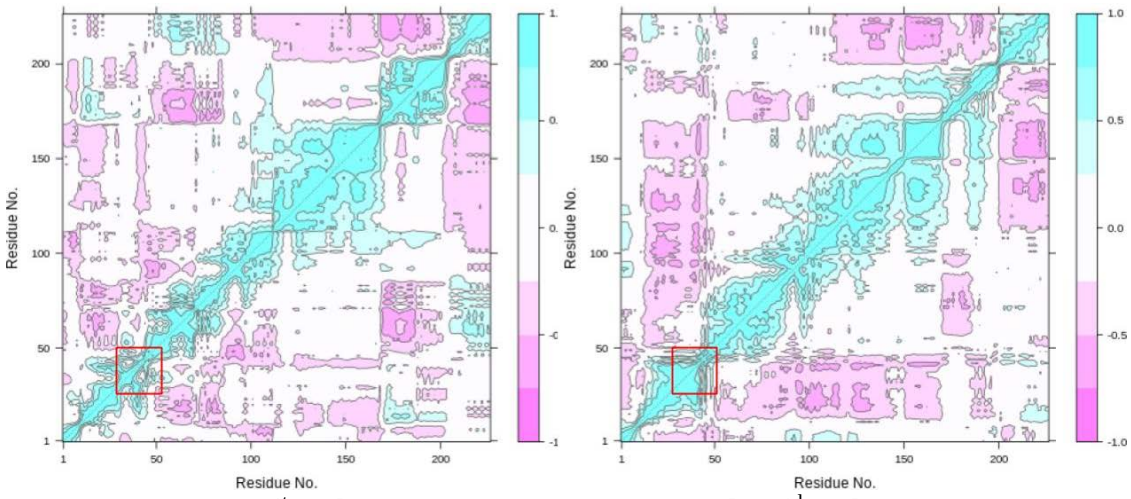
(a) 1st replicate **(b)** 2nd replicate

Figure F.1: DCCM plot of WT. The color bar gradient shows positive and negative correlation using blue color and pink color, respectively. White color corresponds to the absence of correlation.



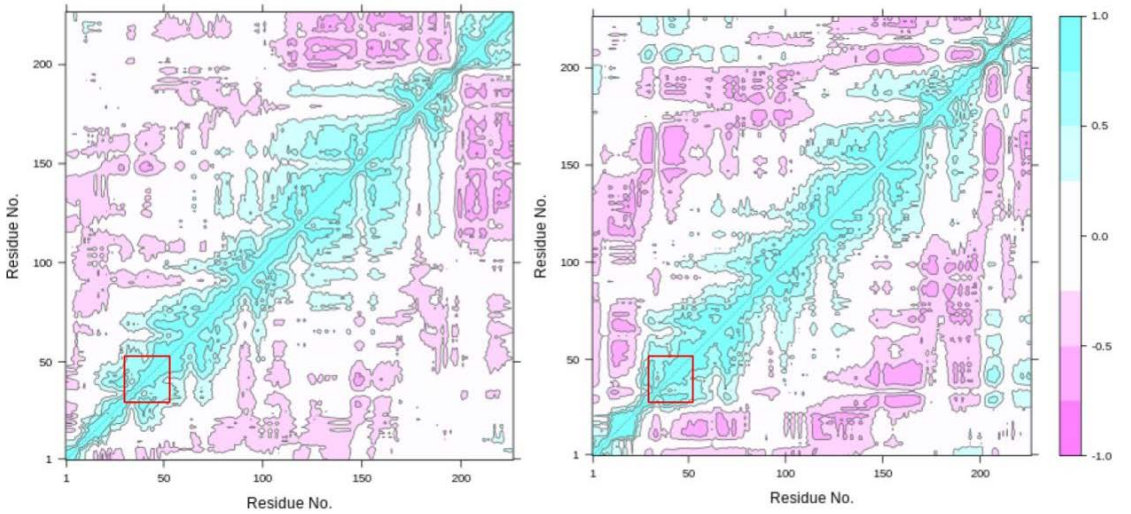
(a) 1st replicate **(b)** 2nd replicate

Figure F.2: DCCM plot of R339W. The color bar gradient shows positive and negative correlation using blue color and pink color, respectively. White color corresponds to the absence of correlation.



(a) 1st replicate **(b)** 2nd replicate

Figure F.3: DCCM plot of K344E. The color bar gradient shows positive and negative correlation using blue color and pink color, respectively. White color corresponds to the absence of correlation.



(a) 1st replicate **(b)** 2nd replicate

Figure F.4: DCCM plot of H345R. The color bar gradient shows positive and negative correlation using blue color and pink color, respectively. White color corresponds to the absence of correlation.

APPENDIX G

Residue-Residue Networks

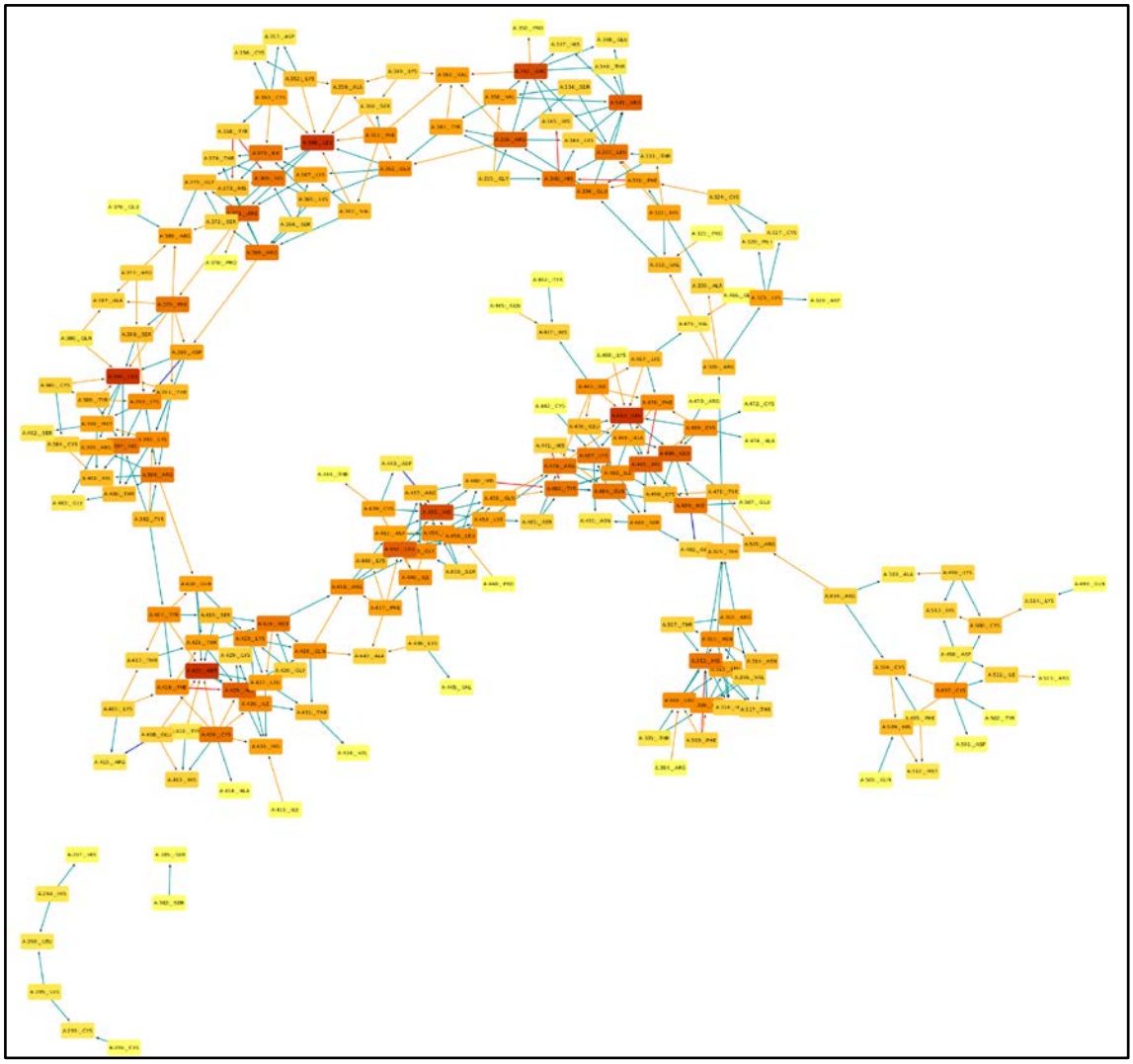
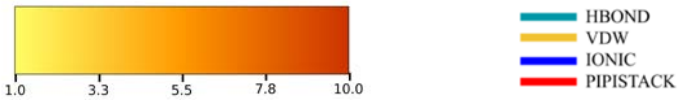
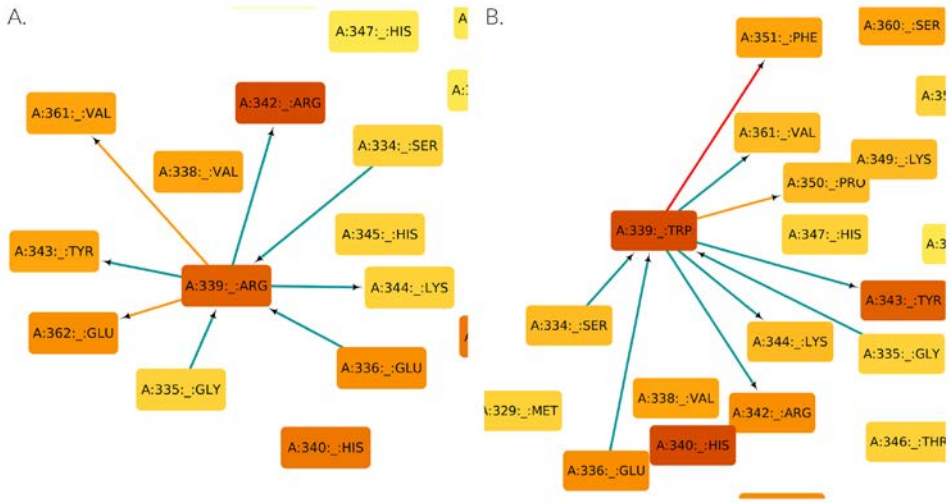


Figure G.1: RING 2.0-based residue-residue network analysis of CTCF wild-type protein. Color of nodes correspond to degree (number of edges) of nodes.



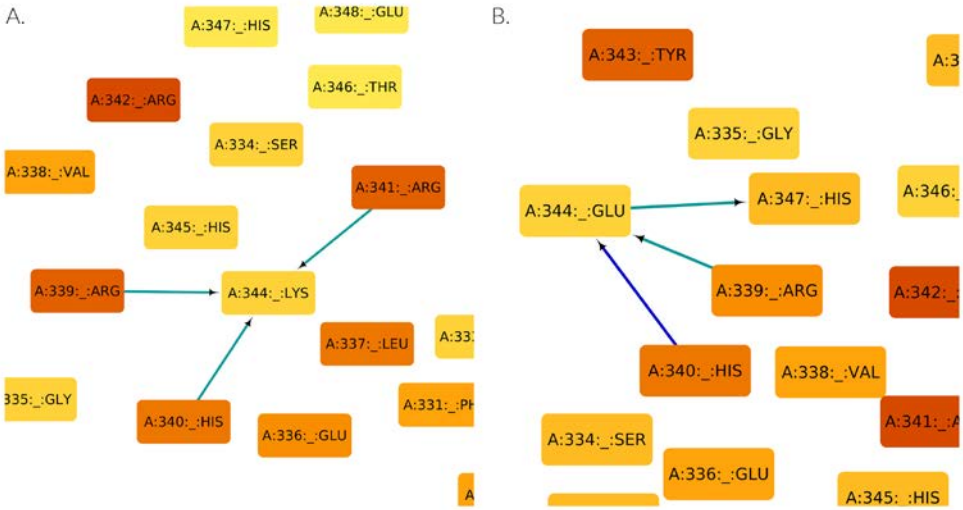
(a) Nodes **(b)** Edge

Figure G.2: Color map of RING 2.0 **a)** node color bar shows node degree as gradient from low (yellow) to high (red), **b)** edge color showing interaction type.



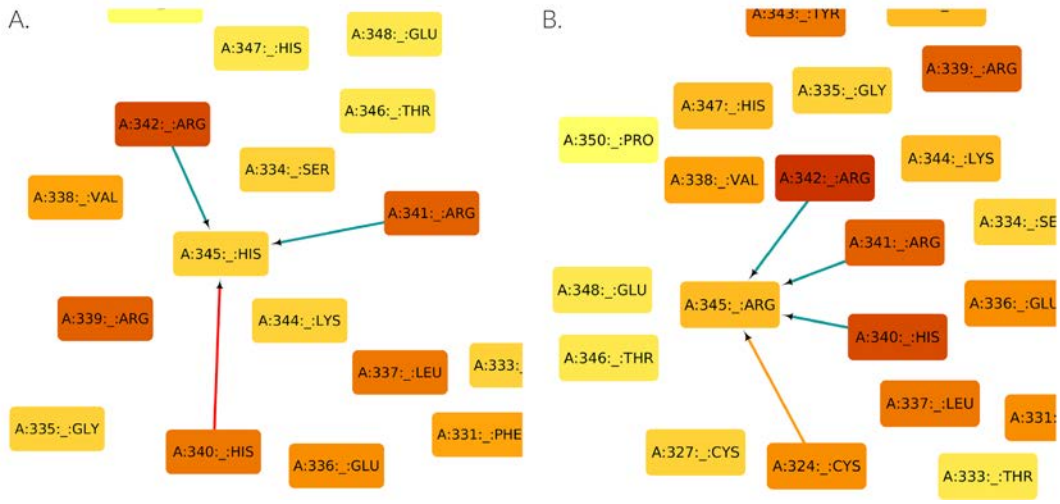
(a) **(b)**

Figure G.3: Residue-residue network analysis of the Arg (R) to Trp (W) substitution of the residue number 339. **a)** R339 (WT); **b)** W339. Color of nodes correspond to degree (number of edges) of nodes. Edges color as shown above.



(a) **(b)**

Figure G.4: Residue-residue network analysis of the Lys (K) to Glu (E) substitution of the residue number 344. **a)** K344; **b)** E344. Color of nodes correspond to degree (number of edges) of nodes. Edges color as shown above.



(a) **(b)**

Figure G.5: Residue-residue network analysis of the His (H) to Arg (R) substitution of the residue number 345. **a)** H345; **b)** R345. Color of nodes correspond to degree (number of edges) of nodes. Edges color as shown above.

APPENDIX H

Hydrogen Bond Analysis

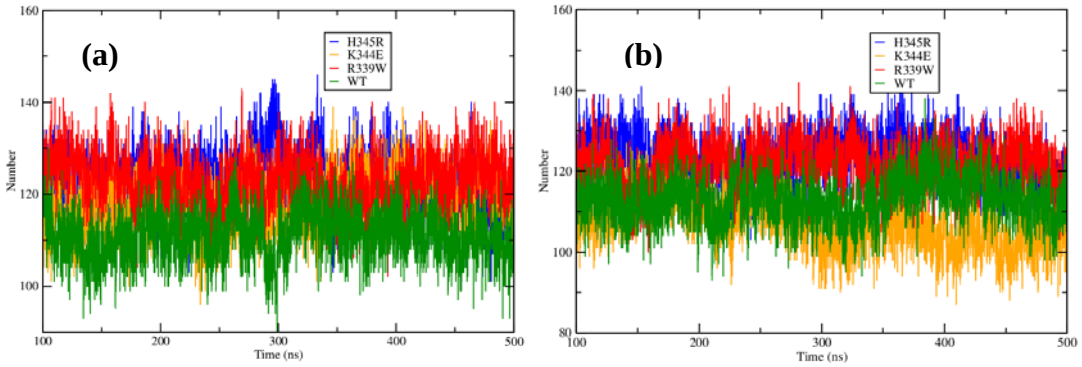


Figure H.1: Comparison based on number of hydrogen bonds within the protein. **a)** 1st replicate , **b)** 2nd replicate.

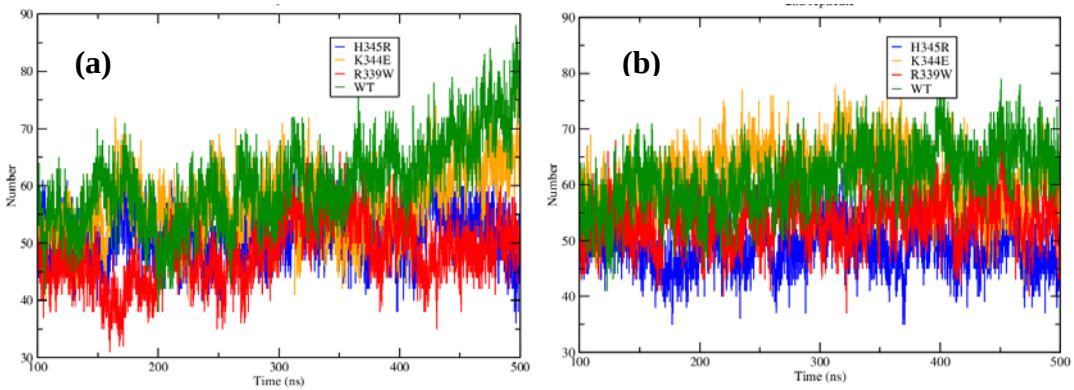


Figure H.2: Comparison based on number of hydrogen bonds within the DNA. **a)** 1st replicate , **b)** 2nd replicate.

Table H.1: Mean and median of hydrogen bonds of the 1st replicate.

	Protein-Protein		**Protein-DNA**		**ΔMeans (Mutant - Wild Type)**	
	Mean	**Median**	**Mean**	**Median**	**Protein-Protein**	**Protein-DNA**
WT	110,9	111	60,3	60	-	-
H345R	122,7	123	51,2	51	11,8	-9,1
K344E	118,6	119	56,8	57	7,7	-3,5
R339W	123,1	123	48,4	48	12,2	-11,9

Table H.2: Mean and median of hydrogen bonds of the 2nd replicate.

	Protein-Protein		**Protein-DNA**		**ΔMeans (Mutant - Wild Type)**	
	Mean	**Median**	**Mean**	**Median**	**Protein-Protein**	**Protein-DNA**
WT	112,9	113	61,5	61	-	-
H345R	123	123	48,9	49	10,1	-12,6
K344E	107,6	107	61,2	61	-5,3	-0,3
R339W	121,8	122	53,8	54	8,9	-7,7

APPENDIX I

Analysis of Distance between the DNA-binding Residues and their Contact Sites in DNA

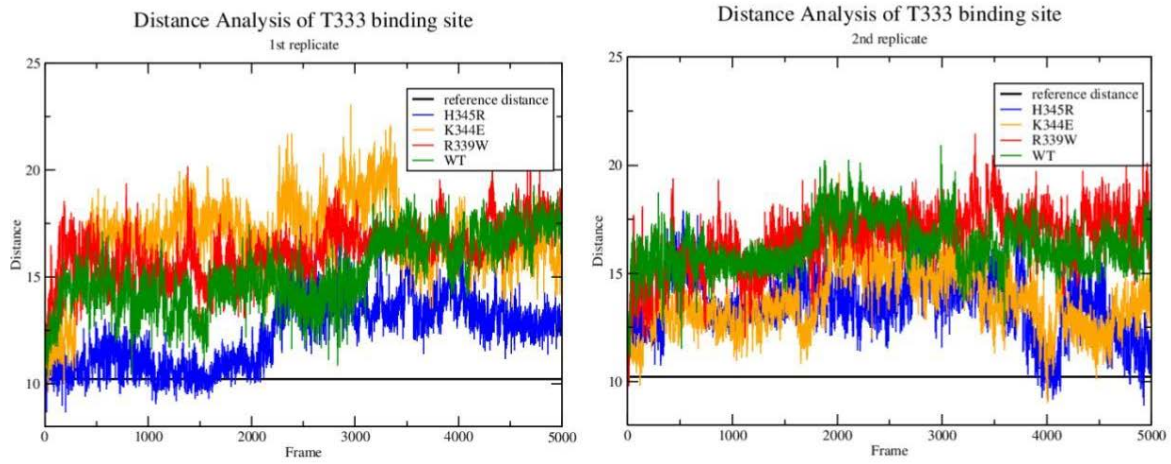


Figure I.1: Distance changes during 500ns the T333 residue and its DNA binding site. Herein, the change in the "THR333:CA - THY28:P" distance was compared between the investigated structures.

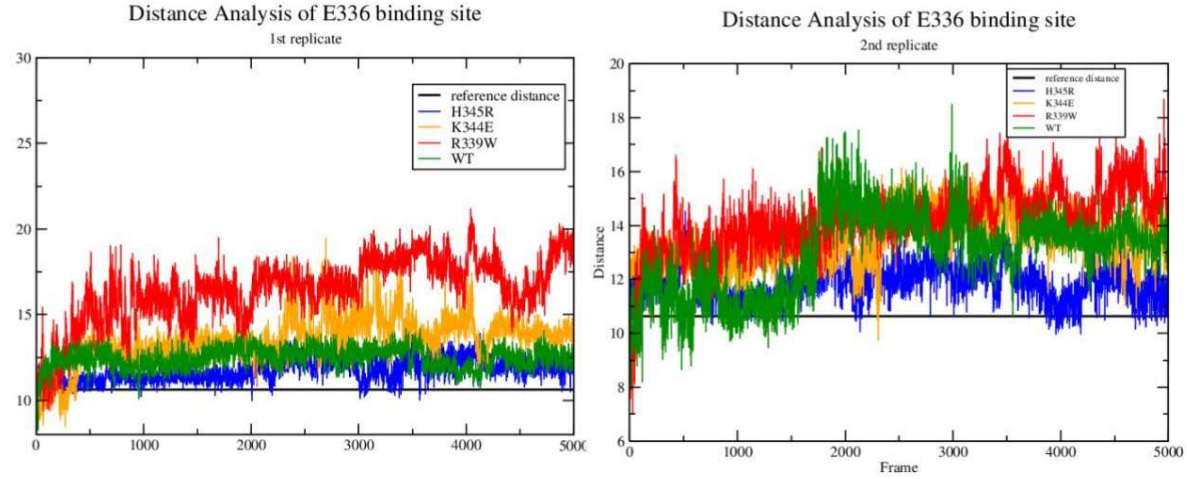


Figure I.2: Distance changes during 500ns the E336 residue and its DNA binding site. Herein, the change in the "GLU336:CA - CYT27:P" distance was compared between the investigated structures.

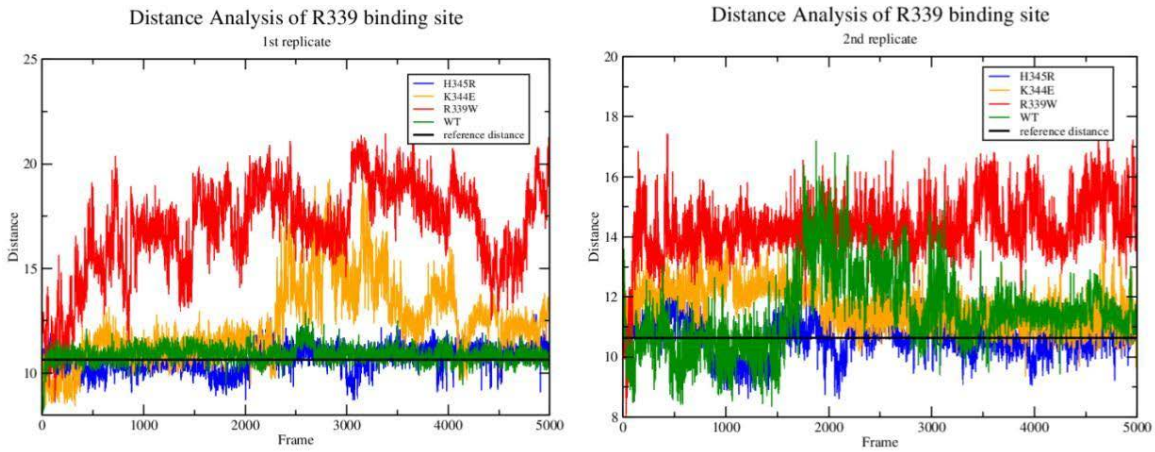
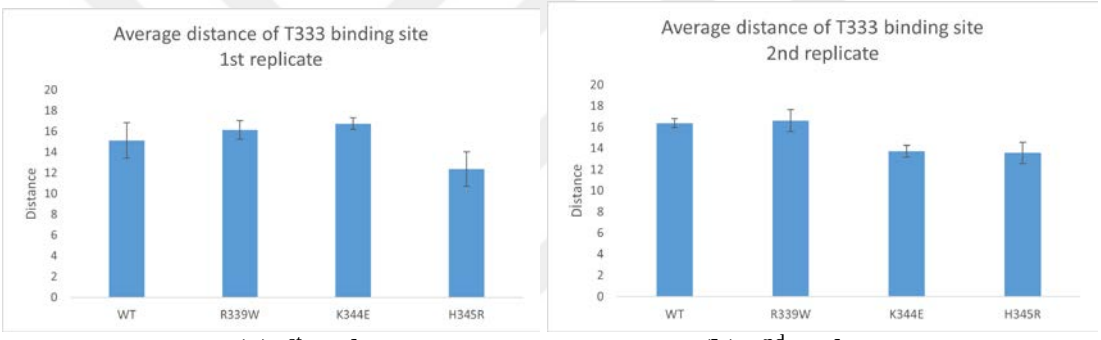
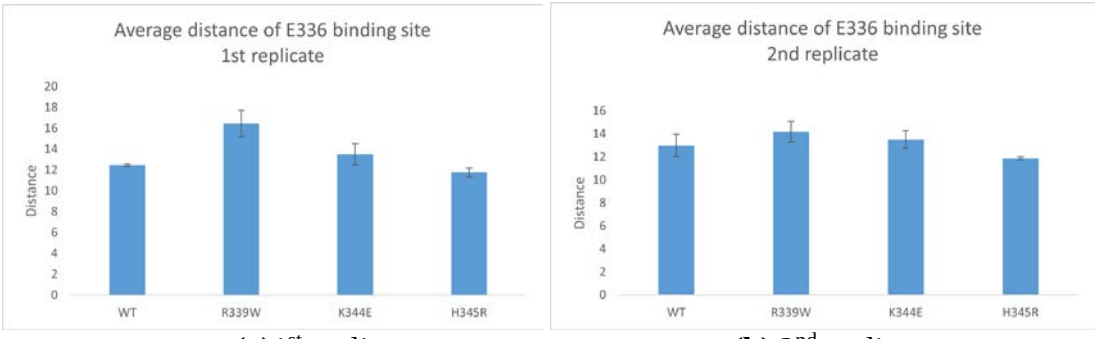


Figure I.3: Distance changes during 500ns the R339 residue and its DNA binding site. Herein, the change in the "ARG339:CA - GUA26:P" distance was compared between the investigated structures.



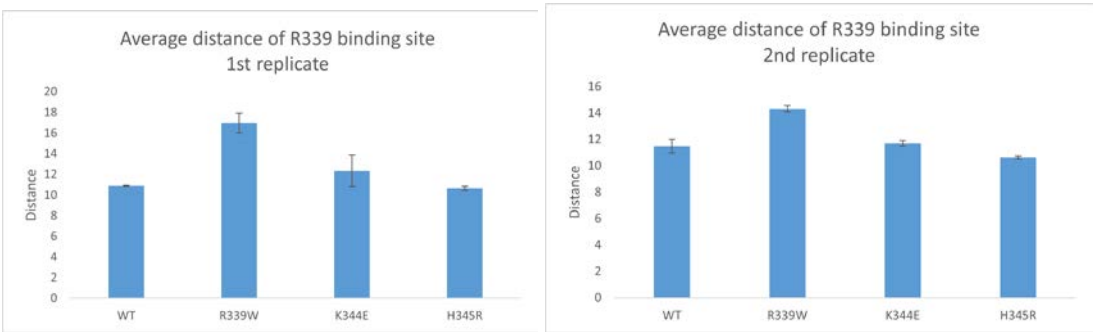
(a)1st replicate **(b)** 2nd replicate

Figure I.4: Average distance between the T333 residue and its DNA binding site for 1st replicate **a)** and 2nd replicate **b)**. Error bars represent Standard Error of the Mean.



(a)1st replicate **(b)** 2nd replicate

Figure I.5: Average distance between the E336 residue and its DNA binding site for 1st replicate **a)** and 2nd replicate **b)**. Error bars represent Standard Error of the Mean.



(a) 1st replicate **(b)** 2nd replicate

Figure I.6: Average distance between the R339W residue and its DNA binding site for 1st replicate **a)** and 2nd replicate **b)**. Error bars represent Standard Error of the Mean.

APPENDIX J

Program Commands

Gromacs commands for analysis of MD simulations

RMSD

```
#calculation of RMSD
gmx rms -f file.trr -s step5_production.tpr -n index.ndx -tu ns -o rmsd
```

RMSF

```
#calculation of RMSF
gmx rmsf -f file.trr -s step5_production.tpr -n index.ndx -o rmsf -res
```

HBOND

```
#calculation of number of hydrogen bonds
gmx hbond -f file.trr -s step5_production.tpr -n index.ndx -tu ns -num hbondnum
```

PCA

```
#calculation of PCA
gmx covar -s calpha.pdb -f calpha.trr
gmx anaeig -s calpha.pdb -f calpha.trr -extr extreme1.pdb -first 1 -last 1 -nframes 30
gmx anaeig -s calpha.pdb -f calpha.trr -rmsf -first 1 -last 2
```

Python script for Cumulative Fluctuation

```
import matplotlib.pyplot as plt

eigenvector, eigenvalue = [], []
with open ("CF.txt") as f:
  for line in f:
    cols = line.split()
    if len(cols) == 2:
      eigenvalue.append(float(cols[1]))

file=open("CumulativeFluctuation.txt","w")
sumofeigenvalues=sum(eigenvalue)
i=0
aegv=0
while i < len(eigenvalue):
          aegv+=eigenvalue[i]
          fluctuation=(aegv/sumofeigenvalues)*100
          file.write("%f %5.3f\n" % (i+1,fluctuation))
          i+=1

index, fluct = [], []
with open ("CumulativeFluctuation.txt") as n:
  for line in n:
    cols = line.split()
    if len(cols) == 2:
      fluct.append(float(cols[1]))
      index.append(float(cols[0]))

plt.plot(index, fluct)
```

```
plt.show()
```

R script for DCCM analysis

```
# This recalls functions from bio3d library.
library(bio3d)

# This parses your PDB file.
pdb <- read.pdb("system.pdb")

# This parses your DCD file.
dcd <- read.dcd ("dcd.dcd")

# Calpha selection
ca.inds <- atom.select(pdb, elety="CA")
xyz <- fit.xyz(fixed=pdb$xyz, mobile=dcd,
       fixed.inds=ca.inds$xyz,
       mobile.inds=ca.inds$xyz)

# Check
dim(xyz) == dim(dcd)
## If everything is okay, you will see "[1] TRUE" in the screen.

# This generates Dynamic Cross Correlation Map (DCCM).
cij<-dccm(xyz[,ca.inds$xyz])
plot(cij)
write.csv(cij, file = "file.csv")
# This saves your DCCM as pgn.
dev.copy(png, "file.png")
dev.off()
```

CURRICULUM VITAE

Name Surname : Sinem Darwish

EDUCATION:

B.Sc. : 2018, Üsküdar University, Faculty of Engineering and Natural Sciences, Molecular Biology and Genetics

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2018, Graduated with honor degree (3.21/4.00), Üsküdar University

OTHER PUBLICATIONS, PRESENTATIONS, AND PATENTS:

- A. Ayaz, B. Abak, B. Süzek and S. Darwısh, "Whole Exome Sequencing Analysis Method Offering Multiple Views for Clinical Diagnosis", TR Patent, In submission.
- A. Ayaz, B. Abak, B. Süzek and S. Darwısh, "Next-Generation Sequencing in Cancer Genetic Testing", presented at the 3rd International Cancer and Ion Channels Congress, Istanbul, Turkey, Sep. 17, 2021.

STRUCTURAL BIOINFORMATICS ANALYSIS OF THE CANDIDATE TUMOR SUPPRESSOR PROTEIN CTCF

ORIGINALITY REPORT

11%	7%	8%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Higher Education Commission Pakistan Student Paper	1%
2	www.nature.com Internet Source	1%
3	academic.oup.com Internet Source	<1%
4	Maolu Yin, Jiuyu Wang, Min Wang, Xinmei Li, Mo Zhang, Qiang Wu, Yanli Wang. "Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites", Cell Research, 2017 Publication	<1%
5	www.annualreviews.org Internet Source	<1%
6	Simranjeet S. Narang, Suniba Shuaib, Deepti Goyal, Bhupesh Goyal. "Assessing the effect of D59P mutation in the DE loop region in amyloid aggregation propensity of β2 -	<1%