Molecular Modeling and Phylogeny of the Krüppel-like Factor 4 (cKLF4) Protein from the Arabian Camel, *Camelus dromedarius*

Abdullah O. Alawad[1,*], Sultan N. Alharbi[1,*], Othman A. Alhazzaa[1], Faisal S. Alagrafi[1], Mohammad N. Alkhrayef[1], Ziyad A. Alhamdan[1], Abdullah D. Alenazi[1], Mohamed Hammad[1,2], Sami A. Alyahya[3], Hasan A. AlJohi[4] and Ibrahim O. Alanazi[4]

[1]National Center for Stem Cell Technology, King Abdulaziz City for Science and Technology (KACST), Riyadh, Kingdom of Saudi Arabia. [2]SAAD Research and Development Center, Clinical Research Laboratory and Radiation Oncology, SAAD Specialist Hospital, Al Khobar, Kingdom of Saudi Arabia. [3]National Center for biotechnology, King Abdulaziz City for Science and Technology (KACST), Riyadh, Kingdom of Saudi Arabia. [4]National Center for Genomic Technology, King Abdulaziz City for Science and Technology, Riyadh, Kingdom of Saudi Arabia. *Contributed equally to this work.

ABSTRACT: Krüppel-like factor 4 (KLF4) is a pluripotency transcription factor that helps in generating induced pluripotent stem cells (iPSCs). We sequenced for the first time the full coding sequence of *Camelus dromedarius* KLF4 (cKLF4), which is also known as the Arabian camel. Bioinformatics analysis revealed the molecular weight and the isoelectric point of cKLF4 protein to be 53.043 kDa and 8.74, respectively. The predicted cKLF4 protein sequence shows high identity with some other species as follows: 98% with Bactrian camel and 89% with alpaca KLF4 proteins. A three-dimensional (3D) structure was built based on the available crystal structure of the *Mus musculus* KLF4 (mKLF4) of 82 residues (PDB: 2 WBS) and by predicting 400 residues using bioinformatics software. The comparison confirms the presence of the zinc finger domains in cKLF4 protein. Phylogenetic analysis showed that KLF4 from the Arabian camel is grouped with the Bactrian camel, alpaca, cattle, and pig. This study will help in the annotation of KLF4 protein and in generating camel-induced pluripotent stem cells (CiPSCs).

KEYWORDS: KLF4, pluripotency, 3D structure, *Camelus dromedarius,* sequencing, camel-induced pluripotent stem cells (CiPSCs)

CITATION: Alawad et al. Molecular Modeling and Phylogeny of the Krüppel-like Factor 4 (cKLF4) Protein from the Arabian Camel, *Camelus dromedarius. Bioinformatics and Biology Insights* 2016:10 291–300 doi: 10.4137/BBI.S40782.

TYPE: Original Research

RECEIVED: August 26, 2016. **RESUBMITTED:** October 13, 2016. **ACCEPTED FOR PUBLICATION:** October 20, 2016.

ACADEMIC EDITOR: J. T. Efird, Associate Editor

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1731 words, excluding any confidential comments to the academic editor.

FUNDING: The study was sponsored by the National Center for Stem Cell Technology (NCSCT), King Abdulaziz City for Science and Technology (KACST). Special thanks are due to SAAD Research & Development Center at SAAD Specialist Hospital for the support. This work was supported by a 2012 National Science, Technology and Innovation Plan (NSTIP) Translational Stem Cell Research (TSCR) grant (no. 33-837). The authors also gratefully acknowledge the financial support from KACST under NSTIP grant (no. 32-685). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: snharbi@kacst.edu.sa

Introduction

KLF4 is one of the significant reprogramming factors that is involved in remodeling cell fate.[1] KLF4 is a transcription factor that is related to the specificity protein/Krüppel-like factor (Sp/KLF) superfamily of gene regulatory proteins.[2] It is expressed at high levels in erythroid tissues, and it is also obtained in other tissues, most notably in the brain tissues.[3] The function of KLF4 has been carefully studied in normal homeostasis, cell differentiation, and cancer formation. However, the biochemical and biophysical properties of KLF4 protein and its three-dimensional (3D) structure are not fully understood.

The protein structure of KLF4 is characterized by three highly conserved C_2 H_2 -type zinc finger domains at its C-terminus, which bind to GC/GT-rich regions of DNA in order to influence either activation or repression of transcription.[4] They contain three characteristic C–C (cysteine–cysteine): H–H (histidine–histidine) Krüppel-like zinc fingers and recognize CACCC motifs of DNA.

Despite sequencing the whole genomes of both the Arabian camel and the Bactrian camel[5–7], there is a lack of sequencing information studies that target genes of camel. In the present study, we sequenced cDNA and used bioinformatics approaches to characterize KLF4 protein of *Camelus dromedarius*, which is commonly known as the Arabian camel. Henceforth, for easy mention, the Arabian camel KLF4 will be referred to as cKLF4. The present study aimed to (1) sequence the mRNA of cKLF4, (2) predict its amino acid sequence, (3) utilize the homology-based method to identify homology in the regulatory domains for cKLF4 and KLF4 in six different species, (4) model and compare its predicted 3D structure with the available mammalian homologs, and (5) construct

the phylogenetic tree of cKLF4 with six mammalian species using KLF4 proteins.

Materials and Methods

Ethics statement. All procedures for animal anesthesia, euthanasia, and sample collection were carried out in strict accordance with the recommendations by King Abdulaziz City for Science and Technology (KACST) National Committee of BioEthics for rules of experimentation on live animals and others.[8]

Sample collection. Camel brain tissue was obtained from an adult male camel, immediately after slaughtering at the Southern Riyadh Main Slaughterhouse. Brain tissue samples were taken from different parts of the camel brain, which was then submerged in RNAlater solution (Qiagen) to avoid RNA degradation, and stored at –20 °C. If not otherwise stated, *Escherichia coli* strains were grown in Luria-Bertani medium supplemented with 100 mg/mL of ampicillin.

Oligonucleotide design. Primers were designed according to the data from the Arabian camel genome project at KACST and using Primer-BLAST tool at GenBank website (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). β-Actin was used as an endogenous control. Combinations between primer pairs were tested using AmplifX 1.7.1 (http://crn2m.univ-mrs.fr/pub/amplifx-dist) in order to determine the optimized annealing temperatures to yield specific Polymerase Chain Reaction (PCR) products representing either the full coding sequence or the partial coding sequence that was subjected to sequencing. The sequence and amplification product length of each primer couples are listed in Table 1.

RNA extraction and cDNA synthesis. About 50 mg of brain tissue from male camels was homogenized in RTL lysis buffer (Qiagen) supplemented with 1% 2-mercaptoethanol. Total RNA was extracted using E.Z.N.A. kit (Omega Bio-Tek), according to the manufacturer's instructions. Then, the sample was quantified at 260 nm using nanodrop spectrophotometer (NanoDrop; Thermo Scientific), and the integrity of RNA sample was assessed using denaturing formaldehyde agarose gel (1%) electrophoresis. A total of 2 µg of total RNAs was reverse transcribed to single-stranded cDNA using ImProm-II Reverse Transcription System (Promega), as recommended by the manufacturer.

PCR. Gradient PCR was performed using annealing temperatures that ranged from 50 °C to 60 °C in a final volume of 25 µL as follows: 12.5 µL of GoTaq Green Master Mix (Promega), 5 µL of cDNA, and 1 µL of each forward and reverse primers (5 pmol, Table 1) in a final volume of 25 µL adjusted with nuclease-free water. The PCR condition used was as follows: one cycle at 95 °C for 2 minutes followed by 25 cycles at 95 °C for 30 seconds, 50 °C–60 °C for 45 seconds, and 72 °C for 105 seconds. Final extension was carried out at 72 °C for five minutes. PCR products were analyzed by electrophoresis using a 1.2% agarose gel (Supplementary Fig. 2).

DNA sequencing and prediction of amino acid sequence. The full-length coding sequence of cKLF4 was obtained using the 3730XL series platform Sequencer (Applied Biosystems). cDNA fragments of 877 and 1,560 bp were amplified by PCR using the primer couple cKLF4F1/cKLF4R1 and cKLF4F2/cKLF4R2 (Table 1, Supplementary Fig. 2), which were then sequenced using 3730XL DNA Sequencer (Applied Biosystems) using the same PCR primers; nucleotide sequences were determined in both forward and reverse directions, and the sequences were analyzed using Geneious 7.1.7 software (http://www.geneious.com).[9] The similarity of the obtained sequence was examined in the GenBank database using the BLASTN algorithm on the NCBI BLAST server (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Multiple sequence alignment and analysis of phylogenetic relationship. The sequenced mRNA was translated using Geneious 7.1.7 software, and the deduced cKLF4 amino acid sequence was then compared with the existing sequences in the NCBI Protein Database using the BLASTP algorithm on the NCBI BLAST server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The predicted amino acid sequence of cKLF4 was used as a template to identify homologous mammalian sequences in PSI-BLAST searches in the NCBI Protein Database. Six homologous sequences from different mammals were used for multiple sequence alignment by ClustalW alignment using Geneious 7.1.7 software. The amino acid sequences of KLF4 from camel and other mammalian species were aligned, and a phylogenetic tree was constructed using BLOSUM62 matrix.

Secondary and 3D structures of cKLF4 protein. The secondary structure of cKLF4 was predicted using Geneious 7.1.7 software, while the 3D structure was predicted using

Table 1. List of primers used for the amplification and sequencing studies.

PRIMER COUPLE	PRIMER	PRIMER SEQUENCE	PRODUCT (bp)
Internal Primer	cKLF4F1	CTCCTGCCTCTGCTCCCTAC	877
	cKLF4R1	GGAGACAGCCTCCTGCTTG	
Full-Coding Region	cKLF4F2	TGGCCCTCTCTCTAACTCCT	1560
	cKLF4R2	GCCTCTTCATGTGTAAGGCG	
	Camel-β-Actin F	GATATTGCTGCGCTCGTGGT	1139
	Camel-β-Actin R	TGGAACGTAACTAAGTCCGCCT	

SWISS-MODEL server based on homology structure modeling[10] and Protean 3D (Lasergene 12; DNASTAR).

Globular and disordered regions in cKLF4 protein. In order to identify ordered *globular* and disordered regions of cKLF4 protein, we used GlobPlot 2.3 server[11] at globplot.embl.de website. The Russell/Linding set was chosen in which structures of α-helices and β-sheets are assigned as globular regions (GlobDoms), whereas random coils and turns structure as disordered regions (Disorder).

ANCHOR analysis. In order to predict binding sites within disordered regions of cKLF4, ANCHOR web server[12] at http://anchor.enzim.hu has been used. A list of all the software programs and websites used for the analysis is provided in Table 2.

Table 2. List of software programs and websites used for the analysis.

FUNCTION	WEBSITE AND SOFTWARE
To search nucleotide sequence	http://www.ncbi.nlm.nih.gov/nuccore/
To find conserved sequence	Geneious 7.1.7 software
To find open reading frame	Geneious 7.1.7 software
To analysis amino sequence	Geneious 7.1.7 software
To analysis identity of amino acid	http://blast.ncbi.nlm.nih.gov/Blast.cgi
To predict 3D structure	http://swissmodel.expasy.org/Protean
To identify ordered and disordered regions	1) globplot.embl.de 2) http://www.disprot.org/metapredictor.php
To search potential binding sites in disordered regions	http://anchor.enzim.hu

Results

Sequence identity of cKLF4. The similarity of the obtained sequence was examined in the GenBank database using the BLASTN algorithm on the NCBI BLAST server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The cKLF4 cDNA sequence encoded a cKLF4 protein of 501 amino acids. The nucleotide BLAST analysis for cKLF4 mRNA showed that it shared high identity (99%–86%) with the *KLF4* mRNAs from other mammals: 99% with alpaca (*Vicugna pacos*), 98% with wild Bactrian camel (*Camelus ferus*), 94% for cattle (*Bos taurus*), 93% with pig (*Sus scrofa*), 89% with human (*Homo sapiens*), and

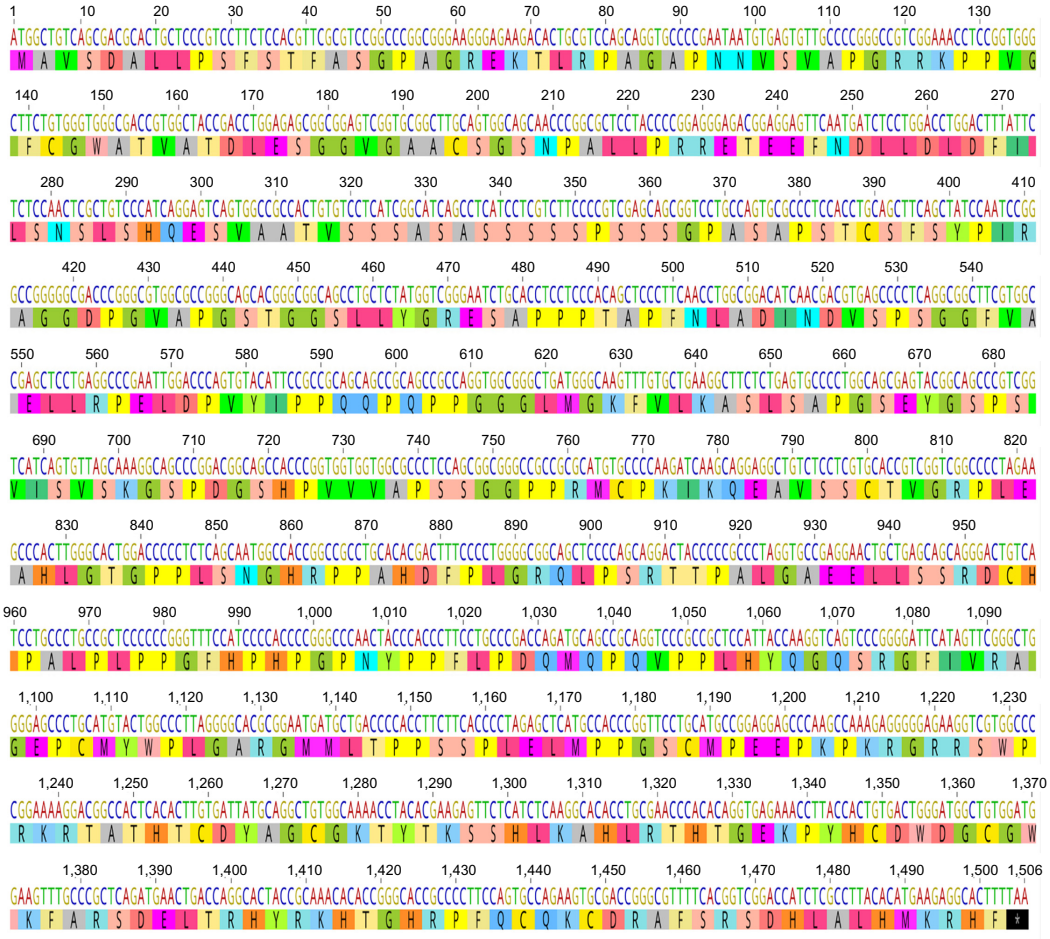

Figure 1. Complete nucleotide sequence encoding cKLF4 and its predicted amino acids. *Is the termination codon.

86% with house mouse (*Mus musculus*). The complete sequence consisted of 1,506 nucleotides (Fig. 1) and represents the first cKLF4 cDNA sequence. It has been reported that the whole genomes of Arabian, Bactrian, and Alpaca camels consist of 28.4%, 30.4%, and 32.1% repeat sequences, respectively, lower by ~10% than human (46.1%) and cattle (42.5%), which makes it easier to identify a particular gene within camelid genomes than other species.[7]

Multiple sequence alignment and phylogenetic analysis. Comparison of the predicted amino acid sequence of the whole cKLF4 protein and most similar sequences from six species (Table 3) using the BLASTP algorithm on the NCBI BLAST server (http://blast.ncbi.nlm.nih.gov/Blast.cgi) indicated relative percentage identities (98%–83%) as follows: 98% for *C. ferus*, 93% for *S. scrofa*, 89% for *V. pacos*, 87% for *B. taurus*, 89% for *H. sapiens*, and 83% for *M. musculus* (Fig. 2).

The phylogenetic tree for the predicted amino acid sequence of whole cKLF4 protein and six of the mostly similar mammalian KLF4 is shown in (Fig. 3). Our result revealed that cKLF4 is grouped together with KLF4 from pig, cattle, alpaca, and Bactrian camel and separated in the early evolution from human and mouse.

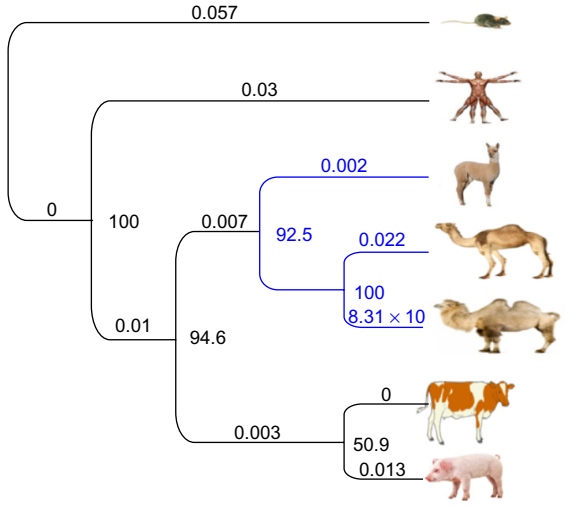

Figure 3. The rooted phylogenetic tree of cKLF4 and other six species.

Domains and structure analysis of cKLF4 protein. *C_2H_2-type zinc finger domains*. C-terminus of KLF4 protein is

Table 3. Comparison of cKLF4 protein and other KLF4 proteins from various, mostly similar, mammals.

SPECIES	(REF. SEQ)	NUMBER OF AMINO ACIDS	COVERAGE (%)	E-VALE	IDENTITY
Camelus ferus (Bactrian Camel)	XP_006180844	475	90%	0.0	98%
Vicugna pacos (Alpaca)	XP_006213443	485	100%	0.0	89%
Bos Taurus (Cattle)	NP_001098855	477	100%	0.0	87%
Sus scrofa (Pig)	AAY16111	510	100%	0.0	93%
Homo sapiens (Human)	AAH30811	504	100%	0.0	89%
Mus musculus (Mouse)	AAH10301	474	100%	0.0	83%

Note: The comparison included number of amino acid residues, percent identity, and E-value.

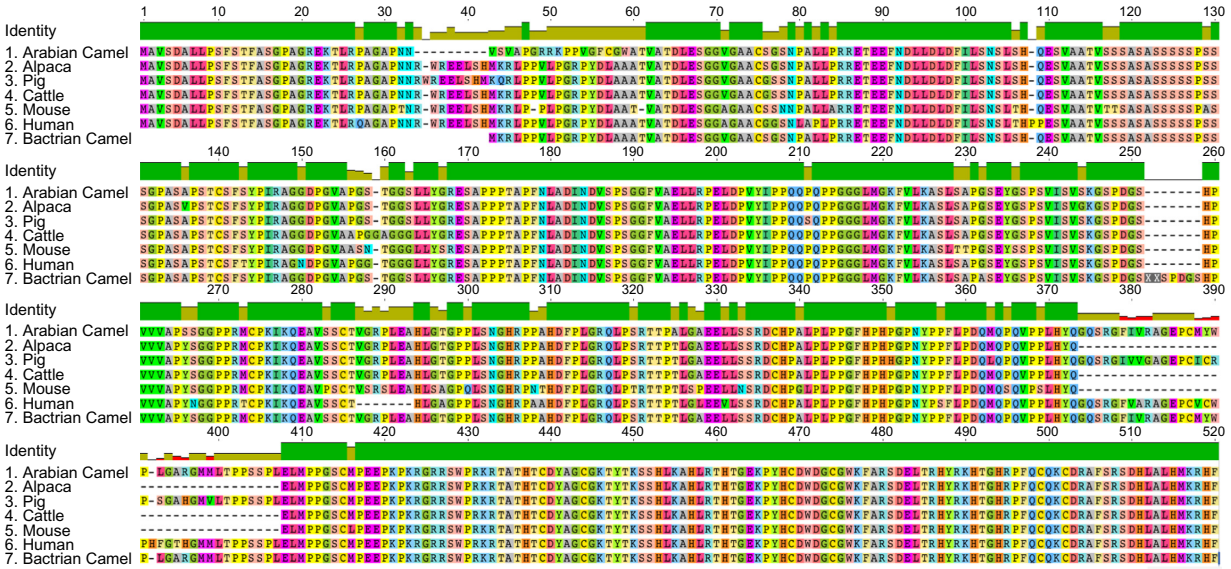

Figure 2. Amino acid sequence alignment of cKLF4 protein with KLF4 proteins of six species. The alignment was generated with Geneious 7.1.7 Multiple Sequence Alignment software. Residues were color coded according to their conservancy.

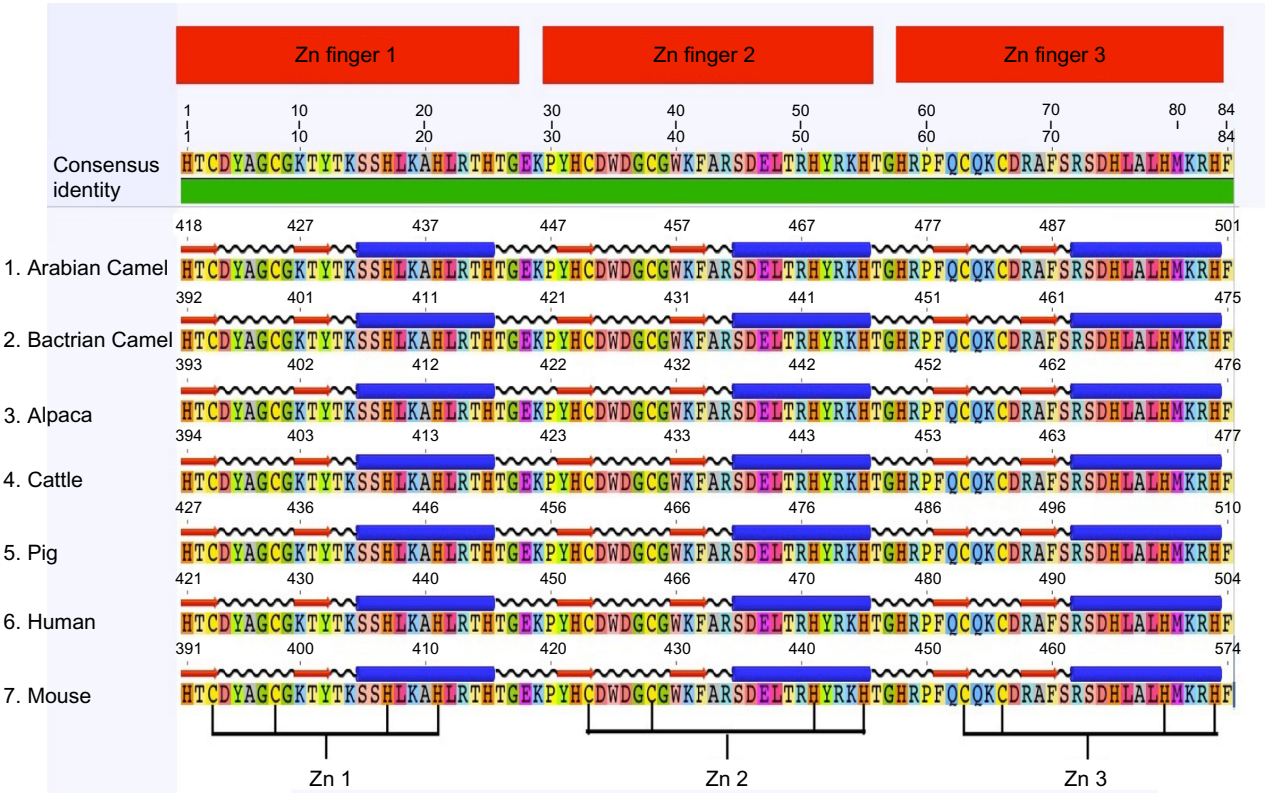

Figure 4. Three highly conserved C_2H_2-type zinc finger domains of cKLF4 protein aligned with those from other species.

characterized by three highly conserved $C_2 H_2$ -type zinc finger domains, which bind to GC/GT-rich regions of DNA in order to influence activation and/or repression of transcription.[4] They contain three characteristic C–C (cysteine–cysteine): H–H (histidine–histidine) Krüppel-like zinc fingers and recognize CACCC motifs of DNA.

Predicted globular and disordered regions in cKLF4 protein. The Russell/Linding set was chosen in which structures of α-helices and β-sheets are assigned as globular regions (GlobDoms), whereas random coils and turns structure as disordered regions (Disorder). Residue ranges for disordered regions (blue) and globular regions (green) are shown at the bottom of the graph (Fig. 5). Furthermore, we used JRONN method,[13,14] which is based on regional order neural network (RONN) analysis method using Protean program (Lasergene 12; DNASTAR) in order to predict disordered regions of cKLF4 (Fig. 6).

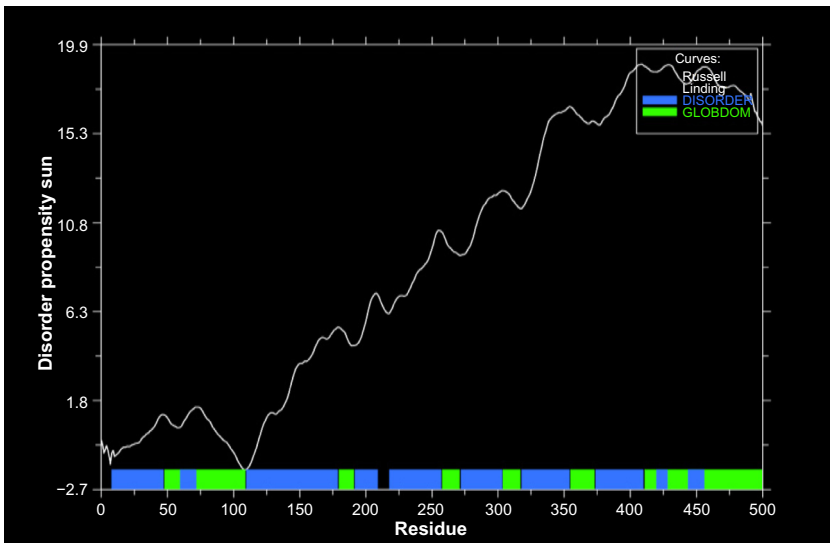

Figure 5. GlobPlot analysis. Blue boxes are disordered regions and green boxes are ordered regions in cKLF4 protein.

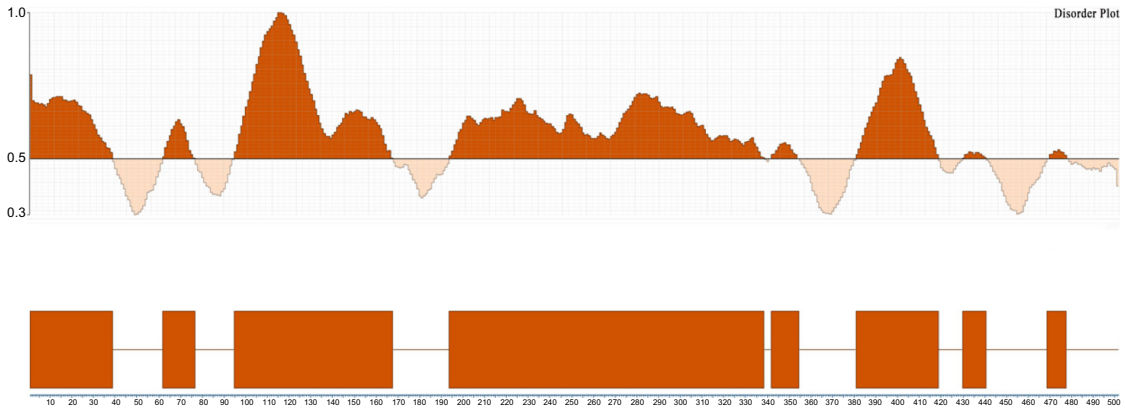

Figure 6. The disorder (JRONN) method predicts disordered regions of cKLF4 protein. The value is between (0.3, 1.0), where disordered regions are separated from ordered one, regions with values above 0.5 are considered as disordered, and those below 0.5 are ordered regions. Disordered regions are labeled orange boxes. Disorder predictions for 501 amino acids of cKLF4.

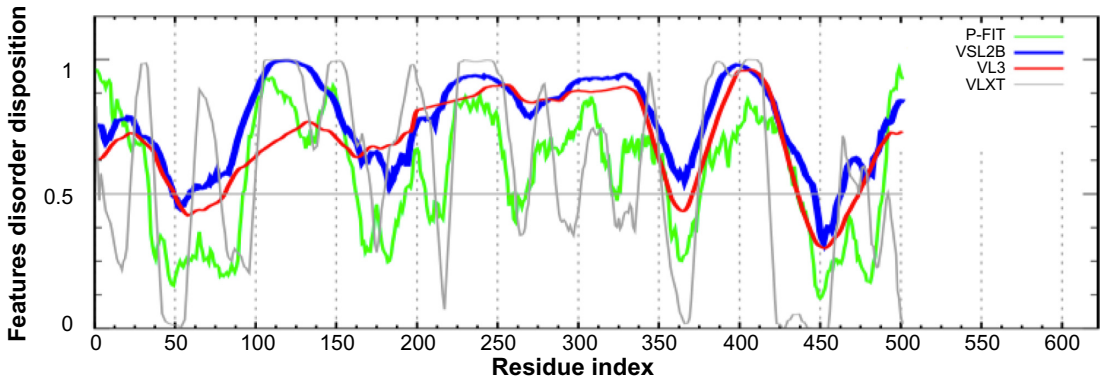

Figure 7. The green line is disorder prediction from P-FIT; blue line is prediction of VSL2B; the red line is prediction of VL3; and the gray line is the prediction of VLXT.

We also used four different predictors in order to determine ordered (structured) and disordered (unstructured) regions within cKLF4 protein (Fig. 7). These predictors are VLXT,[15] VL3,[16] VSL2B,[16] and P-FIT,[17] which use amino acids sequence as inputs and give a structure order or disorder as output.

In order to search potential binding sites in disordered regions of cKLF4 protein, we used ANCHOR algorithm[18] at (http://anchor.enzim.hu). (Fig. 8) Shows 11 potential binding sites within cKLF4 proteins (blue boxes) as follows: 1–14, 46–55, 83–96, 129–138, 150–156, 167–193, 206–219, 227–232, 256–280, 291–326, and 354–379. VLXT predictor and ANCHOR-indicated binding sites are often completely or partially overlapping with each other. Eight of 11 potential binding sites from ANCHOR predictor overlapped with VLXT predictor (Figs. 7 and 8).

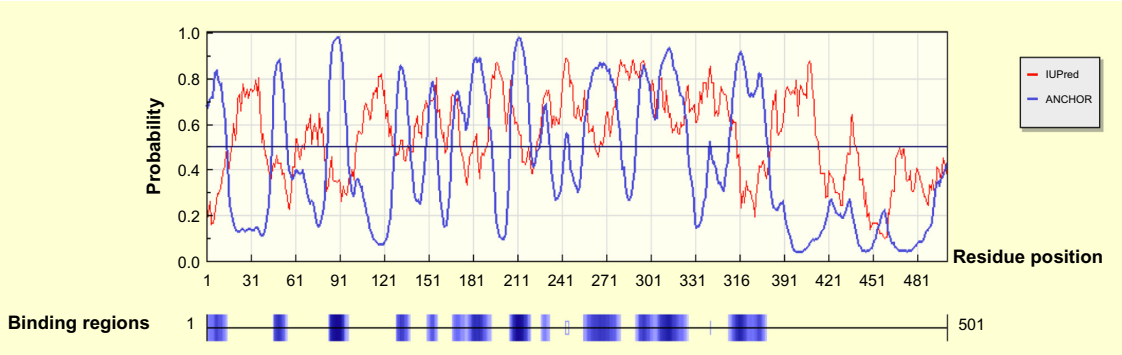

Figure 8. ANCHOR plot. Prediction of protein-binding regions in disordered cKLF4 protein. Blue boxes are binding regions.

Another server used for predicting intrinsically unstructured/disordered region of cKLF4 is Intrinsically Unstructured Proteins (IUP) server red line, as shown in (Fig. 8). Regions above 0.5 thresholds are predicted as disorder. Based on the assumption by IUP, the sequences do not fold due to their inability to form sufficient stabilizing interresidue interactions, which classified cKLF4 protein as an unstable protein. Moreover, we examined the stability and instability of cKLF4 using the approach of Guruprasad et al.[19], which revealed that most of the cKLF4 proteins consisted of unstable regions (Supplementary Fig. 4).

Secondary and 3D structure modeling of cKLF4 compared with mouse KLF4. The prediction of the secondary structure of cKLF4 was done using Geneious 7.1.7 software and compared with mouse KLF4 (Fig. 9). The predicted structure suggested that cKLF4 protein is almost the same like mouse counterpart. The predicted structure also suggested that cKLF4 protein is composed of 12 α-helices, 29 β-sheets, 58 coils, and 49 turns.

The main secondary structure elements were three zinc finger domains of both cKLF4 and mKLF4 encompassing residues 418–501 and 395–474, respectively. We generated de novo 3D model of cKLF4 protein (Fig. 10). The 3D predicted structure of the cKLF4 protein predicted the N-terminus to be unfolded. It still stays ambiguous that which C-terminal part of the remaining Klf4 protein is responsible for the observed induction of self-renewal and inhibition of differentiation. The quality of the predicted structure of cKLF4 and mouse KLF4 was TM-score = 0.40 and 0.32, respectively. TM-score has the value in [0, 1], in which the value 1 indicates a perfect match between the two structures.

Similarities between structure of cKLF4 and mouse KLF4. The similarities between cKLF4 and mKLF4 were studied by superimposing their structures using Protean 3D (Lasergene 12; DNASTAR). The overall folds of predicted cKLF4 structure at its C-terminus were highly similar to C-terminus of mouse KLF4 (Fig. 11). The overall root-mean-square deviation [RMSD]) between cKLF4 protein and mouse KLF4 protein structures was 3.493.

Discussion

The predicted isoelectric point (pI) was found to be 8.74. The N-terminal of the sequence was considered M (Met). The chemical composition of the predicted protein is illustrated in Figure 1 and Supplementary Table 1. The cKLF4 protein is rich in proline (P) and serine (S) residues, which constitute 13.8% and 12% of the total amino acids, respectively. It has been shown that transcription factors are rich in these two amino acids.[20] The molecular analysis of the whole cKLF4 protein using the program Protean (Lasergene 12; DNASTAR) showed that it contains 133 charged amino acids (26.5%), 133 hydrophobic amino acids (26.5%), 39 acidic amino acids (7.7%), 49 basic amino acids (9.7%), and 130 polar amino acids (25.9%). Moreover, the distribution of hydrophilic and hydrophobic amino acids of cKLF4 using Eisenberg's method (Supplementary Fig. 3)[21] showed that C-terminus of cKLF4 is more hydrophobic than its N-terminus.

In general, the amino acid alignment of the cKLF4 protein and those from six mammalian species has shown that the C-terminus is more conserved than the N-terminus (Fig. 2). The phylogenetic tree showed that Arabian and Bactrian camels shared a more recent common ancestor with each other

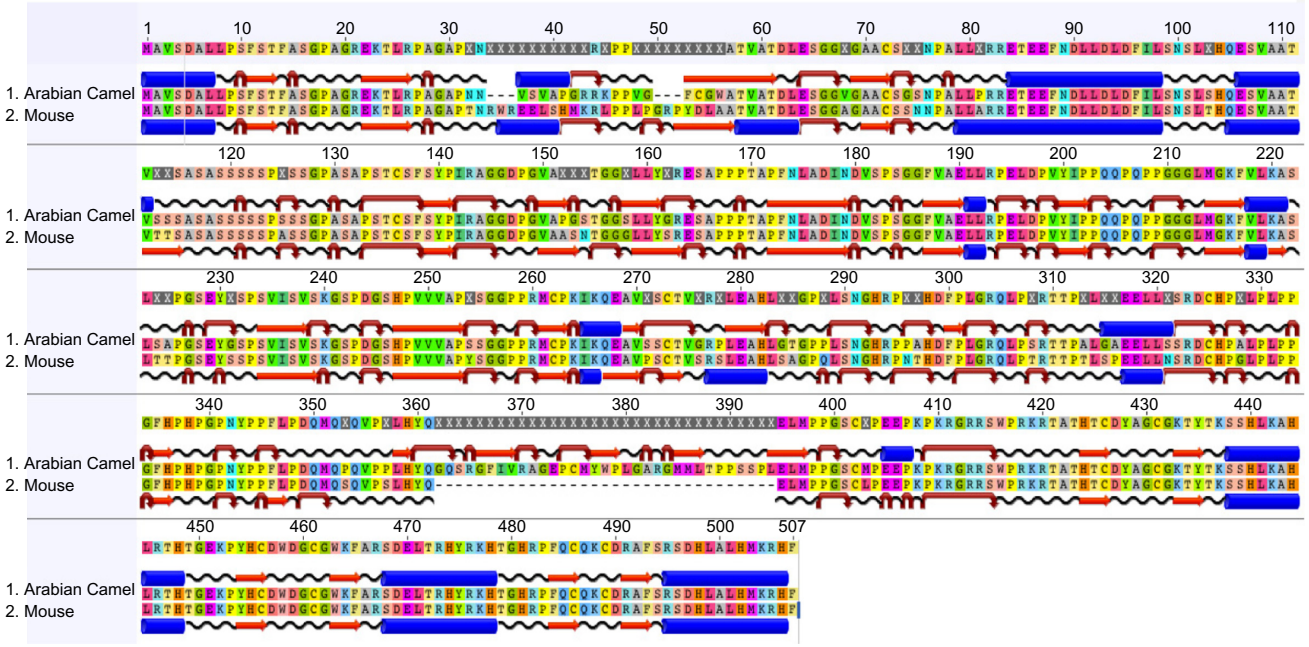

Figure 9. The predicted secondary structure sites of the cKLF4 and mKLF4 protein sequences. Blue cylinders and red arrows indicate α-helix and β-sheets, respectively.

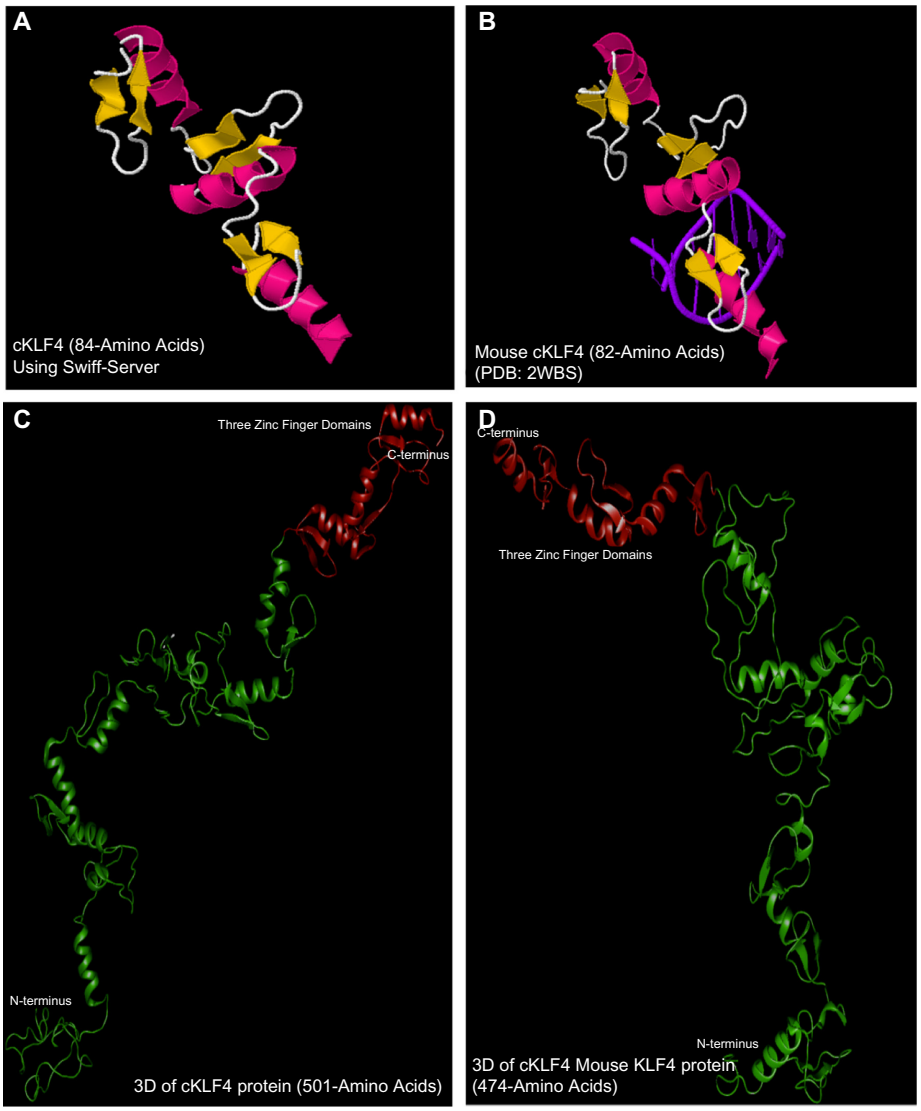

Figure 10. 3D structure models. (**A**) Predicted for 84 residues of cKLF4. (**B**) Experimental model (PDB: 2WBS) for mouse KLF4 that best matches with cKLF4. (**C**) The fully predicted 3D structure model for cKLF4 protein and (**D**) for mouse KLF4 protein. Red regions indicate three zinc finger domains.

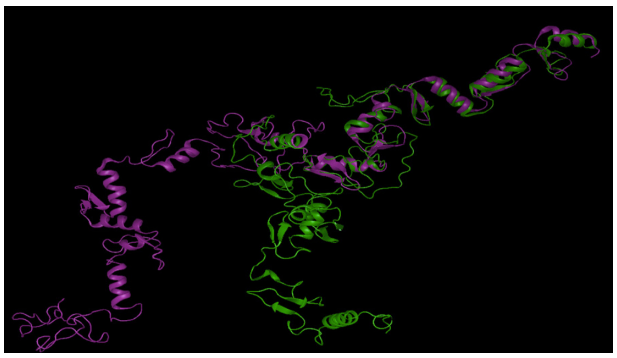

Figure 11. Stereo ribbon representation of the predicted 3D structure of cKLF4 (violet) and the superimposition with mouse KLF4 (green).

than other species (Fig. 3). Arabian and Bactrian camels and alpaca form a clade (monophyletic group) that consists of an ancestral species. It has been estimated that the divergence time between camels and cattle is 42.7 million years ago, and the estimated divergence time of the ancestors of alpaca and the two camels is 16.3 million years.[7] In Figure 3, branches represent evolutionary lineages changing over time. Ancestor of Arabian, Bactrian, and Alpaca camels exists prior to ancestor of cattle and pig, and time is approximately flowing from left to right. The numbers next to each node represent a measure of support for the node (confidence level for the obtained phylogenetic tree). These are generally given as percentages where 100% represents maximal support. A high value means that there is strong evidence that the sequences to the right of the node cluster together to the exclusion of any other. We used Bootstrap resampling, which was repeated 1,000 times for each test, and obtained >90% confidence except for the node clustering cattle and pig, which was 50.9%, indicating that both cattle and pig can make clustering with any other understudied species.

In Arabian camel, the amino acid residues 418–501 form the three zinc finger domains (Fig. 4). These three C_2H_2 zinc fingers are located at residues 418–442, 448–472, and 478–501. The two H/C links are TGEKP and TGHRP. The degree of sequence identity in the 84 amino acids of zinc finger domains of cKLF4 to the other six understudied species is 100%, as shown in Figure 4. Each zinc finger has a conserved ββα structural sequence and binds one zinc ion that is inserted between the two-stranded antiparallel β-sheets and the one α-helix. Each zinc ion tetrahedral coordinates to the side chains of two cysteines at the end of the second β-sheet and two histidines in the C-terminal of the α-helix (Supplementary Fig. 1). This zinc binding site increases the thermal and conformational stability of KLF4 domains but may not be directly involved in its function.[22,23]

As shown in Figure 5, the N-terminal part of cKLF4 protein is predicted to be disordered, whereas its C-terminal part is predicted to be ordered. The cKLF4 protein contains more disordered regions than a globular one; there are 10 predicted disordered regions within cKLF4 protein as follows: 8–47, 60–72, 110–179, 192–209, 218–257, 272–303, 318–354, 374–410, 420–428, and 444–456, and there are six predicted globular domains as follows: 48–109, 180–191, 258–271, 304–317, 355–373, and 411–501. As a result, cKLF4 can be classified as an intrinsically disordered protein. It has been suggested that most transcription factors contain high amounts of disordered regions, indicating that there is an intrinsic requirement for these transcription factors to be highly flexible, and thus to be able to interact with other proteins and DNA partners.[24]

In Figure 7, all amino acids/regions with disorder disposition higher than 0.5 score are predicted to be disordered. We used four meta-predictors because they use different predictive approaches and emphasize different features of the sequence. In general, the graph reveals that of the 501 amino acids of cKLF4 protein, more than 50% are in the disordered regions (above the threshold of 0.5). The VLXT predictor (gray), whose accuracy reached 70%[25] and integrated three different predictors, clearly shows that 11 regions within cKLF4 protein, namely, 10–20, 40–55, 85–100, 170–180, 215–225, 260–265, 290–300, 325–335, 350–375, 425–455, and 475–501, have a higher tendency of being structured and flanked by disordered regions. The accuracy of this predictor is low to predict short regions (<10 amino acids) of disorder. However, this predictor has significant advantages in finding potential binding sites in proteins.[25]

VL3 predictor has higher accuracy in predicting longer unstructured regions.[25] The predictor (red) predicts the whole cKLF4 protein to be mostly disordered except for its C-terminal that located three zinc finger domains. Similarly, VSL2B predictor (blue) uses the result of sequence alignments from PSI-blast and second structure prediction from PHD and PSI-Pred; hence, it is the most accurate predictor. Both VL3 and VSL2B predicted ~82% of cKLF4 protein to be disordered.

P-FIT predictor, also known as meta-predictor, is a combination of several individual predictors. This predictor uses a collection of results from many individual predictors as its input. The general trend of P-FIT predictor (green) is very similar to VLXT predictor except at N- and C-terminal regions in which VLXT predictor shows a stronger effect of termini. The most differences between P-FIT and VLXT predictors are nine sharp dips found by VLXT predictor at AA14 (F), AA47 (F), AA96 (L), AA175 (V), AA217 (L), AA264 (V), AA294 (F), AA365 (A), and AA487 (F); usually these dips indicate Molecular Recognition Feature (MoRF) region within protein which in general has much higher content of aliphatic and aromatic amino acids than disordered regions.[26] To summarize the results in Figure 7 and Supplementary Figure 5, all predictors accurately predicted the ordered C-terminal of cKLF4 protein, whereas its N-terminal is disordered. cKLF4 is classified as an intrinsically disordered protein.

The sequence-to-structure-to-function concept does not seem suitable for cKLF4 protein since it has more disordered than ordered regions within its structure. Little is known about the cellular function of the different domains of Klf4 protein. Moreover, no structural information about the DNA-binding properties of the zinc finger domain of Klf4 proteins is available. A major problem with experimental structure (eg, X-ray structure) is that it is restricted to what is found in the Protein Data Bank (PDB). Many of these structures are incapable of representing protein explicitly owing to the regions that are cut out are disordered/unfolded or highly flexible regions of a protein. To date, only crystal structures of related C2H2-type zinc finger proteins are for mouse KLF4 (PDB: 2WBS) with 82 residues. The 3D structure of cKLF4 was modeled using homology structure modeling on the Swiss model server. To predict the 3D modeling structure of cKLF4, a 3D structure at 2.7°A of mouse KLF4 (PDB: 2WBS), which shared 59% sequence identity, was applied.

Our results based on full-length mRNA, homology, read sequencing quality, and comparative genetic analysis suggest that we have successfully achieved KLF4 mRNA in the Arabian camel that matched known coding sequences in other mammalian species. Our work based on comparative cKLF4 mRNA will have a significant impact on induced pluripotent stem cells (iPSCs) research since we have sequenced and described one of the reprogramming transcriptional factors, which is the backbone of the iPSCs technology. To the best of our knowledge, the data presented here represent novel Arabian camel KLF4 mRNA sequence data as well as its 3D modeling protein, and we believe that this genetic and structural information will become a helpful resource for the annotation.

Author Contributions

Conceived and designed the experiments: AOA, SNA. Analyzed the data: SNA. Wrote the first draft of the

manuscript: SNA. Contributed to the writing of the manuscript: AOA, SNA, OAA, FSA, MNA, ZAA, ADA, MH, SAA, HAA, IOA. Agree with manuscript results and conclusions: AOA, SNA, OAA, FSA, MNA, ZAA, ADA, MH, SAA, HAA, IOA. Jointly developed the structure and arguments for the paper: SNA, OAA. Made critical revisions and approved final version: SNA, FSA. All authors reviewed and approved of the final manuscript.

Supplementary Material

Supplementary Table 1.
Supplementary Figure 1.
Supplementary Figure 2.
Supplementary Figure 3.
Supplementary Figure 4.
Supplementary Figure 5.

REFERENCES

1. Simara P, Motl JA, Kaufman DS. Pluripotent stem cells and gene therapy. *Transl Res*. 2013;161:284–92.
2. Shields JM, Yang VW. Identification of the DNA sequence that interacts with the gut-enriched Krüppel-like factor. *Nucleic Acids Res*. 1998;3:796–802.
3. Crossley M, Whitelaw E, Perkins A, Williams G, Fujiwara Y, Orkin SH. Isolation and characterization of the cDNA encoding BKLF/TEF-2, a major CACCC-box-binding protein in erythroid cells and selected other cells. *Mol Cell Biol*. 1996;16:1695–705.
4. Pearson R, Fleetwood J, Eaton S, Crossley M, Bao S. Krüppel-like transcription factors: a functional family. *Int J Biochem Cell Biol*. 2008;40:1996–2001.
5. Al-Swailem AM, Shehata MM, Abu-Duhier FM, et al. Sequencing, analysis, and annotation of expressed sequence tags for *Camelus dromedarius*. *PLoS One*. 2010;5:e10720.
6. Bactrian Camels Genome Sequencing and Analysis Consortium, Jirimutu, Wang Z, et al. Genome sequences of wild and domestic bactrian camels. *Nat Commun*. 2012;3:1202.
7. Wu H, Guang X, Al-Fageeh MB, et al. Camelid genomes reveal evolution and adaptation to desert environments. *Nat Commun*. 2014;5:5188.
8. Almubarak AI. Effects of some anaesthetics and chemical restraints on blood clotting in camels. *J Anim Vet Adv*. 2007;6:33–5.
9. Kearse M, Moir R, Wilson A, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647–9.
10. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 2006;22:195–201.
11. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. 2003;31:3701–8.
12. Dosztanyi Z, Meszaros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*. 2009;25:2745–6.
13. Troshin PV, Procter JB, Barton GJ. Java bioinformatics analysis web services for multiple sequence alignment – JABAWS:MSA. *Bioinformatics*. 2011;27:2001–2.
14. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*. 2005;21:3369–76.
15. Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. *J Mol Graph Model*. 2001;19:26–59.
16. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 2006;7:208.
17. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta*. 2010;1804:996–1010.
18. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol*. 2009;5:e1000376.
19. Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng*. 1990;4:155–61.
20. Nakamura T, Alder H, Gu Y, et al. Genes on chromosomes 4, 9, and 19 involved in 11q23 abnormalities in acute leukemia share sequence homology and/or common motifs. *Proc Natl Acad Sci U S A*. 1993;90:4631–5.
21. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci U S A*. 1984;81:140–44.
22. Laity JH, Lee BM, Wright PE. Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol*. 2001;11:39–46.
23. Wolfe SA, Grant RA, Elrod-Erickson M, Pabo CO. Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. *Structure*. 2001;9:717–23.
24. Xue B, Oldfield CJ, Van Y-Y, Dunker AK, Uversky VN. Protein intrinsic disorder and induced pluripotent stem cells. *Mol Biosyst*. 2012;8:134–50.
25. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. Mining α-helix-forming molecular recognition features with cross species sequence alignments†. *Biochemistry*. 2007;46:13468–77.
26. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). *J Mol Biol*. 2006;362:1043–59.