ProTG4: A Web Server to Approximate the Sequence of a Generic Protein From an in Silico Library of Translatable G-Quadruplex (*TG*4)-Mapped Peptides

Bioinformatics and Biology Insights
Volume 15: 1–7

DOI: 10.1177/11779322211045878
SAGE

Siddhartha Kundu
Department of Biochemistry, All India Institute of Medical Sciences, New Delhi, India.

ABSTRACT: An RNA G-quadruplex in the protein coding segment of mRNA is translatable $(TG4)$ and may potentially impact protein translation. This can be consequent to staggered ribosomal synthesis and/or result in an increased frequency of missense translational events. A mathematical model of the peptides that encompass the substituted amino acids, ie, the *TG*4 -mapped peptidome, has been previously studied. However, the significance and relevance to disease biology of this model remains to be established. ProTG4 computes a confidence-of-sequence-identity (γ)-score, which is the average weighted length of every matched *TG*4 -mapped peptide in a generic protein sequence. The weighted length is the product of the length of the peptide and the probability of its non-random occurrence in a library of randomly generated sequences of equivalent lengths. This is then averaged over the entire length of the protein sequence. ProTG4 is simple to operate, has clear instructions, and is accompanied by a set of ready-to-use examples. The rationale of the study, algorithms deployed, and the computational pipeline deployed are also part of the web page. Analyses by ProTG4 of taxonomically diverse protein sequences suggest that there is significant homology to *TG*4 -mapped peptides. These findings, especially in potentially infectious and infesting agents, offer plausible explanations into the aetiology and pathogenesis of certain proteopathies. ProTG4 can also provide a quantitative measure to identify and annotate the canonical form of a generic protein sequence from its known isoforms. The article presents several case studies and discusses the relevance of ProTG4-assisted peptide analysis in gaining insights into various mechanisms of disease biology (mistranslation, alternate splicing, amino acid substitutions).

KEYWORDS: Bioinformatics, mathematical and computational biology, peptidome and peptide analysis, protein sequences, translatable G-quadruplex

RECEIVED: March 23, 2021. **ACCEPTED:** August 13, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by an early career intramural grant awarded to S.K. (Code A-766) by the All India Institute of Medical Sciences, New Delhi, INDIA.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Siddhartha Kundu, Department of Biochemistry, All India Institute of Medical Sciences, Ansari Nagar, New Delhi 110029, India. Emails: siddhartha_kundu@yahoo.co.in; siddhartha_kundu@aiims.edu

Background

Non-canonical or atypical translation is the synthesis of proteins by one or more mechanism(s) that depart significantly from the usual molecular biology. This includes the synthesis of short peptides $(<100\,aa)$ from upstream open reading frames (uORF), short open reading frames (sORF)-encoded polypeptides (SEPs), and translation from atypical start sites.[1-4] A translatable G-quadruplex (*TG*4) is defined as the presence of Hoogsteen or reverse-Hoogsteen pairing between interspersed repeats of Guanosine residues in the protein coding segment (PCS) of a messenger RNA (mRNA).[5-9] The presence of one or more *TG*4 s may stall the ribosomal machinery and/or facilitate missense translations. In fact, the stress-independent amino acid substitutions that result from the latter can account for errors rates $\approx 10^{-4}-10^{-3}$ for proteins of approximately 300 amino acids.[5,10,11] Although most of these substitutions favour misfolding and thence disorder, the same may also contribute to alternate splicing and a compensatory adaptation to missing tRNA molecules under perturbed conditions.[11-15]

The combinatorial superset of protein/peptide sequences that will result under the assumptions of the 'wobble'- and 'superwobble'-hypotheses is referred to as the 'statistical'-proteome/peptidome.[5,10,11] While the myriad of forms has precluded a comprehensive empirical validation, reverse-mapping, ie, populating the 'empirical'-peptidome (mass spectrometry, microarrays) and thence characterizing the same might be an alternative strategy. In fact, protocols such as the data dependent acquisition (DDA) and quantification of rare amino acid substitutions (QRAS) have made much progress.[10] In addition, generic software which are platform-dependent and -independent may utilize advanced data analytics to compare and infer biological relevance.[16-21] However, a web-based tool which can scan the sequence of a generic protein for the occurrences of this 'statistical'-peptidome and thereby establish biological relevance is not available.

A mathematical model (codon-association with third-base wobble) of a short *TG*4 was able to establish a map with a subset of the 'statistical'-peptidome $(TG4 \sim \boldsymbol{PTG4})$.[5] Interestingly, the model was able to distinguish and quantitate the differential occurrence of *TG*4 -mapped peptides in a 6-frame translation model of empirically validated protein sequences.[5] This study also established a strong association (co-occurrence, correlation) between *TG*4 -mapped peptides and disorder-favouring short linear motifs (*SLiMS*) within and across taxa.[5,22] Despite these preliminary studies, a detailed assessment of the biomedical relevance of *TG*4 -mapped peptides (***PTG*4**) is awaited. ProTG4 will compare the full complement of *TG*4 -mapped peptides with a user-defined/generic protein sequence, assign a weight to the selected matches and compute a confidence-of-sequence-identity (γ)-score.

The impact of ProTG4 on basic and disease biology will be gauged by comparing the distribution of $TG4$-mapped peptides across taxonomically diverse protein sequences. In particular, ProTG4 will examine protein sequences from the non-haem iron(II)- and 2-oxoglutarate-dependent dioxygenase (i2OGdd) and the carbohydrate active enzyme (CAZy) enzyme superfamilies.[23-27] Members are present in all kingdoms of life, possess a conserved active site, are well characterized (in silico, laboratory), and are clinically relevant.[27-29] Furthermore, the γ-score generated by ProTG4 will be parameterized and evaluated as indices to identify and annotate the canonical form of a generic protein sequence from its known isoforms. This objective will be accomplished by evaluating known isoforms of protein sequences from a diverse set of curated and reference proteomes.

Rationale, Mathematical Derivation, and Algorithm Deployed by ProTG4

A detailed account of the mathematics (formulation, derivation, enumeration) involved in constructing a model of a short $TG4$ in the protein coding segment (PCS) of a mRNA and its mapping has already been described.[5] Briefly, SEPs with molecular weights $(\sim 0.8-2.3\text{KDa};\sim 7-20\,aa)$ were used to define the boundaries of the peptides that comprise $\boldsymbol{PTG4}$.[5] The corresponding $TG4$ was then modelled as an intra-strand subsequence of the mRNA $(\sim 20-60\,Mer)$ of the gene of a hypothetical protein.[5]

$$TG4=\left(\begin{pmatrix}\left(G_{t,k}\right)_{3\le t\le 9}\left(N_{h,k}\right)_{2\le h\le 7}\\ \left(G_{t,k}\right)_{3\le t\le 9}\end{pmatrix}_{k=3}\right)_{m=1} \qquad \text{Def. (1)}$$

t = *Number of Guanines per G – rich cluster*
h = *Number of loop – forming generic intervening nucleotides*
k = *Cluster index*
m = *Number of strands*
G = *Guanine*
A = *Adenine*
T = *Thymine*
C = *Cytosine*
N = *Any nucleotide*

A codon-association with third-base 'wobble' model was deployed to annotate a subset of vertebrate codons which was then used to identify amino acids for the Guanosine stretches $(y\in\boldsymbol{Y})$ and loops $(z\in\boldsymbol{Z})$,[5]

$$pTG4_{ij}=\left(\begin{pmatrix}\begin{pmatrix}\left(y_{i,k}\right)_{1\le i\le 3}\left(z_{i,k}\right)_{1\le i\le 2}\end{pmatrix}_{k=3}\\ \left(y_i\right)_{1\le i\le 3}\end{pmatrix}\left(z_i\right)_{1\le i\le 2}\right)_j \qquad \text{Def. (2)}$$

$\boldsymbol{PTG4}$ = *Peptidome corresponding to TG4*
$pTG4_{ij}$ = *j^{th} canonical amino acid form of PTG4 with "i" amino acids*
i = *Number of amino acids that comprise the modelled PTG4*
J = *Maximum number of canonical pTG4 for "i" amino acids*
$y\in$ $\boldsymbol{Y}$ *(set of specialized amino acids)*
$z\in$ $\boldsymbol{Z}$ *(set of all amino acids)*

The $TG4$-mapped peptidome $\left(pTG4_{ij}\in\boldsymbol{PTG4}\right)$ is then,

$$\boldsymbol{PTG4}=\bigcup_{i=7}^{i=20}\bigcup_{j=1}^{j=J}\left|\,pTG4_{ij}\,\right| \qquad (1)$$

The association $f(TG4,\boldsymbol{PTG4})$ is clearly a surjection, $f:TG4\twoheadrightarrow\boldsymbol{PTG4}$.[5]

Assessing the occurrence of *TG4*-mapped peptides in a generic protein sequence

ProTG4 is a web server that queries a protein sequence for the full complement of $TG4$-mapped peptides. The occurrences of every matched peptide sequence $\left(pTG4_{ij}\in\boldsymbol{pTG4}\subset\boldsymbol{PTG4}\right)$ are then computed in a library of randomly generated sequences $(pTG4_{ij}\in\boldsymbol{V})$ of equivalent lengths.

$$\phi_{pTG4_{ij}}=\sum pTG4_{ij} \qquad (2)$$

The weight (ω) of each positive match in a generic protein sequence is the probability that this occurrence is not due to chance, ie,

$$\omega=1-\frac{\phi_{pTG4_{ij}}}{\#\boldsymbol{V}}=1-\frac{\left(\sum pTG4_{ij}\right)}{\#\boldsymbol{V}} \qquad (3)$$

Here,

$$\#\boldsymbol{V}=\begin{cases}10^4, L\in[7,50]\,aa\\ 10^5, L\in[50,\infty)\,aa\end{cases} \qquad (4)$$

L = *Length of generic target protein sequence*
V = *Library of randomly generated sequences*
aa = *Amino acids*

The weighted length of a positive match in a generic protein sequence is then,

$$\left(\omega.l\right)_{k\in[1,\infty)}=\omega_k.l_k \qquad (5)$$

The confidence-of-sequence-identity (γ) over the entire length of the protein is,

$$\gamma=\left(\frac{1}{L}\right).\left(\sum_{k=1}\omega_k.l_k\right) \qquad (6)$$

Clearly, γ can be computed for a single protein sequence or be utilized to derive the corresponding data from entire proteomes (Table 1).

Implementation and usage of ProTG4

The computations outlined, *vide supra*, are dependent on the length of protein sequences. A reasonable $(\sim 10-20s)$ for the results may be obtained by restricting the number of user-defined sequences $(\sim 5-10)$. The server is simple to use and provides the user with a brief description of the rationale, algorithm and pipeline deployed. There are also several important instructions and precautions that the user must adhere to for relevant and timely feedback. Several ready-to-use examples (radio button) are provided to explore and comprehend the functioning of ProTG4. If ProTG4 finds suitable peptides in the user-defined sequence(s), it outputs these as tables to independent files which can be downloaded (summary, details). While the summarized data includes the sequence ID, length of the protein, number of positive matches, and confidence-of-sequence-identity (γ), details mention the amino acid sequence, start and end positions (Figure 1). ProTG4 gives consistent results when tested in three independent browsers (Chrome, FireFox, Microsoft Edge). The coding is done using in-house developed PERL scripts along standard HTML for the design and layout.

Biomedical relevance of ProTG4

The translatable G-quadruplex $(TG4)$ may be an important cause of abnormal amino acid substitutions in de novo

Table 1. ProTG4-derived confidence-of-sequence-identity scores in sets of taxonomically diverse protein sequences.

	DATA SETS	SEQUENCES	$\bar{\gamma}^{Prot}$	$\bar{\gamma}^{Prot} \in [\bar{\gamma}^{Prot}_{min}, \bar{\gamma}^{Prot}_{max}]$	REFERENCE(S)
1.	i2OGdd sequences				
	Generic	3,429	0.94	0.44-1.00	Supplementary Texts 2a and 2b
	Human	17	0.93	0.81-0.99	
	Animals	357	0.92	0.75-0.99	
	Plants	1,531	0.93	0.76-1.00	
	Fungi	324	0.94	0.44-1.00	
	Bacteria (gram positive)	205	0.96	0.83-1.00	
	Bacteria (gram negative)	995	0.95	0.62-1.00	
	Archaea	2	0.91	0.88-0.95	
2.	CAZy sequences				
	Generic	8,616	0.94	0.44-1.00	Supplementary Texts 3a and 3b
	Humans	293	0.92	0.78-0.99	
	Plants	896	0.91	0.77-0.99	
	Fungi	685	0.89	0.58-0.99	
	Bacteria	3,226	0.94	0.52-1.00	
	Virus	266	0.85	0.64-0.99	
	Metazoa	179	0.91	0.69-0.98	
	Protists	30	0.82	0.53-0.95	
3.	*H. sapiens* (UP000005640)				
	Generic	20,349	0.93	0.39-1.00	Supplementary Texts 4a and 4b
	Combined	31,897			
	Isoforms (canonical & curated)	10,175			
	Misc.	10,174			

(Continued)

Table 1. (Continued)

	DATA SETS	SEQUENCES	$\bar{\gamma}^{Prot}$	$\bar{\gamma}^{Prot} \in [\bar{\gamma}^{Prot}_{min}, \bar{\gamma}^{Prot}_{max}]$	REFERENCE(S)
4.	*C. trachomatis* (UP000000431)				
	Generic	1992	0.92	0.63-1.00	Supplementary Texts 5a and 5b
	Combined	1993			
	Isoforms (canonical & curated)	742			
	Misc.	1,251			
5.	*C. albicans* (UP000000559)				
	Generic	1447	0.88	0.40-1.00	Supplementary Texts 6a and 6b
	Combined	1478			
	Isoforms (canonical & curated)	264			
	Misc.	1,214			
5.	*D. rerio* (UP000000437)				
	Generic	3216	0.92	0.58-1.00	Supplementary Texts 7a and 7b
	Combined	3456			
	Isoforms (canonical & curated)	441			
	Misc.	3015			
	D. melanogaster (UP000000803)				
6.	Generic	4886	0.92	0.33-1.00	Supplementary Texts 8a and 8b
	Combined	6442			
	Isoforms (canonical & curated)	2335			
	Misc.	4107			
7.	*G. gallus* (UP000000539)				
	Generic	2605	0.92	0.41-1.00	Supplementary Texts 9a and 9b
	Combined	2867			
	Isoforms (canonical & curated)	414			
	Misc.	2453			
8.	*C. elegans* (UP000001940)				
	Generic	4747	0.92	0.34-1.00	Supplementary Texts 10a and 10b
	Combined	6880			
	Isoforms (canonical & curated)	3259			
	Misc.	3621			
9.	*B. taurus* (UP000009136)				
	Generic	6904	0.92	0.00-1.00	Supplementary Texts 11a and 11b
	Combined	7353			
	Isoforms (canonical & curated)	791			
	Misc.	6562			

(Continued)

Table 1. (Continued)

	DATA SETS	SEQUENCES	$\bar{\gamma}^{Prot}$	$\bar{\gamma}^{Prot} \in [\bar{\gamma}^{Prot}_{min}, \bar{\gamma}^{Prot}_{max}]$	REFERENCE(S)
10.	*R. norvegicus* (UP000002494)				
	Generic	8222	0.93	0.00-1.00	Supplementary Texts 12a and 12b
	Combined	9891			
	Isoforms (canonical & curated)	2635			
	Misc.	7256			
11.	*A. thaliana* (UP000006548)				
	Generic	16329	0.91	0.00-1.00	Supplementary Texts 13a and 13b
	Combined	18613			
	Isoforms (canonical & curated)	4155			
	Misc.	14458			

Abbreviations: CAZy, carbohydrate active enzymes; Combined, curated non-canonical and canonical sequences; i2OGdd, non-haem iron(II)- and 2-oxoglutarate-dependent dioxygenases; Misc., non-canonical and non-curated canonical sequences; *TG4* , translatable G-quadruplex; UP, Uniprot ID; $\bar{\gamma}^{Prot}$, average confidence-of-sequence-identity score for sets of generic protein sequences.

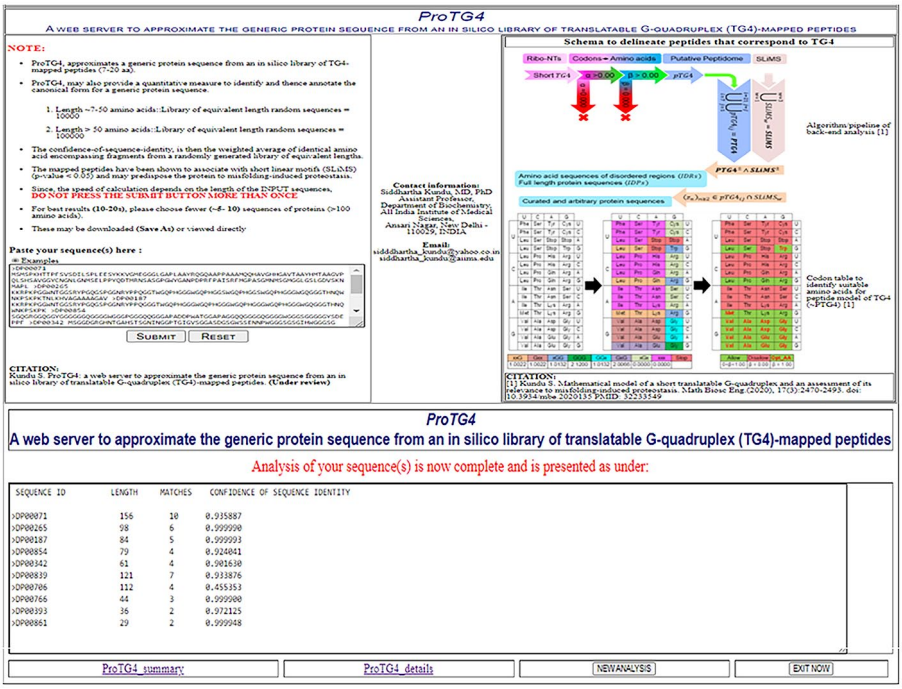

Figure 1. Implementation and usage of ProTG4: ProTG4, computes a confidence-of-sequence-identity (γ)-score, which is the average weighted length of every matched *TG4* -mapped peptide in a generic protein sequence. The weighted length is the product of the length of the peptide and the probability of its non-random occurrence in a library of randomly generated sequences of equivalent lengths. This is then averaged over the entire length of the protein sequence. ProTG4, is simple to operate, has clear instructions and is accompanied by a set of ready-to-use examples. The rationale of the study, algorithms deployed and the computational pipeline deployed are also part of the web page. The output of ProTG4 is in tabular format and written to independent files which can be downloaded (summary, details). The summarized data include the sequence ID, length of the protein, number of positive matches, and the γ -score. Details include the amino acid sequence of each matching pattern, start and end positions. mRNA indicates messenger ribonucleic acid; PCS, protein coding segment; ***PTG4*** , peptidome associated with *TG4* ; ***SLiMS*** , short linear motifs; ***TG4*** , translatable G-quadruplex.

protein synthesis.[5] The persistent of these modifications reflect 'escape' mechanisms from the proof reading machinery and are implicated in the proteopathies or diseases associated with dysfunctional proteostasis.[5,10,11,12,22,30] An interesting finding of the previous study was the redundancy of specific amino acids in proteins across taxa.[5] These were

Table 2. ProTG4-derived and confidence-of-sequence-identity based predictors of the canonical form of reference protein sequences with known isoforms.

	ORGANISM	$\#\,iprot$	γ_{min}^{iprot}	γ_{max}^{iprot}	$\gamma_{min}^{iprot} \vee \gamma_{max}^{iprot}$	TP	FN	$R=\left(\frac{TP}{TP+FN}\right).100$	REFERENCE(S)
1.	*H. sapiens*	10175	+	–	–	3230	6945	≈ 32%	Supplementary Texts 4c, 4d and 4e
			–	+	–	4104	6071	≈ 40%	
			–	–	+	7334	2841	≈ 72%	
2.	*C. trachomatis*	201	+	–	–	26	175	≈ 13%	Supplementary Texts 5c, 5d and 5e
			–	+	–	39	162	≈ 19%	
			–	–	+	65	136	≈ 32%	
3.	*C. albicans*	114	+	–	–	37	77	≈ 33%	Supplementary Texts 6c, 6d and 6e
			–	+	–	39	75	≈ 34%	
			–	–	+	76	38	≈ 67%	
4.	*D. rerio*	201	+	–	–	84	117	≈ 42%	Supplementary Texts 7c, 7d and 7e
			–	+	–	106	95	≈ 53%	
			–	–	+	190	11	≈ 79%	
5.	*D. melanogaster*	759	+	–	–	303	456	≈ 95%	Supplementary Texts 8c, 8d and 8e
			–	+	–	293	466	≈ 39%	
			–	–	+	596	163	≈ 79%	
6.	*G. gallus*	152	+	–	–	60	92	≈ 40%	Supplementary Texts 9c, 9d and 9e
			–	+	–	71	81	≈ 47%	
			–	–	+	131	21	≈ 86%	
7.	*C. elegans*	1126	+	–	–	399	727	≈ 35%	Supplementary Texts 10c, 10d and 10e
			–	+	–	490	636	≈ 44%	
			–	–	+	889	237	≈ 79%	
8.	*B. taurus*	342	+	–	–	162	180	≈ 47%	Supplementary Texts 11c, 11d and 11e
			–	+	–	159	183	≈ 47%	
			–	–	+	321	21	≈ 94%	
9.	*R. norvegicus*	966	+	–	–	349	617	≈ 36%	Supplementary Texts 12c, 12d and 12e
			–	+	–	475	491	≈ 49%	
			–	–	+	824	142	≈ 85%	
10.	*A. thaliana*	1871	+	–	–	732	1139	≈ 39%	Supplementary Texts 13c, 13d and 13e
			–	+	–	984	887	≈ 52%	
			–	–	+	1716	155	≈ 92%	

Abbreviations: FN, false negative; R, recall or sensitivity analysis of predictor; TP, true positive; $\#\,iprot$, cardinal number of the set of unique protein sequences with one or more curated isoforms; γ_{min}^{iprot}, minimum value of the confidence-of-sequence-identity among isoforms of a generic protein sequence; γ_{max}^{iprot}, maximum value of the confidence-of-sequence-identity among isoforms of a generic protein sequence; $\gamma_{min}^{iprot} \vee \gamma_{max}^{iprot}$, minimum or maximum value of the confidence-of-sequence-identity among isoforms of a generic protein sequence.

purported to be the reason for molecular mimicry, a mechanistic explanation for the secondary proteopathies.[5] Here, an offline analysis by ProTG4 of protein sequences in taxonomically diverse protein sequences (i2OGdd, CAZy) suggests that there is significant homology of *TG*4 -mapped peptides (Table 1; Supplementary Texts 1-3).[23,27,29] These findings, especially in potentially infectious and infesting (bacteria, virus, fungi, helminths) agents, offer plausible explanations into the aetiology and pathogenesis of certain proteopathies.[5] Isoforms of proteins $(iprot)$ share similar functionality and may be arise from gene duplications, re-insertional events (retrotransposition), polyploidy (aneuploidy, polyploidy) and atypical recombination. The canonical form of a generic protein sequence is annotated on the basis of several sequence-dependent and sequence-independent strategies. Here, too, the γ -score generated by ProTG4 can be parameterized $\left(\gamma_{min}^{iprot}, \gamma_{max}^{iprot}, \gamma_{min}^{iprot} \vee \gamma_{max}^{iprot}\right)$ and may offer a quantitative measure to identify and thereby annotate the canonical form of a generic protein sequence from its known isoforms (Tables 1 and 2; Supplementary Texts 4-13).

Conclusions

ProTG4 is a web server that examines the distribution of short stretches of a specialized subset of amino acids that correspond to a translatable G-quadruplex, ie, the *TG*4 -mapped peptidome in a generic protein sequence. Here, the implementation, usage and scope of ProTG4 are presented. The article also discusses the relevance of ProTG4-assisted peptide analysis in gaining insights into probable mechanisms (mistranslation, alternate splicing, amino acid substitution) of disease causation and/or progression.

Author Contributions

S.K. designed the study, formulated and developed the algorithms, collated data and conducted the analysis, and wrote all the code and the manuscript.

Availability and Implementation

The web server is available at the following URL (http://204.152.217.16/ProTG4.html), is free and does not require a login ID.

ORCID iD

Siddhartha Kundu https://orcid.org/0000-0003-3962-776X

Data Availability

All data described in the article is available as supplementary material.

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Wethmar K, Barbosa-Silva A, Andrade-Navarro MA, Leutz A. uORFdb: a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res*. 2014;42:D60-67.
2. Ma J, Ward CC, Jungreis I, et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res*. 2014;13:1757-1765.
3. Slavoff SA, Mitchell AJ, Schwaid AG, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol*. 2013;9:59-64.
4. Frith MC, Forrest AR, Nourbakhsh E, et al. The abundance of short proteins in the mammalian proteome. *PLoS Genet*. 2006;2:e52.
5. Kundu S. Mathematical model of a short translatable G-quadruplex and an assessment of its relevance to misfolding-induced proteostasis. *Math Biosci Eng*. 2020;17:2470-2493.
6. Agarwala P, Pandey S, Maiti S. The tale of RNA G-quadruplex. *Org Biomol Chem*. 2015;13:5570-5585.
7. Millevoi S, Moine H, Vagner S. G-quadruplexes in RNA biology. *Wiley Interdiscip Rev RNA*. 2012;3:495-507.
8. Hoogsteen K. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallographica*. 1963;16:907-916.
9. Garant JM, Luce MJ, Scott MS, et al. G4RNA: an RNA G-quadruplex database. *Database (Oxford)*. 2015;2015:bav059.
10. Garofalo R, Wohlgemuth I, Pearson M, et al. Broad range of missense error frequencies in cellular proteins. *Nucleic Acids Res*. 2019;47:2932-2945. doi:10.1093/nar/gky1319.
11. Ou X, Cao J, Cheng A, et al. Errors in translational decoding: tRNA wobbling or misincorporation? *PLoS Genet*. 2019;15:e1008017.
12. Hutchinson S, Furger A, Halliday D, et al. Allelic variation in normal human FBN1 expression in a family with Marfan syndrome: a potential modifier of phenotype? *Hum Mol Genet*. 2003;12:2269-2276.
13. Davey NE, Trave G, Gibson TJ. How viruses hijack cell regulation. *Trends Biochem Sci*. 2011;36:159-169.
14. Jucker M, Walker LC. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature*. 2013;501:45-51.
15. Piovesan D, Tabaro F, Micetic I, et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res*. 2017;45:D219-D227.
16. Valikangas T, Suomi T, Elo LL. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief Bioinform*. 2018;19:1344-1355.
17. Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. *BMC Syst Biol*. 2014;8(Suppl. 2):S3.
18. Efstathiou G, Antonakis AN, Pavlopoulos GA, et al. ProteoSign: an end-user online differential proteomics statistical analysis platform. *Nucleic Acids Res*. 2017;45:W300-W306. doi:10.1093/nar/gkx444.
19. Tyanova S, Temu T, Sinitcyn P, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*. 2016;13:731-740.
20. Rainer J, Sanchez-Cabo F, Stocker G, et al. CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res*. 2006;34:W498-W503.
21. Colaert N, Helsens K, Impens F, Vandekerckhove J, Gevaert K. Rover: a tool to visualize and validate quantitative proteomics data from different sources. *Proteomics*. 2010;10:1226-1229.
22. van der Lee R, Buljan M, Lang B, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev*. 2014;114:6589-6631.
23. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*. 2009;37:D233-D238.
24. Martinez S, Hausinger RP. Catalytic mechanisms of Fe(II)- and 2-oxoglutarate-dependent oxygenases. *J Biol Chem*. 2015;290:20702-20711.
25. Islam MS, Leissing TM, Chowdhury R, et al. 2-oxoglutarate-dependent oxygenases. *Annu Rev Biochem*. 2018;87:585-620.
26. Kundu S. Distribution and prediction of catalytic domains in 2-oxoglutarate dependent dioxygenases. *BMC Res Notes*. 2012;5:410.
27. UniProt. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49:D480-D489.
28. Kundu S. Insights into the mechanism(s) of digestion of crystalline cellulose by plant class C GH9 endoglucanases. *J Mol Model*. 2019;25:240.
29. Kundu S. Fe(2)OG: an integrated HMM profile-based web server to predict and analyze putative non-haem iron(II)- and 2-oxoglutarate-dependent dioxygenase function in protein sequences. *BMC Res Notes*. 2021;14:80.
30. Chang YF, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*. 2007;76:51-74.