Mathematical Biosciences and Engineering

MBE, 17(3): 2470–2493.
DOI: 10.3934/mbe.2020135
Received: 17 November 2019
Accepted: 14 February 2020
Published: 25 February 2020

http://www.aimspress.com/journal/MBE

Research article

Mathematical model of a short translatable G-quadruplex and an assessment of its relevance to misfolding-induced proteostasis

Siddhartha Kundu*

Department of Biochemistry, All India Institute of Medical Sciences, Ansari Nagar, New Delhi 110029, INDIA

* **Correspondence: Email:** siddhartha_kundu@yahoo.co.in; **Tel:** +91-9818054915.

Abstract: G-quadruplexes can form in protein coding and non-coding segments such as the untranslated regions and introns of the mRNA transcript of several genes. This implies that amino acid forms of the G-quadruplex may have important consequences for protein homeostasis and the diseases caused by their alterations thereof. However, the absence of a suitable model and multitude of predicted physical forms has precluded a comprehensive enumeration and analysis of potential translatable G-quadruplexes. In this manuscript a mathematical model of a short translatable G-quadruplex ($TG4$) in the protein coding segment of the mRNA of a hypothetical gene is presented. Several novel indices (α, β) are formulated and utilized to categorize and select codons along with the amino acids that they code for. A generic algorithm is then iteratively deployed which computes the entire complement of peptide members that $TG4$ corresponds to, i.e., $\boldsymbol{PTG4}{\sim}TG4$. The presence, distribution and relevance of this peptidome to protein sequence is investigated by comparing it with disorder promoting short linear motifs. In frame termination codon, co-occurrence, homology and distribution of overlapping/shared amino acids suggests that $TG4$ ($\boldsymbol{\sim PTG4}$) may facilitate misfolding-induced proteostasis. The findings presented rigorously argue for the existence of a unique and potentially clinically relevant peptidome of a short translatable G-quadruplex that could be used as a diagnostic- or prognostic-screen of certain proteopathies.

Keywords: algorithm; codon; G-quadruplex; peptidome; misfolding induced proteopathy; translatable G-quadruplex

1. Introduction

G-quadruplex or G-tetrad ($G4$), is a thermodynamically stable structural element that is formed between clusters/stretches/tracts of Guanine (G) residues ($|G| \geq 3$) and is intra- or inter-molecular [1–3]. The intervening loops whence applicable are composed of one or more nucleotide(s) ($N \in \{A, U, T, G, C\}$) (Figure 1). $G4$ is found in DNA (telomeres, double-strand break

sites, transcription start sites) and in the untranslated region(s) (5'-, 3'-UTR, introns) of mRNA [4,5]. *In vivo,* $G4$ may function to preserve the telomeric ends of chromosomes, repress or promote transcription and regulate translation [4,5]. The generic representation of an intra-strand $G4$ may be described as follows:

$$\left(\left((G_{t,k})_{t\geq 3}(N_{h,k})_{h\geq 1}\right)_{k=3}((G_{t,k})_{t\geq 3})_{k=1}\right)_{m=1} \quad \text{(Def. 1)}$$

$$\begin{aligned} t &:= Number\ of\ Guanines\ per\ G-rich\ cluster \\ h &:= Number\ of\ loop-forming\ generic\ intervening\ nucleotides \\ k &:= Cluster\ index \\ m &:= Number\ of\ strands \\ G &:= Guanine \\ A &:= Adenine \\ T &:= Thymine \\ C &:= Cytosine \\ N &:= Any\ nucleotide \end{aligned}$$

The high melting temperature ($T_m{\sim}60^0C$) of $G4$ implies that the mature quadruplex is stable and refractory to unfolding. This is partly due to stabilizing Hoogsteen ($N7^{gu1}-N2^{gu2}; O6^{gu1}-N1^{gu2}$) and reverse Hoogsteen ($N7^{gu1}-N1^{gu2}; O6^{gu1}-N2^{gu2}$) hydrogen bonding as well as π-orbital stacking between the purine rings of non-contiguous guanine pairs ($gu1, gu2$) (Figure 1) [6,7]. Additionally, the presence of Adenine residues in the intervening loops, variable loop length ($h{\sim}1-30\ Mer$) and permutation have all been shown to contribute to the stability and thence persistence of the mature quadruplex [8–11].

$$T_m \propto (\#Adenine/h) = \tau.(\#Adenine/h) \quad (1)$$

$$\begin{aligned} T_m &:= Melting\ temperature \\ \tau &:= Constant\ of\ proportionality \\ h &:= Length\ of\ intervening\ loops \end{aligned}$$

Despite the wide range of methods available that can predict $G4$ formation in DNA/RNA, there is poor agreement between sequence-based motif locators and empirically derived biophysical data [12–15]. Motif-independent methods such as those that directly measure the GC-content or the GC-/AT-skew of a query sequence and utilize this data to train machine learning algorithms may address some of these discrepancies [16–18].

Investigations into transcribed RNA suggests that secondary and tertiary forms (5'- and 3'-UTRs) may not only coexist with stretches of unfolded ribonucleotides, but can also be read by the ribosomal machinery. Non-canonical translation is described as: a) translation from atypical start sites ($AUG \rightarrow \{CUG, GUG\}$) or b) peptides ($\leq 100\ aa$) of short open reading frames (sORF)-encoded polypeptides (SEPs) and upstream open reading frames (uORFs) [19–22]. The latter are rarely silent and can function as modulators of metabolism (S-Adenosylmethionine decarboxylase, *AMD1*) or transcription (activating transcription factor, *ATF4*, *H19*; yeast AP-1 like, *YAP1*) and as generic transcription factors (general control protein, *GCN4*) [19]. $G4$ has also been observed in one or more exons of the prion protein (*PRNP*, exon 2), zinc finger protein (*ZNF669*, exon 1), β-amyloid secretase (*BACE1*, exon 3) and the estrogen receptor 1 (*ESR1*, exon 4) among several others [16,23–29].

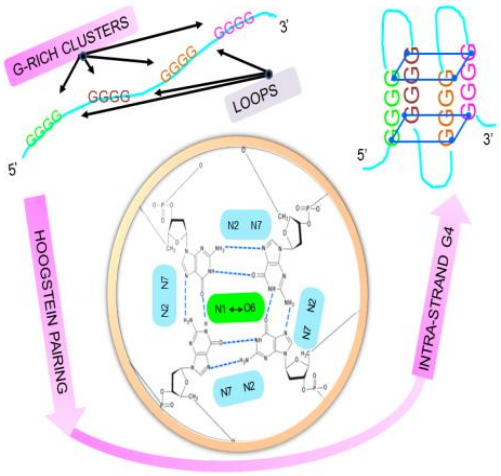

Figure 1. Definition, delineation and identity of a short translatable G-quadruplex. G-quadruplexes are stable structural elements in DNA/RNA and are characterized by Hoogsteen and reverse Hoogsteen pairing along with π-bond purine stacking of non-contiguous Guanine residues. Here, a short intra-strand G-quadruplex $(20 \leq N \leq 60; N \in \{A, U, G, C\})$ is modeled in the PCS of the mRNA of a hypothetical gene and is translatable $(TG4)$. The modeled $TG4$ is represented as a sequence of codons $\left(\left(COD_q\right)_{q\in\mathbb{N}}^{L}; COD \in \boldsymbol{COD}\right)$. Whilst, an arbitrary Guanine-rich cluster/stretch/tract is represented by suitably scored Guanine containing codon(s), the selection of codons for the intervening loops is without constraint. Abbreviations: $\boldsymbol{COD}$, set of vertebrate codons; L, total number of codons used to model the translatable G-quadruplex; N, Any generic ribonucleotide; PCS, protein coding segment; $TG4$, translatable G-quadruplex; q, numerical index of codon.

Whilst the presence of segments of folded mRNA may have a significant influence on the yield of the protein product(s), the effect on sequence whence part of the protein coding segment (PCS) is largely unknown [4,5,30–33]. Proteopathies are diseases that result directly from aggregates of truncated and misfolded proteins. These may occur secondary to a faulty translation machinery such as a ribosome that has stalled on encountering a secondary or tertiay folded mRNA sub segment. Recent data suggests ~45% of the human genome may code for proteins that are either intrinsically disordered $(IDPs)$ or comprise one or more sub-segments that are disordered $(IDRs)$ [34]. The absence of delineable structural features notwithstanding, disordered regions are characterized by short linear motifs (*SLiMS*) and/or molecular recognition features (MoRFs) [34,35]. The improper folding and heightened degradation rates could lead to perturbed proteostasis and thence contribute to the pathogenesis of proteopathies [34,35]. Primary proteopathies are likely to result directly from mutations (point, chromosomal translocations) in the PCS of a gene. These include sickle cell disease ($\beta^{E6\rightarrow V6}$-mediated defective polymerization), amylin-based type II Diabetes Mellitus, Cystic Fibrosis (cystic fibrosis transmembrane conductance regulator), Alzheimer's disease (Amyloid β-peptide) and Parkinson's disease ($\alpha-$synuclein) [36,37]. Secondary proteopathies, in contrast, result from motif or molecular mimicry of a host protein(s) by a pathogen. These are further classified into acute and chronic variants depending on the onset, genesis and/or resolution of the resultant infection or infestation [34,35].

$G4$ is known to stall the ribosome during translation and the resultant protein is truncated and/or degraded at an accelerated rate. The manuscript subsumes ribosomal read-through of mRNA with a G-quadruplex and assesses influence of the translated product to proteostasis. Here, I present a mathematical model of a short $G4\ (20-60\ Mer)$ in the PCS, *i.e.*, translatable G-quadruplex $(TG4)$, in the mRNA of a hypothetical gene. The mapping uses several novel indices to annotate, classify and select suitable Guanine-containing codons (α) and amino acids (β). A generic algorithm then computes and validates, as proof-of-principle, possible peptides $(pTG4_{ij})$ that correspond to the modeled $TG4$ $(pTG4_{ij} \in \boldsymbol{PTG4} \sim TG4)$. Co-occurrence, homology and the distribution of overlapping/shared amino acids between $\boldsymbol{PTG4}$ and the disorder promoting $\boldsymbol{SLiMS}$ are used to infer probable mechanisms of $TG4 \sim \boldsymbol{PTG4}$ facilitated misfolding. Standard bioinformatics indices (accuracy, precision, recall, $p-value$) are used to arrive at these conclusions.

2. Materials and methods

2.1. Mathematical expression for the canonical peptidome of a short translatable G-quadruplex ($\boldsymbol{PTG4}$)

The objective of this investigation is to model a short $G4$ in an arbitrary PCS ($TG4$) which when translated will result in a set of peptides ($\boldsymbol{PTG4}$) with an average length that is less than 100 amino acids. The hypothesis explored in this manuscript is that in the event of a ribosomal read-through, the translated mRNA, with its G4 will result in a modified protein product. This protein will then exhibit considerable propensity to misfold on account of the presence of one or more members of the $\boldsymbol{PTG4}$.

2.1.1. Model of a translatable G-quadruplex ($TG4$)

SEPs-derived peptides with the lowest molecular weight $(\sim 2.5\ KDa)$ and with lengths varying from $\sim 7-20\ aa$ were identified and used to define the boundaries of the peptides that comprise $\boldsymbol{PTG4}$ [20,21]. The $TG4\ (m=1)$ is therefore, modeled as an intra-strand sub sequence of the mRNA of a hypothetical gene and has a length of $\sim 20-60\ Mer$. This is represented (with symbols and variable names as explained in $Def.1$) as follows:

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

2.1.2. Codon association as a suitable representation of the $TG4$

Since the Guanine-rich clusters and loops are contiguous, the aforementioned model $(Def.2)$ of the $TG4$ may be approximated with a sequence of codons and is as under:

$$TG4 \coloneqq \left(COD_q \right)_{q \in \mathbb{N}}^{L} | COD \in \boldsymbol{COD} \quad \text{(Def. 3)}$$

The algorithm to compute L, which is the number of codons needed to model $TG4$ is presented and is as follows:

$$
\begin{aligned}
&1: && N \leftarrow \{u \in [20,60)\}\\
&2: && r \leftarrow N \bmod 3\\
&3: && e \leftarrow N - ((N \bmod 3)/3)\\
&4: && If\ e < (\lfloor e \rfloor + \lceil e \rceil)/2\ then\\
&5: && L = \lceil e/3 \rceil\\
&6: && else\ If\ e \geq (\lfloor e \rfloor + \lceil e \rceil)/2\ then\\
&7: && L = \lfloor e/3 \rfloor\\
&8: && end\ If
\end{aligned}
$$

$$
\begin{aligned}
N &:= Number\ of\ ribonucleotides\ required\ to\ model\ TG4\\
L &:= Total\ number\ of\ codons\ \ needed\ to\ model\ \ TG4\ (7 \leq L < 21)\\
q &:= \ qth\ codon\\
r &:= Remainder = \{0,1,2\}\\
e, u &:= Generic\ variables\\
\boldsymbol{COD} &:= Set\ of\ vertebrate\ codons
\end{aligned}
$$

2.1.3. Codon classification and amino acid composition of $\boldsymbol{PTG4}$

The codons selected for modelling $TG4\ (Def.3)$ comprised suitably scored Guanine-containing vertebrate codons $(\boldsymbol{gCOD_n^{\pm}} \subset \boldsymbol{COD})$ for the Guanine-rich clusters/stretches/tracks $(3 \leq t \leq 9; Defs.1\ and\ 2)$ and generic/no-stop codons for the intervening loops (Figures 2 and 3). Briefly, a Guanine-containing codon $(gCOD_n)$ is scored by considering its association with two similar flanking codons, *i.e.*, $gCOD_{n-1}, gCOD_n, gCOD_{n+1}$ such that there is at least one occurrence of $'GGGG'\ (\delta \geq 1.0)$ (Figures 2 and 3). This non-trivial case $(4 \leq t \leq 9)$ is chosen since its trivial equivalent $(t = 3)$, is already subsumed $(Defs.1\ and\ 2)$. Numerically,

$$\alpha_{codon}^{amino} = \gamma.\theta.\delta + \Omega \tag{2}$$

$$
\begin{aligned}
\gamma &:= Probability\ of\ \ codon\ \ occurrence\ \ (\gamma = 1/64 \approx 0.02)\\
\theta &:= Probability\ of\ \ codon\ \ occurrence\ \ within\ \ a\ \ group\ (\theta = \{0.04,0.11,0.33,1\})\\
\delta &:= Distinct\ \ occurrences\ of\ 'GGGG'(\delta = \{0,1,2,6\})\\
\Omega &:= Number\ of\ adjacent\ codons\ with\ \ \delta\ (\Omega = \{0,1,2\})
\end{aligned}
$$

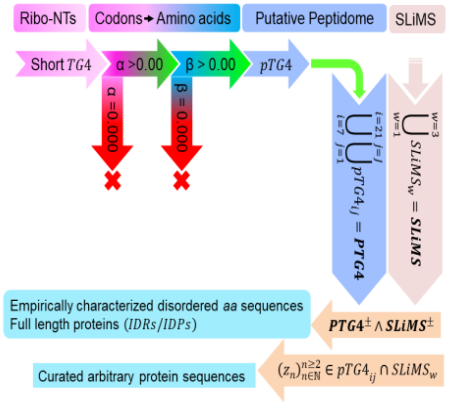

Figure 2. Algorithm to delineate and assess relevance of the peptidome of the modeled translatable G-quadruplex. Sub sections 2.1.1–4 are devoted to constructing the model of

a short translatable G-quadruplex in the PCS of the mRNA of a hypothetical gene. Briefly, Guanine-containing vertebrate codons are scored and selected using codon association as the underlying model. An amino acid is then scored on the basis of the proportion of the $G4$ favoring codons it possesses. This schema is deployed iteratively and results in the complete set of peptides for the modeled translatable G-quadruplex. Sub section 2.1.5 and 2.1.6 are used to assess the relevance of the predicted peptidome to the genesis of misfolding induced proteostasis. This is done by examining the co-occurrence, homology and distribution of overlapping/shared amino acids of members of this peptidome with one or more short linear motifs. The sequences utilized are length adjusted empirically determined disordered regions, full length protein sequences with disordered segments and taxonomically diverse generic protein sequences. Abbreviations: $G4$, G-quadruplex; mRNA, messenger ribonucleic acid; PCS, protein coding segment; $\boldsymbol{PTG4}$, hypothetical peptidome of the modeled short translatable G-quadruplex; $\boldsymbol{SLiMS}$, short linear motifs.

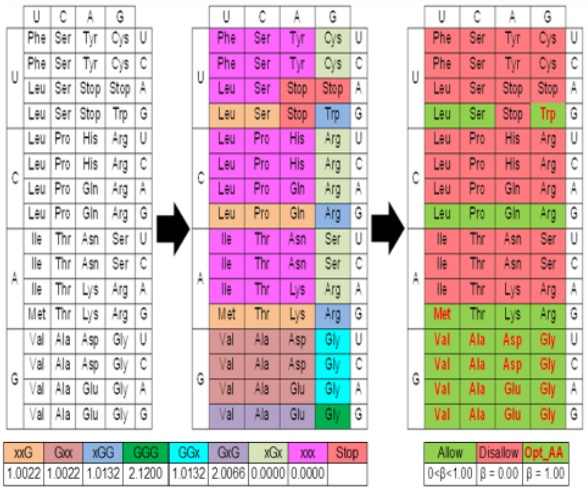

Figure 3. Schema to characterize a short translatable G-quadruplex ($TG4$) and its associated peptidome ($\boldsymbol{PTG4}$). The $TG4$ is modeled as a short sequence $(L = 20)$ of Guanine-containing codons $\left(\left(COD_q\right)_{q=7}^{q=20}; COD \in \boldsymbol{COD}\right)$. Guanine-containing vertebrate codons $(\boldsymbol{COD})$ are scored $(\alpha_{amino}^{codon} > 0.00)$ and selected $(\boldsymbol{gCOD}_{\boldsymbol{amino}}^{+} \subset \boldsymbol{COD})$ using codon association as the underlying model. This partitions vertebrate codons into permissive $(\boldsymbol{xGG}, \boldsymbol{Gxx}, \boldsymbol{xGG}, \boldsymbol{GGG}, \boldsymbol{GGx}, \boldsymbol{GxG}; \alpha_{amino}^{codon} > 0.0000; n = 28)$ and non-permissive $(\boldsymbol{xGx}, \boldsymbol{xxx}; \alpha_{amino}^{codon} = 0.0000; n = 36)$ codons. Interestingly, the amber stop codon $(UAG; \alpha = 1.0022)$ is also selected by these criteria. An amino acid is then scored on the basis of the proportion of the $G4$ favouring codons amongst all the codons that code it $(\beta_{amino} = |\boldsymbol{gCOD}_{\boldsymbol{amino}}^{+}|/|\boldsymbol{COD}_{\boldsymbol{amino}}|)$. An interesting subset of these amino acids $(\beta = 1.00)$ are those which are encoded entirely by codons which may favour $G4$ formation. These include Valine, Alanine, Aspartic- and Glutamic-acids, Methionine and Glycine. Abbreviations: $\boldsymbol{COD}$, set of vertebrate codons; L, Length of codon model of $TG4$; q, numerical indices; x, generic ribonucleotide.

Since the genetic code is degenerate, amino acids mapped from the selected codons are further scored and grouped $(\boldsymbol{g1}, \boldsymbol{g2}, \boldsymbol{g3})$ (Figures 3 and 4).

$$\beta_{amino} = {|\boldsymbol{gCOD}^{+}_{\boldsymbol{amino}}|}\Big/{|\boldsymbol{COD}_{\boldsymbol{amino}}|} \tag{3}$$

$$\boldsymbol{gCOD}^{+}_{\boldsymbol{amino}} := Set\ of\ optimal\ codons\ for\ each\ amino\ acid\ (\alpha^{amino}_{codon} > 0.0000)$$
$$\boldsymbol{COD}_{\boldsymbol{amino}} := Set\ of\ codons\ for\ each\ amino\ acid$$

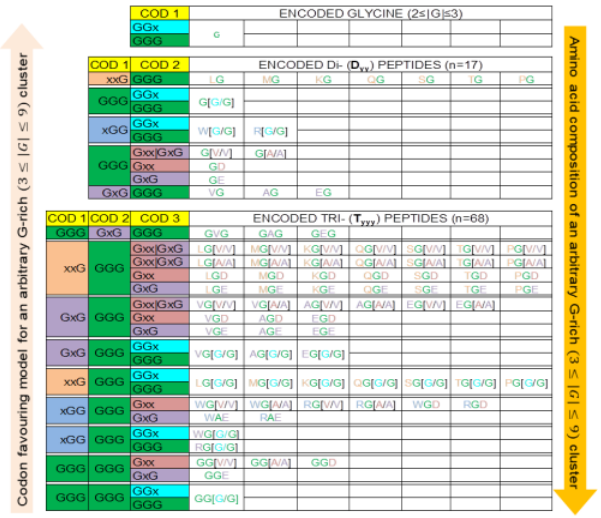

Figure 4. Dissecting the peptidome of the modeled short translatable G-quadruplex. The schema (α, β) outlined in this work is used to construct a peptidome of the modeled short translatable G-quadruplex $(\boldsymbol{PTG4})$. This peptidome is the finite union of several peptides $(pTG4_{ij} \in \boldsymbol{PTG4} | 7 \leq i \leq 20, 1 \leq j \leq J)$. Each such peptide has a subset of amino acids that corresponds to an arbitrary Guanine-rich cluster/stretch/tract and another for the intervening loops. Whilst, the former has a restricted composition of amino acids $(\beta > 0; y \in \boldsymbol{Y})$, the latter is generic $(\beta \geq 0; z \in \boldsymbol{Z})$. The manner in which this predicted peptidome may influence the genesis of primary and secondary misfolding induced proteostasis is next investigated. This is done by examining the co-occurrence of $pTG4_{ij}$ with one or more $SLiMS_w$ $(1 \leq w \leq 3)$ in empirically determined disordered regions and full length protein sequences with disordered segments. These studies are complemented by investigating the distribution of sequences with shared amino acids $\left((z_n)_{n\geq 2} | z \in pTG4_{ij} \cap SLiMS_w, pTG4_{ij} \in \boldsymbol{PTG4}; SLiMS_w \in \boldsymbol{SLiMS}\right)$ across taxa. Interestingly, most of the amino acids that comprise $\boldsymbol{PTG4}$ include those that favour hyperphosphorylation (Serine, Threonine) and non-covalent complex formation (Alanine, Valine, Leucine, Isoleucine, Lysine, Arginine, Aspartic- and Glutamic-acids) and proteolytic cleavage. Abbreviations: $\boldsymbol{g1}, \boldsymbol{g2}, \boldsymbol{g2}$, classification of amino acids used in this study, $\boldsymbol{PTG4}$, hypothetical peptidome of the modeled short translatable G-quadruplex; $\boldsymbol{SLiMS}$; short linear motifs; $\mathbf{Y}$, $\mathbf{Z}$; sets of amino acids.

2.1.4. Characterize the peptidome corresponding to $TG4$ $(\boldsymbol{PTG4})$

Whilst amino acids from groups 1 and 2 $(\beta > 0.00)$ (3) can represent the modeled G-rich clusters $(y \in \boldsymbol{g1} \cup \boldsymbol{g2} = \boldsymbol{Y})$, no constraint was imposed on the amino acids (z) used to model the loops $(z \in \boldsymbol{g1} \cup \boldsymbol{g2} \cup \boldsymbol{g3} = \boldsymbol{Z})$ (Figures 3 and 4). The peptidome $(\boldsymbol{PTG4})$ evaluated by this

study is a combinatorial association of peptides such that the molecular weight is $\sim 0.8-2.3\ KDa$ and length of any arbitrary member is $\sim 7-20\ aa$ (Figure 4). This may be represented as follows:

$$pTG4_{ij} = \left(\left(\left(\left(y_{i,k} \right)_{1 \le i \le 3} \left(z_{i,k} \right)_{1 \le i \le 2} \right)_{k=3} \left(y_i \right)_{1 \le i \le 3} \right) \left(z_i \right)_{1 \le i \le 2} \right)_j \quad \text{(Def. 4)}$$

$$\boldsymbol{PTG4} = \bigcup_{i=7}^{i=20} \bigcup_{j=1}^{j=J} |pTG4_{ij}| \quad \text{(Def. 5)}$$

$$\begin{aligned} \boldsymbol{PTG4} &:= Peptidome\ corresponding\ to\ TG4 \\ pTG4_{ij} &:= j^{th}\ canonical\ amino\ acid\ form\ of\ PTG4\ with\ \text{"i"}\ amino\ acids \\ i &:= Number\ of\ amino\ acids\ that\ comprise\ the\ modelled\ \boldsymbol{PTG4} \\ J &:= Maximum\ number\ of\ canonical\ pTG4\ for\ \text{"}i\text{"}\ amino\ acids \end{aligned}$$

2.1.5. Establish proof-of-principle of biological relevance of $\boldsymbol{PTG4}$

A dataset that comprises experimentally validated $G4$-forming mRNA segments of several genes $(n = 99)$ was downloaded (http://scottgroup.med.usherbrooke.ca/G4RNA/) and used to investigate the distribution of $G4$ [16]. Genes which possess non-redundant RNA ($\boldsymbol{R}$) sub sequences in the PCS are translated in 6 reading frames using an online tool (http://web.expasy.org/translate). The peptides generated are classified as those: i) with one or more uninterrupted stretch of N-terminal amino acids of length $\geq 7\ aa\ (\sim \boldsymbol{A})$, ii) with an in-frame termination signal designated as 'STOP' $(\sim \boldsymbol{B})$ and iii) without any termination signal, i.e., absence of a 'STOP' in their sequence $(\sim \boldsymbol{C})$. The translated peptides are classified as "VALID" $\left((\boldsymbol{B} \cap \boldsymbol{A}) \cup (\boldsymbol{C} \cap \boldsymbol{A}) \right)$ and then queried for matches with $pTG4_{ij}$ $(7 \leq i \leq 20, j \in \mathbb{N})$. The PERL scripts that are required to parse and process the resulting data files have been developed in house and the pseudocode for the same is presented as additional information (Pseudocode, PS1: Supplementary Text 1).

2.1.6. *In silico* assessment of $\boldsymbol{PTG4}$ to misfolding induced proteostasis

This is done by examining the occurrence of $\boldsymbol{PTG4}$ in amino acid/protein sequences of disordered regions $(IDRs)$ and full-length proteins with disordered regions $(IDPs)$. DisProt 7.0 (http://disprot.org), is a database of experimentally validated and non-redundant sequences of $IDRs$ and $IDPs$ [38]. The sequences $(|\boldsymbol{IDR}| = 1445; |\boldsymbol{IDP}| = 800)$ that comprise these are queried for occurrences of $pTG4_{ij}$ $(7 \leq i \leq 20, j \in \mathbb{N})$ (Supplementary Texts 2 and 3). A preliminary partitioning schema divides these datasets into two distinct subsets, *i.e.*, $\#pTG4_{ij} \geq 1$ $(PT^{+} \equiv \boldsymbol{PPOS} \subset \{\boldsymbol{IDR}, \boldsymbol{IDP}\}; (Def.6))$ and $\#pTG4_{ij} = 0$ $(PT^{-} \equiv \boldsymbol{PNEG} \subset \{\boldsymbol{IDR}, \boldsymbol{IDP}\}; (Def.7))$. The extent of co-occurrence of one or more $SLiMS_w \equiv SL$ $(w = \{1,2,3\})$ with $pTG4_{ij}$ $(SL^{\pm} \in \{\boldsymbol{PPOS}, \boldsymbol{PNEG}\})$ $(Defs.8\ and\ 9)$ is then evaluated to infer relevance of $\boldsymbol{PTG4}$ to misfolding induced proteostasis. The distribution of overlapping/shared sequences of amino acids $\left((z_n)_{n \geq 2} \in \left(pTG4_{ij} \cap SLiMS_w \right); z_n \in \boldsymbol{Z}; \right)$ $(Def.10)$, is examined in protein sequences from taxonomically diverse organisms with ScanProsite (https://prosite.expasy.org/scanprosite). The proof behind this rationale is presented:

$$
\begin{aligned}
(z_n)_{n\geq 2} &\in \left(pTG4_{ij} \cap SLiMS_w\right) \\
&= \left((z_n)_{n\geq 2} \in pTG4_{ij}\right) \cap \left((z_n)_{n\geq 2} \in SLiMS_w\right)
\end{aligned}
$$

Let $z_n = z_n^{'}$ *and* $z_n = z_n^{"}$.
Rewriting

$$
\begin{aligned}
&= \left((z_n^{'})_{n\geq 2} \in pTG4_{ij}\right) \cap \left(\left(z_n^{"}\right)_{n\geq 2} \in SLiMS_w\right) \\
&= \left((z_n^{'})_{n\geq 2}, \left(z_n^{"}\right)_{n\geq 2}\right) \\
&= pTG4_{ij} \times SLiMS_w
\end{aligned}
$$

■

$$
\begin{aligned}
\boldsymbol{Z} &:= Set\ of\ amino\ acids\ (z_n \in \boldsymbol{Z}) \\
pTG4_{ij} &:= Canonical\ amino\ acid\ form\ of\ \boldsymbol{PTG4} \\
\boldsymbol{SLiMS} &:= Set\ of\ short\ linear\ motifs\ (SLiMS_w \in \boldsymbol{SLiMS}) \\
i,j,n,w &:= Indices\ of\ members\ of\ \boldsymbol{Z}, \boldsymbol{PTG4}, \boldsymbol{SLiMS}
\end{aligned}
$$

2.2. Statistical measures to compute and assess biological relevance of $TG4$

The indices utilized by this study to establish relevance of matched instances of various motifs/co-motifs in the peptide/protein sequences of interest include the accuracy (A), precision (P), recall (R) and the $p-value$. A $2X2$ table which represents the categorized data (2.1.4) is constructed and used to compute various bioinformatics indices. This is outlined as under:

	PT^-	PT^+
SL^-	TN	FP
SL^+	FN	TP

$$
\begin{aligned}
TN &:= True\ negative\ (|\boldsymbol{sNEG} \cap \boldsymbol{PNEG}| = |\boldsymbol{sNEG}|) \equiv SL^-PT^- \quad (Def.11) \\
FP &:= False\ positive\ (|\boldsymbol{SNEG} \cap \boldsymbol{PPOS}| = |\boldsymbol{SNEG}|) \equiv SL^-PT^+ \quad (Def.12) \\
FN &:= False\ negative\ (|\boldsymbol{sPOS} \cap \boldsymbol{PNEG}| = |\boldsymbol{sPOS}|) \equiv SL^+PT^- \quad (Def.13) \\
TP &:= True\ positive\ (|\boldsymbol{SPOS} \cap \boldsymbol{PPOS}| = |\boldsymbol{SPOS}|) \equiv SL^+PT^+ \quad (Def.14)
\end{aligned}
$$

The equations may then be written as:

$$(A) = \left({}^{TN+TP}/_{TN+FP+FN+TP}\right)X100 \tag{4}$$

$$(P) = \left({}^{TP}/_{FP+TP}\right)X100 \tag{5}$$

$$(R) = \left({}^{TP}/_{FN+TP}\right)X100 \tag{6}$$

The $p-values$ for these analyses are computed by comparing the frequency of occurrence of all $pTG4_{ij}$ in a test sequence $(\phi_{pTG4_{ij}})$ with the same in randomly-generated $(v \in \boldsymbol{V})$ sequences of similar lengths $(\phi_{pTG4_{vij}})$, *i.e.*, $7-50\ aa\ (1 \leq v \leq 10000)$ and $>50\ aa\ (1 \leq v \leq 100000)$ (Pseudocode, PS2: Supplementary Text 1):

$$
\begin{aligned}
p-value &= \phi_{pTG4_{vij}}/\phi_{pTG4_{ij}} \\
&= \left(\sum_{v=1}^{v=|V|}\sum_{i=7}^{i=21}\sum_{j=1}^{j=J} pTG4_{vij}\right)\Big/\left(\sum_{i=7}^{i=21}\sum_{j=1}^{j=J} pTG4_{ij}\right) \\
&= \left(\sum_{v=1}^{v=|V|}\sum_{i=7}^{i=21}\sum_{j=1}^{j=J} pTG4_{vij}\Big/\sum_{i=7}^{i=21}\sum_{j=1}^{j=J} pTG4_{ij}\right)
\end{aligned} \tag{7}
$$

The frequency of occurrence of overlapping sequences of amino acids $\left((z_n)_{n\geq 2} \in \left(pTG4_{ij} \cap SLiMS_w\right); z_n \in \boldsymbol{Z}\right)$ in pre-compiled and curated protein sequences $(\phi_{(z_n)})$ across taxa is compared with randomly chosen sequences of comparable lengths $(\phi_{(vz_n)}; n = 5000)$. These are used to estimate statistical significance, i.e., $p-value = \phi_{(vz_n)}/\phi_{(z_n)}$ (8).

3. Results

The data presented discusses implementation of a model of short intra-strand $TG4$ for various values of α and β, populates $\boldsymbol{PTG4}$ and establishes the equivalence $TG4{\sim}\boldsymbol{PTG4}$. Co-occurrence and homology studies between $\boldsymbol{PTG4}$ and the $\boldsymbol{SLiMS}$ in $IDRs/IDPs$ and generic protein sequences across taxa are used to infer probable mechanisms of $TG4{\sim}\boldsymbol{PTG4}$ facilitated misfolding-induced proteostasis.

3.1. Suitability of Guanine-containing codons as a model for an arbitrary G-rich cluster of $TG4$

An association-competent codon not only takes into account the presence of a Guanine residue, but also gives weightage to its position (Figures 1–3, Table 1). This schema partitions standard vertebrate codons into those with a high- $(Ranks\ 1-4; \alpha > 0.0000)$ or low- $(Rank\ 5; \alpha = 0.0000)$ propensity to form a contiguous cluster of Guanine residues (Figures 3 and 4, Table 1). Whilst, $'\boldsymbol{GGG}'$ $(Rank\ 1; \alpha = 2.12)$ can associate with $(\{\boldsymbol{GGG}, \boldsymbol{GxG}, \boldsymbol{xGG}, \boldsymbol{GGx}, \boldsymbol{Gxx}, \boldsymbol{xxG}\})$ bilaterally $(\delta = 6; \Omega = 2)$, $'\boldsymbol{GxG}'$ $(Rank\ 2;\ \alpha = 2.0066)$ can do so only with $'\boldsymbol{GGG}'$ $(\delta = 1; \Omega = 2)$. On the other hand, the codon subsets $'\boldsymbol{GGx}'$ and $'\boldsymbol{xGG}'$ $(Rank\ 3; \alpha = 1.0132)$ can form two clusters of contiguous Guanine residues with $'\boldsymbol{GGG}'$ and $'\boldsymbol{xGG}'/'\boldsymbol{GGx}'$ unilaterally $(\delta = 2; \Omega = 1)$. Similarly, the subsets $'\boldsymbol{xxG}'$ or $'\boldsymbol{Gxx}'$ $(Rank\ 4; \alpha = 1.0022)$, can form contiguous Guanines with a single occurrence of $'\boldsymbol{GGG}'$ $(\delta = 1; \Omega = 1)$ (Figures 3 and 4, Table 1). Conversely, codons with either a single occurrence of a central Guanine residue $'\boldsymbol{xGx}'$ or no Guanine residues $'\boldsymbol{xxx}'$ $(Rank\ 5; \alpha = 0.0000)$ are unable to form the $'GGGG'$ and are excluded from this study (Figures 3 and 4, Table 1).

3.2. Validation studies of $\boldsymbol{PTG4}$ in known $G4$ -forming exons to establish equivalence ($TG4{\sim}\boldsymbol{PTG4}$)

An estimate of the possible combinations of the simplest peptide $(\sum_{i=7}\sum_{j=1}^{j=J} pTG4_{ij} = 8.00E+03; GlyzGlyzGlyzGly; J = (20)^3; i = length\left(pTG4_{ij}\right) = 7\ aa; z \in \boldsymbol{Z})$ (Figures 3 and 4, Table 2). This justifies usage of $\boldsymbol{PTG4}$ $(pTG4_{ij} \in \boldsymbol{PTG4})$ as a generic representation of the putative peptidome encoded by the $TG4$ $(\boldsymbol{PTG4})$. Approximately $\sim 12\%$ $(n = 11)$ of *in silico* translated amino acid sequences from exon-derived $TG4$ possesses one of more "STOP" signals and include

ESR1, longer RNA variants of *PRNP* $(85\,nt)$ and *BCL2* $(29\,n\,t, 33\,nt, 34\,nt)$ (Table 3; Supplementary Table 1, Supplementary Text 2). With the exceptions of *KCNH2*/*ZNF669* and the shorter variants of *PRNP* $(14\,nt, 15\,nt, 20\,nt, 24\,nt)$, "VALID" sub sequences are found for *BACE1*, *BCL2*, *ESR1*, *PRNP* (long) and *TERF2* (Table 3; Supplementary Tables 1A and 1C). Interestingly, all the genes considered possessed at least one occurrence of $\boldsymbol{PTG4}$ $(P = 100\%, n = 6)$ (Table 3; Supplementary Table 1B). This finding, despite the small sample size is proof-of-principal that the $TG4$ can be mapped to definite peptide sequences, *i.e.*, $TG4{\sim}PTG4$. Since this can occur only after a ribosomal read through of the $G4$ containing mRNA, it raises the intriguing possibility that $PTG4$ whence part of a larger protein may increase its propensity to undergo misfolding. This notion is investigated in non-redundant sequences of $IDRs$ $(\boldsymbol{PTG4}{\sim}10\%, n = 145; 0.00 \leq p - value \leq 0.20)$ and $IDPs$ $(\boldsymbol{PTG4}{\sim}34\%, n = 269; 0.00 \leq p - value < 0.5)$ (Table 4; Supplementary Tables 2 and 3).

Table 1. Rank wise arrangement of codon scores for the non-trivial $(4 \leq |\boldsymbol{G}| \leq 9)$ $TG4$.

Rank	Codon set, Cardinality	Codon	$\boldsymbol{\gamma}$	$\boldsymbol{\theta}$	$\boldsymbol{\delta}$	$\boldsymbol{\Omega}$	$\boldsymbol{\alpha = \gamma.\theta.\delta{+}\Omega}$	aa
1	**GGG, 1**	GGG	0.02	1.00	6	2	2.1200	Gly
2	**GxG, 3**	GUG	0.02	0.33	1	2	2.0066	Val
		GCG	0.02	0.33	1	2	2.0066	Ala
		GAG	0.02	0.33	1	2	2.0066	Glu
3	**xGG, 3**	UGG	0.02	0.33	2	1	1.0132	Trp
		CGG	0.02	0.33	2	1	1.0132	Arg
		AGG	0.02	0.33	2	1	1.0132	Arg
3	**GGx, 3**	GGU	0.02	0.33	2	1	1.0132	Gly
		GGC	0.02	0.33	2	1	1.0132	Gly
		GGA	0.02	0.33	2	1	1.0132	Gly
4	**xxG, 9**	UUG	0.02	0.11	1	1	1.0022	Leu
		UCG	0.02	0.11	1	1	1.0022	Ser
		UAG	0.02	0.11	1	1	1.0022	*Ter*
		CUG	0.02	0.11	1	1	1.0022	Leu
		CCG	0.02	0.11	1	1	1.0022	Pro
		CAG	0.02	0.11	1	1	1.0022	Gln
		AUG	0.02	0.11	1	1	1.0022	Met
		ACG	0.02	0.11	1	1	1.0022	Thr
		AAG	0.02	0.11	1	1	1.0022	Lys
4	**Gxx, 9**	GUU	0.02	0.11	1	1	1.0022	Val
		GCU	0.02	0.11	1	1	1.0022	Ala
		GAU	0.02	0.11	1	1	1.0022	Asp
		GUC	0.02	0.11	1	1	1.0022	Val
		GCC	0.02	0.11	1	1	1.0022	Ala
		GAC	0.02	0.11	1	1	1.0022	Asp
		GUA	0.02	0.11	1	1	1.0022	Val
		GCA	0.02	0.11	1	1	1.0022	Ala
		GAA	0.02	0.11	1	1	1.0022	Glu
5	**xGx, 9**	UGU	0.02	0.11	0	0	0.0000	Cys
		UGC	0.02	0.11	0	0	0.0000	Cys

Continued on next page

Rank	Codon set, Cardinality	Codon	γ	θ	δ	Ω	$\alpha = \gamma.\theta.\delta+\Omega$	aa
		UGA	0.02	0.11	0	0	0.0000	*Ter*
		CGU	0.02	0.11	0	0	0.0000	Arg
		CGC	0.02	0.11	0	0	0.0000	Arg
		CGA	0.02	0.11	0	0	0.0000	Arg
		AGU	0.02	0.11	0	0	0.0000	Ser
		AGC	0.02	0.11	0	0	0.0000	Ser
		AGA	0.02	0.11	0	0	0.0000	Arg
5	**xxx, 27**	UUU	0.02	0.04	0	0	0.0000	Phe
		UCU	0.02	0.04	0	0	0.0000	Ser
		UAU	0.02	0.04	0	0	0.0000	Tyr
		UUC	0.02	0.04	0	0	0.0000	Phe
		UCC	0.02	0.04	0	0	0.0000	Ser
		UUA	0.02	0.04	0	0	0.0000	Leu
		UCA	0.02	0.04	0	0	0.0000	Ser
		UAA	0.02	0.04	0	0	0.0000	*Ter*
		CUU	0.02	0.04	0	0	0.0000	Leu
		CCU	0.02	0.04	0	0	0.0000	Pro
		CAU	0.02	0.04	0	0	0.0000	His
		CUC	0.02	0.04	0	0	0.0000	Leu
		CCC	0.02	0.04	0	0	0.0000	Pro
		CAC	0.02	0.04	0	0	0.0000	His
		CUA	0.02	0.04	0	0	0.0000	Leu
		CCA	0.02	0.04	0	0	0.0000	Pro
		CAA	0.02	0.04	0	0	0.0000	Gln
		AUU	0.02	0.04	0	0	0.0000	Ile
		ACU	0.02	0.04	0	0	0.0000	Thr
		AAU	0.02	0.04	0	0	0.0000	Asn
		AUC	0.02	0.04	0	0	0.0000	Ile
		ACC	0.02	0.04	0	0	0.0000	Thr
		AAC	0.02	0.04	0	0	0.0000	Asn
		AUA	0.02	0.04	0	0	0.0000	Ile
		ACA	0.02	0.04	0	0	0.0000	Thr
		AAA	0.02	0.04	0	0	0.0000	Lys

Abbreviations

γ: General probability of a codon ($\gamma = 1/64 \cong 0.02$)
θ: Probability of codon within a group ($\theta = \{0.04,0.33,0.11,1.00\}$)
δ: Number of distinct codon sets that could complete 'GGGG' ($\delta = \{0,1,2,6\}$)
Ω: Number of adjacent positions that contain δ ($\Omega = \{0,1,2\}$)
α: Threshold for selecting codons that may favour G-quadruplex formation
x: Codon specific generic ribonucleotide $\{A, G, U, C\}$
aa: Amino acid
Ter: Stop codons $\{UAG, UGA, UAA\}$

Table 2. Codon-based classification of amino acids.

		aa	COD_{amino}	$gCOD^{+}_{amino}$	β
Group 1	$(n = 7)$	Ala	4	4	1.00
		Val	4	4	1.00
		Asp	2	2	1.00
		Glu	2	2	1.00
		Trp	1	1	1.00
		Met	1	1	1.00
		Gly	4	4	1.00
Group 2	$(n = 7)$	Leu	6	2	0.3333
		Gln	2	1	0.5
		Arg	6	2	0.3333
		Lys	2	1	0.5
		Ser	6	1	0.1667
		Thr	4	1	0.25
		Pro	4	1	0.25
Group 3	$(n = 6)$	Cys	2	0	0.00
		Asn	2	0	0.00
		Ile	3	0	0.00
		His	2	0	0.00
		Phe	2	0	0.00
		Tyr	2	0	0.00

Abbreviations

$gCOD^{+}_{amino}$: Guanine-containing optimal codons excluding STOP (UAG) $(\alpha > 0.000)$

COD^{-}_{amino}: Non-optimal codon excluding STOP (UGA, UAA) $(\alpha = 0.000)$

$COD_{amino} = gCOD^{+}_{amino} + COD^{-}_{amino}$: All codons for an amino acid

Table 3. Genes (*Homo sapiens*) with G-quadruplex forming mRNA segments derived from one or more exons (Ex).

GENE	NAME	G4 (nt)	Ex	STOP ($\boldsymbol{n = 11}$)	VALID ($\boldsymbol{n = 59}$)	$\boldsymbol{\lvert PTG4 \rvert}$
BACE1	Beta-secretase 1	33	3	$n = 0$	$n = 6$	$n = 2$
BCL2	B-cell lymphoma 2	33	2	$n = 1$	$n = 6$	$n = 1$
		23		$n = 0$	$n = 6$	
		28		$n = 0$	$n = 6$	
		29		$n = 1$	$n = 5$	
		34		$n = 1$	$n = 5$	
		33	3	$n = 2$	$n = 5$	
ESR1	Estrogen receptor alpha (ERα)	36	4	$n = 1$	$n = 5$	$n = 2$
KCNH2	Potassium Voltage-Gated Channel sub family H	18	12	$n = 0$	$n = 0$	NA
ZNF669	Member 2 Zinc Finger Protein 669		1			
PRNP	Prion protein	14	2	$n = 0$	$n = 0$	$n = 1$
		15		$n = 0$	$n = 0$	
		20		$n = 0$	$n = 0$	
		24		$n = 0$	$n = 6$	
		85		$n = 6$	$n = 3$	
TERF2	Telomeric repeat-binding factor 2	55	1	$n = 0$	$n = 6$	$n = 1$

3.3. The peptidome of the translatable G-quadruplex may trigger misfolding of the encompassing protein

The amino acids that comprise the peptide members of $\boldsymbol{PTG4}$ and the short linear motifs ($\boldsymbol{g1, g2}$ *vs* $\boldsymbol{SLiMS}$) are well conserved. The co-occurrence of $\boldsymbol{PTG4}$ with $\boldsymbol{SLiMS}$ in the $IDRs$ ($A{\sim}85-89\%; 0.00 < p-value \leq 0.05$) suggests that this association is non-trivial and may favor all purported mechanisms of misfolding (hyperphosphorylation, proteolytic cleavage, complex formation) (Table 4; Supplementary Tables 2 and 3). However, the higher precision of $\boldsymbol{PTG4}$ with the proteolytic-$SLiMS$ suggests that this may predominate (Table 4; Supplementary Tables 2 and 3). The data with the $IDPs$ suggests a similar predilection for proteolytic cleavage ($A{\sim}40-77\%; P{\sim}99\%; 0.00 < p-value < 0.05$, although hyperphosphorylation ($P{\sim}60\%; 0.00 < p-value < 0.05$) and complex-promotion ($P{\sim}30\%; 0.00 < p-value < 0.05$) may constitute viable alternatives to the genesis of misfolding (Table 4; Supplementary Tables 2 and 3). The presence of overlapping sequences of amino acids between $\boldsymbol{PTG4}$ and the $\boldsymbol{SLiMS}$ when examined in protein sequences from taxonomically diverse organisms is degenerate for $SLiMS_1$ ($number\ of\ matches = 6251$) and $SLiMS_3$ ($number\ of\ matches = 1480$) (Table 5; Supplementary Table 4). In contrast, the corresponding data for $SLiMS_2$ ($number\ of matches = 3759; 0.00 < p-value < 0.05$) is statistically significant (Table 5). The taxonomic spread includes archaea ($n = 150$), bacteria ($n = 1735$), viruses ($n = 84$), green land plants ($n = 199$),

fungi ($n = 182$), eukaryotic invertebrates ($n = 43$) and vertebrates ($n = 700$) (Supplementary Table 4).

4. Discussion

The significant association and homology between $\boldsymbol{PTG4}$ and the $\boldsymbol{SLiMS}$ along with the equivalence data ($\boldsymbol{PTG4}{\sim}TG4$) suggest that $TG4$ may influence proteostasis in a multitude of ways (Tables 1–5; Supplementary Tables 1–4, Supplementary Text 2–4).

4.1. $TG4$ may effect stability of mRNA and indirectly influence proteostasis

The short $TG4$ modeled in this study has an average loop length ($h{\sim}2\,Mer$) which may contribute to thermodynamic stability by restricting the mobility of the participating strands (1) [8–11]. The physical presence of $TG4$ will result in a stalled ribosome and translation which is prolonged, inefficient and incomplete [31–33]. Interestingly, this analysis also includes UAG (Amber; $\alpha >$ 0.0000), which when present in-frame will prematurely terminate translation and result in a truncated protein (Table 1) [39]. Whilst nonsense-mediated mRNA decay may be triggered if the stop codon is within ±50 Mer of the exon-junction complex (EJC), a read-through may occur nonetheless. The resulting protein sequences may be modified which in tandem with one or more occurrences of $\boldsymbol{PTG4}$ and/or $\boldsymbol{SLiMS}$ would predispose the same to aggregate and result in a proteopathy [39,40].

4.2. Mechanism(s) of $\boldsymbol{PTG4}$-mediated misfolding

Whilst the preponderance of Glycine (Gly) might impart heightened flexibility and limit the formation of stabilizing secondary structural elements in the hypothetical protein, Proline (Pro) confers rigidity and may retard proper folding. There is also remarkable conservation between the amino acids that comprise $\boldsymbol{PTG4}$ and the $\boldsymbol{SLiMS}$. These include the complex-promoting hydrophobic (Ala, Val, Met, Trp) and ionic (Asp, Glu, Lys, Arg) residues, along with nucleophile-favoring Serine and Threonine (Figures 3 and 4, Tables 2–5). Whilst, the former may favor aggregation by non-covalent interactions, the latter may promote phosphorylation-mediated charge imbalance and thence misfolding. Interestingly, the loops of G4 whence modeled by Adenine-containing codons ($\boldsymbol{Axx}$) are translated to Lysine (K), Arginine (R), Serine (S), Threonine (T) and Isoleucine (I); all of which may also promote misfolding (Figures 3 and 4, Tables 2–5) [8–11,34,35]. The distribution of $\boldsymbol{PTG4}$ amongst physiologically relevant proteins further suggests that the peptide-mediated misfolding may influence/regulate signal transduction, cytoskeleton organization, metabolism, synaptic transmission and transcription/translation (Table 6; Supplementary Table 5).

Table 4. Co-occurrence data for $\boldsymbol{PTG4}$ and known $\boldsymbol{SLiMS}$.

	Disordered regions ($\boldsymbol{IDRs; n = 1445; 0.00 \leq p - value < 0.05}$)										
	$\boldsymbol{SL^-PT^-}$	$\boldsymbol{SL^-PT^+}$	$\boldsymbol{SL^+PT^-}$	$\boldsymbol{SL^+PT^+}$	$\boldsymbol{R_1T}$	$\boldsymbol{R_2T}$	$\boldsymbol{C_1T}$	$\boldsymbol{C_2T}$	**A (%)**	**P (%)**	**R (%)**
$\boldsymbol{SLiMS_1}$	1078	64	58	9	1142	67	1136	73	89.90	12.32	13.43
$\boldsymbol{SLiMS_2}$	749	18	121	29	767	150	870	47	84.84	61.70	19.33
$\boldsymbol{SLiMS_3}$	1212	108	34	9	1320	43	1246	117	89.58	7.69	20.93
	Proteins with disordered segments ($IDPs; n = 800; 0.00 \leq p - value < 0.05$)										
	$\boldsymbol{SL^-PT^-}$	$\boldsymbol{SL^-PT^+}$	$\boldsymbol{SL^+PT^-}$	$\boldsymbol{SL^+PT^+}$	$\boldsymbol{R_1T}$	$\boldsymbol{R_2T}$	$\boldsymbol{C_1T}$	$\boldsymbol{C_2T}$	**A (%)**	**P (%)**	**R (%)**
$\boldsymbol{SLiMS_1}$	86	12	28	18	98	46	114	30	72.22	60.00	39.10
$\boldsymbol{SLiMS_2}$	1	1	96	66	2	162	97	67	40.85	98.50	40.74
$\boldsymbol{SLiMS_3}$	250	57	26	25	307	51	276	82	76.81	30.48	49.01

Abbreviations

$IDRs$: Intrinsically disordered regions
IDPs: Intrinsically disordered proteins
z: Any amino acid
$SLiMS_1$: $[ST]PzR$
$SLiMS_2$: $[ED]zz[DE][AGS]$
$SLiMS_3$: $[KR]zPzzP$
SL^-PT^-: $|\boldsymbol{SNEG} \cap \boldsymbol{PNEG}|$
SL^-PT^+: $|\boldsymbol{SNEG} \cap \boldsymbol{PPOS}|$
SL^+PT^-: $|\boldsymbol{SPOS} \cap \boldsymbol{PNEG}|$
SL^+PT^+: $|\boldsymbol{SPOS} \cap \boldsymbol{PPOS}|$
R_1T: $SL^-PT^- + SL^-PT^+$
R_2T: $SL^+PT^- + SL^+PT^+$
C_1T: $SL^-PT^- + SL^+PT^-$
C_2T: $SL^-PT^+ + SL^+PT^+$
A: Accuracy

P: Precision
R: Recall

Table 5. Occurrence of overlapping amino acids of $\boldsymbol{PTG4}$ and known $\boldsymbol{SLiMS}$ in curated full length protein sequences.

$\boldsymbol{SLiMS}$	Sample	$(z_n)_{n\geq 2} \in \boldsymbol{pTG4_{ij}} \cap \boldsymbol{SLiMS_w}$ $(\boldsymbol{p-value})$	
$\boldsymbol{SLiMS_1 = [ST]PzR}$	Pz	PG $(n=1)$	PG (Degenerate)
$\boldsymbol{SLiMS_2 = [ED]zzD[AGS]}$	$z[DE]$	$G[DE]$ $(n=2)$	$[WGRVAELMKQSTP][AG][DE]z(2)EG[VADE]$
	$[DE]z$	$[DE]G$ $(n=1)$	$(p-value = 0.00069)$
	$zz[DE]$	$[LMKQSTP]G[DE]$ $(n=14)$	
		$[VAE]G[DE]$ $(n=6)$	
		$[WGR][AG][DE]$ $(n=6)$	
	$[DE]zz[DE]$	$[VAE]G[DE]zzEG[VADE]$ $(n=28)$	
		$[WGR][AG][DE]zzEG[VADE]$ $(n=24)$	
		$[WGR][AG][DE]zzEG$ $(n=6)$	
		$GEzzEG[VADE]$ $(n=4)$	
		$GEzzEG$ $(n=1)$	
$\boldsymbol{SLiMS_3 = [KR]zPzzP}$	Pzz	$PG[VADE]$ $(n=4)$	PGV (Degenerate)
			PGA (Degenerate)
			PGD (Degenerate)
			PGE (Degenerate)

Abbreviations

$pTG4_{ij}$: Members of putative peptidome $(pTG4_{ij} \in \boldsymbol{PTG4})$
$SLiMS_w$: Short linear motifs $(SLiMS_w \in \boldsymbol{SLiMS})$
z_n: Shared sequence(s) of amino acids between $\boldsymbol{PTG4}$ and $\boldsymbol{SLiMS}$
i, j, w, n: Indices to characterize members of $\boldsymbol{PTG4}, \boldsymbol{SLiMS}, \boldsymbol{Z}$

Table 6. Proteins encompassing $\boldsymbol{PTG4}$ as candidates for motif mimicry. *

	Cellular function	Disordered regions of proteins
1.	Signal transduction	DP00274, DP00224, DP00141, DP00332, DP01063, DP00506, DP00418, DP00341, DP00435, DP00613, DP00463, DP00954, DP00959, DP01104, DP00611, DP00519, DP00086, DP00707, DP00712
2.	Endocytosis	DP01073, DP01065, DP01066, DP00225
3.	Calcium-calmodulin	DP00092, DP00132, DP00561, DP00118, DP00253
4.	Myofibril assembly	DP01090
5.	Cytoskeleton	DP01056 DP00240, DP01022, DP00169, DP00716, DP00717, DP01100, DP00122
6.	Nuclear pore	DP01075, DP01077, DP01079
7.	Phototransduction	DP00768, DP00347
8.	Targeting	DP00893, DP00609, DP00610, DP01058
9.	Transcription	DP00062, DP00177, DP00633, DP00348, DP00786, DP00049, DP00231, DP00873, DP00720, DP00217, DP00081
10.	Translation	DP00082, DP00164, DP00229 DP00949, DP00134
11.	Synaptic transmission	DP00943
12.	Supercoiling	DP00076
13.	Binding	DP00539, DP00854, DP01052, DP00659, DP00656
14.	Peptide bond formation	DP00944
15.	Enzymes	DP00557, DP00032, DP00095, DP00337, DP00379, DP00787, DP00427, DP00429
16.	Bacterial/parasitic virulence	
	Secreted toxins	DP00345, DP00591
	Cytoadherence	DP00025, DP00065, DP01096
17.	Viral infectivity	
	Cyclophilin interaction	DP00615, DP01031
	Chaperones	DP00699, DP00700, DP00674
	Capsid assembly	DP00133, DP00876
	Membrane fusion	DP01043
	Latency	DP01060
18.	Unknown	DP00119

Note: $\boldsymbol{DP \coloneqq DisProt\ ID}$

4.3. Degeneracy of $\boldsymbol{PTG4}$ *with* $\boldsymbol{SLiMS}$ *in non-vertebrate taxa may favor development of secondary proteopathies*

The distribution of overlapping/shared amino acids in protein sequences of non-vertebrates suggests that $\boldsymbol{PTG4}$ is either completely degenerate with the $\boldsymbol{SLiMS}$ or present in proportions that is statistically significant (Tables 5 and 6; Supplementary Tables 4 and 5). These data imply that motif-mimicry too, might constitute a probable cause (tropism, oncogenic potential, virulence) of infection/infestation-mediated acute/chronic proteopathies [34,35,41,42]. The contribution(s) of misfolding to the pathogenesis of secondary proteopathies is however, debatable. Whilst, there is evidence that mislocalization of proteins can precipitate misfolding, mimicry itself may result in exonuclease-mediated proteolytic cleavage and thence trigger an infective proteopathy [43,44]. Additionally, the presence of sequences of amino acids such as Proline and Threonine in viral or fungal proteins may be responsible for creating and/or maintaining a milieu conducive to the genesis of infective/transmissible proteopathies, viz., a high charge density and imbalance of electrostatic interactions [43,44].

5. Conclusions

The coexistence of potentially translatable G-quadruplexes ($TG4$) with unfolded ribonucleotides in the PCS of an mRNA transcript may have important consequences for protein homeostasis. Here, I have investigated the contribution of a short intra-strand translatable G-quadruplex and its associated peptidome ($TG4{\sim}\boldsymbol{PTG4}$) to the genesis of misfolding-induced proteostasis. The co-occurrence, homology and distribution of overlapping/shared amino acids of $\boldsymbol{PTG4}$ with the $\boldsymbol{SLiMS}$ suggests that this may occur by truncation, complex formation, increased charge density and/or accelerated degradation. An additional mechanism that is also supported is motif-mimicry by pathogens which may trigger the development of infective proteopathies. The putative peptidome ($\sim 7-20\ aa$) that corresponds to the short translatable G-quadruplex delineated by this investigation may be utilized as novel markers of both the primary and secondary proteopathies.

Author's contribution

SK outlined and designed the study, designed and conceptualized the algorithm(s) and formulae for prediction, wrote mathematical proofs to establish rigor, collated the data, constructed the models, formulated the filters, carried out the computational analysis, wrote all necessary code and the manuscript.

Conflict of interest

The author declares no conflict of interest.

References

1. E. Y. Lam, D. Beraldi, D. Tannahill, S. Balasubramanian, G-quadruplex structures are stable and detectable in human genomic DNA, *Nat. Commun.*, **4** (2013), 1796.

2. P. Agarwala, S. Pandey, S. Maiti, The tale of RNA G-quadruplex, *Org. Biomol. Chem.*, **13** (2015), 5570–5585.
3. A. Y. Zhang, S. Balasubramanian, The kinetics and folding pathways of intramolecular G-quadruplex nucleic acids, *J. Am. Chem. Soc.*, **134** (2012), 19297–19308.
4. D. Rhodes, H. J. Lipps, G-quadruplexes and their regulatory roles in biology, *Nucleic Acids Res.*, **43** (2015), 8627–8637.
5. S. Millevoi, H. Moine, S. Vagner, G-quadruplexes in RNA biology, *Wiley Interdiscip. Rev. RNA*, **3** (2012), 495–507.
6. K. Hoogsteen, The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine, *Acta Crystallograph.*, **16** (1963), 907–916.
7. J. Amato, A. Pagano, S. Cosconati, G. Amendola, I. Fotticchia, N. Iaccarino, et al., Discovery of the first dual G-triplex/G-quadruplex stabilizing compound: A new opportunity in the targeting of G-rich DNA structures?, *Biochim. Biophys. Acta Gen. Subj.*, **1861** (2017), 1271–1280.
8. M. Cheng, Y. Cheng, J. Hao, G. Jia, J. Zhou, J. L. Mergny, et al., Loop permutation affects the topology and stability of G-quadruplexes, *Nucleic Acids Res.*, **46** (2018), 9264–9275.
9. S. Pandey, P. Agarwala, S. Maiti, Effect of loops and G-quartets on the stability of RNA G-quadruplexes, *J. Phys. Chem. B*, **117** (2013), 6896–6905.
10. F. Hao, Y. Ma, Y. Guan, Effects of central loop length and metal ions on the thermal stability of G-quadruplexes, *Molecules*, **24** (2019).
11. B. A. Tucker, J. S. Hudson, L. Ding, E. Lewis, R. D. Sheardy, E. Kharlampieva, et al., Stability of the Na(+) form of the human telomeric G-quadruplex: Role of adenines in stabilizing G-quadruplex structure, *ACS Omega*, **3** (2018), 844–855.
12. A. K. Todd, M. Johnston, S. Neidle, Highly prevalent putative quadruplex sequence motifs in human DNA, *Nucleic Acids Res.*, **33** (2005), 2901–2907.
13. L. Q. Gu, Y. Wang, Biomedical diagnosis perspective of epigenetic detections using alpha-hemolysin nanopore, *AIMS Mater. Sci.*, **2** (2015), 448–472.
14. V. T. Mukundan, A. T. Phan, Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences, *J. Am. Chem. Soc.*, **135** (2013), 5017–5028.
15. A. Guedin, J. Gros, P. Alberti, J. L. Mergny, How long is too long? Effects of loop size on G-quadruplex stability, *Nucleic Acids Res.*, **38** (2010), 7858–7868.
16. J. M. Garant, M. J. Luce, M. S. Scott, J. P. Perreault, G4RNA: An RNA G-quadruplex database, *Database (Oxford)*, **2015** (2015), bav059.
17. A. Bedrat, L. Lacroix, J. L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter, *Nucleic Acids Res.*, **44** (2016), 1746–1759.
18. J. M. Garant, J. P. Perreault, M. S. Scott, Motif independent identification of potential RNA G-quadruplexes by G4RNA screener, *Bioinformatics*, **33** (2017), 3532–3537.
19. K. Wethmar, A. Barbosa-Silva, M. A. Andrade-Navarro, A. Leutz, uORFdb—a comprehensive literature database on eukaryotic uORF biology, *Nucleic Acids Res.*, **42** (2014), D60–67.
20. J. Ma, C. C. Ward, I. Jungreis, S. A. Slavoff, A. G. Schwaid, J. Neveu, et al., Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue, *J. Proteome Res.*, **13** (2014), 1757–1765.
21. S. A. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, et al., Peptidomic discovery of short open reading frame-encoded peptides in human cells, *Nat. Chem. Biol.*, **9** (2013), 59–64.

22. M. C. Frith, A. R. Forrest, E. Nourbakhsh, K. C. Pang, C. Kai, J. Kawai, et al., The abundance of short proteins in the mammalian proteome, *PLoS Genet.*, **2** (2006), e52.
23. C. Weldon, J. G. Dacanay, V. Gokhale, P. V. L. Boddupally, I. Behm-Ansmant, G. A. Burley, et al., Specific G-quadruplex ligands modulate the alternative splicing of Bcl-X, *Nucleic Acids Res.*, **46** (2018), 886–896.
24. C. K. Kwok, S. Balasubramanian, Targeted detection of G-quadruplexes in cellular RNAs, *Angew. Chem. Int. Ed. Engl.*, **54** (2015), 6751–6754.
25. G. Mirihana Arachchilage, M. J. Morris, S. Basu, A library screening approach identifies naturally occurring RNA sequences for a G-quadruplex binding ligand, *Chem. Commun. (Camb.)*, **50** (2014), 1250–1252.
26. R. C. Olsthoorn, G-quadruplexes within prion mRNA: The missing link in prion disease?, *Nucleic Acids Res.*, **42** (2014), 9327–9333.
27. T. Endoh, Y. Kawasaki, N. Sugimoto, Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor alpha, *Nucleic Acids Res.*, **41** (2013), 6222–6231.
28. J. F. Fisette, D. R. Montagna, M. R. Mihailescu, M. S. Wolfe, A G-rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing, *J. Neurochem.*, **121** (2012), 763–773.
29. D. Gomez, A. Guedin, J. L. Mergny, B. Salles, J. F. Riou, M. P. Teulade-Fichou, et al., A G-quadruplex structure within the 5'-UTR of TRF2 mRNA represses translation in human cells, *Nucleic Acids Res.*, **38** (2010), 7187–7198.
30. P. Agarwala, S. Pandey, K. Mapa, S. Maiti, The G-quadruplex augments translation in the 5' untranslated region of transforming growth factor β2, *Biochemistry*, **52** (2013), 1528–1538.
31. A. Arora, B. Suess, An RNA G-quadruplex in the 3' UTR of the proto-oncogene PIM1 represses translation, *RNA Biol.*, **8** (2011), 802–805.
32. M. J. Morris, S. Basu, An unusually stable G-quadruplex within the 5'-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells, *Biochemistry*, **48** (2009), 5313–5319.
33. S. Kumari, A. Bugaut, J. L. Huppert, S. Balasubramanian, An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation, *Nat. Chem. Biol.*, **3** (2007), 218–221.
34. R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, et al., Classification of intrinsically disordered regions and proteins, *Chem. Rev.*, **114** (2014), 6589–6631.
35. N. E. Davey, G. Trave, T. J. Gibson, How viruses hijack cell regulation, *Trends Biochem. Sci.*, **36** (2011), 159–169.
36. M. Jucker, L. C. Walker, Self-propagation of pathogenic protein aggregates in neurodegenerative diseases, *Nature*, **501** (2013), 45–51.
37. M. Goedert, M. G. Spillantini, K. Del Tredici, H. Braak, 100 years of Lewy pathology, *Nat. Rev. Neurol.*, **9** (2013), 13–24.
38. D. Piovesan, F. Tabaro, I. Micetic, M. Necci, F. Quaglia, C. J. Oldfield, et al., DisProt 7.0: A major update of the database of disordered proteins, *Nucleic Acids Res.*, **45** (2017), D219–D227.
39. S. Hutchinson, A. Furger, D. Halliday, D. P. Judge, A. Jefferson, H. C. Dietz, et al., Allelic variation in normal human FBN1 expression in a family with Marfan syndrome: A potential modifier of phenotype?, *Hum. Mol. Genet.*, **12** (2003), 2269–2276.
40. Y. F. Chang, J. S. Imam, M. F. Wilkinson, The nonsense-mediated decay RNA surveillance pathway, *Annu. Rev. Biochem.*, **76** (2007), 51–74.

41. M. M. Klein, A. G. Gittis, H. P. Su, M. O. Makobongo, J. M. Moore, S. Singh, et al., The cysteine-rich interdomain region from the highly variable plasmodium falciparum erythrocyte membrane protein-1 exhibits a conserved structure, *PLoS Pathog.*, **4** (2008), e1000147.
42. J. A. Corcoran, R. Syvitski, D. Top, R. M. Epand, R. F. Epand, D. Jakeman, et al., Myristoylation, a protruding loop, and structural plasticity are essential features of a nonenveloped virus fusion peptide motif, *J. Biol. Chem.*, **279** (2004), 51386–51394.
43. K. A. Morrow, C. D. Ochoa, R. Balczon, C. Zhou, L. Cauthen, M. Alexeyev, et al., Pseudomonas aeruginosa exoenzymes U and Y induce a transmissible endothelial proteinopathy, *Am. J. Physiol. Lung Cell Mol. Physiol.*, **310** (2016), L337–353.
44. A. L. Woerman, S. A. Kazmi, S. Patel, A. Aoyagi, A. Oehler, K. Widjaja, D. A. Mordes, et al., Familial Parkinson's point mutation abolishes multiple system atrophy prion replication, *Proc. Natl. Acad. Sci. U S A*, **115** (2018), 409–414.
45. J. D. Beaudoin, J. P. Perreault, Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening, *Nucleic Acids Res.*, **41** (2013), 5898–5911.
46. J. D. Beaudoin, R. Jodoin, J. P. Perreault, New scoring system to identify RNA G-quadruplex folding, *Nucleic Acids Res.*, **42** (2014), 1209–1223.
47. R. Shahid, A. Bugaut, S. Balasubramanian, The BCL-2 5' untranslated region contains an RNA G-quadruplex-forming motif that modulates protein expression, *Biochemistry*, **49** (2010), 8300–8306.
48. S. Saxena, D. Miyoshi, N. Sugimoto, Sole and stable RNA duplexes of G-rich sequences located in the 5'-untranslated region of protooncogenes, *Biochemistry*, **49** (2010), 7190–7201.
49. J. D. Beaudoin, J. P. Perreault, 5'-UTR G-quadruplex structures acting as translational repressors, *Nucleic Acids Res.*, **38** (2010), 7022–7036.
50. R. Jodoin, L. Bauer, J. M. Garant, A. Mahdi Laaref, F. Phaneuf, J. P. Perreault, The folding of 5'-UTR human G-quadruplexes possessing a long central loop, *RNA*, **20** (2014), 1129–1141.
51. D. Bhattacharyya, P. Diamond, S. Basu, An Independently folding RNA G-quadruplex domain directly recruits the 40S ribosomal subunit, *Biochemistry*, **54** (2015), 1879–1885.
52. A. Cammas, A. Dubrac, B. Morel, A. Lamaa, C. Touriol, M. P. Teulade-Fichou, et al., Stabilization of the G-quadruplex at the VEGF IRES represses cap-independent translation, *RNA Biol.*, **12** (2015), 320–329.
53. M. J. Morris, Y. Negishi, C. Pazsint, J. D. Schonhoft, S. Basu, An RNA G-quadruplex is essential for cap-independent translation initiation in human VEGF IRES, *J. Am. Chem. Soc.*, **132** (2010), 17831–17839.
54. J. Christiansen, M. Kofod, F. C. Nielsen, A guanosine quadruplex and two stable hairpins flank a major cleavage site in insulin-like growth factor II mRNA, *Nucleic Acids Res.*, **22** (1994), 5709–5716.
55. S. Lammich, F. Kamp, J. Wagner, B. Nuscher, S. Zilow, A. K. Ludwig, et al., Translational repression of the disintegrin and metalloprotease ADAM10 by a stable G-quadruplex secondary structure in its 5'-untranslated region, *J. Biol. Chem.*, **286** (2011), 45063–45072.
56. P. Agarwala, S. Pandey, S. Maiti, Role of G-quadruplex located at 5' end of mRNAs, *Biochim. Biophys. Acta*, **1840** (2014), 3503–3510.
57. M. Subramanian, F. Rage, R. Tabet, E. Flatter, J. L. Mandel, H. Moine, G-quadruplex RNA structure as a signal for neurite mRNA targeting, *EMBO Rep.*, **12** (2011), 697–704.

58. A. von Hacht, O. Seifert, M. Menger, T. Schutze, A. Arora, Z. Konthur, et al., Identification and characterization of RNA guanine-quadruplex binding proteins, *Nucleic Acids Res.*, **42** (2014), 6630–6644.
59. S. Balaratnam, S. Basu, Divalent cation-aided identification of physico-chemical properties of metal ions that stabilize RNA G-quadruplexes, *Biopolymers*, **103** (2015), 376–386.
60. S. Stefanovic, G. J. Bassell, M. R. Mihailescu, G quadruplex RNA structures in PSD-95 mRNA: potential regulators of miR-125a seed binding site accessibility, *RNA*, **21** (2015), 48–60.
61. J. C. Grigg, N. Shumayrikh, D. Sen, G-quadruplex structures formed by expanded hexanucleotide repeat RNA and DNA from the neurodegenerative disease-linked C9orf72 gene efficiently sequester and activate heme, *PLoS One*, **9** (2014), e106449.
62. P. Fratta, S. Mizielinska, A. J. Nicoll, M. Zloh, E. M. Fisher, G. Parkinson, et al., C9orf72 hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes, *Sci. Rep.*, **2** (2012), 1016.
63. H. Y. Weng, H. L. Huang, P. P. Zhao, H. Zhou, L. H. Qu, Translational repression of cyclin D3 by a stable G-quadruplex in its 5' UTR: implications for cell cycle regulation, *RNA Biol.*, **9** (2012), 1099–1109.
64. H. H. Woo, T. Baker, C. Laszlo, S. K. Chambers, Nucleolin mediates microRNA-directed CSF-1 mRNA deadenylation but increases translation of CSF-1 mRNA, *Mol. Cell Proteomics*, **12** (2013), 1661–1677.
65. H. Dhayan, A. R. Baydoun, A. Kukol, G-quadruplex formation of FXYD1 pre-mRNA indicates the possibility of regulating expression of its protein product, *Arch. Biochem. Biophys.*, **560** (2014), 52–58.
66. S. G. Rouleau, J. D. Beaudoin, M. Bisaillon, J. P. Perreault, Small antisense oligonucleotides against G-quadruplexes: specific mRNA translational switches, *Nucleic Acids Res.*, **43** (2015), 595–606.
67. J. Kralovicova, A. Lages, A. Patel, A. Dhir, E. Buratti, M. Searle, et al., Optimal antisense target reducing INS intron 1 retention is adjacent to a parallel G quadruplex, *Nucleic Acids Res.*, **42** (2014), 8161–8173.
68. M. Faudale, S. Cogoi, L. E. Xodo, Photoactivated cationic alkyl-substituted porphyrin binding to g4-RNA in the 5'-UTR of KRAS oncogene represses translation, *Chem. Commun. (Camb.)*, **48** (2012), 874–876.
69. M. M. Ribeiro, G. S. Teixeira, L. Martins, M. R. Marques, A. P. de Souza, S. R. Line, G-quadruplex formation enhances splicing efficiency of PAX9 intron 1, *Hum. Genet.*, **134** (2015), 37–44.
70. E. P. Booy, R. Howard, O. Marushchak, E. O. Ariyo, M. Meier, S. K. Novakowski, et al., The RNA helicase RHAU (DHX36) suppresses expression of the transcription factor PITX1, *Nucleic Acids Res.*, **42** (2014), 3346–3361.
71. Y. Zhang, C. M. Gaetano, K. R. Williams, G. J. Bassell, M. R. Mihailescu, FMRP interacts with G-quadruplex structures in the 3'-UTR of its dendritic target Shank1 mRNA, *RNA Biol.*, **11** (2014), 1364–1374.
72. H. Martadinata, A. T. Phan, Formation of a stacked dimeric G-quadruplex containing bulges by the 5'-terminal region of human telomerase RNA (hTERC), *Biochemistry*, **53** (2014), 1595–1600.

73. K. Hirashima, H. Seimiya, Telomeric repeat-containing RNA/G-quadruplex-forming sequences cause genome-wide alteration of gene expression in human cancer cells in vivo, *Nucleic Acids Res.*, **43** (2015), 2022–2032.
74. Y. Katsuda, S. Sato, L. Asano, Y. Morimura, T. Furuta, H. Sugiyama, et al., A Small Molecule That Represses Translation of G-Quadruplex-Containing mRNA, *J. Am. Chem. Soc.*, **138** (2016), 9037–9040.
75. C. K. Kwok, A. B. Sahakyan, S. Balasubramanian, Structural Analysis using SHALiPE to Reveal RNA G-Quadruplex Formation in Human Precursor MicroRNA, *Angew. Chem. Int. Ed. Engl.*, **55** (2016), 8958–8961.
76. P. Agarwala, S. Kumar, S. Pandey, S. Maiti, Human telomeric RNA G-quadruplex response to point mutation in the G-quartets, *J. Phys. Chem. B*, **119** (2015), 4617–4627.
77. V. Marcel, P. L. Tran, C. Sagne, G. Martel-Planche, L. Vaslin, M. P. Teulade-Fichou, et al., G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms, *Carcinogenesis*, **32** (2011), 271–278.
78. A. Decorsiere, A. Cayrel, S. Vagner, S. Millevoi, Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3'-end processing and function during DNA damage, *Genes Dev.*, **25** (2011), 220–225.
79. W. Huang, P. J. Smaldino, Q. Zhang, L. D. Miller, P. Cao, K. Stadelman, et al., Yin Yang 1 contains G-quadruplex structures in its promoter and 5'-UTR and its expression is modulated by G4 resolvase 1, *Nucleic Acids Res.*, **40** (2012), 1033–1049.
80. A. Arora, M. Dutkiewicz, V. Scaria, M. Hariharan, S. Maiti, J. Kurreck, Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif, *RNA*, **14** (2008), 1290–1296.

Supplementary information

Supplementary Tables 1, 2, 3, 4, 5.
Supplementary Texts 1, 2, 3, 4, 5.

AIMS Press