Molecular sequence evolutionary analyses of the EGLN1 gene in *Capra falconeri* (Markhor), a gene involved in high altitude adaptation

By

Jaweria Hafeez

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-I-Azam University

Islamabad, Pakistan

2023

Molecular sequence evolutionary analyses of the EGLN1 gene in *Capra falconeri* (Markhor), a gene involved in high altitude adaptation

By

Jaweria Hafeez

Reg # 02282113013

A thesis submitted in the partial fulfillment of the requirements for the degree of

MASTER OF PHILOSIPHY

In

BIOINFORMATICS

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-I-Azam University,

Islamabad, Pakistan

2023

CERTIFICATE

This thesis is submitted by **Jaweria Hafeez** from the National Center for Bioinformatics, Faculty of Biological Sciences, Quaid-I-Azam University, Islamabad, Pakistan, and is accepted in its present form as satisfying the thesis requirements for the Degree of Master of Philosophy in Bioinformatics.

Internal Examiner:
Dr Amir Ali Abbasi
Professor & Supervisor
Quaid-I-Azam University, Islamabad.

External Examiner:
Dr Peter John
Professor
National University of Sciences and Technology, Islamabad.

Chairperson:
Dr. Syed Sikander Azam
Professor
Quaid-I-Azam University, Islamabad.

Dated: 1st September,2023

DECLARATION

The work reported in this study, entitled "Molecular sequence evolutionary analyses of the EGLN1 gene in *Capra falconeri* (Markhor), a gene involved in high altitude adaptation." was carried out by Jaweria Hafeez under the supervision of Dr Amir Ali Abbasi, from National Centre for Bioinformatics, Faculty of Biological Sciences, Quaid-I-Azam University, Islamabad, Pakistan. Hereby, I declare that the title of the thesis and all the contents presented in the following study are a product of my effort and no part has been copied from any published source (except the references, standard mathematical or genetic models /equations /formulas /protocols, etc.). None of the work has been submitted for the award of any other degree /diploma. The University may act if the information provided is found inaccurate at any stage.

Date: _____ _____

Jaweria Hafeez

Dedicated
To
My Father, Hafeez Ur Rehman

&

My Mother Irfana Yasmeen

ACKNOWLEDGEMENTS

In the name of ALLAH, Most Merciful and Beneficial.

Thanks to **ALMIGHTY ALLAH**, for the wisdom, he bestowed upon me, peace of mind and good health to finish this research, and Darood upon **Prophet Muhammad (PBUH)**, Without His will, the successful completion of my work would not have been possible. With the kind support and assistance of many people, this thesis becomes a reality. I would like to express my heartfelt gratitude to each one of them.

First and foremost, my gratitude goes out to my supervisor, **Dr Amir Ali Abbasi**, for his moral support, encouragement, motivational attitude, and guidance throughout the research.

Special acknowledgements go to my beloved father **Hafeez Ur Rehman**, mother **Irfana Yasmeen** and my brother **Muhammad Tayyab Rehman** for their never-ending affection, support, motivation and prayers. I owe all of my achievements to you. Thank you for being there whenever I needed you.

I am highly thankful to my family for being the ultimate source of happiness and support in my life. Nothing of what I am today would have been possible without having you all around me.

I am greatly thankful to all the Genomics lab mates for their guidance, cooperation and moral support throughout my stay at the lab.

Also, I would like to extend my sincere thanks to my friends and classmates. Especially I would mention **Ayesha Ikram** and **M. Faizan Malik** as they stood by my side through every thick and thin.

May Allah immensely bless all of you.

Jaweria Hafeez

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

EGLN1	Egl-9 family hypoxia-inducible factor 1
PHD2	Prolyl hydroxylase domain-containing protein 2
bp	base pair
Kb	Kilobase
HIF- α	Hypoxia Inducible Factor-alpha
EPO	Erythropoietin
Hb	Hemoglobin
RMSD	Root Mean Square Deviation
NCBI	National Center for Biotechnology Information
ODD	Oxygen Dependent Degradation domain
VHL	Von Hippel-Lindau
UV	Ultra violet
HAPE	high-altitude pulmonary edema
CITES	Convention on International Trade in Endangered Species of Wild Flora and Fauna
wt	Wild type
ORF	Open reading frame
SLAC	Single likelihood ancestor counting
ECYT3	Familial erythrocytosis type 3
DNA	Deoxyribonucleic acid

ABSTRACT

The Markhor (*Capra falconeri*), Pakistan's national animal, faces critical endangerment challenges despite its remarkable resilience in inhabiting high-altitude terrains. Distinguished by its majestic horn morphology, elongated shoulder structure, and unique habitat preferences, the Markhor stands apart from conventional domestic goats and other caprine species. This study delves into the pivotal role of the EGLN1 gene in mediating high-altitude adaptation, extending its investigation to other species with comparable traits, including the snow leopard. EGLN1 encodes the PHD2 protein (Prolyl hydroxylase domain containing protein 2), a component of the oxygen-sensing system that expedites the degradation of HIF-α (Hypoxia-Inducible Factor alpha). In hypoxic environments characterized by low oxygen levels, EGLN1 activity diminishes, permitting HIF-α to evade hydroxylation at critical proline sites, consequently leading to HIF-α's stabilization.To unravel the genetic underpinnings of this intricate adaptive mechanism, we conducted a comprehensive evolutionary analysis of PHD2 protein homologs across diverse sarcopterygian species, encompassing both invertebrates and high-altitude adapted animals like the Tibetan Human, Markhor, and snow leopard. By subjecting EGLN1 coding DNA data to evolutionary rate analysis, we deciphered prevailing selection constraints. Our findings underscore the prevalence of negative selection acting upon the EGLN1 gene, signifying its functional significance in maintaining high-altitude adaptability.Employing MEGA7, ancestral sequences of EGLN1 were reconstructed, and subsequent multiple sequence alignments facilitated the identification of species-specific amino acid substitutions in the snow leopard and Markhor. Notably, while no distinct amino acid changes were observed in the Markhor, a solitary alteration was identified in the goat's EGLN1 protein compared to the Markhor's protein.This investigation not only enhances our understanding of the adaptive traits characterizing the Markhor and other high-altitude dwellers but also uncovers the intricate genetic foundations that underlie their unique capabilities. The intricate interplay between EGLN1 and HIF-α opens avenues for further research into the molecular mechanisms orchestrating high-altitude adaptation across diverse species.

Chapter 1
Introduction

1 Introduction

1.1 EGLN1 gene association to Hypoxia:

EGLN1 is located on human chromosome 1 (Chromosome 1: 231,363,751-231,422,287 reverse strand , GRCh38:CM000663.2). It spans 58.3 kb and contains five exons (Dupuy et al. 2000). EGLN1 gene is involved in cellular hypoxic response. (Majmundar et al. 2010). Prolyl hydroxylase (PHD2) encoded by EGLN1, is oxygen sensor that functions to control the protein levels of the Hypoxia Inducible Factor-α (HIF-α) by hydroxylating specific proline residues. HIF-α is the master transcriptional regulator of the hypoxic response (Semenza GL et al. 2009). PHD2 protein contains 426 amino acids residues with molecular weight of 46.02097 KDa and an isoelectric point of 8.6181. EGLN1 is responsible for post-translational modulation of EPAS1 and HIF-α through oxygen-dependent hydroxylation of specific proline residues (Jaakkola et al. 2001). Under circumstances of low oxygen levels, the hydroxylation process experiences a notable reduction, leading to the stabilization of EPAS1 and HIF-α. (Jaakkola et al. 2001). Genetic alterations in the EGLN1 gene have been linked to conditions such as erythrocytosis, a condition marked by an over production of red blood cells and increase in blood density. Under normal oxygen conditions, PHD2 is consistently active and performs hydroxylation on crucial proline sites within oxygen-sensitive degradation regions of HIF-α. Following this hydroxylation process, HIF-α binds to the VHL protein, forming an E3 ubiquitin ligase complex. This complex facilitates the ubiquitination of HIF-α, leading to its subsequent breakdown through the proteosome pathway. However, in conditions of oxygen deficiency, once evasion of ubiquitin-mediated degradation occurs, HIF-α forms a heterodimer with aryl hydrocarbon nuclear translocator. This heterodimer functions as a transcription factor, subsequently boosting the expression of EPO gene (Figure 1.1) (Barradas, J. et al. 2018).

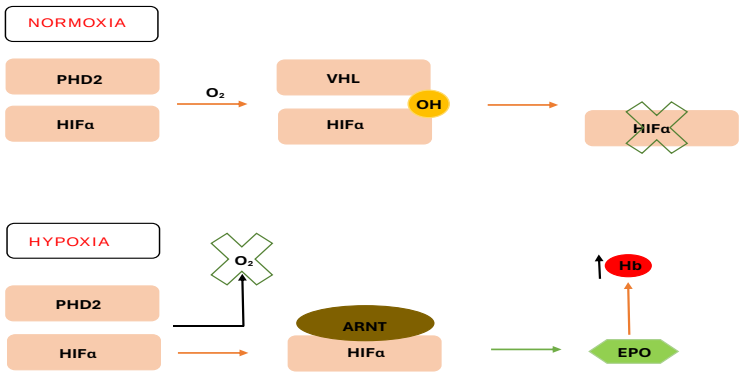
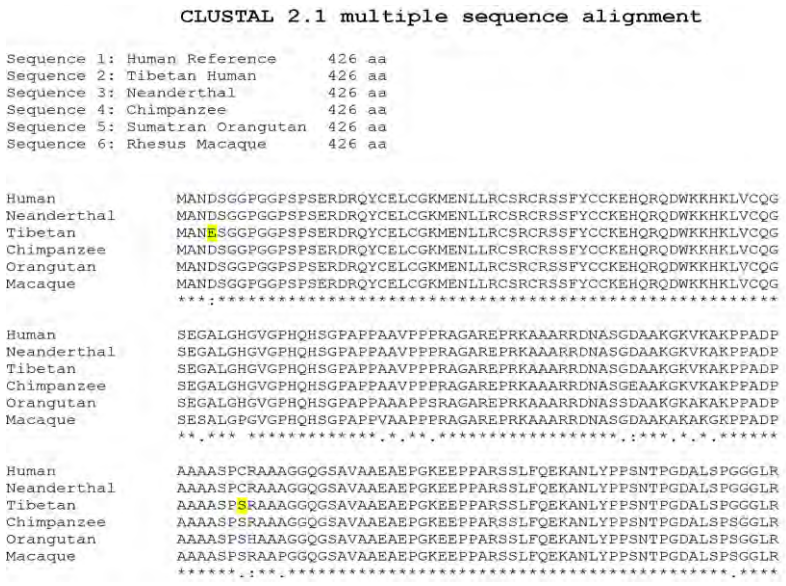


Figure 1.1 Schematic representation of PHD2 protein functional association with oxygen availability.

1.2 Role of EGLN1 in high altitude adaptation:

Genetic adaptation to a new environment is necessary for the survival and adaptation of species. In humans, most recent example is adaptation to high altitudes, such as the Tibetan highlands. The Tibetan Plateau (TP; also known as the Qinghai–Tibet Plateau in China) has an average elevation of ~4,000 m above sea level, where the oxygen concentration is ~40% lower (Beall, C. M. et al. 2007) and UV radiation is ~30% stronger (Dahlback, A. 2007) than at sea level. The native Tibetan population has evolved unique physiological traits in response to the challenging environmental conditions in high-altitude regions. (Beall, C. M. et al. 2007). At elevations around 3000 to 4500 meters, Tibetans display elevated resting ventilation, yet they exhibit lower arterial oxygen levels and reduced oxygen saturation compared to the three groups that have resided in high-altitude areas for numerous generations; among these populations, Tibetans experienced the most pronounced hypoxia (Beall, C. M. et al. 2006). Certain Tibetans demonstrate heightened levels of hemoglobin concentration, but this occurrence is observed exclusively at altitudes exceeding 4000 meters (Beall, C. M. et al. 2006). Genetic investigations conducted on populations have revealed that genetic variations at the EGLN1 loci have undergone favorable natural selection. These genetic mutations are linked to differences in hemoglobin concentration (HB) among Tibetans (Yi, X., Liang, Y. 2010). EGLN1 performs degradation of HIF, leading to the stimulation of red blood cell synthesis. A point mutation of EGLN1 leads to decrease in hemoglobin levels, especially in high-altitude

regions (Lorenzo, F. R. et al. 2014). Xiang, K. et al. 2013 reported a total of 166 SNPs located in exon 1. Within these 166 SNPs, two distinct non-synonymous SNPs were found in Tibetan individuals. One is a G to C mutation (rs186996510) resulting in an amino acid change from aspartic acid to glutamic acid (D4E), and the other one is a G to C mutation (rs12097901) causing a shift from cysteine to serine (S127C). Remarkably, one of these non-synonymous mutations, specifically rs186996510 (D4E), displays a remarkable contrast between Tibetans and non-Tibetans. It is notably prevalent among Tibetans (63.27%) while being exceptionally rare in Han Chinese (1.03%), Japanese (0.56%), Europeans (0.59%) and Africans (2.27%). This discrepancy represents the most significant allelic variation observed between Tibetans and non-Tibetans, strongly suggesting its potential significance in the context of adapting to high altitude environments. (Xiang, K. et al. 2013). Multiple sequence alignment of PHD2 proteins of various primates is shown in figure 1.2. Variations in EGLN1 gene patterns and levels of expression influence an individual's vulnerability to disorders such as high-altitude pulmonary edema (HAPE) and chronic mountain sickness. These disorders arise when individuals encounter high-altitude environments, causing an imbalance between oxygen availability and requirements (Mishra, A. et al. 2013).



```
                    CLUSTAL 2.1 multiple sequence alignment

Sequence 1: Human Reference     426 aa
Sequence 2: Tibetan Human       426 aa
Sequence 3: Neanderthal         426 aa
Sequence 4: Chimpanzee          426 aa
Sequence 5: Sumatran Orangutan  426 aa
Sequence 6: Rhesus Macaque      426 aa

Human             MANDSGGPGGPSPSERDRQYCELCGKMENLLRCSRCRSSFYCCKEHQRQDWKKHKLVCQG
Neanderthal       MANDSGGPGGPSPSERDRQYCELCGKMENLLRCSRCRSSFYCCKEHQRQDWKKHKLVCQG
Tibetan           MANESGGPGGPSPSERDRQYCELCGKMENLLRCSRCRSSFYCCKEHQRQDWKKHKLVCQG
Chimpanzee        MANDSGGPGGPSPSERDRQYCELCGKMENLLRCSRCRSSFYCCKEHQRQDWKKHKLVCQG
Orangutan         MANDSGGPGGPSPSERDRQYCELCGKMENLLRCSRCRSSFYCCKEHQRQDWKKHKLVCQG
Macaque           MANDSGGPGGPSPSERDRQYCELCGKMENLLRCSRCRSSFYCCKEHQRQDWKKHKLVCQG
                  ***:************************************************************

Human             SEGALGHGVGPHQHSGPAPPAAVPPPRAGAREPRKAAARRDNASGDAAKGKVKAKPPADP
Neanderthal       SEGALGHGVGPHQHSGPAPPAAVPPPRAGAREPRKAAARRDNASGDAAKGKVKAKPPADP
Tibetan           SEGALGHGVGPHQHSGPAPPAAVPPPRAGAREPRKAAARRDNASGDAAKGKVKAKPPADP
Chimpanzee        SEGALGHGVGPHQHSGPAPPAAVPPPRAGAREPRKAAARRDNASGEAAKGKVKAKPPADP
Orangutan         SEGALGHGVGPHQHSGPAPPAAAPPSRAGAREPRKAAARRDNASSDAAKGKAKAKPPADP
Macaque           SESALGPGVGPHQHSGPAPPVAAPPPRAGAREPRKAAARRDNASGDAAKAKAKGKPPADP
                  **.*** ************.*.**.******************.:***.*.*.******

Human             AAAASPCRAAAGGQGSAVAAEAEPGKEEPPARSSLFQEKANLYPPSNTPGDALSPGGGLR
Neanderthal       AAAASPCRAAAGGQGSAVAAEAEPGKEEPPARSSLFQEKANLYPPSNTPGDALSPGGGLR
Tibetan           AAAASPSRAAAGGQGSAVAAEAEPGKEEPPARSSLFQEKANLYPPSNTPGDALSPGGGLR
Chimpanzee        AAAASPSRAAAGGQGSAVAAEAEPGKEEPPARSSLFQEKANLYPPSNTPGDALSPSGGLR
Orangutan         AAAASPSHAAAGGQGSAVAAEAEPGKEEPPARSSLFQEKANLYPPSNTPGDALSPSGGLR
Macaque           AAAASPSRAAPGGQGSAVAAEAEPGKEEPPARSSLFQEKANLYPPSNTPGDALSPSGGLR
                  ******.:**.************************************************.****
```

Figure 1.2 Multiple sequence alignment of PHD2 protein from various primate species (residues 1-180). Tibetan human substitutions D4E and C127S are highlighted in yellow color.

1.3 *Capra falconeri* (Markhor) adaptation to high altitude:

Capra falconeri (Markhor) inhabits the north eastern regions of Afghanistan, the northen parts of India (Southwest Jammu and Kashmir), the northern and central areas of Pakistan, the southern regions of Tajikistan, the southwestern parts of Turkmenistan, and the southern areas of Uzbekistan (Grubb, P. et al.2005). The markhor, a wild goat belonging to the Bovidae family and Caprinae subfamily, underwent a shift from Appendix II to Appendix I of the Convention on International Trade in Endangered Species of Wild Flora and Fauna (CITES) in 1992. This alteration marked the cessation of the trophy hunting initiative for markhor, which was initiated by the North West Frontier Province Wildlife Department (NWFP WD) in 1983. Subsequently, in 1993, the NWFP WD engaged local communities in wildlife conservation by implementing Community Game Reserve Rules under the Wildlife Act of 1975. In 1997, with special authorization from CITES, the NWFP WD introduced a community-centered markhor trophy hunting program within the province. A significant portion (80%) of the permit fee is allocated to a Village Conservation Fund (VCF) as an incentive to motivate local community participation in safeguarding markhor and related wildlife species. This approach has effectively transformed local attitudes towards wildlife, fostering a rise in markhor populations within community-managed conservation areas (CMCAs). The success of the markhor conservation program in CMCAs parallels that of government-managed protected areas. This achievement can be attributed to the NWFP WD's involvement of the local community in preserving natural resources. Despite success, the markhor in NWFP confront various threats, encompassing habitat fragmentation, community reliance on natural resources, lack of awareness, poaching, and insufficient conservation funding. Thus, safeguarding markhor remains a complex endeavor for both governmental bodies and local communities. The triumph of the community-based markhor conservation initiative in NWFP can be attributed to the economic incentives it provides. (Ali, S. et al. 2008).

The term “Markhor” seemingly originated from Persian language and translates to “snake eater”. However, it is more commonly believed to have its roots in the Pashto language as “Mar Akhkar”, where “Mar” refers to snake and ‘Akhkar” refers to horn. The markhor,

possess horns that twist akin to a snake, earned its name as “Mar Akhkar”. As time went on, this evolved into “Markhor” (Roberts, T. J. et al. 1977).

Markhor are robust creatures characterized by powerful and relatively stubby legs, along with wide hooves. (Roberts, T. J. et al. 1977). The color of markhor’s coat ranges from shades of brown to deep black-brown and gray (Malik, M. M. et al. 1987). A typical fully grown male flared-horned markhor has a shoulder height of 99-104 cm and a complete body length of 132-185 cm (Malik, M. M. et al. 1987). Mature females are approximately half the size of adult males. (Malik, M. M. et al. 1987). Male flare-horned markhor can weigh between 100-110kg (220-242 lbs), while females weigh around 32-50 kg (70.5-110 lbs) (Ranjitsinh, M. K. 2005). The markhor inhabits the Suleiman Range, ranging from approximately 700 meters to 1000 meters on the lower slopes. In winter, they can be found at altitudes of up to 2700 meters, and during the summer, their range extends to 4000 meters in the Chitral valley. Markhor inhabit challenging landscapes characterized by steep cliffs, rocky inclines and mountains. (Schaller, G. B. et al. 1977). Markhor has developed particular modifications in their lung structure to excel in elevated altitudes. These alterations encompass enlarged alveoli, and an enhanced mechanism for oxygen exchange. Markhor possess elevated levels of red blood cell in their circulatory system. Red blood cells contain hemoglobin that binds oxygens and effectively convey a larger quantity of oxygen to body tissues (Eng, J. T., & Aldenderfer, M. et al.2017).

1.4 Aims and objectives:

The study aimed at unraveling the evolutionary and structural perspectives of EGLN1 gene with following objectives:

1. To report complete tertiary structure of PHD2 protein.
2. To recognize significant motifs and domains across the protein structure from different orthologs.
3. To figure out specie-specific changes in the PHD2 protein of distinct groups of species.
4. To comprehend the structural implications of disease-causing as well as evolutionary mutations in the PHD2 protein.

Chapter 2
Materials and Methods

2 Materials and Methods

This study was conceptually planned to look at the phylogenetic and evolutionary characteristics of the EGLN1/PHD2 protein associated to high altitude adaptation. It involves investigating different aspects of EGLN1, including its evolutionary pattern across different lineages, the selection pressures that are influencing the protein's evolution, structural comparison of the mutated and wild-type forms of the protein as well as its interaction with HIF1α protein.

Figure 2.1 Schematic overview of EGLN1 gene analysis. Each step used in this study is depicted in this flow chart.

2.1 Protein sequences retrieval:

Sequence of Human EGLN1 protein (also known as PHD2) was retrieved using Ensembl genome browser (Fernández & Birney, 2010) under the protein ID ENSP00000355601.3. Blastp (Altschul et al. 1990) program available at National Centre for Biotechnology information (NCBI) (Wheeler et al., 2005) and Blast program integrated in Ensembl genome browser was used to identify putative orthologous proteins against Human EGLN1 protein from different species.

The genomic sequences of *Capra falconeri* (Markhor) EGLN1, EGLN2 and EGLN3 genes were retrieved through a query process using the NCBI tool. The genetic material of Capra falconeri, a specie of particular interest, underwent a Genome Assembly procedure by a colleague. Subsequently, the sequences related to the aforementioned goat genes were extracted via a BLASTN comparison against the assembled genome of Capra falconeri. The DNA sequence of gene was then given to ORF finder as an input. ORF finder program (Rombel, I. T., Sykes, K. F., Rayner, S., & Johnston, S. A. et al. 2002) returned us several ORFs and their protein translations. The protein translation identical to Capra hircus EGLN1 protein was selected.

2.2 Phylogenetic analysis using MEGA7:

MEGA7 was used for Phylogenetic analysis of EGLN1 proteins. Species used for phylogenetic tree construction are: *Homo Sapiens* (Human), *Callorhinchus milii* (Elephant Shark), *Danio rerio* (Zebrafish), *Pan troglodytes* (Chimpanzee), *Mus musculus* (Mouse), *Bos taurus* (Cow), *Lepisosteus oculatus* (Spotted Gar), *Capra falconeri* (Markhor), *Ovis aries* (Sheep), *Capra hircus* (Goat), *Panthera uncia* (Snow leopard), *Cervus hanglu yarkandensis* (Yarkand deer), *Latimeria chalumnae* (Coelacanth) and *Drosophila melanogaster* (Fruit fly), *Octopus bimaculoides* (California two-spot octopuses), *Helobdella robusta* (Freshwater leech), *Caenorhabditis elegans* (Nematode).

Multiple sequence analysis was performed by using ClustalW (default parameters) integrated in MEGA7 program. Neighbor-joining tree was constructed to infer evolutionary history. Bootstrap value of 1000 was applied to validate resulting tree

branches. P-distance method and complete deletion options were used in the construction of NJ tree. Maximun likelihood tree was also constructed using MEGA7.

2.3 Evolutionary rate analysis using MEGA7 and Hyphy:

Protein coding DNA sequences were collected from Ensembl Genome browser. This coding data was used to analyze the difference between the rate of non-synonymous substitutions and the rate of synonymous substitutions.

Z-test implemented in MEGA7 was used to investigate selection acting within Primates (Human, Neanderthal, Chimpanzee, Gorilla, Sumatran orangutan, Rhesus macaque), Even-toed ungulates (Goat, Sheep, Markhor, Siberian musk deer) and Rodents (Mouse, Rat, Chinese hamster, Arctic ground squirrel, Naked mole rat).

Hyphy software was used to analyze rate of non-synonymous and synonymous substitutions (dN/dS) by Goldman and Yang (GY-94) method. GY-94 method uses codon-based model to estimate dN and dS rate.

2.4 Comparative domain organization:

Domains and motifs were mapped to Human reference EGLN1 protein and its orthologs from different species (Tibetan Human, Snow leopard, Markhor, Goat, and Mouse) on their respective positions. This mapping was performed on the basis of multiple sequence alignment, literature review and analysis through databases including InterPro, UniProt and MyHits Motif Scan (Pagni, M. et al. 2007). Specie-specific substitutions were identified by performing MSAs and these substitutions were also mapped. Erthrocytosis Familial 3 related substitutions are mapped on Human EGLN1 protein. Two substitutions responsible for high altitude adaptation in Tibetan Human reported in literature are also mentioned.

2.5 Protein secondary structure prediction:

SYMPRED uses statistical algorithms in order to predict secondary structure of protein. The algorithms used by SYMPRED are trained on a dataset of known protein structures that are experimentally determined. SYMPRED was used to predict protein secondary

structure. PROF, SSPRO, YASPIN, JNET and PSIPRED algorithms well-known for protein secondary structure prediction were applied.

2.6 Template for Protein tertiary structure prediction:

Protein blast (Blastp) from NCBI was performed against PDB database in order to retrieve template for EGLN1 protein tertiary structure prediction. Protein with PDB ID 4bqw having e-value 0.0, length 252 amino acids, query coverage 57% and percentage identity 100% was selected as a template for EGLN1 protein tertiary structure prediction. 4bqw template structure is determined by X-Ray diffraction and has a resolution of 1.79 Å.

2.7 Protein tertiary structure prediction:

Complete 3D structure of EGLN1 is not available in Protein Data Bank (PDB), we used the best match available in PDB as a template for Comparative modelling. Homology modelling predicts the structure of protein based on the known structure of template.

Homology modelling comprises these steps: (1) Alignment of protein sequence of target and template to identify regions of similarity and divergence. (2) Alignment of target 3D structure to template 3D structure in order to generate spatial arrangement between atoms of target and template residues. (3) Create a preliminary 3D model of the target protein using the aligned template structure as a starting point. (4) Model refinement and energy minimization improves the geometry and removes steric clashes in the model.

Robetta server predicts protein structures computationally and uses cutting-edge algorithms for precise 3D protein modeling. To produce accurate predictions of protein structure, it makes use of homology modeling, ab initio approaches, and refining methods. Robetta server was used for comparative modelling of EGLN1 protein. PDB ID 4bqw was used as a template for protein 3D modelling.

2.8 Protein structure validation and refinement:

The SAVES (Structure Analysis and Verification Server) provide comprehensive tools to evaluate the quality and reliability of protein structure models. SAVES was used to evaluate protein model geometry, energetics, and stereochemistry. Quality assessment

programs like ERRAT, VERIFY 3D and PROCHECK were run in order to evaluate our Predicted protein model. Further refinement and optimization of model is performed if the initial model does not satisfy quality requirements.

The YASARA Energy Minimization Server applies sophisticated algorithms to minimize energy and rectify steric clashes. YASARA energy minimization server was used to improve structure of PHD2 protein predicted model and then this model was again validated through SAVES server.

2.9 Primate protein tertiary structure prediction:

Predicted structure of Human Reference PHD2 protein was taken as a template to model Tibetan Human PHD2, Hominin ancestor PHD2, Chimpanzee PHD2 and Hominid ancestor PHD2. These models were then optimized to lowest energy by using YASARA energy minimization server. Each of these structures was analyzed by SAVES server programs: PROCHECK, VERIFY 3D and ERRAT.

2.10 Mutant protein tertiary structure prediction:

Predicted structure of Human Reference PHD2 protein was taken as a template to model mutant proteins that are involved in Erythrocytosis Familial 3. The mutant proteins sequence was generated by manually changing the amino acid residues at specified positions. For instance in Mutant P317R, at position 317 we changed amino acid Proline with Arginine. This sequence was then given as an input to Robetta for comparative modelling with Wild type (Human Reference) PHD2 protein as a template. Following the same procedure other two mutant proteins (R371H PHD2 and H374R PHD2) were modelled. These models were then optimized to lowest energy by using YASARA energy minimization server. Each of these structures was analyzed by SAVES server programs: PROCHECK, VERIFY 3D and ERRAT.

2.11 Carnivores and Artiodactyls Proteins structure prediction:

Snow leopard PHD2 protein was modelled by using Robetta server. The predicted model was analyzed by programs like PROCHECK, VERIFY 3D and ERRAT. Energy

minimization was performed by YASARA energy minimization server and all the Quality factors were again analyzed using SAVES server. The model having best evaluation scores and graph was selected and considered as predicted Snow leopard PHD2. This predicted snow leopard PHD2 protein was taken as a template to model tiger PHD2, Panthera ancestor PHD2 and domestic cat PHD2. The predicted models were analyzed by programs like PROCHECK, VERIFY 3D and ERRAT.

Capra Falconeri (Markhor) PHD2 protein was modelled by using Robetta server. The predicted model was analyzed by programs like PROCHECK, VERIFY 3D and ERRAT. Energy minimization was performed by YASARA energy minimization server and all the Quality factors were again analyzed using SAVES server. The model having best evaluation scores and graph was selected and considered as predicted Markhor PHD2. This predicted Markhor PHD2 protein was taken as a template to model Goat PHD2, Caprine ancestor PHD2 and Siberian musk deer PHD2. The predicted models ware analyzed by programs like PROCHECK, VERIFY 3D and ERRAT.

Ancestral protein sequences were inferred by MEGA7 and were verified manually by performing Multiple sequence alignments.

2.12 Superimposition of Proteins:

The predicted Human Reference protein structure was superimposed with Tibetan Human PHD2 in order to identify the effect of the two substitutions D4E and C127S on the tertiary structure of protein. To observe evolutionary changes in PHD2 tertiary structure in primates, Human reference PHD2 was also superimposed with hominin ancestor PHD2, Chimpanzee PHD2 and Hominid ancestor PHD2. Tibetan Human PHD2 protein was also superimposed with Hominin ancestor PHD2 in order to analyze the deviations in structure that occurred in ancestor and adapted individual.

In carnivores, snow leopard PHD2 undergone two substitutions (V161A and M228K). No substitutions were observed in PHD2 orthologs belonging to carnivores. PHD2 orthologs in Tiger, lion, leopard contains an insertion or deletion of one or two alanine residue in their protein sequence. Snow leopard PHD2 was superimposed with Tiger PHD2, Panthera

ancestor PHD2(ancestral sequence generated by MEGA7 and analyzed by MSA of carnivore species) and domestic cat PHD2.

Capra falconeri (Markhor) PHD2 protein contain no specific substitution. Goat PHD2 protein contains one substitution that was found to pe goat-specific after MSA of artiodactyl proteins. Markhor PHD2 was superimposed with Goat PHD2, Caprine ancestor PHD2 and Siberian musk deer PHD2.

2.13 HIF1α protein tertiary structure prediction:

Human HIF1α protein sequence was retrieved from UniProt database. Protein-protein interaction was performed between HIF1α and PHD2(wt and disease-causing mutants). For this purpose, HIF1α tertiary structure was modelled by Robetta server. Homology modelling was performed. PDB ID 4zpr was used as a template, this template was selected on basis of Blastp results against PDB database. The structure model obtained from Robetta server was then refined by using GalaxyRefine (Heo, L., Park, H., & Seok, C. et al. 2013) web server. The predicted model was analyzed by programs like PROCHECK, VERIFY 3D and ERRAT.

2.14 Interaction of proteins:

Protein-protein interactions of EGLN1 protein (wild type and mutants) with HIF1α was performed by ClusPro (Alekseenko, A., Ignatov, M., Jones, G., Sabitova, M., & Kozakov, D. et al. 2020). Information about the interacting residues between EGLN1 and HIF1α residues was not available in literature (Blind docking was performed). The best docking results were selected on the basis of binding affinity (calculated using ProDigy server) (Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M., & Vangone, A. et al.2016), Lowest energies of models and the number of cluster members. The results of interaction were visualized using UCSF Chimera. For 2D visualization LIGPLOT+ was used.

Table 2.1: Set of databases, web servers and tools used in the study.

Index	Database and server	URL

1.	NCBI	https://www.ncbi.nlm.nih.gov/
2.	Ensembl	https://asia.ensembl.org/index.html
3.	UniProt	https://www.uniprot.org/
4.	RSCB PDB	https://www.rcsb.org/
5.	InterPro	https://www.ebi.ac.uk/interpro/
6.	myhits Motif Scan	https://myhits.sib.swiss/cgi-bin/motif_scan
7.	Clustal Omega	https://www.ebi.ac.uk/Tools/msa/clustalo/
8.	SYMPRED	https://www.ibi.vu.nl/programs/sympredwww/
9.	Robetta	https://robetta.bakerlab.org/
10.	Galaxy Refine	https://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE
11.	YASARA	http://www.yasara.org/minimizationserver.htm
12.	SAVES	https://saves.mbi.ucla.edu/
13.	ClusPro	https://cluspro.bu.edu/login.php
14.	PRODIGY	https://wenmr.science.uu.nl/prodigy/
15.	PDBsum	http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/
16.	ORF finder	https://www.ncbi.nlm.nih.gov/orffinder/
17.	Datamonkey SLAC	https://www.datamonkey.org/slac

Chapter 3
Results

3 Results

3.1 Phylogenetic Analysis of Prolyl hydroxylase domain Protein family:

Neighbor-joining (NJ) tree and Maximum likelihood (ML)tree were generated from MEGA7 (Figure 3.1, Figure 3.2). The evolutionary history of EGLN protein family was analyzed by collecting the orthologous sequences from various vertebrate and invertebrate species. Phylogenetic tree of EGLN protein family showed two duplication events that resulted in 3 paralogs: EGLN1, EGLN2 and EGLN3.

Maximum likelihood tree generated by MEGA7.Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and Bio NJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 38 amino acid sequences. All positions containing gaps and missing data were eliminated. There were total of 211 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.

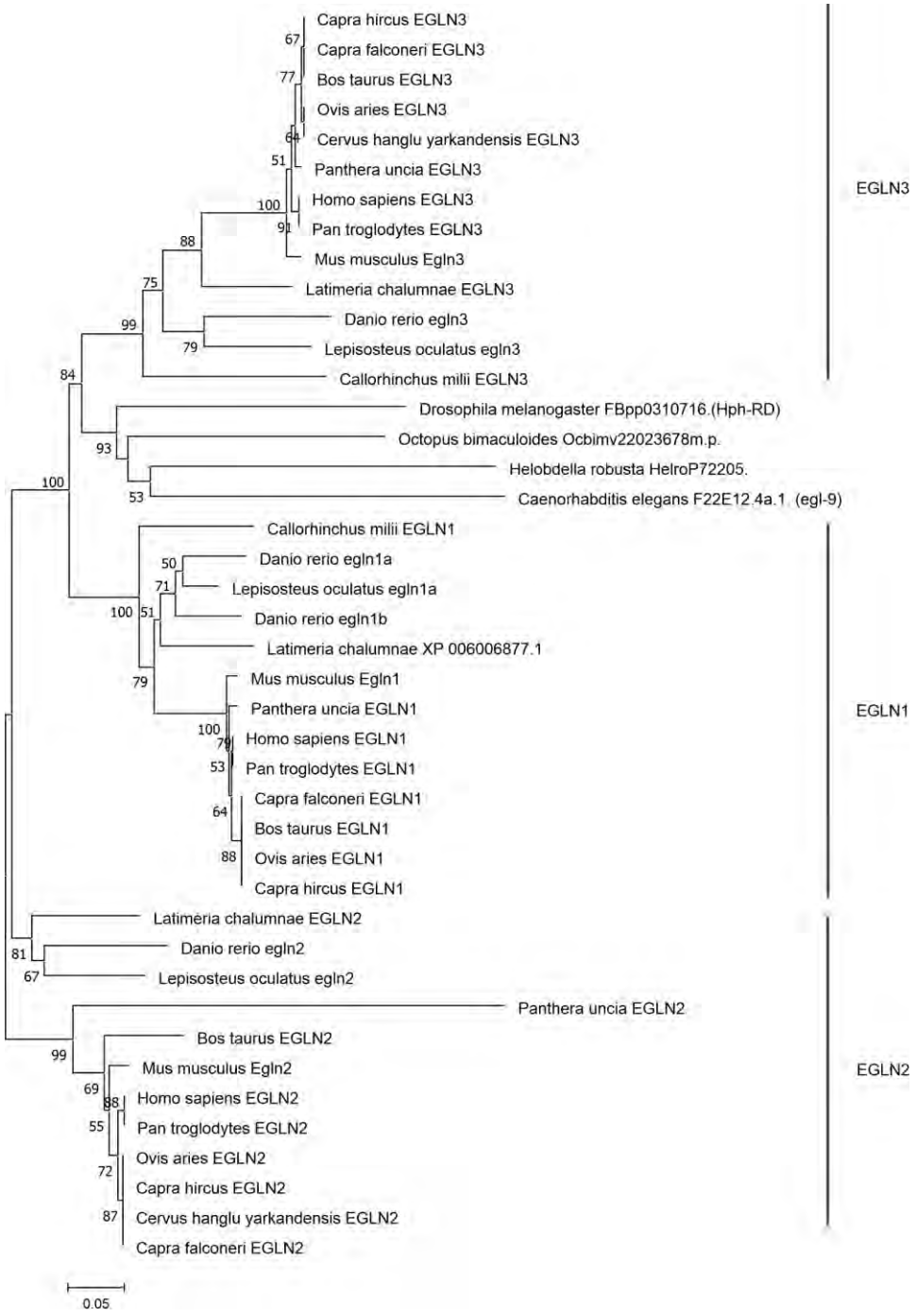


Figure 3.1 Neighbor-Joining tree of the Egl-Nine (EGLN) protein family. Uncorrected p-distance was used. Complete-deletion option was used. Numbers on branches represent bootstrap values (based on 1000 replications) supporting that branch; only the values ≥ 50% are presented here. Scale bar shows amino acid substitution per site.

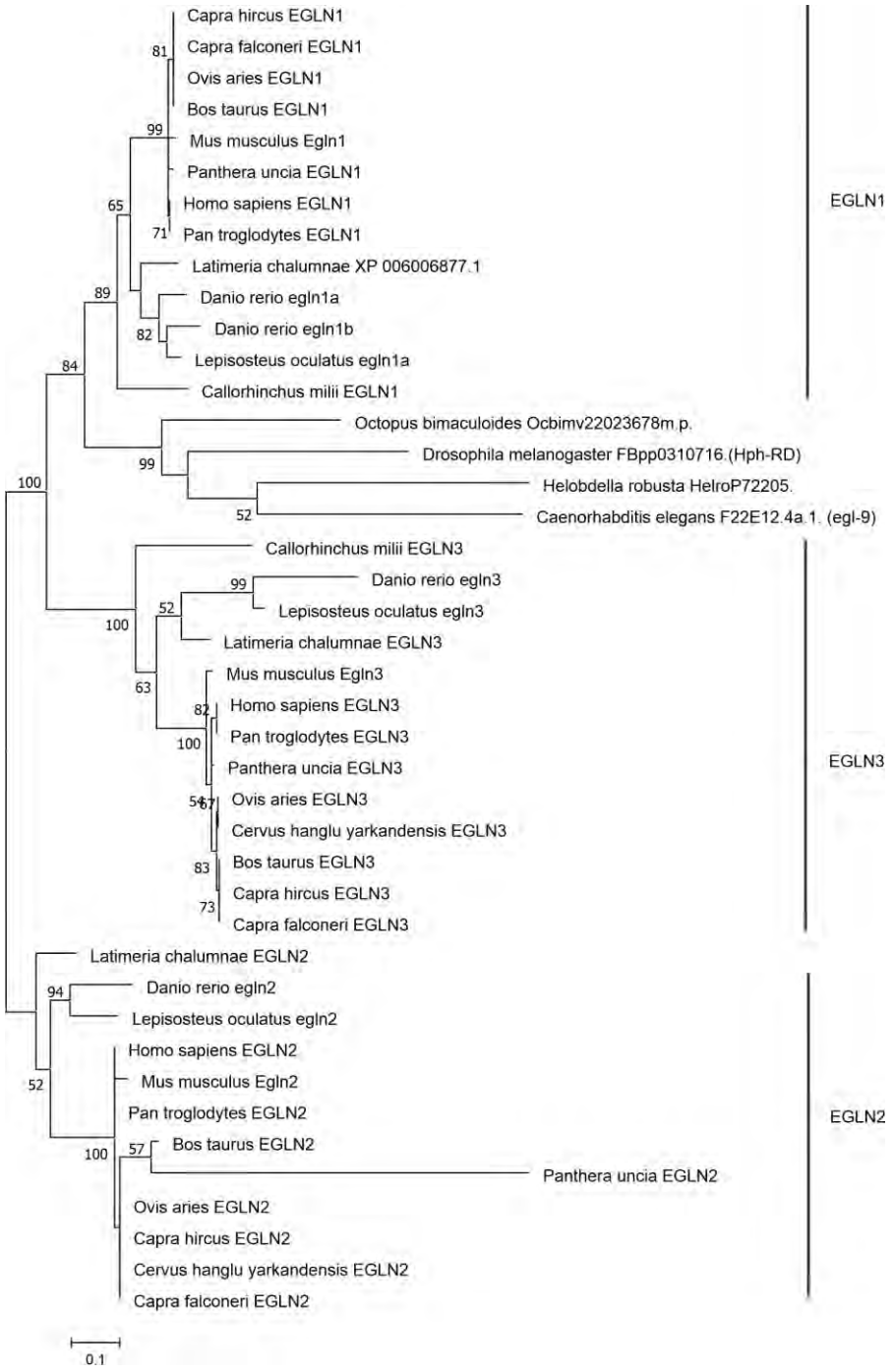


Figure 3.2 The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan and Goldman model. The tree with the highest log likelihood (-4525.57) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches.

3.2 Evolutionary rate analysis:

To understand selection pressure on EGLN1 gene, three evolutionary diverged clads were analyzed. These include primates (human, neanderthal, chimpanzee, gorilla, Sumatran orangutan, rhesus macaque), even-toed ungulates (Goat, sheep, markhor, Siberian musk deer) and rodents (Mouse, rat, Chinese hamster, arctic ground squirrel, naked mole rat).

Z-test implemented in MEGA7 was used to estimate the evolutionary rates in these three groups. Z-test estimates the overall rate of non-synonymous substitutions (dN) and synonymous substitutions (dS) with a probability value. If rate of non-synonymous substitutions is greater as compared to the rate of synonymous substitutions, it means the gene is under positive/diverging evolution. Similarly, if rate of synonymous substitutions is greater as compared to the rate of non-synonymous substitutions, it means the gene is under negative/purifying evolution.

The dN-dS difference was -2.817 for primates, -2.813 for even-toed ungulates and -9.715 for rodents. Probability value for all the selected group of species was less than 0.05, suggesting that null hypothesis is rejected for all three groups (Protein is evolving either positively or negatively). The negative z-statistic value suggested that *EGLN1* has evolved under negative selection during evolution (Table 3.1).

Goldman and Yang (GY-94) method which uses codon-based model for rate estimation was employed using Hyphy. The results also indicated purifying selection in all three group of species (Table 3.2).

Table 3.1 Analysis of selection pressure acting on the EGLN1 gene using Z-test statistics within MEGA program.

Data set	probability	z-statistic
Primates Human, neanderthal, chimpanzee, gorilla, Sumatran orangutan, rhesus macaque	0.006	-2.817
Even-toed ungulates Goat, sheep, markhor, Siberian musk deer	0.006	-2.813
Rodents Mouse, rat, Chinese hamster, arctic ground squirrel, naked mole rat	0.000	-9.715

This table shows the results obtained by Z-test for neutrality implemented in MEGA7. Z-test was performed to analyze the selection pressure acting upon EGLN1 gene. Three different data sets were analyzed in this study which include Primates, Even-toed ungulates and Rodents. The probability value less than 0.05 rejects null hypothesis and suggests that some selective constraints are acting upon this gene. Negative z-statistic value suggests negative selection acted upon EGLN1 gene.

Table 3.2 Estimation of number of synonymous substitutions per synonymous site (dS), number of non-synonymous substitutions per non-synonymous site (dN) with Hyphy.

Substitution Model/Method: Codon Based (GY-94)		
Species-EGLN1	**dN/dS (dN,dS)**	**dN-dS**
	Primates	
Human	0(0,0)	-0.5062
Neanderthal	0(0,0)	
Chimpanzee	17.92(0.0233,0.0013)	
Gorilla	0(0,0.1380)	
Sumatran orangutan	0.256(0.0583, 0.2278)	
Rhesus macaque	0.26(0.0773,0.2980)	
	Even-toed ungulates	
Goat	0(0.0108,0.0000)	0.1626
Sheep	2.93(0.0331,0.0113)	
Markhor	0(0,0)	
Siberian musk deer	3.549(0.1810,0.0510)	
	Rodents	
Mouse	0.098(0.0131,0.1336)	-2.0486
Rat	0.129(0.0392,0.3029)	
Chinese hamster	0.119(0.0393,0.3308)	
Arctic ground squirrel	0.2016(0.1894,0.9395)	
Naked mole rat	0.466(0.5434,1.1662)	

[1]**dN-dS<0 implies negative selection constraint on EGLN1 within sarcopterygian lineage.** This table shows the results of evolutionary rate analysis performed by Hyphy (GY-94 model). The test was performed on three different data sets. 1st column enlists all the species that were included in the specified dataset, 2nd column represents the ratio of non-synonymous substitutions per site to synonymous substitutions per site(dN/dS). The 3rd column indicates overall difference between non-synonymous substitutions and synonymous substitutions in the specified dataset.

3.3 Residual constraints detection:

SLAC (Single likelihood ancestor counting) analysis was performed on the Hominoids (Human, Neanderthal, Chimpanzee, Gorilla Gibbon, Bonobo and orangutan), non-hominoids (macaque, bushbaby and marmoset), non-primate placental mammals (Cow, goat, cat, elephant and mouse) and birds (common canary, chicken, duck, white throated sparrow). SLAC reconstructs most probable ancestor sequences and identifies synonymous and non-synonymous substitutions that occurred at codon level in the alignment. It uses weighing scheme to calculate number of substitutions. SLAC analysis identified 29 negatively constrained sites in the *EGLN1* protein (Table 3.3). SLAC analysis plot is shown in Figure 3.3.

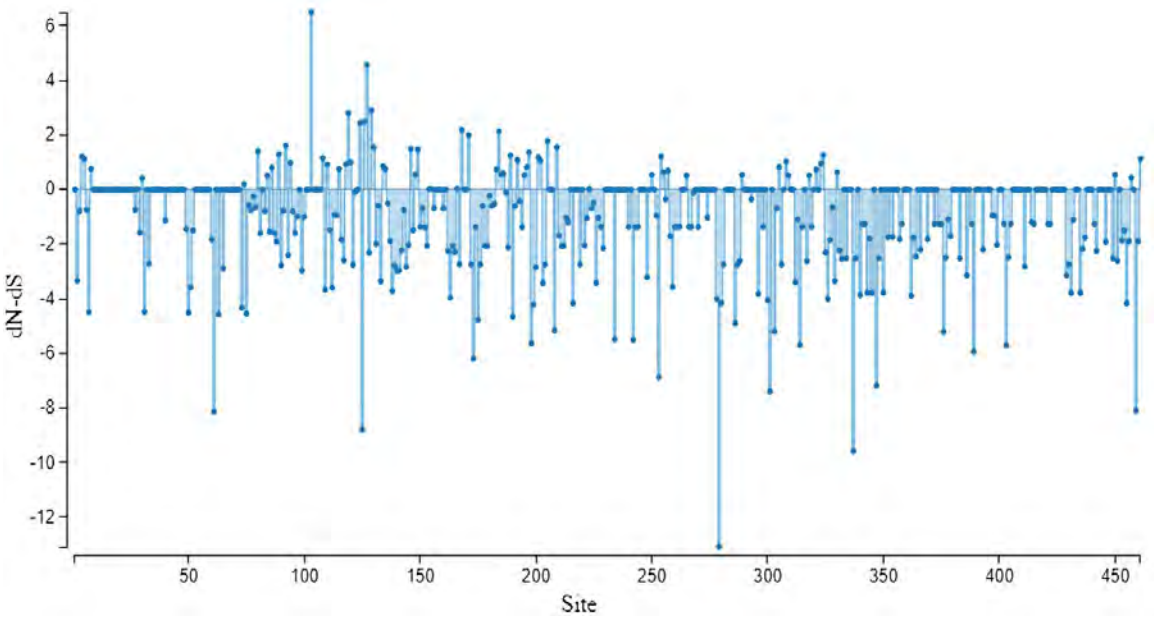


Figure 3.3 Window displaying the sites under the negative selection constraint in EGLN1 among sarcopterygians. Results are generated with Hyphy by implementing SLAC method which uses global codon model and maximum likelihood to reconstruct the evolutionary history. X-axis represents amino acid sites and y-axis represents the rate of non-synonymous substitution to synonymous substitutions at the specific residue site.

Table 3.3 Identification of negatively constrained sites in EGLN1 among sarcopterygians at 0.1 significance level with Hyphy.

Index	Residue no	dN-dS	P-value
1	279	-13.1	0.00039
2	337	-9.58	0.00094
3	253	-6.88	0.0042
4	347	-7.20	0.0054
5	234	-5.50	0.012
6	198	-5.64	0.012
7	61	-8.13	0.016
8	173	-6.19	0.018
9	389	-5.94	0.020
10	301	-7.40	0.020
11	403	-5.72	0.022
12	459	-8.09	0.022
13	314	-5.70	0.026
14	376	-5.21	0.026
15	286	-4.91	0.025
16	242	-5.52	0.030
17	303	-5.20	0.031
18	431	-3.79	0.037
19	435	-3.79	0.037
20	345	-3.79	0.037
21	7	-4.5	0.037
22	343	-3.79	0.038
23	73	-4.33	0.042
24	340	-3.87	0.047
25	350	-3.77	0.050
26	362	-3.90	0.060

27	248	-3.22	0.078
28	175	-4.78	0.086
29	455	-4.17	0.096

This table gives the results of SLAC analysis. Abbreviations: dS, synonymous substitutions per synonymous site; dN, non-synonymous substitutions per non-synonymous site. 2nd column depicts the sites under negative constraints within sarcopterygians. 3rd column gives the value of dN-dS at the specific residue. 4th column depicts p-value ($p<0.1$) suggesting putative negatively constrained sites.

3.4 Comparative domain organization:

Domains, motifs and sub-motifs of EGLN1 protein was identified through literature survey and Myhit Motif Scan. These domains were mapped on the EGLN1 protein orthologs from different species (Human ref., Tibetan human, snow leopard, markhor, goat and mouse). Comparative domain analysis revealed two highly conserved domains: Zinc-finger MYND-type domain (21-58), and Fe (+2) dioxygenase domain (291-392). The beta2beta3 'finger-like' loop domain is required for hypoxia inducible factors selectivity (CODD or NODD). The comparative domains are represented in schematic view (Figure 3.3)

Substitutions are displayed with the help of an arrow head at their respective position. Erythrocytosis familial 3 associated substitutions that were reported previously are displayed by orange arrowheads onto Human reference EGLN1 protein on their respective sites. Two substitutions that were involved in high altitude adaptation and changes in hemoglobin concentrations (D4E and C127S) reported in Tibetan Human are mapped onto Tibetan EGLN1 protein by yellow arrow heads. Two substitutions in Snow leopard EGLN1 were identified in current study by performing Multiple sequence alignment within Carnivores. These substitutions are mapped by red arrow heads onto Snow leopard EGLN1 protein. Only one residue difference is identified between Capra Falconeri (Markhor) EGLN1 protein and Capra Hircus (Goat)EGLN1 protein. After performing MSA of EGLN1 proteins from various Artiodactyls, it was concluded that this substitution was Goat-specific (S416L indicated by purple arrow head on Goat EGLN1 protein).

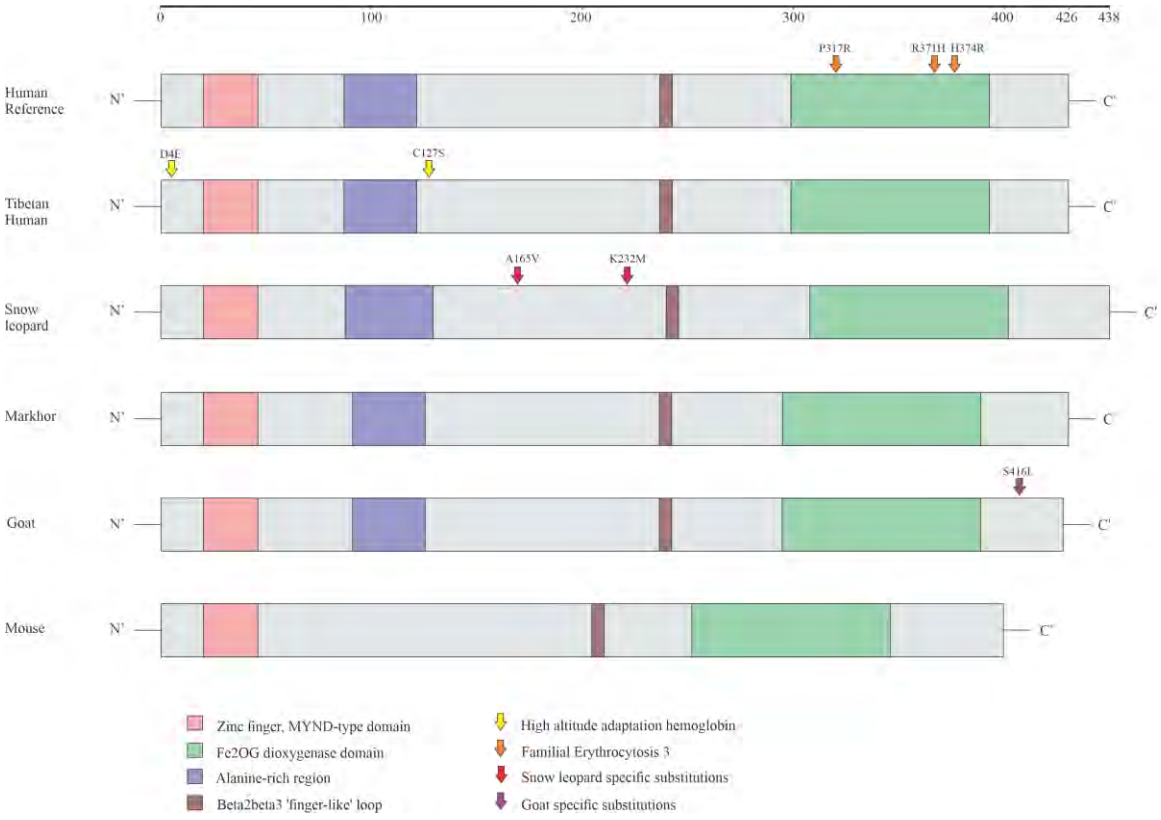


Figure 3.4 Domain organization of PHD2 protein. Schematic view of comparative organization of key functional domains and motifs of PHD2 across orthologous proteins from phylogenetically distant species. Protein lengths are drawn approximately to scale. Domains and motifs are color-coded. Numbers on scale represents amino acids.

3.5 Secondary structure analysis:

Secondary structure of EGLN1 protein was predicted by its primary amino acids sequence. EGLN1 protein sequence was submitted to SYMPRED server. The consensus of different algorithms that predict secondary structures predicted 8 alpha helices and 12 beta sheets. Most of the beta sheet were predicted from amino acids 251-420. Results of secondary structure prediction are shown in Table 3.4.

Table 3.4 Table depicts results obtained from SYMPRED. Dynamic programming with no weighting was used by SYMPRED to predict secondary structures.

```
Ruler      .........11........21........31........41........51........61.......
Sequence   MANDSGGPGGPSPSERDRQYCELCGKMENLLRCSRCRSSFYCCKEHQRQDWKKHKLVCQGSEGALGHGVG
PROF                     HHHHHHH    HHHHHH       EE     HHH     HHHHH
SSPRO                      HHHHH    HHHHHHH      EE     H H     HHHHH     H
YASPIN
JNET                       EEEE       EE      EEEEE
PSIPRED                           HHH          EEEE   HHHHHHHHHHHH
           ----------------------------------------------------------------------
SYMPRED                    HHHH     HHHHHH       EE     HHH     HHHH

Ruler      .........81........91........101.......111.......121.......131.......
Sequence   PHQHSGPAPPAAVPPPRAGAREPRKAAARRDNASGDAAKGKVKAKPPADPAAAASPCRAAAGGQGSAVAA
PROF
SSPRO
YASPIN
JNET
PSIPRED
           ----------------------------------------------------------------------
SYMPRED

Ruler      .........151.......161.......171.......181.......191.......201.......
Sequence   EAEPGKEEPPARSSLFQEKANLYPPSNTPGDALSPGGGLRPNGQTKPLPALKLALEYIVPCMNKHGICVV
PROF                                                     HHHHHHHHHHHHHHH    EEE
SSPRO                    HHHHHH                          HHHHHHHHHHHHHHH    EEE
YASPIN
JNET                                                     HHHHHHHHHHHHHH    EEEE
PSIPRED                    HHHHH                   HHH   HHHHHHHHHHHHHH    EEE
           ----------------------------------------------------------------------
SYMPRED                                                  HHHHHHHHHHHHHH    EEEE

Ruler      .........221.......231.......241.......251.......261.......271.......
Sequence   DDFLGKETGQQIGDEVRALHDTGKFTDGQLVSQKSDSSKDIRGDKITWIEGKEPGCETIGLLMSSMDDLI
PROF       E    HHHHHHHHHHHHHHHH                    EEEEE      HHHHHHHHHHHHHHH
SSPRO           HHHHHHHHHHHHHHHH        E           EEEEE      HHHHHHHHHHHHHHH
YASPIN
JNET            HHHHHHHHHHHHHHHHH                   EEEEE      HHHHHHHHHHHHHHH
PSIPRED         HHHHHHHHHHHHHHHH                    EEEEE     HHHHHHHHHHHHHHH
           ----------------------------------------------------------------------
SYMPRED         HHHHHHHHHHHHHHHH                    EEEEE      HHHHHHHHHHHHHHH

Ruler      .........291.......301.......311.......321.......331.......341.......
Sequence   RHCNGKLGSYKINGRTKAMVACYPGNGTGYVRHVDNPNGDGRCVTCIYYLNKDWDAKVSGGILRIFPEGK
PROF       HHHHHH   EEEE   EEEEEEE     EEEEEE      EEEEEEEEE          EEEEE
SSPRO      HHHHHH   EEE    EEEEEEE     EEEEE        EEEEEEEE          EEEEE
YASPIN
JNET       HHHHHH    EE     EEEEEE     EEEEEEE      EEEEEEEE          EEEEE
PSIPRED    HHHHHHH          EEEEEEE      EEE        EEEEEEEEE         EEEEEE
           ----------------------------------------------------------------------
SYMPRED    HHHHHH    EE     EEEEEEE     EEEEE        EEEEEEEE          EEEEE
```

```
Ruler       .........361.......371.......381.......391.......401.......411.......
Sequence    AQFADIEPKFDRLLFFWSDRRNPHEVQPAYATRYAITVWYFDADERARAKVKYLTGEKGVRVELNKPSDS
PROF         EEEEEE     EEEEEEE                EEEEEEE   HHHHHHHHHHHH       EEE
SSPRO        EEEEE      EEEEEE                 EEEEEEE   HHHHHHHHHHHH       EEE
YASPIN
JNET         EEEEE      EEEEEE                 EEHHHHHHHHHHHHHHHHHH       EEEE
PSIPRED      EEEEE      EEEEEE          EE     EEEEEEEEE HHHHHHHHHHHHHH
             ------------------------------------------------------------------------
SYMPRED      EEEEE      EEEEEE                 EEEEEEE   HHHHHHHHHHHH       EEE

Ruler       ......
Sequence    VGKDVF
PROF
SSPRO
YASPIN
JNET
PSIPRED
            ------
SYMPRED
```

3.6 Protein tertiary structure:

Human reference PHD2 protein tertiary structure was predicted using Robetta server by homology modelling. The best template available template (4bqw) selected for homology modelling of Human EGLN1/PHD2 protein have a query coverage 57% and percentage identity 100%. Robetta predicted protein tertiary structure model is shown in figure 3.5. Human disease-causing (Erythrocytosis Familial 3) mutants were also modelled using Human reference PHD2 predicted protein as a template. Similarly, other primates (Tibetan human, hominin ancestor, Chimpanzee and hominid ancestpr) PHD2 proteins were also modelled. Ancestral sequences were obtained from MEGA7.

Snow leopard and Markhor PHD2 proteins were also modelled by using Robetta server and are shown in Figure 3.6 and Figure 3.7 respectively.

Figure 3.5 Predicted structure of Human Reference PHD2 protein. Protein is represented in Cyan color. Domains are color-coded. Zinc-finger MYND type domain (21-58 amino acids) is yellow colored. Alanine-rich region is shown in green color. Fe (+2) OG dioxygenase domain (291-392) is orange colored.

Figure 3.6 Predicted structure of Snow leopard PHD2 protein. Protein is represented in Cyan color. Domains are color-coded. Zinc-finger MYND type domain (21-58 amino acids) is yellow colored. Alanine-rich region (78-165 amino acids) is shown in green color. Fe (+2) OG dioxygenase domain (306-404 amino acids) is orange colored.

Figure 3.7 Predicted structure of Markhor PHD2 protein. Protein is represented in Cyan color. Domains are color-coded. Zinc-finger MYND type domain (21-58 amino acids) is yellow colored. Alanine-rich region (83-141 amino acids) is shown in green color. Fe (+2) OG dioxygenase domain (290-388 amino acids) is orange colored.

3.7 Structure evaluation:

Tertiary structures of modelled proteins were evaluated using SAVES server.

Verify 3D gives us the average 3D-1D score (in percentage) of residues in a protein. ERRAT indicates the overall quality of protein. And Ramachandran plots obtained from PROCHECK indicates the number of residues present in allowed/ favorable regions as well as the number of residues that are present in disallowed/ unfavorable regions. Results are shown in Table 3.5.

Table 3.5 Results of protein tertiary structures evaluated from Verify 3D, ERRAT and PROCHECK.

Protein evaluated	Verify 3D (Residues	ERRAT (overall quality score)	Residues in allowed region	Residues in disallowed region

	average 3D-1D score)			
Human	91.55%	96.154	304 (88.6%)	1 (0.3%)
Human mutant (P317R)	86.38%	96.875	311 (90.4%)	1 (0.3%)
Human mutant (R371H)	93.43%	96.403	304 (88.6%)	1 (0.3%)
Human mutant (H374R)	95.31%	98.082	301 (87.8%)	0 (0.0%)
Tibetan Human	92.49%	97.122	302 (88%)	1 (0.3%)
Hominin ancestor	90.147%	95.683	301 (87.2%)	2 (0.6%)
Hominid ancestor	88.73%	93.510	304 (88.1%)	1 (0.3%)
Chimpanzee	90.38%	96.86	304 (88.1%)	2 (0.6%)
Snow leopard	89.27%	94.614	303 (85.4%)	1 (0.3%)
Tiger	91.32%	94.86	306 (86.2%)	2 (0.6%)
Panthera ancestor	91.32%	94.159	299 (84.2%)	1 (0.3%)
Domestic cat	92.26%	94.172	309 (86.6%)	1 (0.3%)
Markhor	92.18%	93.947	300 (88%)	1 (0.3%)
Goat	93.13%	95.122	300 (88%)	1 (0.3%)

Caprinae ancestor	88.86%	93.128	305 (87.4%)	2 (0.6%)
Siberian musk deer	90.49%	92.654	307 (87.5%)	1 (0.3%)

Ramachandran plots of all modelled proteins are shown in figure 3.8.

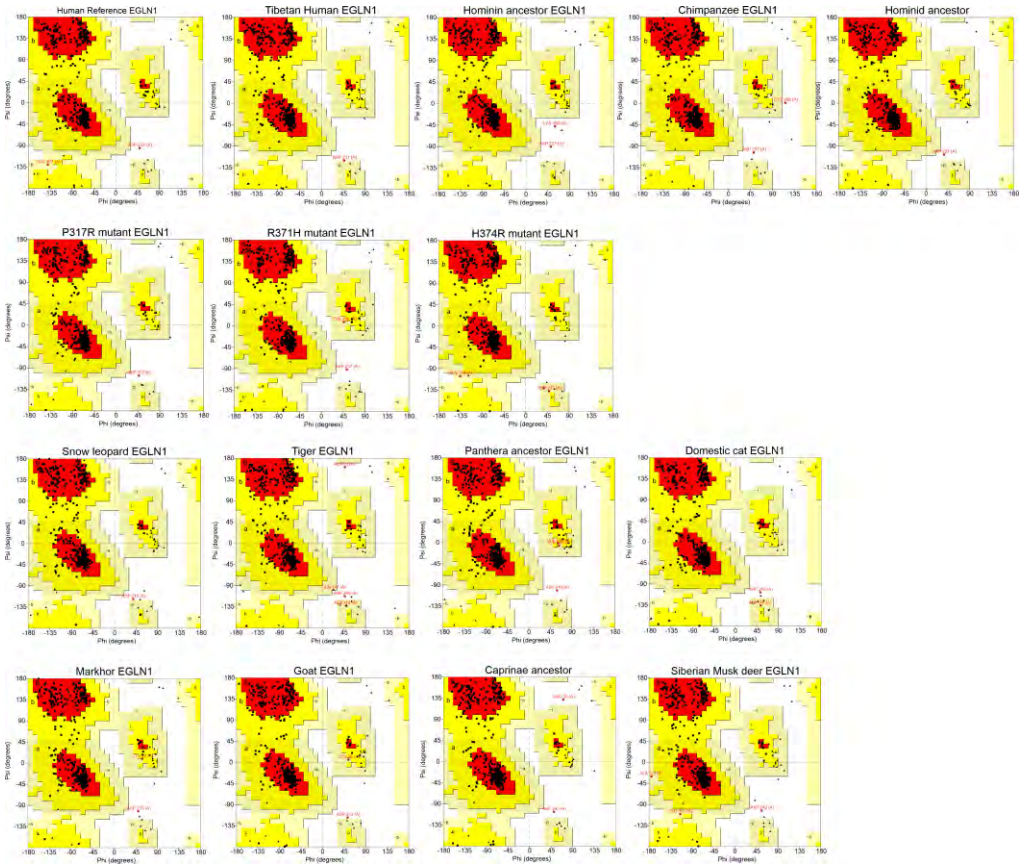


Figure 3.8 Ramachandran plots of all proteins predicted by Robetta server and energy minimized by YASARA energy minimization server. 1st panel shows Ramachandran plots of predicted structures of primate proteins. 2nd panel shows Ramachandran plots of predicted structures of human disease-causing mutants involved in Erythrocytosis Familial 3. 3rd panel shows Ramachandran plots of predicted carnivore proteins. 4th panel shows Ramachandran plots of predicted artiodactyls proteins. Result statistics are mentioned in Table 3.5

ERRAT plots predicted primate protein structures are shown in figure 3.9. ERRAT plots predicted human EGLN1 mutant protein structures are shown in figure 3.10. ERRAT plots predicted carnivore protein structures are shown in figure 3.11. ERRAT plots predicted artiodactyl protein structures are shown in figure 3.12.

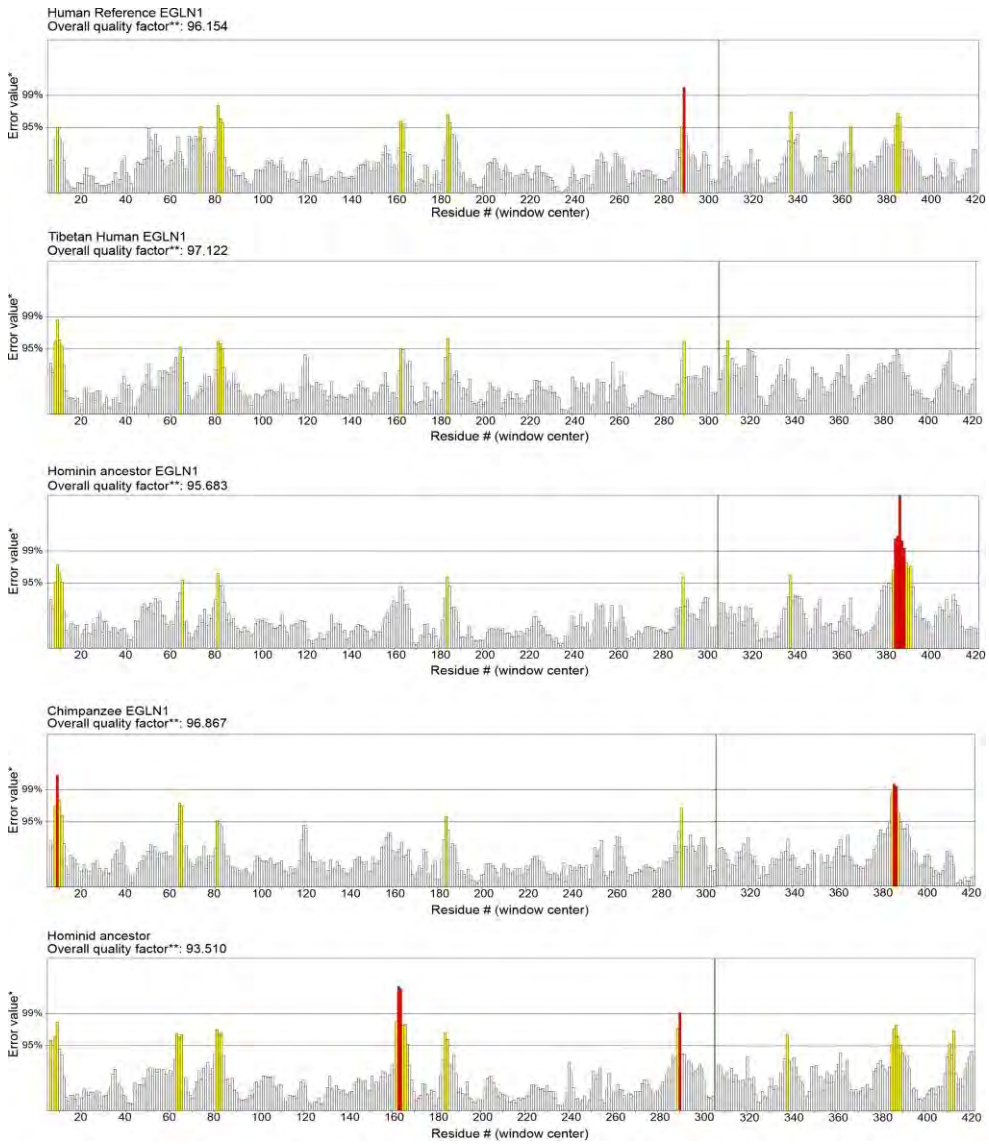


Figure 3.9 ERRAT plots of primate proteins predicted by Robetta server and energy minimized by YASARA energy minimization server. Overall quality factor of all these proteins is > 90%.

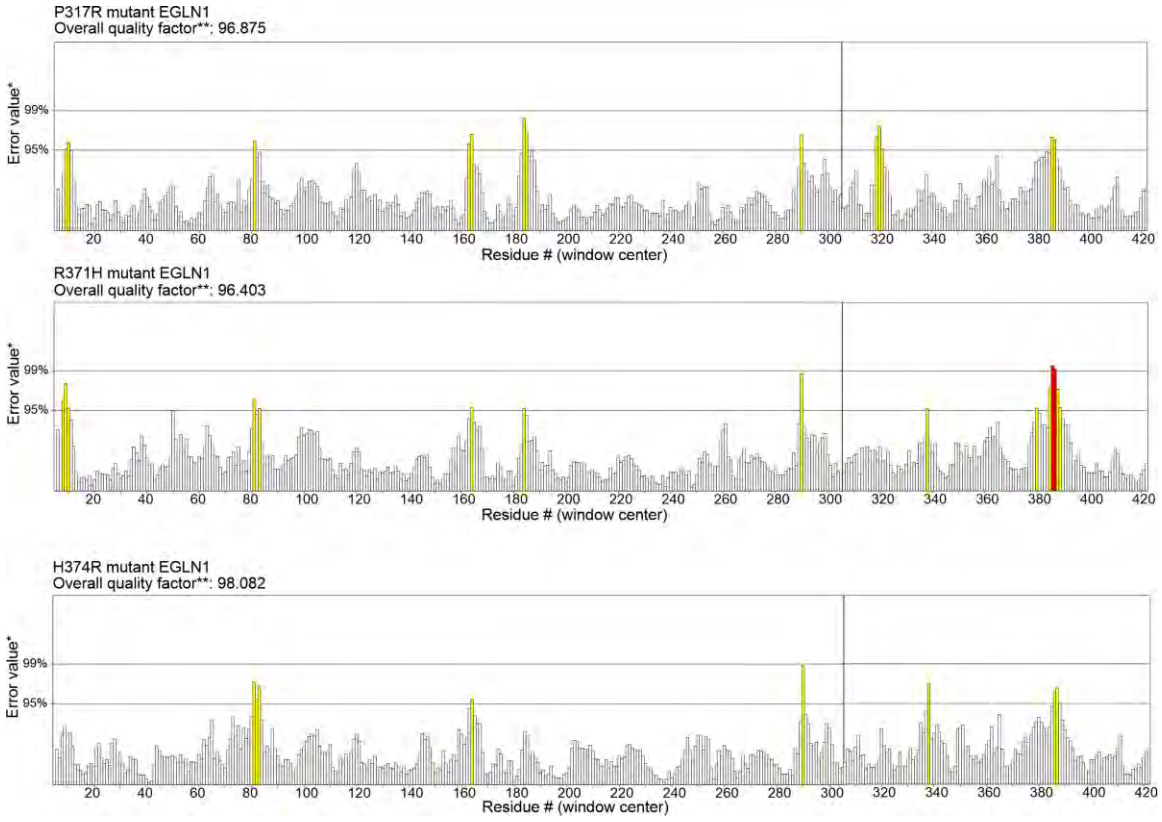


Figure 3.10 ERRAT plots of Human mutants (Erythrocytosis Familial 3) proteins predicted by Robetta server and energy minimized by YASARA energy minimization server. Overall quality factor of all these proteins is > 95%.

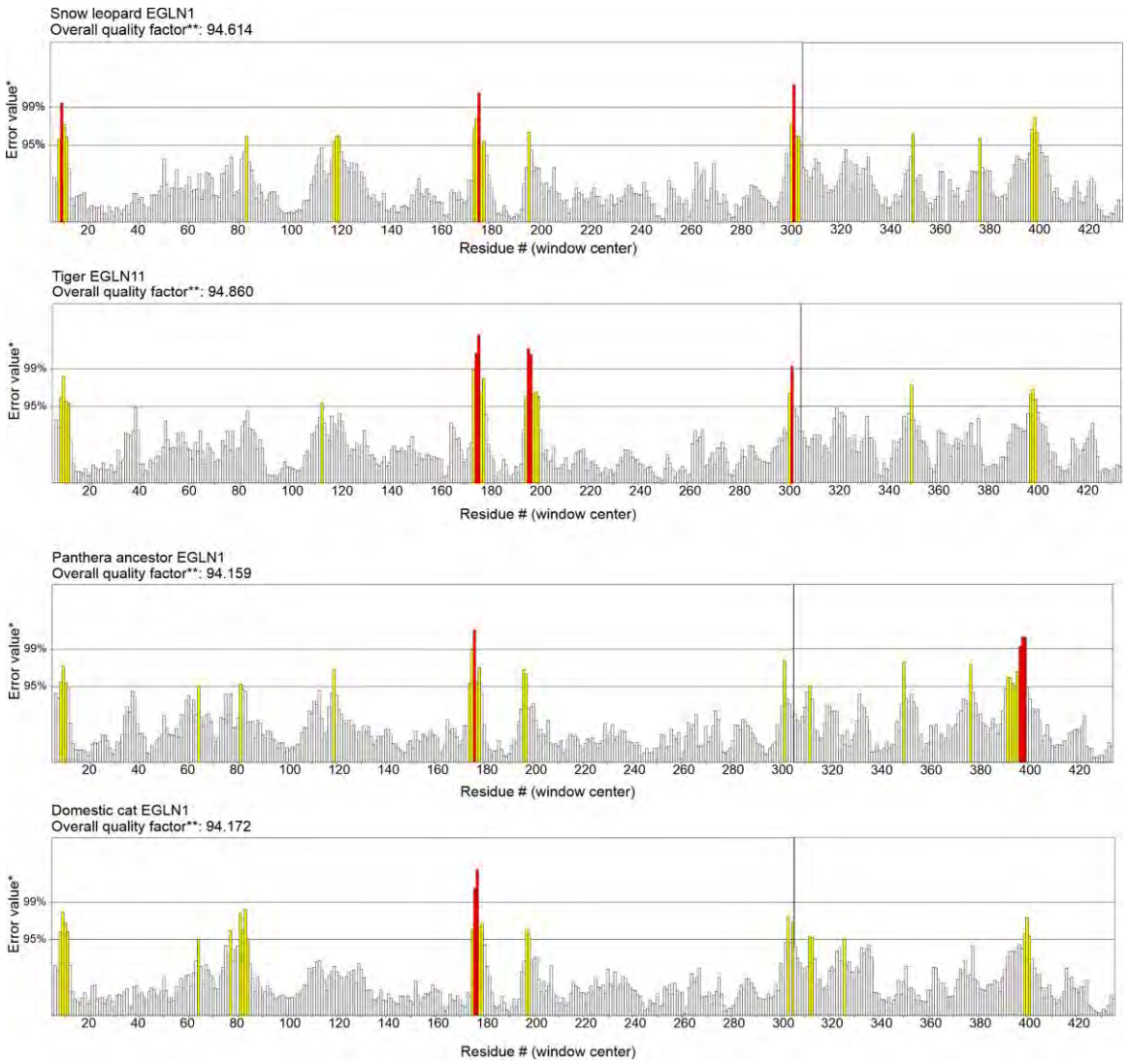


Figure 3.11 ERRAT plots of carnivore proteins predicted by Robetta server and energy minimized by YASARA energy minimization server. Overall quality factor of all these proteins is > 90%.

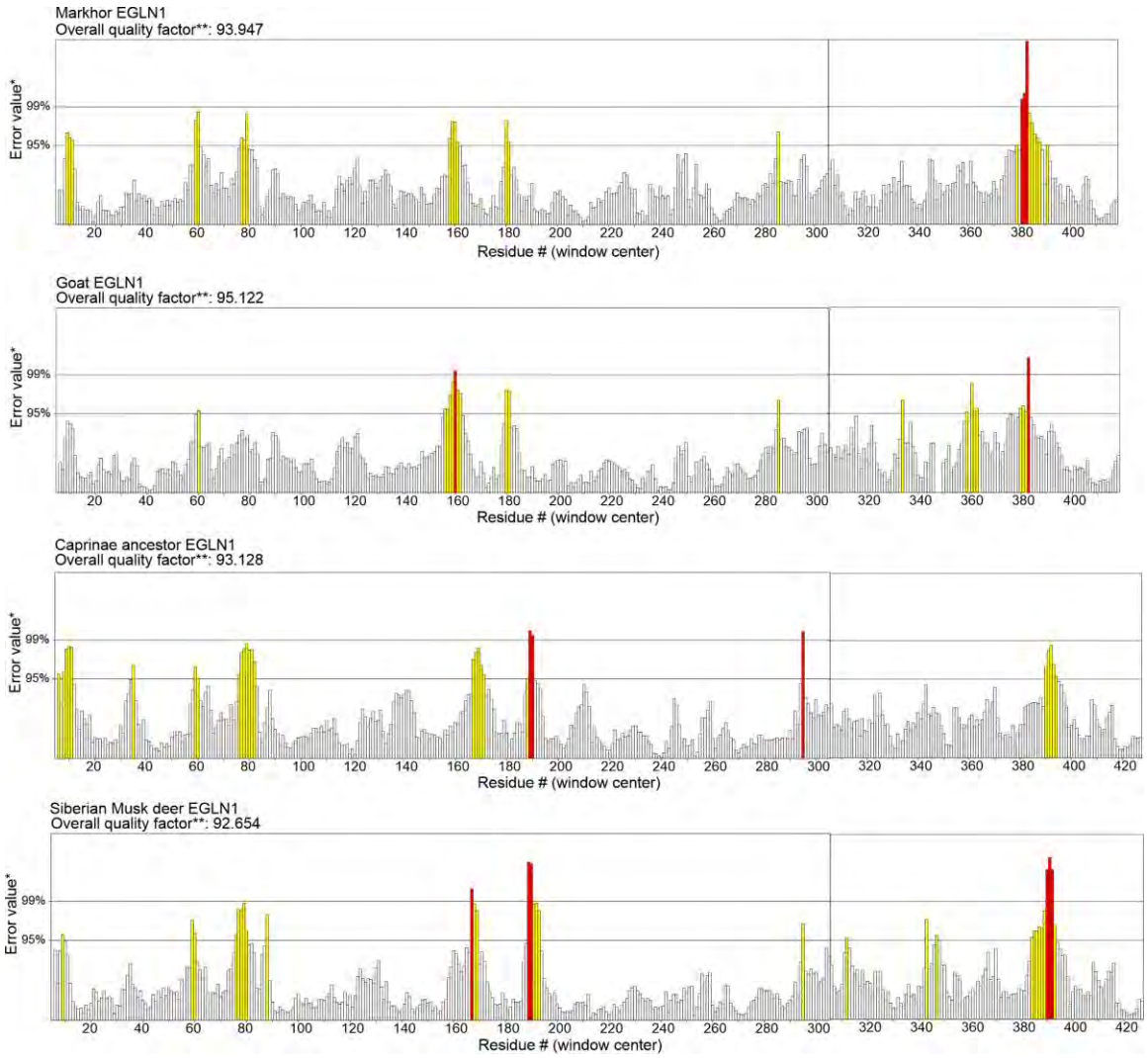


Figure 3.12 ERRAT plots of artiodactyl proteins predicted by Robetta server and energy minimized by YASARA energy minimization server. Overall quality factor of all these proteins is > 90%.

3.8 Superimposition of tertiary structures:

Superimposition of protein tertiary models was performed in order to evaluate the structural deviations caused by changes in proteins primary structure. Superimposition was performed in four steps. In first step, primate protein structures were superimposed to evaluate the changes that occurred in protein structures specifically in primates. In second step, Human reference protein was superimposed with three mutants (that were reported to be a cause of Erythrocytosis Familial 3 disease) to evaluate the changes that occurred in

protein structures by these point mutations. In third step, protein structures from various carnivores were superimposed with snow leopard predicted EGLN1 protein to evaluate the changes in snow leopard protein structure as compared to other carnivores. In last step, protein structures from various artiodactyl species were superimposed with Capra falconeri (Markhor) predicted EGLN1 protein to evaluate the changes in Markhor protein structure as compared to other artiodactyls. Protein structures are superimposed by UCSF chimera.

Figure 3.14 shows superimposition of primate protein structures with human reference (human ref.) Domains of Human reference protein structure are depicted by different colors. Zinc-finger MYND type domain (21-58 amino acids) is yellow colored. Alanine-rich region is shown in green color. Fe (+2) OG dioxygenase domain (291-392) is orange colored. Deviated residues were depicted with red color. The major changes in torsion angles (residue number), changes in secondary structures and RMSD of each superimposed structure is given in Table 3.6.

Figure 3.15 shows superimposition of human reference (wild type) PHD2 protein structures with human PHD2 mutant proteins (P317R, R371H, H374R). The major changes in torsion angles (residue number), changes in secondary structures and RMSD of each superimposed structure is given in Table 3.7.

Figure 3.16 shows superimposition of carnivore predicted protein structures with snow leopard PHD2 protein. Two snow leopard specific substitutions have occurred but no specific substitution was observed in other carnivores. Insertion of one or two alanine residues has occurred in other carnivore PHD2 proteins. Domains of Snow leopard protein structure are depicted by different colors. Zinc-finger MYND type domain (21-58 amino acids) is yellow colored. Alanine-rich region (78-165 amino acids) is shown in green color. Fe (+2) OG dioxygenase domain (306-404 amino acids) is orange colored. The major changes in torsion angles (residue number), changes in secondary structures and RMSD of each superimposed structure is given in Table 3.8.

Figure 3.17 shows superimposition of artiodactyl predicted protein structures with Markhor PHD2 protein. One substation S416L is observed between goat and markhor. After performing multiple sequence alignments of artiodactyl proteins and ancestral sequences prediction from MEGA7, it is concluded that this substitution is specific to goat PHD2 protein. In markhor, no substitution occurred as compared to other artiodactyls.

Domains of Markhor PHD2 protein structure is depicted by different colors. Domains are color-coded. Zinc-finger MYND type domain (21-58 amino acids) is yellow colored. Alanine-rich region (83-141 amino acids) is shown in green color. Fe (+2) OG dioxygenase domain (290-388 amino acids) is orange colored. The major changes in torsion angles (residue number), changes in secondary structures and RMSD of each superimposed structure is given in Table 3.9.

A tree (Figure 3.13) representing substitutions in primate, artiodactyls and carnivores was constructed. Ancestral substitutions and species-specific substitutions are verified by Multiple sequence alignments and Prediction of ancestral sequences from MEGA7. Multiple sequence alignment of primates is performed in order to find the differences between proteins. No difference was observed between Neanderthal and Human Reference proteins. Only one Chimpanzee specific substitution was observed. Multiple sequence alignment of carnivores was performed to find differences between proteins. Snow leopard EGLN1 protein has two specific substitutions that are not observed in any other carnivore. Multiple sequence alignments of Artiodactyls suggested that Markhor as compared to Goat has only one substitution. This substitution was observed to be Goat specific.

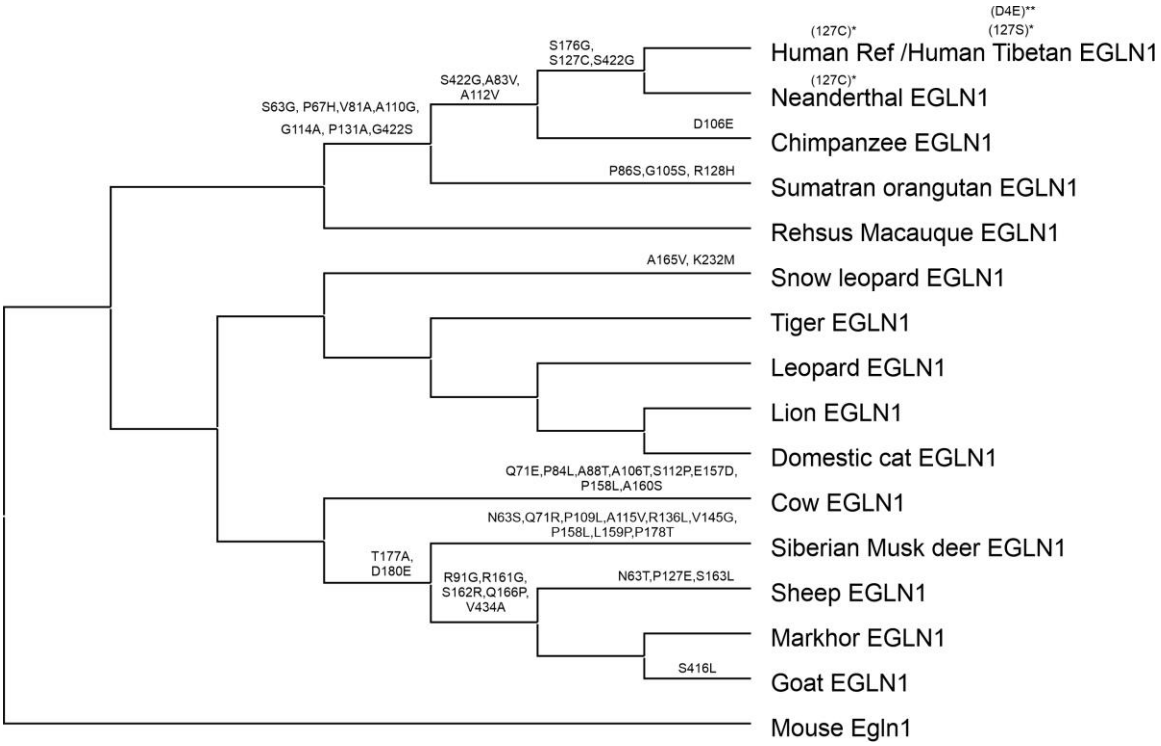


Figure 3.13 Substitution tree representing structural evolution of EGLN1 protein among three groups: Primates, Artiodactyls and Carnivores. No difference was observed between Neanderthal and Human Reference proteins. The two substitutions in Tibetan human EGLN1 protein reported in previous studies were observed when compared to Human Reference protein. These substitutions are shown by D4E** and C127S*. Ancestral substitutions are mentioned at the nodes and specie-specific substitutions are mentioned on their respective branch.

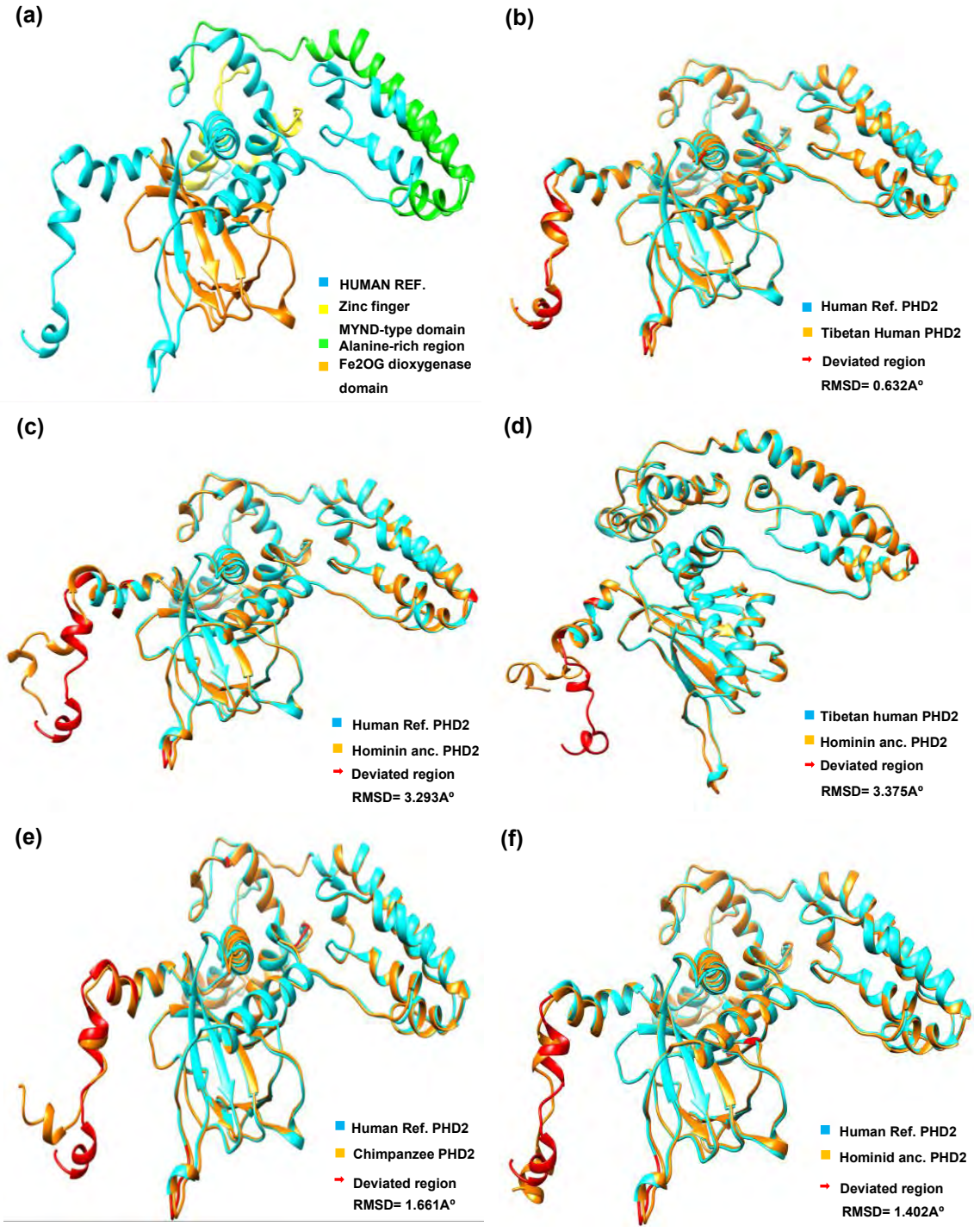


Figure 3.14 Structural evolution of PHD2 protein in primates (a) Predicted structure of Human Ref. PHD2 protein. Domains are color-coded. **(b)**Superimposed Human ref. and Tibetan human structure. **(c)** Superimposed Tibetan human and Hominin anc. structure. **(d)** Superimposed Human Ref. and Hominin anc. structure. **(e)** Superimposed Human Ref. and Chimpanzee structure. **(f)** Superimposed Human Ref. and Hominid anc. structure.

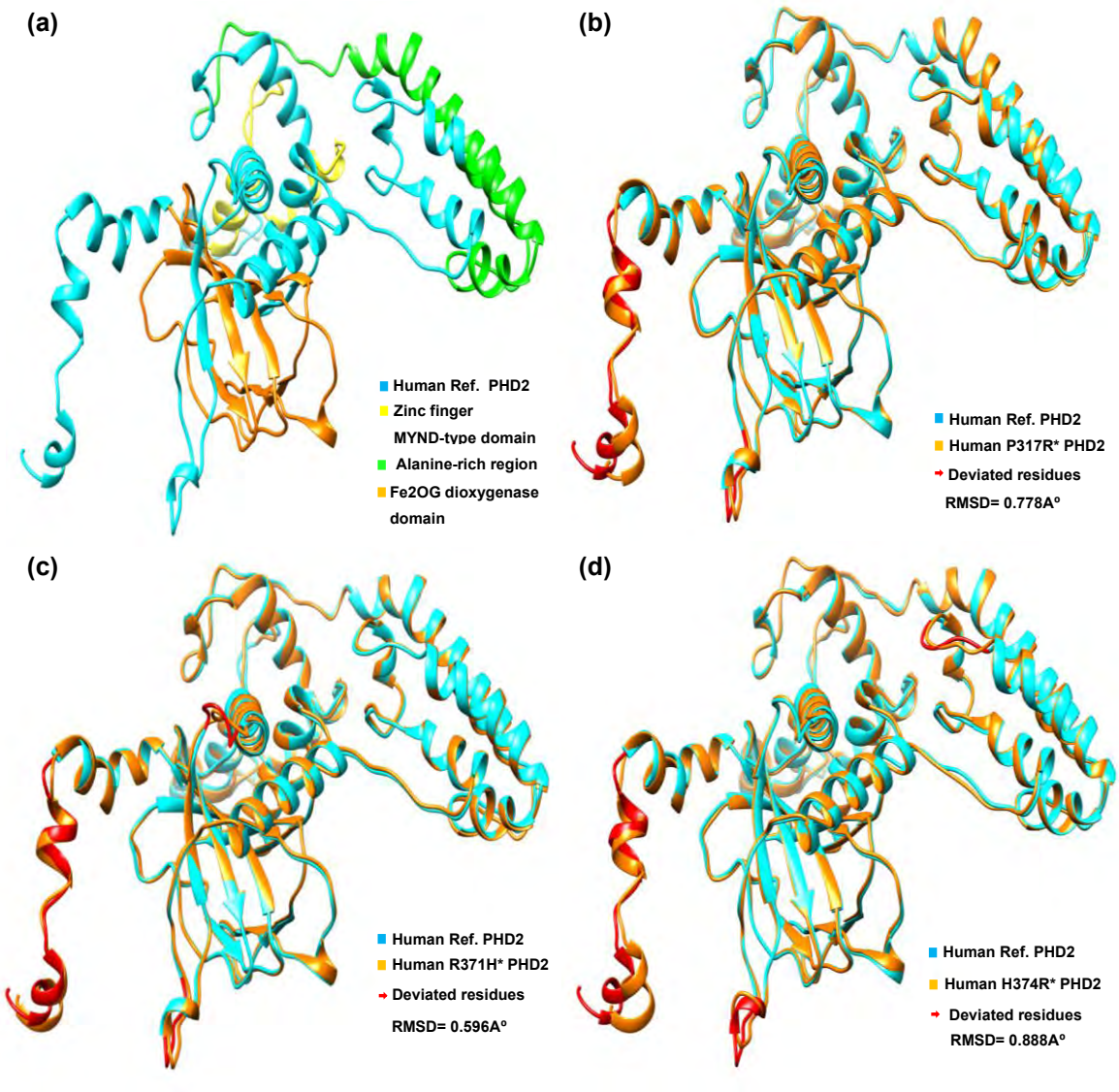


Figure 3.15 Protein structural deviations in Familial erythrocytosis 3 associated mutant versions of EGLN1. a) Structure of Human reference EGLN1/PHD2. **(b)** Structural superimposition of Human reference and P317R mutated model**. (c)** Structural superimposition of Human reference and R371H mutated model. (d) Structural superimposition of Human reference and H374R mutated model. Deviated residues are depicted in red color. Mutant sequences were generated by manually changing the amino acid at a specific site. Proteins tertiary structures are predicted by Robetta server.

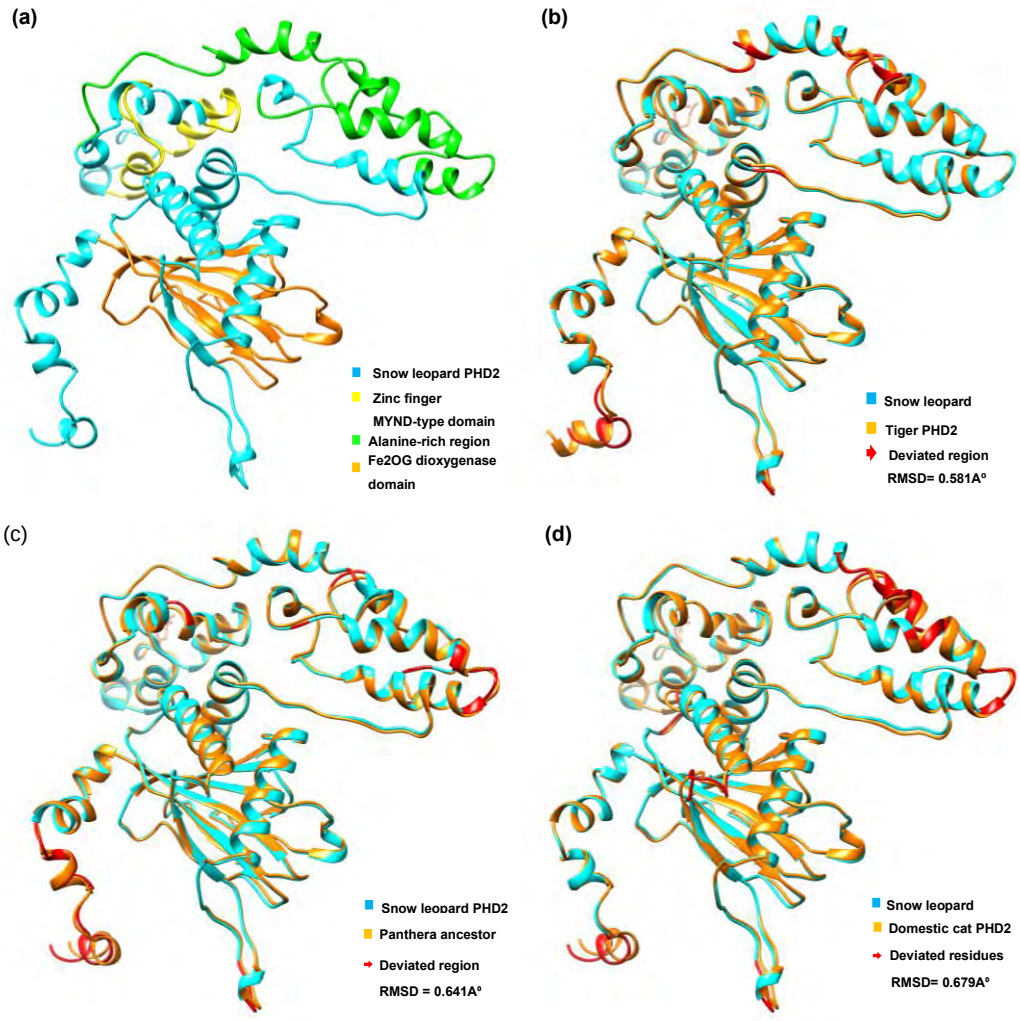


Figure 3.16 Structural evolution of PHD2 protein in carnivores. Structural divergence of snow leopard PHD2 from other carnivores. **(a)** Predicted structure of Snow leopard PHD2 protein. Domains are color-coded. **(b)** Superimposed Snow leopard and Tiger structure. **(c)** Superimposed Snow leopard and Panthera ancestor structure. **(d)** Superimposed Snow leopard and Domestic cat structure. Deviated residues are depicted in red color.

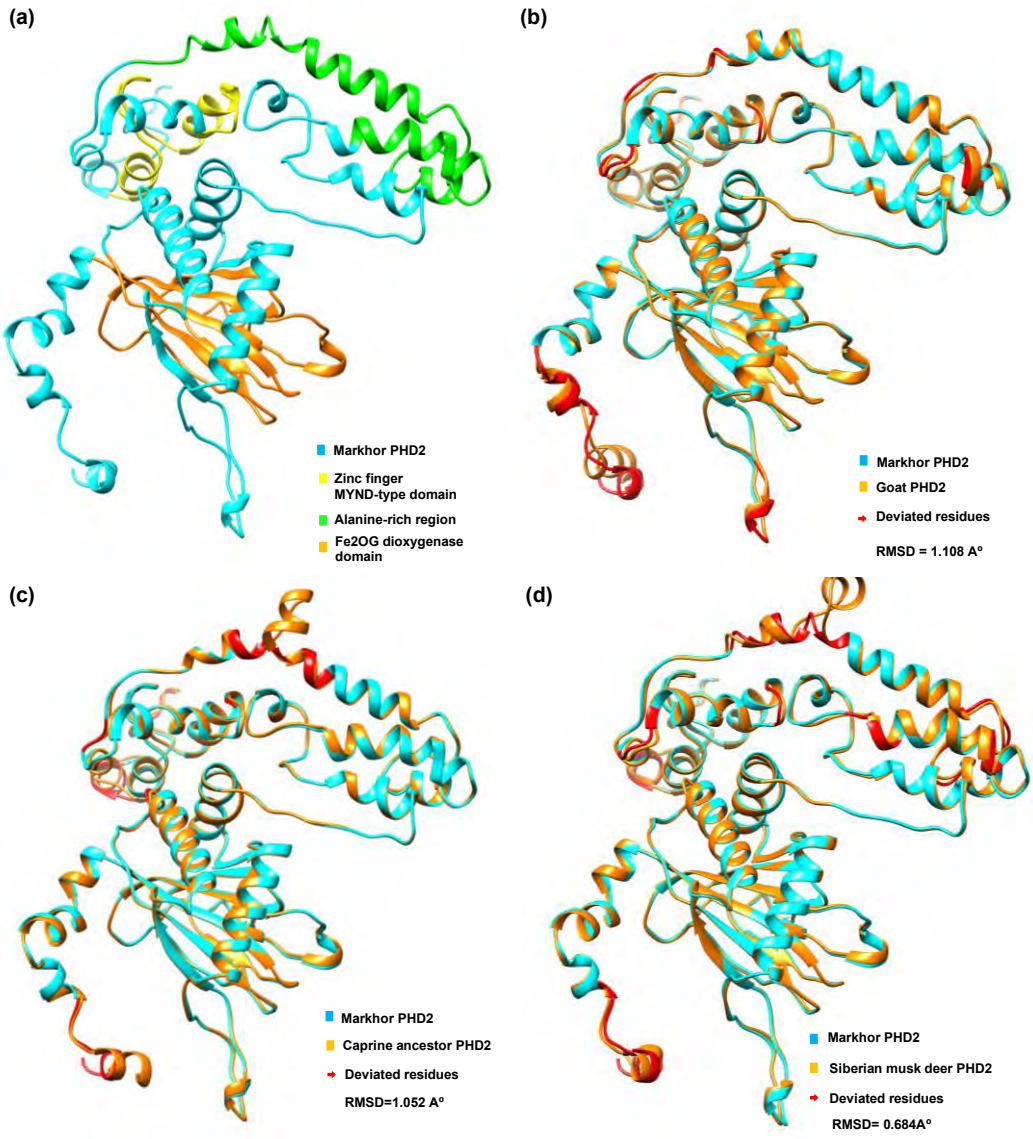


Figure 3.17 Structural evolution of PHD2 protein in Artiodactyls. Structural divergence of Capra falconeri (Markhor) PHD2 from other artiodactyls. **(a)** Predicted structure of Markhor PHD2 protein. Domains are color-coded. **(b)** Superimposed Markhor and Goat structure. **(c)** Superimposed Markhor and Caprine ancestor structure. **(d)** Superimposed Markhor and Siberian musk deer structure. Deviated residues are depicted in red color.

Table 3.6 Structural comparisons of predicted PHD2 proteins in primates.

Comparison between lineages	Amino acid Differences	RMSD value	Changes in secondary structure	Changes in backbone torsion angles(residues)	Critical regions
Human Ref. ↔ Human Tibetan	D4E, C127S.	0.632A◦	**Helix↔loop:** Gly409, Val410.	241-246, 407-426.	241-246($beta_2beta_3$ 'finger-like' loop.
Human Tibetan↔ Hominin ancestor	E4D, G176S, G422S.	3.375A◦	**Helix↔loop:** Gly133, Ala397, Arg398, Lys400, Arg411, Val412, Lys 423, Asp 424. **Loop↔Sheet:** Asp211. **Loop↔Helix:** Asn415	1, 50-56, 210-215, 132-133, 189-190, 204-205.241-246, 399-426.	50-56(MYND-type zinc finger domain), 241-246($beta_2beta_3$ 'finger-like' loop.
Human Ref.↔ Hominin ancestor	C127S, G176S, G422S.	3.293A◦	**Helix↔loop:** Gly133, Ala397, Arg398, Lys400, Gly409, Val 410, Arg411, Val412, Lys 423, Asp 424. **Loop↔Sheet:** Asp211. **Loop↔Helix:** Glu407, Asn415	243-246, 406-426.	243-246($beta_2beta_3$ 'finger-like' loop.

Human Ref.↔Chimpanzee	D106E, C127S, G176S, G422S.	1.661A◦	**Helix↔loop:** Val69, Gly406, Gly409, Val410, Leu414, Lys423, Asp424.	54-56, 241-246, 401-426.	54-56(MYND-type zinc finger domain), 241-246($beta_2beta_3$ 'finger-like' loop.
Human Ref.↔Hominid ancestor	V83A, V112A, C127S, G176S, G422S.	1.402A◦	**Helix↔loop:** Asn203, Lys204, Glu413, Leu414. **Sheet↔loop:** Arg362. **Loop↔Helix:** Val425.	241-246,407-426.	241-246($beta_2beta_3$ 'finger-like' loop.

Table 3.7 Structural comparisons of disease-causing mutant versions of PHD2 protein.

Comparison between lineages	Amino acid Differences	RMSD value	Changes in secondary structure	Changes in backbone torsion angles(residues)	Critical regions
Human Ref. ↔ Human P317R	P317R.	0.778A◦	-	241-246, 407-426.	241-246($beta_2beta_3$' finger-like' loop.
Human Ref. ↔ Human R371H	R371H.	0.596A◦	**Helix↔loop:** Gly177. **Loop↔Helix:** Val425.	261-266, 241-246, 407-426.	241-246($beta_2beta_3$' finger-like' loop.
Human Ref. ↔ Human H374R	H374R.	0.888A◦	**Helix↔loop:** Val69, Gly73, His74, Ser75, Ser247, Ser248, Lys249.	149-153, 242-249, 407-426.	242-249($beta_2beta_3$' finger-like' loop.

			Loop↔Helix: Val425.		

Table 3.8 Structural comparisons of predicted PHD2 proteins in carnivores.

Comparison between lineages	**Amino acid Differences**	**RMSD value**	**Changes in secondary structure**	**Changes in backbone torsion angles(residues)**	**Critical regions**
Snow leopard↔ Tiger	V161A, M228K	0.581A◦	**Helix↔loop:** Val109, Glu110, Asn111, Ala142. **Loop↔Helix:** Arg94, Ala106, Ala107, Gln147, Val437.	1-5, 89-95, 106-112, 117-118, 200-201, 255-258, 427-438.	255-258($beta_2beta_3$ 'finger-like' loop.
Snow leopard↔ Panthera ancestor	V161A, M228K	0.641A◦	**Helix↔loop:** Gly63, Val109, Glu110, Asn111, Glu160. **Loop↔Helix:** Gln147, Ala409, Arg410.	1-5, 63-64, 108, 117-118, 131-134, 142-148, 163, 255-258, 419-438.	255-258($beta_2beta_3$ 'finger-like' loop.

Snow leopard↔ Domestic Cat	153(Insertion A), P101Q, V162A, M228K.	0.679A◦	**Helix↔loop:** Val109, Glu110, Asn111. **Loop↔Helix:** Gln147.	1-4, 92-93, 106-111, 117-125, 142-147, 152-155, 255-258, 272-278, 296-297, 300-301, 430-438.	255-258($beta_2beta_3$ 'finger-like' loop, 296-297(Fe2OG dioxygenase domain).

Table 3.9 Structural comparisons of predicted PHD2 proteins in artiodactyls.

Comparison between lineages	Amino acid Differences	RMSD value	Changes in secondary structure	Changes in backbone torsion angles(residues)	Critical regions
Markhor↔ Goat	S416L	1.108A◦	**Helix↔loop:** Leu400, Leu410. **Loop↔Helix:** Ala93, Ala94, Gly128, Pro413, Ser414, Ala421. **Loop↔Sheet:** Asp207. **Sheet↔loop:** Arg318, Lys319, Phe387.	2-5, 54-56, 70-76, 81-82, 86-87, 130-132, 207-209, 241-246.	54-56(MYND-type zinc finger domain), 241-246($beta_2beta_3$'finger-like' loop.
Markhor↔ Caprinae ancestor	PRRDRAVEA Insertion (94-102).	1.052A◦	**Helix↔loop:** Tyr20. **Loop↔Helix:** Ala86, Gly87, Ala103, Gly95, Ala421.	1-5, 14-21, 37,54, 78-79, 90-99, 281-282, 411-422.	54(MYND-type zinc finger domain).

Markhor↔ Siberian musk deer	PRRDRAV EA Insertion (94-102), N63S, Q71P, G87R, P145L, L146P, G148A, R149D, D153Q, P165T, A421V.	0.684A◦	**Helix↔loop:** Thr69, Ala92, Glu95, Pro96, Gln129, Gly130, Arg131.	15-20, 54-56, 69-73, 86-96, 109-119, 130-132, 142-145, 410-422.	54-56(MYND-type zinc finger domain),

3.9 Interaction analysis:

The structure of HIF1α protein was predicted by comparative modelling. PDB ID (4zpr) was selected as a template for prediction of HIF1α tertiary structure. The overall quality factor of predicted HIF1α protein was 88 (calculated by ERRAT). Ramachandran plot indicated 91.5% (677) residues in most favored regions and 0.3% (2) residues in disallowed regions. Predicted protein tertiary structure of HIF1α is shown in Figure 3.18. Domains and motifs are color-coded. bHLH (17-70 amino acids) domain is shown in cyan color. PAS1 (85-158 amino acids) domain is shown in yellow color. PAS2 (228-298 amino acids) domain is shown in green. PAC (302-345 amino acids) domain is shown in Dark blue color. ODD (401-603 amino acids) domain is shown in pink color. Nuclear localization signal (718-721 amino acids) is shown in red. Image is retrieved from UCSF Chimera.

Ramachandran plot and overall quality factor plot is shown in Figure 3.19.

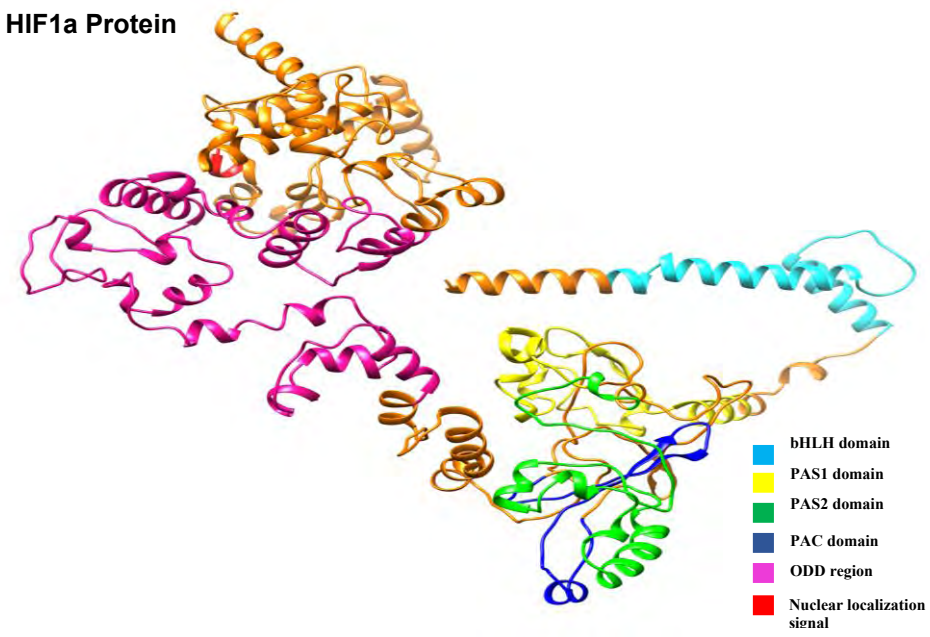


Figure 3.18 Predicted structure of HIF1α (Predicted by Robetta server using 4ZPR as a template). Protein is represented in orange color. Domains and motifs are color-coded.

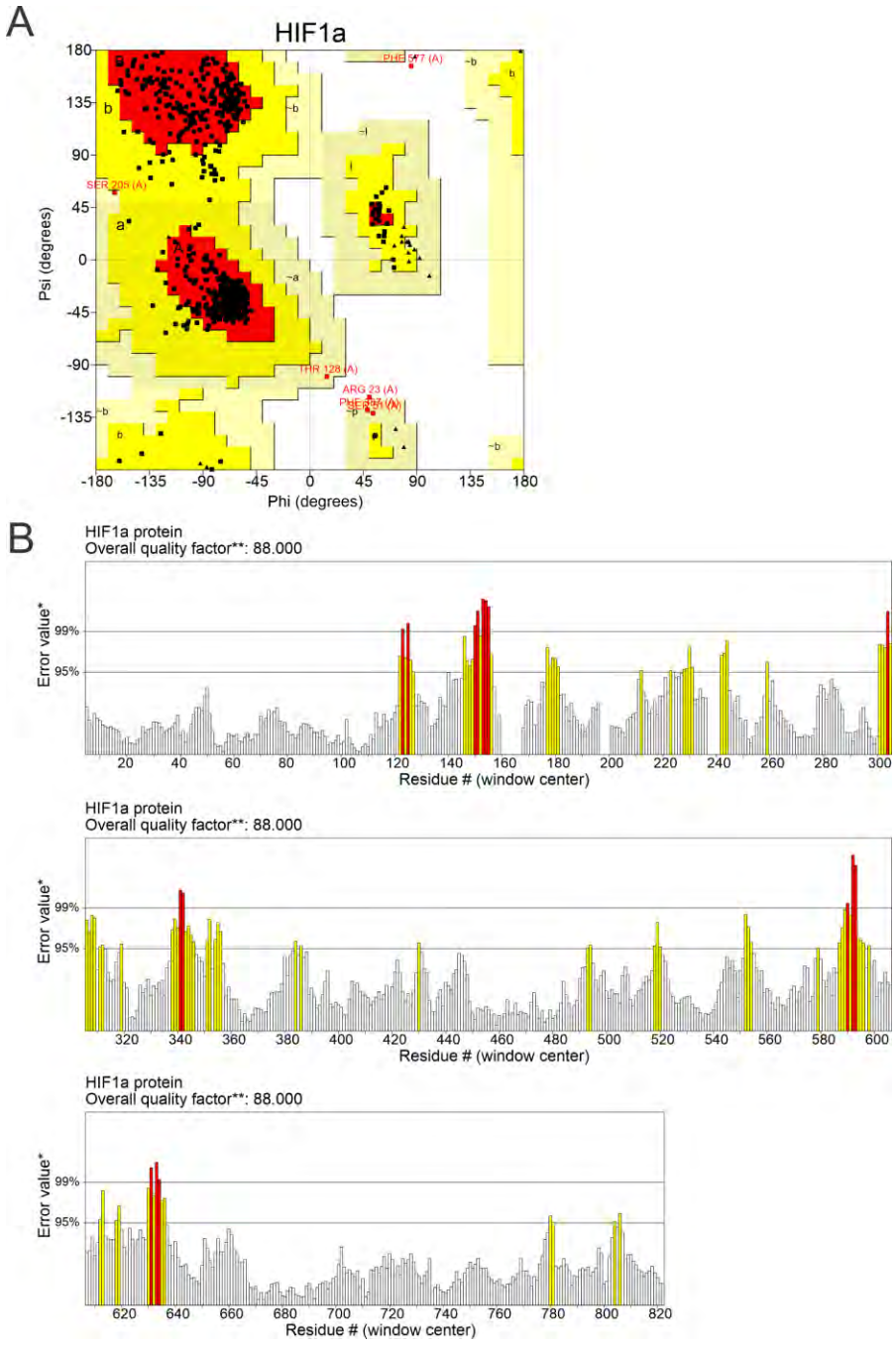


Figure 3.19 Evaluation of 3D models of HIF1α protein. (A) Ramachandran plot of Human HIF1α. **(B)** Quality of model examined with the help of ERRAT. Overall quality factor is expressed as percentage of the protein for which the calculated error value falls below 95% rejection limit, calculated by ERRAT.

Protein-protein interactions of EGLN1 protein (wild type and mutants) with HIF1α was performed by ClusPro. The interaction among EGLN1 (wild type and mutant H374R) exhibited same residues involved in hydrogen bonding with HIF1α protein. The interaction among EGLN1 (mutant P317R and mutant R371H) and HIF1α protein exhibited significant shift in hydrogen bonded residues when compared to wild type protein. The interactions are represented in Figure 3.20.

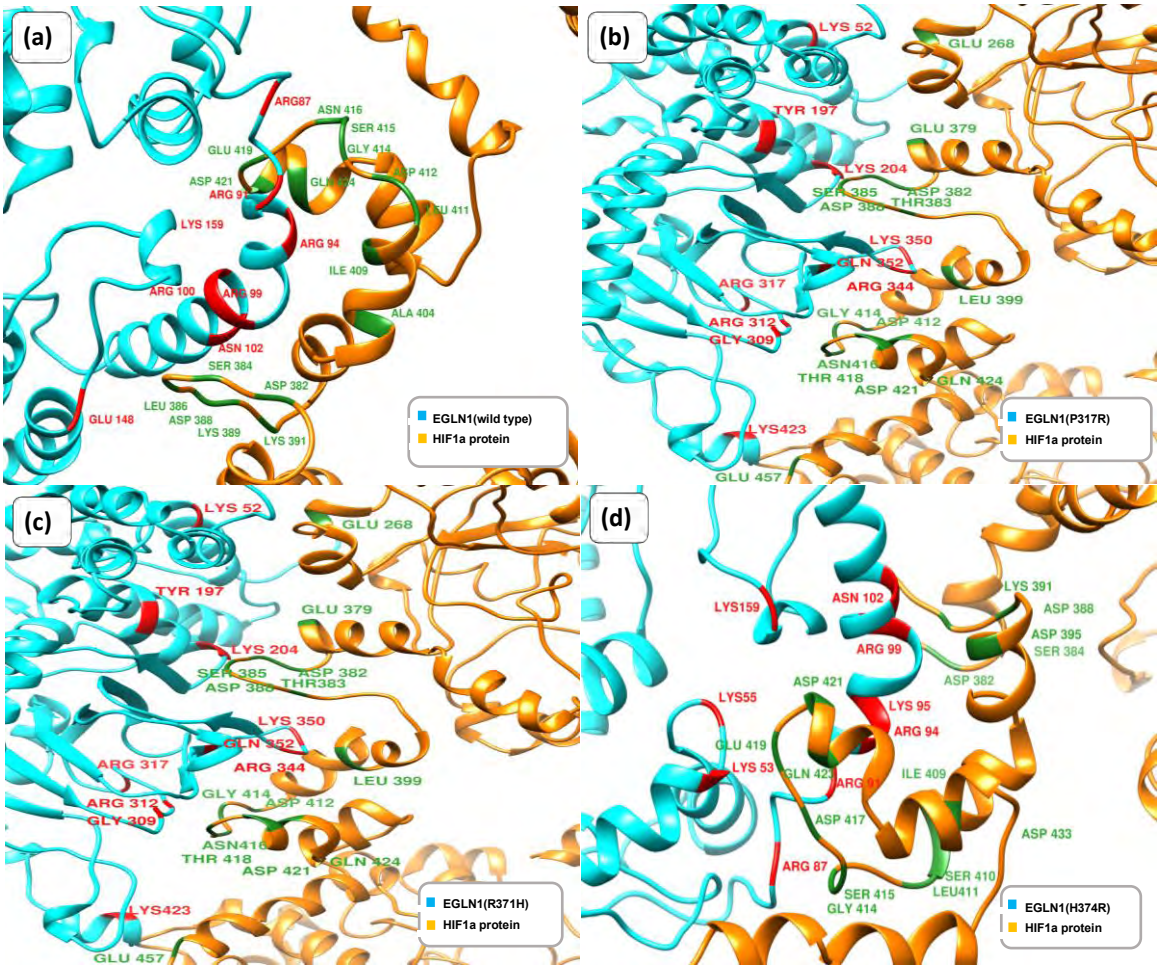


Figure 3.20 Interaction analysis of wild type EGLN1/PHD2 protein and its mutants with HIF1α. (a) Interactions among Human (wt) EGLN1 and HIF1α. **(b)** Interactions among EGLN1 P317R and HIF1α. **(c)** Interactions among EGLN1 R371H and HIF1α. **(d)** Interactions among EGLN1 H374R and HIF1α. Human EGLN1 protein is shown in cyan color and HIF1α is shown in orange. The residues of EGLN and HIF1α involved in hydrogen bonding are represented with red and dark green color respectively.

Binding affinity of docked complexes calculated by Prodigy is in negative. This indicated strong interaction between docked molecules. Similarly final energy of docked molecules is also negative value which indicates stable interactions. Cluster members mentioned in table shows number of clusters having the same docking pose. Table represents binding affinity value, final energy of docked molecule and cluster member of each of the interaction study.

Table 3.10 Calculated Binding affinity of EGLN1(wild type and mutants) with HIF1α and total energy of docked molecules.

Docked molecule	Binding affinity (Kcal/mol)	Final energy of Docked molecules	Cluster members
EGLN1 wt↔HIF1α	-12.7	-864.2	27
EGLN1 P317R↔HIF1α	-14.8	-930.9	45
EGLN1 R371H↔HIF1α	-13.4	-973.1	42
EGLN1 H374R↔HIF1α	-11.1	-1016.1	75

2D images of interactions were obtained from LIGPLOT+ to carefully identify the residues involved in hydrogen bonding. LIGPLOT image of EGLN1(wild type and HIF1α) is given in Figure 3.21. LIGPLOT image of EGLN1(mutant P317R and HIF1α) is given in Figure 3.22. LIGPLOT image of EGLN1(mutant R371H and HIF1α) is given in Figure 3.23. LIGPLOT image of EGLN1(mutant H374R and HIF1α) is given in Figure 3.24.

Residues of EGLN1 and HIF1α are shown in blue and green color respectively. Green dotted lines represent hydrogen bonds with bond lengths mentioned. Red dotted lines represent salt bridges. Hydrophobic residues of EGLN1 and HIF1α are also represented in red and pink clouds respectively.

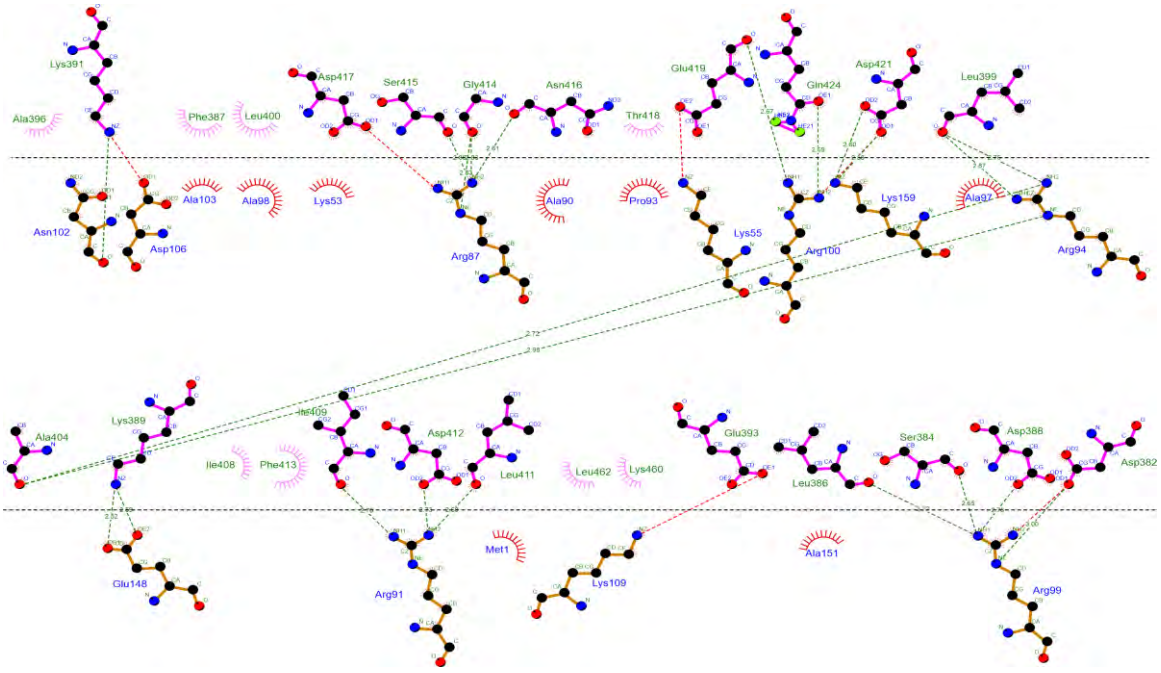


Figure 3.21 2D diagram of interactions among wild type human EGLN1 protein and human HIF1α protein retrieved using LIGPLOT.

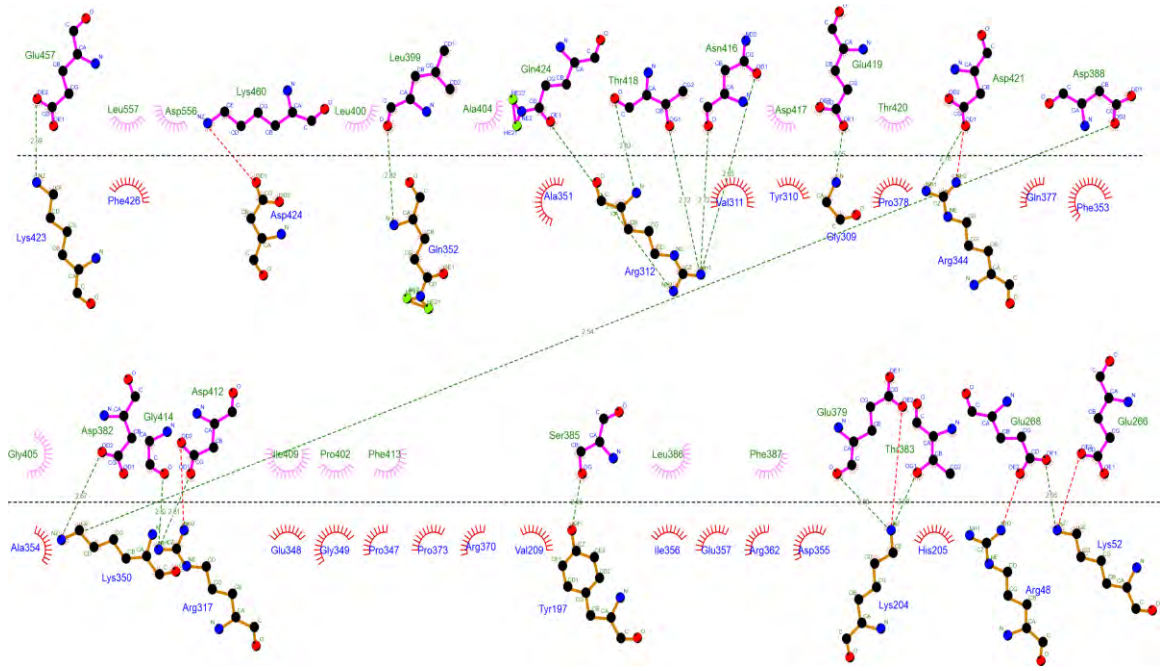


Figure 3.22 2D diagram of interactions among human EGLN1 mutant P317R protein and human HIF1α protein retrieved using LIGPLOT.

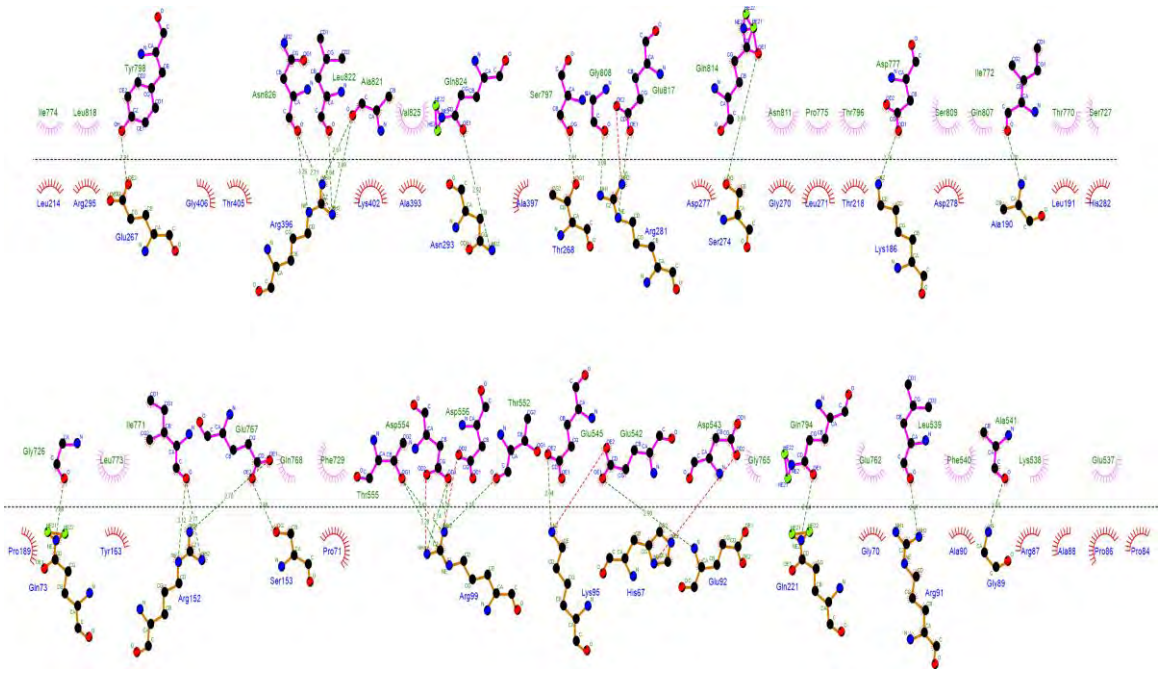


Figure 3.23 2D diagram of interactions among human EGLN1 mutant R371H protein and human HIF1α protein retrieved using LIGPLOT.

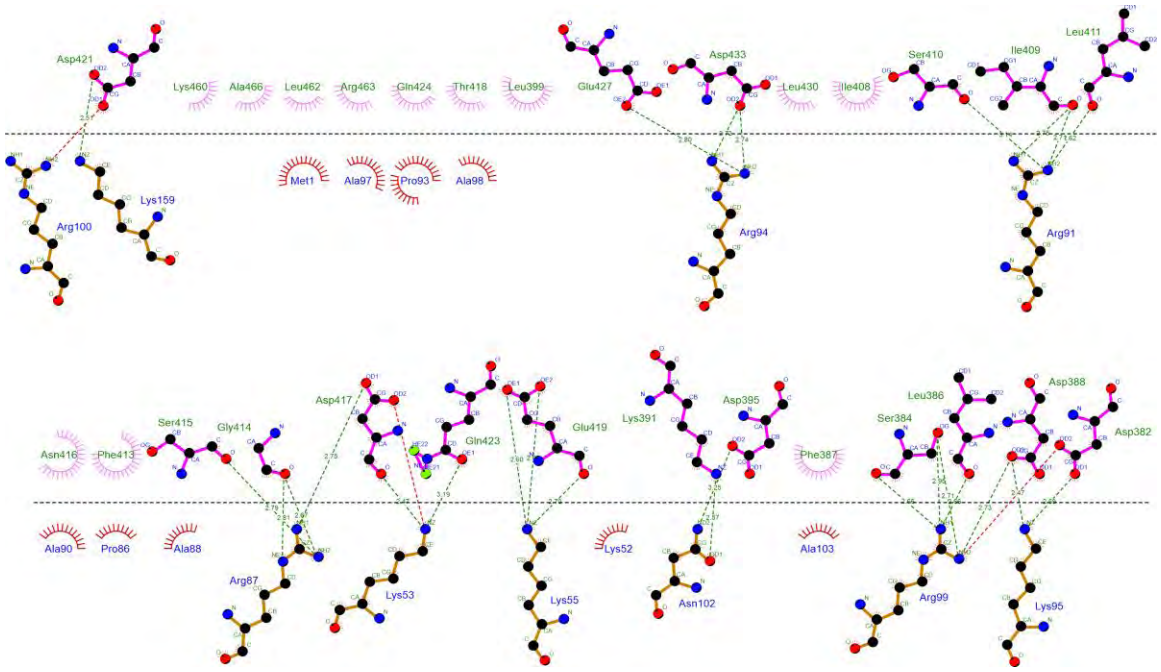


Figure 3.24 2D diagram of interactions among human EGLN1 mutant H374R protein and human HIF1α protein retrieved using LIGPLOT.

The hydrogen bonding residues of EGLN1(wild type and mutants) and HIF1α protein involved in interaction are listed in Table 3.11. The atoms involved and hydrogen bond distance (in angstrom) is also mentioned.

Table 3.11 Analysis of hydrogen bonding in angstrom in each docked molecule in corresponding to each interacting residue of both proteins.

Docked complex	Interaction residues of EGLN1		Interaction residues of HIF1a		Hydrogen bond length(A^0)
	Residue (Name, Number)	Atom (Name, Number)	Residue (Name, Number)	Atom (Name, Number)	
EGLN1 wt↔HIF1α	Lys 159	NZ, 133	Asp 421	OD2, 107	2.604
	Lys 159	NZ, 133	Asp 421	OD1, 106	2.577
	Arg 100	NH2, 55	Gln 424	OE1, 200	2.694
	Arg 100	NH1, 54	Glu 419	O, 59	2.668
	Arg 99	NH1, 10	Asp 388	OD2, 124	2.765
	Arg 99	NH1, 10	Leu 386	O, 206	2.771
	Arg 99	NH1, 10	Ser 384	O, 175	2.653
	Arg 99	NE, 8	Asp 382	OD1, 140	3.003
	Arg 94	NH2, 44	Ala 404	O, 181	2.72
	Arg 94	NH2, 44	Leu 399	O, 95	2.762
	Arg 94	NH1, 43	Leu 399	O, 95	2.872
	Arg 94	NE, 41	Ala 404	O, 181	2.979
	Arg 91	NH2, 22	Asp 412	OD2, 226	2.731
	Arg 91	NH2, 22	Leu 411	O, 256	2.684
	Arg 91	NH1, 21	Ile 409	O, 214	2.697
	Arg 87	NH2, 33	Asn 416	O, 264	2.609
	Arg 87	NH2, 33	Ser 415	O, 272	2.852
	Arg 87	NH2, 33	Gly 414	O, 154	2.826

	Arg 87	NE, 30	Gly 414	O, 154	2.832
	Glu 148	OE2, 163	Lys 389	NZ, 150	2.59
	Glu 148	OE1, 162	Lys 389	NZ, 150	2.521
	Asn 102	O, 76	Lys 391	NZ, 116	3.110
EGLN1	Lys 423	NZ, 257	Glu 457	OE2, 258	2.692
P317R↔HIF1α	Gln 352	N, 20	Leu 399	O, 15	2.823
	Lys 350	NZ, 68	Asp 388	OD2, 99	2.542
	Lys 350	NZ, 68	Asp 382	OD2, 275	2.667
	Arg 344	NH1, 134	Asp 421	OD1, 114	2.765
	Arg 317	NH1, 58	Gly 414	O, 294	2.821
	Arg 317	NH1, 58	Asp 412	OD1, 142	2.810
	Arg 312	NH2, 11	Gln 424	OE1, 247	2.602
	Arg 312	NH1, 10	Thr 418	OG1, 82	2.722
	Arg 312	NH1, 10	Asn 416	OD1, 107	2.647
	Arg 312	NH1, 10	Asn 416	O, 103	2.722
	Arg 312	N, 1	Thr 418	O, 79	2.826
	Gly 309	N, 177	Glu 419	OE1, 38	3.051
	Lys 204	NZ, 98	Thr 383	OG1, 75	2.691
	Lys 204	NZ, 98	Glu 379	O, 119	2.662
	Tyr 197	OH, 234	Ser 385	OG, 186	2.654
	Lys 52	NZ, 91	Glu 268	OE1, 167	2.564
EGLN1	Arg 396	NH2, 22	Asn 826	O, 57	2.711
R371H↔HIF1α	Arg 396	NH2, 22	Leu 822	O, 301	2.635
	Arg 396	NH2, 22	Ala 821	0, 205	2.693
	Arg 396	NH1, 21	Ala 821	0, 205	2.666
	Arg 396	NE, 19	Asn 826	0, 57	3.259
	Asn 293	ND2, 296	Gln 824	OE1, 365	2.918
	Arg 281	NH1, 89	Gly 808	O, 428	3.093
	Arg 281	NE, 87	Glu 817	OE1, 242	3.300
	Ser 274	OG, 201	Gln 814	OE1, 124	2.809

	Thr 268	OG1, 312	Ser 797	OG, 325	2.808
	Glu 267	OE2, 161	Tyr 798	OH, 152	2.908
	Gln 221	NE2, 277	Gln 794	OE1, 252	2.836
	Ala 190	N, 374	Ile 772	O, 165	3.004
	Lys 186	NZ, 215	Asp 777	OD1, 222	2.756
	Ser 153	OG, 373	Glu 767	OE2, 140	2.875
	Arg 152	NH2, 33	Ile 771	O, 65	2.765
	Arg 152	NH1, 32	Glu 767	OE1, 139	2.704
	Arg 152	NE, 30	Ile 771	O, 65	3.116
	Arg 99	NH2, 11	Thr 555	OG1, 176	2.630
	Arg 99	NH2, 11	Asp 5540	OD1, 114	2.777
	Arg 99	NH2, 11	Thr 552	O, 386	2.742
	Arg 99	NH1, 10	Asp 554	OD1, 114	2.738
	Arg 99	NE, 8	Thr 555	OG1, 176	3.280
	Lys 95	NZ, 107	Glu 545	OE2, 289	2.638
	Glu 92	N, 333	Glu 542	OE1, 52	2.900
	Arg 91	NH2, 44	Leu 539	O, 345	2.667
	Gly 89	N, 429	Ala 541	O, 130	3.000
	Gln 73	NE2, 184	Gly 726	O, 382	2.895
EGLN1	Lys 159	NZ, 185	Asp 421	OD2, 129	2.513
H374R↔H	Asn 102	OD1, 145	Lys 391	NZ, 214	2.568
IF1α	Asn 102	ND2, 144	Asp 395	OD2, 300	3.248
	Arg 99	NH2, 11	Asp 388	OD2, 78	2.735
	Arg 99	NH2, 11	Ser 384	OG, 113	2.712
	Arg 99	NH1, 10	Leu 386	0, 189	2.664
	Arg 99	NH1, 10	Ser 384	OG, 113	2.949
	Arg 99	NH1, 10	Ser 384	O, 111	2.662
	Lys 95	NZ, 107	Asp 388	OD2, 78	2.470
	Lys 95	NZ, 107	Asp 382	OD1, 136	2.594
	Arg 94	NH2, 44	Asp 433	OD2, 153	2.743

	Arg 94	NH2, 44	Glu 427	OE2, 280	2.804
	Arg 94	NH1, 43	Asp 433	OD2, 153	2.723
	Arg 91	NH2, 22	Leu 411	O, 245	2.622
	Arg 91	NH2, 22	Ser 410	O, 284	3.143
	Arg 91	NH2, 22	Ile 409	O, 117	2.707
	Arg 91	NH1, 21	Ile 409	O, 117	2.751
	Arg 87	NH2, 33	Ser 415	0, 290	2.794
	Arg 87	NH2, 33	Gly 414	O, 165	2.666
	Arg 87	NH1, 32	Asp 417	OD1, 60	2.746
	Arg 87	NE, 30	Gly 414	O, 165	2.809
	Lys 55	NZ, 87	Glu 419	OE2, 70	2.623
	Lys 55	NZ, 87	Glu 419	OE1, 69	2.599
	Lys 55	NZ, 87	Glu 419	O, 65	2.696
	Lys 53	NZ, 53	Gln 423	OE1, 258	3.191
	Lys 53	NZ, 53	Asp 417	O, 57	2.470

Chapter 4
Discussion

4 Discussion

The increasing accessibility of genetic information and efficient annotation of genes of various species have empowered bioinformaticians to conduct analysis on the genes of their interest. These progressions have also furnished opportunities to deeply examine how a species evolved and how the changes that occurred in a gene are responsible for adaptation or disease (Pervaiz, N., & Abbasi, A. A. et al. 2016).

The homologs of EGLN1 gene have been identified in vertebrates as well as invertebrates. EGLN1 homologous proteins sequences from various species were collected through NCBI (Wheeler, D. L. et al. 2007) and Ensembl genome browsers (Fernández, X. M., & Birney, E. et al. 2010). Evolutionary analysis indicated two duplication events in vertebrates which resulted into three paralogous genes EGLN1, EGLN2 and EGLN3. The gathered sequences display a significant degree if similarity among them, especially the Fe (+2) OG dioxygenase domain is observed in all homologous proteins.

To explore the evolutionary history of a protein, phylogenetic tree construction emerges as a crucial investigative aspect. Two methods are used for tree construction. Distance based methods rely on quantifying evolutionary mutations among homologs and character-based methods encompass all accessible evolutionary information. Both distance-based (Neighbor-Joining) and character-based (Maximum likelihood) methods were applied for tree construction.

For tree construction complete deletion option was used and a bootstrap test was conducted with a bootstrap value of 1000. The anticipated count of nucleotide substitutions per site was 0.05.

One of the core aspects in the process of evolution is the influence of new mutations that accumulate over time. Each amino acid site in a protein can experience varying selection constraints, which lead to different ratios of non-synonymous to synonymous changes ($\omega = dN/dS$). This ratio is referred as the “acceptance rate”. If the value of dN/dS is greater than 1, it indicates positive selection/diversifying evolution and if dN/dS value is less than 1, it indicates negative selection/purifying evolution. The dN/dS value = 1, indicates neutrality (Yang, Z. et al. 2000). Negative selection occurs when natural selection removes

changes that affect protein's function, while positive selection favors and accumulates beneficial changes during evolutionary process (Nawaz, et al. 2020).

To assess the impact of Darwinian selection on EGLN1 gene, we focused on protein coding DNA sequences within three groups of species (primates, artiodactyls and rodents). The acceptance rate of all three groups signifies negative selection (the acceptance rate was consistently below 1). Utilization of a z-test implemented in MEGA7 produced similar results identifying negative selection. Moreover, identification of sites under selection pressure was conducted through the SLAC (Single likelihood ancestor counting) method. A total of 29 sites with significant negative constraints were identified.

Ancestral reconstruction involves the computation of ancient amino acid sequences based on extant protein sequences (Merkl, R. et al. 2016).

Ancestral sequence reconstruction was employed using primate, artiodactyl and carnivore proteins sequences of EGLN1 to examine potential alterations over time in these three groups. Multiple sequence Alignments by Clustal omega (Sievers, F. et al. 2014).

The protein's domain organization was analyzed and a comparative examination of EGLN1/PHD2 protein was conducted in diverse organisms (mouse, goat, human) as well as those adapted to high altitudes (markhor, Tibetan Human and Snow leopard). Two conserved domains were identified across all EGLN1 orthologs: Zinc-finger MYND type domain and Fe (+2) OG dioxygenase domain.

We broadened our research scope to encompass a structural standpoint, aiming to achieve a clear picture of effect of changes that occurred in protein structure level during evolution in these three groups of species and the effect of point mutations that were reported to be a cause of Erythrocytosis familial 3 (Bento, C. et al. 2018).

EGLN1 performs oxygen-sensitive hydroxylation of HIF1α, controlling its stability. This interaction is pivotal for cells to adjust to varying oxygen conditions, modulating HIF1α's actions on gene expression downstream and enabling adaptive responses. To study the interaction between HIF1α and PHD2 proteins, a protein- protein interaction was performed between the two proteins (EGLN1 wild type and HIF1α) as well as between EGLN1 mutants and HIF1α protein to analyze the changes in interactions caused by point mutations.

4.1 Conclusion:

This study unveils EGLN1's pivotal role in high-altitude adaptation, shedding light on its significance in Markhor and related species like the snow leopard. EGLN1's genetic variations influence oxygen saturation, hemoglobin levels, and crucial physiological adjustments essential for thriving at high altitudes. These insights extend to broader attributes governed by EGLN1, potentially aiding conservation efforts and enhancing our understanding of endangered species like the Markhor. By grasping the genetic foundations driving high-altitude adaptation, we empower conservation initiatives and deepen our appreciation for nature's resilience in challenging environments.

Chapter 5
References

5 References

Alekseenko, A., Ignatov, M., Jones, G., Sabitova, M., & Kozakov, D. (2020). Protein–protein and protein–peptide docking with ClusPro server. *Protein Structure Prediction*, 157-174.

Ali, S. (2008). Conservation and status of markhor (capra falconeri) in the northen parts of north west frontier province, Pakistan.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Barradas, J., Rodrigues, C. D., Ferreira, G., Rocha, P., Constanço, C., Andrade, M. R., & Silva, H. M. (2018). Congenital erythrocytosis–discover of a new mutation in the EGLN1 gene. Clinical Case Reports, 6(6), 1109.

Beall, C. M. (2006). Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. Integrative and comparative biology, 46(1), 18-24.

Beall, C. M. (2007). Two routes to functional adaptation: Tibetan and Andean high-altitude natives. Proceedings of the National Academy of Sciences, 104(suppl_1), 8655-8660.

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., & Zheng, Y. T. (2010). Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. Proceedings of the National Academy of Sciences, 107(25), 11459-11464.

Bento, C. (2018). Genetic basis of congenital erythrocytosis. *International Journal of Laboratory Hematology*, *40*, 62-67.

Dahlback, A., Gelsor, N., Stamnes, J. J., & Gjessing, Y. (2007). UV measurements in the 3000–5000 m altitude region in Tibet. Journal of Geophysical Research: Atmospheres, 112(D9).

Dupuy, D., Aubert, I., Dupérat, V. G., Petit, J., Taine, L., Stef, M., & Arveiler, B. (2000). Mapping, characterization, and expression analysis of the SM-20 human homologue, c1orf12, and identification of a novel related gene, SCAND2. Genomics, 69(3), 348-354.

Eng, J. T., & Aldenderfer, M. (2017). Bioarchaeological profile of stress and dental disease among ancient high altitude Himalayan communities of N epal. *American Journal of Human Biology*, *29*(4), e22998.

Fernández, X. M., & Birney, E. (2010). Ensembl genome browser. In Vogel and Motulsky's Human Genetics (pp. 923-939). Berlin, Heidelberg: Springer Berlin Heidelberg.

Grubb, P. (2005). Order artiodactyla. Mammal Species of the World. A Taxonomic and Geographic Reference 3rd edition, 1, 637-722.

Heo, L., Park, H., & Seok, C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic acids research*, *41*(W1), W384-W388.

Jaakkola, P., Mole, D. R., Tian, Y. M., Wilson, M. I., Gielbert, J., Gaskell, S. J., & Ratcliffe, P. J. (2001). Targeting of HIF-α to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. Science, 292(5516), 468-472.

Lay, D. M. (1978). Roberts, TJ The mammals of pakistan. Ernest Benn Ltd., London, xxvi+ 361 p., 4 color plates, 90 figures, 118 distribution maps, 1977. Price,£ 35 English Pounds, $66.

Lorenzo, F. R., Huff, C., Myllymäki, M., Olenchock, B., Swierczek, S., Tashi, T., & Prchal, J. T. (2014). A genetic mechanism for Tibetan high-altitude adaptation. Nature genetics, 46(9), 951-956.

Majmundar, A. J., Wong, W. J., & Simon, M. C. (2010). Hypoxia-inducible factors and the response to hypoxic stress. Molecular cell, 40(2), 294-309.

Malik, M. M. (1987). Management plan for wild artiodactyls in North West Frontier Province, Pakistan.

Merkl, R., & Sterner, R. (2016). Ancestral protein reconstruction: techniques and applications. *Biological chemistry*, *397*(1), 1-21.

Mishra, A., Mohammad, G., Thinlas, T., & Pasha, M. Q. (2013). EGLN1 variants influence expression and S ao2 levels to associate with high-altitude pulmonary oedema and adaptation. *Clinical Science*, *124*(7), 479-489.

Nawaz, M. S., Asghar, R., Pervaiz, N., Ali, S., Hussain, I., Xing, P., & Abbasi, A. A. (2020). Molecular evolutionary and structural analysis of human UCHL1 gene demonstrates the relevant role of intragenic epistasis in Parkinson's disease and other neurological disorders. *BMC Evolutionary Biology*, *20*, 1-15.

Pagni, M., Ioannidis, V., Cerutti, L., Zahn-Zabal, M., Jongeneel, C. V., Hau, J., & Falquet, L. (2007). MyHits: improvements to an interactive resource for analyzing protein sequences. *Nucleic acids research*, *35*(suppl_2), W433-W437.

Pervaiz, N., & Abbasi, A. A. (2016). Molecular evolution of WDR62, a gene that regulates neocorticogenesis. *Meta gene*, *9*, 1-9.

Ranjitsinh, M. K., Seth, C. M., Ahmad, R., Bhatnagar, Y. V., & Kyarong, S. S. (2005). Goats on the border: A rapid assessment of the Pir Panjal markhor in Jammu and Kashmir: Distribution, status and the threats. Report by the Wildlife Trust of India, New Delhi, India.

Roberts, T. J., & Bernhard (principe d'Olanda.). (1977). The mammals of Pakistan.

Rombel, I. T., Sykes, K. F., Rayner, S., & Johnston, S. A. (2002). ORF-FINDER: a vector for high-throughput gene identification. *Gene*, *282*(1-2), 33-41.

Schaller, G. B. (1977). Mountain monarchs. Wild sheep and goats of the Himalaya. University of Chicago Press.

Semenza, G. L. (2009). Regulation of oxygen homeostasis by hypoxia-inducible factor 1. Physiology, 24(2), 97-106.

Sievers, F., & Higgins, D. G. (2014). Clustal omega. *Current protocols in bioinformatics*, *48*(1), 3-13.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., & Ge, R. (2010). Genetic evidence for high-altitude adaptation in Tibet. Science, 329(5987), 72-75.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., & Yaschenko, E. (2005). Database resources of the national center for biotechnology information. Nucleic acids research, 33(suppl_1), D39-D45.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., & Yaschenko, E. (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, *35*(suppl_1), D5-D12.

Xiang, K., Ouzhuluobu, Peng, Y., Yang, Z., Zhang, X., Cui, C., & Su, B. (2013). Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. Molecular biology and evolution, 30(8), 1889-1898.

Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M., & Vangone, A. (2016). PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, *32*(23), 3676-3678.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, *155*(1), 431-449.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., & Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. science, 329(5987), 75-78.