Cross-Species Identification of Hydroxylation Sites for ARD and FIH Interaction

Ying-Tsang Lo[1], Tsan-Huang Shih[1], Han-Jia Lin[2,3], Tun-Wen Pai[1,3*], Margaret Dah-Tsyr Chang[4]
[1] Dept.of Computer Science and Engineering, [2] Institute of Bioscience and Biotechnology, [3] Center of Excellence for Marine Bioenvironment and Biotechnology, Nation Taiwan Ocean University, Keelung, Taiwan
[4] Institute of Molecular and Cellular Biology & Department of Medical Science, National Tsing Hua University, Hsinchu 300, Taiwan
[*]twp@mail.ntou.edu.tw

***Abstract*—Ankyrin repeat domain (ARD) proteins contain various numbers of internal repeat units. They are considered as one important factor to influence hypoxia response through hydroxylation interaction with Factor Inhibiting HIF (FIH) enzymes which can repress HIF under normoxia environment. In this study, we adopted sequence based method and applied conserved hydroxylation motif patterns for identifying ASN/ASP/HIS hydroxylation sites on ARDs. First, a set of known ARD proteins was collected, and all corresponding repeat units were manually constructed and verified by removing redundant units. All extracted segments served as fundamental seed units to retrieve all ARDs proteins from 5 different species. Those ARD candidates were automatically segmented and a conserved hydroxylation motif pattern was applied for identifying all hydroxylation sites. As a result, the retrieval performance for ARDs achieved a sensitivity of 82% and a specificity of 98% for human species based on a testing dataset of 1,244 protein sequences. For hydroxylation site prediction, a sensitivity of 72.2% and a positive prediction value of 62% were achieved based on a set of 18 experimentally verified hydroxylation residues.**

Keywords:HIF, FIH, ARD, internal repeat, hydroxylation

I. INTRODUCTION

Hypoxia-inducible factors (HIFs) are transcription factors that play a crucial role in response to hypoxic stress for all metazone organisms. Maintaining constant concentration of oxygen for efficient metabolism in organisms is the first challenge to overcome which leads to continue all life activities. It is well known that HIFs in hypoxia environment are responsible for regulation of metabolism, apoptosis, proliferation, and angiogenesis of new blood cells. The biological mechanism of hypoxia response involves increasing quantity of HIF-α and HIF-β, formation of heterodimers as transcription factors to regulate downstream gene expression [1]. More specifically, both HIF-α and HIF-β are members of bHLH (basic helix loop helix) – PAS (Per/ARNT/Sim) superfamily with high homologous conservation. Interestingly, the HIF-α subunit is sensitive to oxygen but HIF-β is relatively stable and insensitive to oxygen. After heterodimerization, the complexes can recognize and bind to hypoxia responsive elements (HREs) located within the promoter regions of all HIF target genes. In combination with transcriptional co-activating proteins such as CBP and p300 coactivators, HIFs are directly involved in transcriptional activation and regulation of target genes [2, 3]. The up-stream protein interaction activities of HIF-α is mainly controlled by two enzymes, prolyl hydroxylases (PHD) and factor inhibiting HIF (FIH) [4, 5]. Under normoxia conditions, HIFs are hydroxylated by PHD enzymes, followed by ubiquitylation mechanism of being captured by ubiquitin ligase Von-Hippel-Lindau protein (VHL), and then rapidly degraded by the proteasome pathways. The other way of HIF protein degradation through hydroxylation by FIH enzymes at C-TAD structural domain, which inhibit the interaction between HIF-α and CBP/p300 coactivators, and therefore decrease the formation of basal transcription complex. Inversely, under hypoxia conditions, oxygen-dependent PHD and FIH hydroxylases become inactive, and the HIF–α subunit can stably translocate into cell nucleus and heterodimerizes with HIF-β to induce DNA binding with target genes at the HREs. Accordingly, interaction with CBP/p300 initiates the induction or repression of a large number of genes such as vascular endothelial growth factor (VEGF) or erythropoietin (EPO) involved in angiogenesis and glucose metabolism, and through the mechanism, an individual organism is able to physically adapt itself to oxygen changes in the external environments[6]. It has been reported that HIFs affect human cardiovascular related diseases and various cancers since Semenza and Wang's discovery in 1992[8-10]. If the PHD and/or FIH hydroxylases can be controlled under hypoxia environments, it will provide a novel strategy for clinical-pathological research. However, there is another interesting discovery that FIH prefers hydroxylation interactions with ankyrin repeat domain (ARD) proteins compared to HIF-α subunits[7]. Therefore, the hypoxia responses caused by HIFs can be indirectly degraded if the ARD hydroxylation processes can be under control [12]. ARD proteins involve different physiological reactions and are present in abundance within a cell governing various functions such as cytoskeleton integrity, cell cycle control, transcriptional regulation, cell signaling, development and differentiation, apoptosis, inflammatory response, plant defense, and bacterial invasion[8, 9]. The fundamental unit of ankyrin repeat is composed of 30-34 residues consisting of two antiparallel alpha helices separated by loops[10]. The repeat number ranges from one to dozens diversely. Each internal repeat possesses high structure conservation but low homology in sequence. Hence, ARD structural surface may have different amino acid contents at corresponding locations in general [11]. It was experimentally demonstrated that the

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

positions of asparagine hydroxylation for both HIF-α and ARD are quite similar for being recognized by FIH [12]. In addition to asparagines hydroxylation, aspartic acid and histidine hydroxylation were also experimentally discovered in ARDs [17, 18]. Therefore, this study aims at retrieving all possible ARD proteins within a specified organism and identifying all possible hydroxylation sites according to a well conserved motif pattern. In this study, the *in silico* analysis of ARD protein prediction for various organisms including human, mouse, zebrafish, sea urchin, and lancelet was performed for cross-species comparison and the prediction accuracy of hydroxylation sites was also verified based on the experimental results from previous reports.

II. MATERIALS AND METHODS

To retrieve all possible ARD proteins and identify hydroxylation sites from a specified species, a comprehensive genome data of the designated organism is required. Here, three model species and two non-model species were collected from Ensembl, JGI and SpBase databases, respectively. Protein sequences of model species of human, mouse, and zebrafish were obtained from Ensembl in the released version of 62 including *Homo sapiens* (GRCh37), *Mus musculus* (NCBIM37), and *Danio rerio* (Zv9)[12], and the other two non-model species are lancelet collected from JGI database (*Branchiostoma floridae*, v1.0)[13] and sea urchin from SpBase (version SPU2.6)[14].

To increase retrieval accuracy rates, a set of verified ARD proteins should be trained and analyzed. Here we adopted the data from Schmierer's report[15]. Each protein in the collected dataset was initially segmented into fundamental repeat units and stored as the basic query patterns for following retrieval processes. The Schmierer's dataset contains 252 protein sequences which were initially collected from SMART[16], PFAM[17] and UniProt[18], and these collected ARD proteins were further examined by removing redundant sequences. In this study, only 250 proteins were applied since two proteins from Schmierer's dataset cannot be mapped onto UniProt IDs successfully. The first protein ID is "Q7Z6C4" which was already removed by UniProt, and the second one is "Q6UX02" which was replaced by "A6QC64". Based on these representative ARD proteins, segmentation algorithm was performed to identify all individual repeat units automatically. Here, the ARD proteins in Schmierer's dataset were segmented into a total number of 1,505 fundamental units. Accordingly, these repeat units were evaluated by performing multiple sequence alignment for discovering a conserved pattern of hydroxylated segments, especially for protein interaction between FIH and ARD proteins. Based on the conserved pattern and hydroxylation residue information, all hydroxylation sites of ARD proteins within various species were predicted. All detailed processes are described in the following sections.

A. ARD Retrieval

The total number of internal repeat units in ARD proteins varies in a wide range. Besides, the sequence contents of each unit in an ARD protein possess low sequence similarity. For example, the average sequence identity of repeat units within an ARD protein in Schmierer's dataset is about 0.235. According to low sequence variations among intra repeat units and the distinct number of repeat units of an ARD protein, blasting the target protein database based on repeat units could provide better and more complete results than performing matching algorithm based on its full sequence contents.

The fundamental matching tool for cross-species retrieving of ARD proteins adopted BLAST algorithms in this study. The numbers of retrieved ARD proteins using full sequences in Schmierer's dataset were respectively 1,231 entries for human, 659 for zebrafish, 769 for mouse, 886 for sea urchin, and 635 for lancelet. Using segmented units as query sequences instead of full sequences will give more comprehensive result, since remotely evolved ARD proteins might be composed of various fundamental repeat units. In this study, all 1,505 fundamental ARD repeat units were processed to remove redundancy. All pairwise sequence identities higher than 0.8 were removed and only 1,286 fundamental repeat units were remained as a primitive ARD seed unit dataset. Then, these representative ARD seed units were compared to a specifically assigned genome dataset, and those found proteins will be verified again by removing all repeated protein IDs. For example, a dataset of 106,177 non repeated protein IDs were finally retrieved from human protein database in UniProt. This assumption is reasonable for cross-species analysis. As long as only one repeat unit is conserved during evolutionary history, the protein can still be retrieved by the proposed algorithms.

B. Internal repeat identification for ARD proteins

Once an ARD protein candidate was retrieved from the target database, the boundaries of internal repeats of the protein should be identified for hydroxylation site prediction in next step. The internal repeat identification algorithm was adopted using internal repeat identification system (IRIS)[19], a tool for judging whether a protein sequence or a structure contains repeat structures including domain repeats, solenoid repeats, and fibrous repeats, and identification the boundaries of the repeat segments. The IRIS has collected most of repeat proteins and all corresponding repeat units were stored for comparison. Both sequence and structural features were employed in IRIS for automatic identification of repeat units. In this study, the Schmierer's ARD seed units were adopted as a constrained dataset and used to retrieve all sequences containing the ARD repeat seed units. All the ARD protein candidates were fed into IRIS for internal repeat identification. The target units were limited with a length of ± 20% of original ARD seed units. Figure 1 shows the statistics of identified repeat units for various species through IRIS detection. The majority of ARD repeat units in a candidate protein ranges from 2 to 6. The maximum length between two sequential ARD units was set as 10 residues.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

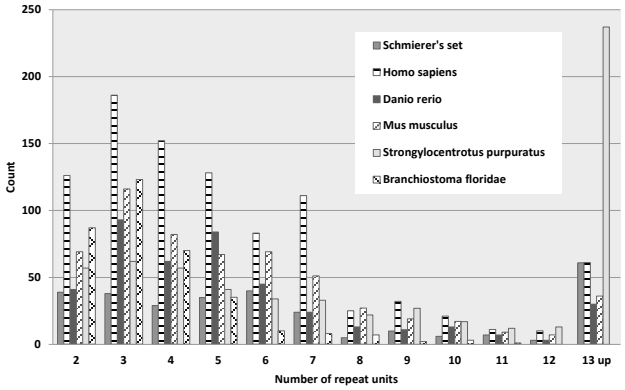

Figure 1. Statistics of number of internal repeat units within the retrieved ARD proteins for various species. Interanl repeat identification was achieved by IRIS.

C. FIH-ARD hydroxylation residues

From previous report, the hydroxylation residues on ARD units were located on β-chain [19] and preserved a consensus motif pattern of "LXXXXXXXX**N**" (L-8N) [15, 20]. Accordingly, this rule was followed to identify the hydroxylation sites from detected ARD repeat units. From previous identified ARD repeat units, the statistical analysis of FIH-ARD binding segments were shown in Figure 2. Beside the pattern limitation, these searched motifs were constrained to locate at the positions near the end of the unit within 10 residues. Comparing the retrieved ARD datasets of various species to Schmierer's dataset, though our retrieved ARD units possesses one thousand units more than Schmierer's dataset, the proportional ratios of the hydroxylation patterns of the FIH-ARD binding segments are well conserved with respect to the human species. In addition, except lancelet, more than 50% of retrieved ARD proteins may possess more than two L-8N patterns, suggesting the possibility of more than one hydroxylation site being interacted between FIH and ARD proteins.

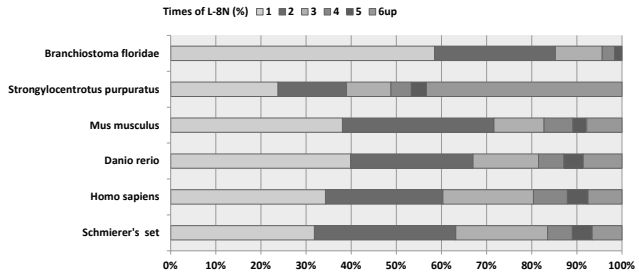

Figure 2. Proportional ratios of detected hydroxylation patterns within all retrieved ARD proteins.

III. RESULT

A. Evaluation of internal repeat segmentation

To evaluate the performance of internal repeat segmentation from all retrieved ARD proteins, we have collected all annotated proteins through keyword searching in UniProt. The keywords of "ankyrin repeat" were applied and totally 872 proteins were retrieved in this study. After removing the 250 ARD proteins from Schmierer's dataset, 622 retrieved ARD proteins were considered as the positive group for further evaluation. In UniProt database, there are 106,799 proteins for human (date: 2011.06.14), among which 106,177 proteins were not annotated with the keyword "ankyrin repeat" in description. Hence, we randomly selected 622 proteins from the non-ankyrin dataset as the negative group. Hence, 1,244 testing protein sequences were applied for ARD repeat detection by our developed IRIS approaches. If the query protein with "ankyrin repeat" description was correctly segmented by IRIS, then the number of true positive (TP) increased by one; if the query protein without "ankyrin repeat" description was misrecognized as an ARD protein and incorrectly segmented by IRIS, then the false positive (FP) increased by one; if the query protein with "ankyrin repeat" description was rejected for segmentation by IRIS, then the false negative (FN) increased by one; if the query protein without "ankyrin repeat" description was rejected for segmentation by IRIS, then the true negative (TN) increased by one. Two parameters of sensitivity and specificity were calculated by the following equations.

$$Sensitivity = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

In this experiment, we have obtained a sensitivity of 82% and a specificity of 98% for human species. The misrecognized proteins in false negative prediction cases might be due to requirement of at least two ARD repeat units in our algorithms. It should be noted that the intermediate loop segments between two consecutive ARD units should be less than 10 amino acids. However, in practical, both assumptions may not be satisfied for few cases.

B. Prediction accuracy of hydroxylation sites

To verify the accuracy of predicted hydroxylation sites, we applied three proteins possessing experimentally demonstrated hydroxylation residue information. The UniProt IDs of these three protein sequences are P46531, Q9H2K2, and P16157 [21-23]. These sequences were not contained in our IRIS template datasets for a fair system evaluation. The original location of ARD repeats within three proteins and the corresponding positions predicted by IRIS was shown in Table II-IV, respectively. After internal repeat segmentation, the hydroxylation pattern motif was scanned backwards from right-hand side. If the patterns were matched, the corresponding hydroxylation residues would be underlined. It is noticed that the predicted boundaries of ARD units were slightly different from the annotated positions from published papers. However, several ARD units not satisfying the length limitation were discarded from our prediction system. The conserved hydroxylation patterns in this study include L-8N, L-8D, and L-6DVH [18], and the total numbers of identified patterns in five different species were shown in Table I. Only the lancelet species cannot find any L-6DVH motif from the 183 retrieved ARD proteins. Since only few papers provide comprehensive experimental results regarding to protein hydroxylation, we can only list the predicted hydroxylation residues for these three representative proteins. For P46531, there are 2

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

experimentally verified asparagine hydroxylation residues; for Q9H2K2, 6 asparagine hydroxylation residues and 2 histidine hydroxylation residues; for P16157, 6 hydroxylation residues. Based on the conserved pattern detection, our predicted hydroxylation sites correctly hit 2 asparagines hydroxylation residues for P46531, 6 hydroxylation residues for Q9H2K2, and 5 hydroxylation residues for P16157. Totally, a sensitivity of 72.2% was achieved (13 hydroxylation sites were correctly predicted from 18 experimental verified residues), and a positive predictive rate of 62% was obtained.

C. *L-8N pattern statistics for various species*

Adopting the same approaches, we have retrieved all possible ARD proteins from four other species. The total numbers of identified ARD proteins were listed in Table 1. There are 653 ARD proteins for human, 423 for mouse, 303 for zebrafish, 508 for sea urchin, and 183 for lancelet. Compared to Schmierer's set, more than four hundred ARD proteins were retrieved and three folds of hydroxylation sites were identified through IRIS analysis. The weblogo plots of identified hydroxylated patterns were shown in Figure 3. It should be noticed that even for distant species, the hydroxylated motifs conserved very well.

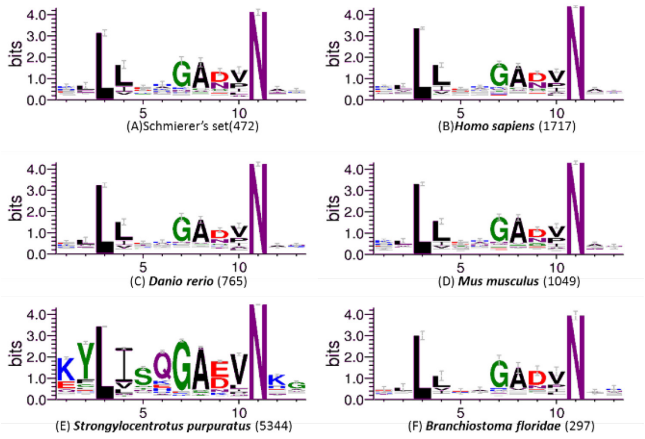

Figure 3. Conserved hydroxylated motifs within identified ARD proteins from various species.

TABLE I. NUMBERS OF IDENTIFIED ARD PROTEINS AND CONSERVED HYDROXYLATION MOTIFS.

	Protein Sequence	Predict L-8N motif	Predict LxxxxxDVN motif
Homo sapiens	653	1717	26
Danio rerio	303	765	12
Mus musculus	423	1049	15
Strongylocentrotus purpuratus	508	5344	21
Branchiostoma floridae	183	297	0

IV. DISCUSION AND CONCLUSION

Identification of hydroxylation sites for ARD and FIH interaction play an important role in HIF transcriptional responses, which is one of the most crucial issues in response to hypoxic stress for all metazone organisms. The quantity of ARD protein influences the hydroxylation processes between ARD and FIH proteins, which indirectly affects the inhibiting ability of FIH on HIF. Increasing hydroxylation processes for ARD-FIH will decrease the inhibition of FIH-HIF such that more HIF–α subunits can translocate into cell nucleus and heterodimerizes with HIF-β to induce DNA binding with target genes at the HREs. Inversely, decreasing the ARD-FIH hydroxylation events will increase the inhibitory effects of FIH such that less HIF-α subunits can translocate into cells to regulate downstream target genes. Hence, it is important to discover all ARD proteins and identify corresponding hydroxylation sites in a species of interest. To predict the hydroxylation sites within an ARD repeat unit, the proposed method adopted sequence matching and conserved pattern analysis. The highly conserved pattern motifs of hydroxylation sites are adopted to identify all possible residues for asparagines (ASN) / aspartic acid (ASP) / histidine (HIS) hydroxylation. These results provide biologists and medical doctors key information for their biological experiments on enzyme activity research and drug discovery. Furthermore, we have successfully retrieved and identified all possible ARD proteins and hydroxylation sites for two other model and two non-model species through cross-species comparison. The retrieved data can facilitate evolutionary analysis of HIF related gene expression divergence on a large scale.

ACKNOWLEDGMENT

This work is supported by Center of Excellence for Marine Bioenvironment and Biotechnology of National Taiwan Ocean University and the National Science Council, Taiwan, R.O.C. (NSC 99-2627-B-019-007 and NSC 100-2321-B-019-004 to T.-W. Pai and NSC 99-2627-B-007-001 to M. D.-T. Chang)

REFERENCES

[1] J. W. Lee, *et al.*, "Hypoxia-inducible factor (HIF-1)alpha: its protein stability and biological functions," *Exp Mol Med,* vol. 36, pp. 1-12, Feb 29 2004.

[2] K. Lisy and D. J. Peet, "Turn me on: regulating HIF transcriptional activity," *Cell Death Differ,* vol. 15, pp. 642-9, Apr 2008.

[3] D. Lando, *et al.*, "FIH-1 is an asparaginyl hydroxylase enzyme that regulates the transcriptional activity of hypoxia-inducible factor," *Genes Dev,* vol. 16, pp. 1466-71, Jun 15 2002.

[4] J. D. Webb, *et al.*, "Hypoxia, hypoxia-inducible factors (HIF), HIF hydroxylases and oxygen sensing," *Cell Mol Life Sci,* vol. 66, pp. 3539-54, Nov 2009.

[5] K. S. Hewitson, *et al.*, "Hypoxia-inducible factor (HIF) asparagine hydroxylase is identical to factor inhibiting HIF (FIH) and is related to the cupin structural family," *J Biol Chem,* vol. 277, pp. 26351-5, Jul 19 2002.

[6] G. L. Semenza, "Hydroxylation of HIF-1: oxygen sensing at the molecular level," *Physiology (Bethesda),* vol. 19, pp. 176-82, Aug 2004.

[7] S. E. Wilkins, *et al.*, "Differences in hydroxylation and binding of Notch and HIF-1alpha demonstrate

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

substrate selectivity for factor inhibiting HIF-1 (FIH-1)," *Int J Biochem Cell Biol,* vol. 41, pp. 1563-71, Jul 2009.
[8] M. A. Andrade, *et al.*, "Protein repeats: structures, functions, and evolution," *J Struct Biol,* vol. 134, pp. 117-31, May-Jun 2001.
[9] P. Forrer, *et al.*, "A novel strategy to design binding molecules harnessing the modular nature of repeat proteins," *FEBS Lett,* vol. 539, pp. 2-6, Mar 27 2003.
[10] L. K. Mosavi, *et al.*, "The ankyrin repeat as molecular architecture for protein recognition," *Protein Sci,* vol. 13, pp. 1435-48, Jun 2004.
[11] L. K. Mosavi, *et al.*, "Consensus-derived structural determinants of the ankyrin repeat motif," *Proc Natl Acad Sci U S A,* vol. 99, pp. 16029-34, Dec 10 2002.
[12] P. Flicek, *et al.*, "Ensembl 2011," *Nucleic Acids Res,* vol. 39, pp. D800-6, Jan 2011.
[13] N. H. Putnam, *et al.*, "The amphioxus genome and the evolution of the chordate karyotype," *Nature,* vol. 453, pp. 1064-71, Jun 19 2008.
[14] R. A. Cameron, *et al.*, "SpBase: the sea urchin genome database and web site," *Nucleic Acids Res,* vol. 37, pp. D750-4, Jan 2009.
[15] B. Schmierer, *et al.*, "Hypoxia-dependent sequestration of an oxygen sensor by a widespread structural motif can shape the hypoxic response--a predictive kinetic model," *BMC Syst Biol,* vol. 4, p. 139, 2010.
[16] I. Letunic, *et al.*, "SMART 6: recent updates and new developments," *Nucleic Acids Res,* vol. 37, pp. D229-32, Jan 2009.
[17] R. D. Finn, *et al.*, "The Pfam protein families database," *Nucleic Acids Res,* vol. 38, pp. D211-22, Jan 2010.
[18] M. Magrane and U. Consortium, "UniProt Knowledgebase: a hub of integrated protein data," *Database (Oxford),* vol. 2011, p. bar009, 2011.
[19] H.-Y. Kao, *et al.*, "A Comprehensive System for Identifying Internal Repeat Substructures of Proteins," presented at the Proceedings of the 2010 International Conference on Complex, Intelligent and Software Intensive Systems, 2010.
[20] J. D. Webb, *et al.*, "MYPT1, the targeting subunit of smooth-muscle myosin phosphatase, is a substrate for the asparaginyl hydroxylase factor inhibiting hypoxia-inducible factor (FIH)," *Biochem J,* vol. 420, pp. 327-33, Jun 1 2009.
[21] M. L. Coleman, *et al.*, "Asparaginyl hydroxylation of the Notch ankyrin repeat domain by factor inhibiting hypoxia-inducible factor," *J Biol Chem,* vol. 282, pp. 24027-38, Aug 17 2007.
[22] M. Yang, *et al.*, "Asparagine and aspartate hydroxylation of the cytoskeletal ankyrin family is catalyzed by factor-inhibiting hypoxia-inducible factor," *J Biol Chem,* vol. 286, pp. 7648-60, Mar 4 2011.
[23] M. Yang, *et al.*, "Factor-inhibiting hypoxia-inducible factor (FIH) catalyses the post-translational hydroxylation of histidinyl residues within ankyrin repeat domains," *FEBS J,* vol. 278, pp. 1086-97, Apr 2011.

TABLE II. P46531_HUMAN

Schmiere' Unit Set location	Ankyrin Repeat Unit Sequences	IRIS Detection location	Hydroxylation residue Predict Sites
1881-1927	GFTPLMIASCSGGGLETGNSEEEEDAPAVISDFIYQGASLHNQTDRT	1880-1926	DGFTPLMIASCSGGGLETGNSEEEEDAPAVISDFIYQGASLHNQTDR
1928-1960	GETALHLAARYSRSDAAKRLLEASADA**N**IQDNM	1926-1959	RTGETALHLAARYSRSDAAKRLLEASADA**N**IQDN
1961-1994	GRTPLHAAVSADAQGVFQILIRNRATDLDARMHD	1960-1993	MGRTPLHAAVSADAQGVFQILIRNRATDLDARMH
1995-2027	GTTPLILAARLAVEGMLEDLINSHADV**N**AVDDL	1994-2026	DGTTPLILAARLAVEGMLEDLINSHADV**N**AVDD
2028-2060	GKSALHWAAAVNNVDAAVVLLKNGANKDMQNNR	2027-2059	LGKSALHWAAAVNNVDAAVVLLKNGANKDMQNN
2061-2093	EETPLFLAAREGSYETAKVLLDHFANRDITDHM	2060-2092	REETPLFLAAREGSYETAKVLLDHFANRDITDH

TABLE III. Q9H2K2_HUMAN

2011 L-DVH' Unit Set location	Ankyrin Repeat Unit Sequences		IRIS Detection location	Hydroxylation residue Predict Sites
57-89	RKSTPLHFAAGFGRKDVVEYLLQNGANVQARDD		55-89	AGRKSTPLHFAAGFGRKDVVEYLLQNGANVQARDD
90-122	GGLIPLHNACSFGHAEVVNLLLRHGADPNARDN		90-122	GGLIPLHNACSFGHAEVVNLLLRHGADP**N**ARDN
123-155	WNYTPLHEAAIKGKIDVCIVLLQHGAEPTIRNT		123-158	WNYTPLHEAAIKGKIDVCIVLLQHGAEPTIRNTDGR
156-209	DGRTALDLADPSAKAVLKMMALLTPLNV**N**CHASDG	TGEYKKDELLESARSGNEE		
210-242	RKSTPLHLAAGYNRVKIVQLLLQHGADV**H**AKDK		208-242	DGRKSTPLHLAAGYNRVKIVQLLLQHGADV**H**AKDK
243-275	GDLVPLHNACSYGHYEVTELLVKHGACV**N**AMDL		243-275	GDLVPLHNACSYGHYEVTELLVKHGACV**N**AMDL
276-308	WQFTPLHEAASKNRVEVCSLLLSYGADPTLLNC		276-311	WQFTPLHEAASKNRVEVCSLLLSYGADPTLLNCHNK
309-362	HNKSAIDLAPTPQLKERLIKKHLSLEMVNFKHPQ	AYEFKGHSLLQAAREADVTR		
363-398	THETALHCAAASKRKQICELLLRKGANINEKTK	PYP	363-399	THETALHCAAASPYPKRKQICELLLRKGANI**N**EKTKE
399-431	EFLTPLHVASEKAHNDVVEVVVKHEAKV**N**ALDN			
432-464	LGQTSLHRAAYCGHLQTCRLLLSYGCDPNIISL		432-466	LGQTSLHRAAYCGHLQTCRLLLSYGCDP**N**IISLQG

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

465-524	QGFTALQMGNENVQQLLTVKKLCTVQSV**N**CRDIEG	QEGISLGNSEADRQLLEAAKAGDVE			
			491-524	EADRQLLEAAKAGDVETVKKLCTVQSVNCRDIEG	
525-557	RQSTPLHFAAGYNRVSVVEYLLQHGADV**H**AKDK		525-557	RQSTPLHFAAGYNRVSVVEYLLQHGADV**H**AKDK	
558-590	GGLVPLHNACSYGHYEVAELLVKHGAVV**N**VADL		558-590	GGLVPLHNACSYGHYEVAELLVKHGAVV**N**VADL	
591-623	WKFTPLHEAAAKGKYEICKLLLQHGADPTKKNR		591-626	WKFTPLHEAAAKGKYEICKLLLQHGADPTKKNRDGN	
624-677	DGNTPLDLVKDGDTDIQRVKKLSSPDNV**N**CRDTQG	DLLRGDAALLDAAKKGCLA			
678-710	RHSTPLHLAAGYNNLEVAEYLLQHGADV**N**AQDK		678-710	RHSTPLHLAAGYNNLEVAEYLLQHGADV**N**AQDK	
711-743	GGLIPLHNAASYGHVDVAALLIKYNACV**N**ATDK		711-743	GGLIPLHNAASYGHVDVAALLIKYNACV**N**ATDK	
744-776	WAFTPLHEAAQKGRTQLCALLLAHGADPTLKNQ		744-779	WAFTPLHEAAQKGRTQLCALLLAHGADPTLKNQEGQ	
777-799	EGQTPLDLVSADDVSALLTAAM				

TABLE IV. P16157_HUMAN

2011 L-8D' Unit Set location	**Ankyrin Repeat Unit Sequences**	**IRIS Detection location**	**Hydroxylation residue Predict Sites**
10-42	DAATSFLRAARSGNLDKALDHLRNGVDINTCNQ		
43-75	NGLNGLHLASKEGHVKMVVELLHKEIILETTTK	43-75	NGLNGLHLASKEGHVKMVVELLHKEIILETTTK
76-108	KGNTALHIAALAGQDEVVRELVNYGANV**N**AQSQ	76-108	KGNTALHIAALAGQDEVVRELVNYGANV**N**AQSQ
109-142	KGFTPLYMAAQENHLEVVKFLLENGANQ**N**VATED	109-142	KGFTPLYMAAQENHLEVVKFLLENGANQ**N**VATED
142-170	DGFTPLAVALQQGHENVVAHLINYGTKG----K		
171-203	VRLPALHIAARNDDTRTAAVLLQNDPNPDVLSK	171-203	VRLPALHIAARNDDTRTAAVLLQNDPNPDVLSK
204-236	TGFTPLHIAAHYENLNVAQLLLNRGASV**N**FTPQ	204-236	TGFTPLHIAAHYENLNVAQLLLNRGASV**N**FTPQ
237-269	NGITPLHIASRRGNVIMVRLLLDRGAQIETKTK	237-269	NGITPLHIASRRGNVIMVRLLLDRGAQIETKTK
270-302	DELTPLHCAARNGHVRISEILLDHGAPIQAKTK	270-302	DELTPLHCAARNGHVRISEILLDHGAPIQAKTK
303-335	NGLSPIHMAAQGDHLDCVRLLLQYDAEIDDITL	303-335	NGLSPIHMAAQGDHLDCVRLLLQYDAEIDDITL
336-368	DHLTPLHVAAHCGHHRVAKVLLDKGAKPNSRAL	336-368	DHLTPLHVAAHCGHHRVAKVLLDKGAKP**N**SRAL
369-401	NGFTPLHIACKKNHVRVMELLLKTGASIDAVTE	369-401	NGFTPLHIACKKNHVRVMELLLKTGASIDAVTE
402-434	SGLTPLHVASFMGHLPIVKNLLQRGASPNVSNVK	402-435	SGLTPLHVASFMGHLPIVKNLLQRGASP**N**VSNVK
435-467	KGFTPLHVAAKYGKVRVAELLLERDAHP**N**AAGK		
468-500	DDQTPLHCAARIGHTNMVKLLLENNANPNLATT		
501-533	AGHTPLHIAAREGHVETVLALLEKEASQACMTK	502-533	GHTPLHIAAREGHVETVLALLEKEASQACMTK
534-566	KGFTPLHVAAKYGKVRVAELLLERDAHPNAAGK	534-566	KGFTPLHVAAKYGKVRVAELLLERDAHP**N**AAGK
567-599	NGLTPLHVAVHHNNLDIVKLLLPRGGSPHSPAW	567-598	NGLTPLHVAVHHNNLDIVKLLLPRGGSPHSPA
600-632	NGYTPLHIAAKQNQVEVARSLLQYGGSA**N**AESV	600-632	NGYTPLHIAAKQNQVEVARSLLQYGGSA**N**AESV
633-665	QGVTPLHLAAQEGHAEMVALLLSKQANGNLGNK	633-665	QGVTPLHLAAQEGHAEMVALLLSKQANG**N**LGNK
666-698	SGLTPLHLVAQEGHVPVADVLIKHGVMVDATTR	666-798	SGLTPLHLVAQEGHVPVADVLIKHGVMVDATTR
699-731	MGYTPLHVASHYGNIKLVKFLLQHQADV**N**AKTK	699-731	MGYTPLHVASHYGNIKLVKFLLQHQADV**N**AKTK
732-764	LGYSPLHQAAQQGHTDIVTLLLKNGASPNEVSS	732-764	LGYSPLHQAAQQGHTDIVTLLLKNGASP**N**EVSS
765-797	DGTTPLAIAKRLGYISVTDVLKVVTDETSFV	765-795	DGTTPLAIAKRLGYISVTDVLKVVTDETSFV

978-1-4577-1666-9/11/$26.00 ©2011 IEEE