Abstract: Tandem repeats of various trinucleotide motifs are frequent entities in transcripts, and RNA structures formed by these sequences depend on the motif type and number of reiterations. The functions performed by normal triplet repeats in transcripts are poorly understood, but abnormally expanded repeats of certain types trigger pathogenesis in several human genetic disorders known as the triplet repeat expansion diseases (TREDs). The diseases caused by expanded non-coding CUG and CGG repeats in transcripts include myotonic dystrophy type 1 and fragile X-associated tremor ataxia syndrome. Another group of disorders in which transcripts containing translated CAG repeats play an auxiliary role in pathogenesis include Huntington’s disease and several spinocerebellar ataxias. In this review, we gathered existing knowledge regarding the structural features of triplet repeats in transcripts and discussed this in the context of various pathogenic mechanisms assigned to toxic RNA repeats. These mechanisms include aberrant alternative splicing, the inhibition of nuclear transport and export, induction of the innate immune response, alteration of a microRNA biogenesis pathway and abnormal activation of an RNA interference pathway. We also provide ideas for future investigations to reveal further mechanisms of pathogenesis directly triggered by mutant RNA repeats in TREDs.

Keywords: CAG repeats; fragile X-associated tremor ataxia; Huntington’s disease; myotonic dystrophy type 1; RNA pathogenesis; spinocerebellar ataxia.

Introduction

Trinucleotide repeats (TNRs) are members of a large family of short tandem repeats (STRs) also known as simple sequence repeats (SSRs) or microsatellites (Toth et al., 2000). The STR family comprises all sequences composed of 1-6-nucleotide motifs generally repeated <30 times (Ellegren, 2004). Altogether, these repeats make a 3% contribution to the entire sequence of the human genome (Subramanian et al., 2003). What makes TNRs distinct from other STRs, e.g., dinucleotide or tetranucleotide repeats, is their abundance in the intergenic regions of the genome and introns and untranslated regions of genes as well as in the translated portions of protein coding sequences. This abundance within the coding regions can occur because the frequent length variation of TNRs, resulting from the very high mutation rate of STR sequences, does not generate deleterious frameshifts in the encoded proteins (Borstnik and Pumpernik, 2002). Functionally, TNRs in DNA may be neutral sequences, have several regulatory roles in gene expression or be the source of rapid evolutionary changes, depending on their type, length and localization in the genome and within genes. Abnormal expansions of certain TNR tracts during their intergenerational transmission induce a number of human neurological diseases known as triplet repeat expansion diseases (TREDs) (Orr and Zoghbi, 2007). The complex pathomechanisms of these disorders have posed serious challenges for researchers in the past two decades and still remain rather poorly recognized.

The abundance of transcripts harboring diverse TNRs may be very high, much higher than the abundance of their genomic templates. The TNRs in transcripts, as in genomic DNA, may be nonfunctional or have some regulatory roles. Additionally, they may give rise to shorter regulatory RNAs after being processed in cells. Mounting evidence has demonstrated that transcripts from an increasing number of the TREDs genes, and specifically the TNR portions of these transcripts, are toxic to cells, for example mutant repeats can directly alter certain normal cellular processes and eventually lead to disease (reviewed...
in Wojciechowska and Krzyzosiak, 2011a; Krzyzosiak et al., 2012). These observations indicate that the TNR RNAs deserve special attention because normal cellular functions and pathological transformations may depend on their status. An important aspect of this status, in addition to factors such as repeat localization and length, is the RNA structure formed by the different types of TNRs.

In this review, we first present and discuss the structures formed by all 20 different types of isolated TNR sequences. We next combine this information with the frequencies of specific TNRs in the human genome and exome and analyze the structures from functional and evolutionary perspectives. Then, we focus on the structures of TNRs in the sequence context of the gene transcripts implicated in TREDs. We compare the structures of the normal variants of these transcripts with those of their mutant forms and evaluate the contribution of the flanking sequences to the overall structure of the repeat region. Finally, we indicate those cellular pathways and processes that have already been demonstrated to be altered by mutant TNR transcripts and discuss possible further investigation of other mechanisms of RNA toxicity in TREDs.

**Structural classes of isolated triplet repeats**

The four common RNA nucleotides may form 64 different combinations of trinucleotide motifs. However, this number is reduced to 20 different TNRs when the four homonucleotide motifs are excluded and the three different phases of the 60 individual sequence motifs are combined. Recently, the ability of these TNRs to adopt higher-order RNA structures has been investigated by

![Figure 1](image-url)
chemical and enzymatic structure probing and biophysical methods (Sobczak et al., 2010). As a result, the TNRs have been divided into four structural classes: (I) very stable G-quadruplex structures, known also as G-quartets; (II) semi-stable hairpins; (III) unstable hairpins; and (IV) not forming any higher order structures (Figure 1A). The G-quadruplex group consists of two members: the G-rich AGG and UGG motifs. A nuclear magnetic resonance study of the (AGG)n quadruplex provided deeper insight into its structure and revealed the existence of the G-tetrad plane and G(:A):G(:A):G hexad plane (Nishikawa et al., 2009). The more numerous group of semi-stable hairpins contains TNRs composed of CGA, CGU and all four CNG motifs (where N is any nucleotide, A, U, G or C). The stem portions of the CNG repeat hairpins are composed of C–G and G–C base pairs and periodic N–N interactions, the detailed character of which cannot be determined by chemical and biochemical structure probing (Sobczak et al., 2003). The terminal loop present in the CNG repeat hairpins typically contains four nucleotides. Three of the four CNG repeats, CUG, CAG and CGG, show the tendency to align into alternative ‘in register’ conformations, i.e., ‘slippery hairpins’. The ‘slippage’ effect could be reduced to align into alternative ‘in register’ conformations, i.e., four CNG repeats, CUG, CAG and CGG, show the tendency to align into alternative ‘in register’ conformations, i.e., ‘slippery hairpins’. The ‘slippage’ effect could be reduced or completely eliminated by ‘clamping’ the repeats at both ends with several canonical G–C and C–G pairs (Sobczak et al., 2003). This result showed that the sequence context of the repeats may strongly influence their structural features and hence biological properties. The class of TNRs forming unstable hairpins under decreased temperature or increased Mg\(^{2+}\) concentration contains the AUG, CAU, CUA, UAG and UUA motifs. The cleavage patterns of these transcripts determined at 37°C do not show any hallmarks of hairpin structures. Finally, as many as seven TNR RNAs, the AAG, CAA, CCA, CCU, CUU, UAA and UUG motifs, appeared unable to form any structural states (Sobczak et al., 2010). The structural variety of the TNR RNAs may reflect their potential to interact with a variety of RNA binding proteins and perform diverse biological functions. The normal functions of TNRs in transcripts are, however, poorly understood.

In recent years, the crystal structures of short oligoribonucleotides composed of the CUG, CAG and CGG repeats have been determined (Mooers et al., 2005; Kiliszek et al., 2009, 2010, 2011; Kumar et al., 2011). All these repeated sequences form duplex structures that stack onto each other and likely represent the stem portions of hairpins formed by long sequences of this type. The common feature of all these duplexes is the formation of RNA-A helices stabilized by the periodic occurrence of two standard C–G and G–C base pairs. The structural variety of these duplexes is determined by the unique geometry and different hydrogen bonding pattern of each nonstandard base pair. Although these non-canonical pairs alter the local duplex architecture to various extents, the disruption of the overall helix structure is nevertheless either low or moderate. In the CUG repeat duplex, the two opposed uracil residues do not create an accommodation problem because they remain well separated from each other and form only a single direct hydrogen bond between the N3 atom of one uracil residue and O4 atom of the other. This pair is called the ‘stretched U–U wobble’ (Kiliszek et al., 2009). A problem occurs when two bulky purine bases oppose each other within the helical structure. The crystal structures of the CAG and CGG repeat duplexes show two different solutions to this problem. Two adenine residues facing each other stay in the anti conformation but are shifted out of the helical axis towards the major groove. This conformation results in a general unwinding of the helical structure. The geometry of the A–A pair allows for a single, weak hydrogen bonding interaction between the carbon atom C2–H2 of one adenine and the nitrogen N1 atom of the other (Kiliszek et al., 2010). This type of A–A wobble has not been observed in any other RNA structure reported thus far. In contrast, two opposing guanine residues resolve the accommodation problem by flipping one of the guanine rings from the anti to the syn conformation, which results in a local unwinding of the helix. Two guanine residues present in this structural arrangement form two direct hydrogen bonds between the O6 and N1 atoms and between N7 and N2H (Kiliszek et al., 2011). This relatively strong hydrogen bonding system makes the CGG repeat duplexes more stable than the duplexes formed by the CUG and CAG repeats and explains why the CGG repeats do not show the slippage effect (Sobczak et al., 2003).

Occurrence and function of triplet repeats in transcripts

Based on our current knowledge of the composition of various human transcriptomes, we can assume that the great majority of the human genome is transcribed and that much of the genome is transcribed in both directions (Mattick, 2005; Batra et al., 2010). This knowledge provides an impression of the complexity of the universe of human TNR-containing transcripts. Despite this complexity, the approximate number of different transcripts bearing TNRs could be estimated from analysis of the human genome sequence performed to identify all TNRs composed of at least six pure repeats. More than 32,000 sequences in the human genome fulfill these criteria.
Ten different TNRs, resulting from the combination of both sense and antisense sequences in all their phases, differed greatly in abundance (Figure IB). The most abundant in the genome are the AAT and AAC repeats, whereas the least abundant are the GAC and CGG repeats. As many as 1030 TNRs were found in the exon sequences of 878 annotated human genes. A comparison of the genomic frequency of repeats with their exonic frequency revealed that AAT and ATT were the most under-represented in exons, whereas CGA and CGG were the most over-represented. Importantly, the same comparison made for TNRs in their sense and antisense orientations separately revealed an interesting correlation between their frequencies and the structures formed by the repeats in the transcripts. The group of TNRs most over-represented in exons is that showing a strong tendency to form a hairpin structure that includes all CNG repeats (Kozlowski et al., 2010). This result might be interpreted as a strong positive selection of hairpin-forming repeats within exons and might be related to some advantageous features of the repeats in the functional regions of human genes. The biological functions of TNRs might be carried out at the transcript or protein level. In mRNAs, the TNR functionality may depend on its type, structure and localization in the different functional regions of mature transcripts. The highest frequency of TNRs occurred in the open reading frame (ORF, 59%), followed by the five prime (5′UTR, 28%) and three prime untranslated region (3′UTR, 13%). These results corresponded well with the results of an earlier survey (Jasinska et al., 2003). The hairpin structure-forming TNRs occur frequently in the 5′UTR and the AT-rich repeats in the 3′UTR. In the ORF, as many as 15 different TNR types are present, and the CAG motif encoding glutamine has the highest representation, at 123 occurrences. RNA repeats may regulate gene expression via the structures they form and through interactions with repeat binding proteins. In the cell nucleus, the repeats were shown to act as splicing enhancers and inducers of transcription slippage (Fabre et al., 2002). In the cytoplasm, TNRs may act as regulators of mRNA localization and translation (Richards et al., 1993; Gay and Babajko, 2000; Raca et al., 2000). Short tracts of TNR sequences are present in the transcriptomes of many organisms, including bacteria and lower eukaryotes, and nature has had sufficient time at the evolutionary scale to design specific repeat binding proteins and develop functions for these sequences (Krzyzosiak et al., 2006). The normal functions of TNRs in RNA poorly understood, however, and no comprehensive study has been performed thus far to address and resolve this issue. At present, TNRs are much better known from the genetic disorders they trigger in humans when the repeats abnormally expand beyond the characteristic ranges in length of normal polymorphic sequences of this type.

### Brief characteristics of triplet repeat expansion diseases

At least 17 human hereditary neurological diseases are caused by the expansion of triplet repeats located in either coding or non-coding regions of single, functionally unrelated genes (Figure 2). These typically late-onset diseases include fragile X syndrome (FXS), fragile X-associated tremor ataxia syndrome (FXTAS), Huntington’s disease (HD), a number of spinocerebellar ataxias (SCAs) and myotonic dystrophy type 1 (DM1). FXS and FXTAS are triggered by different lengths of the CGG repeat expansions present in the 5′UTR of the same gene, **FMR1**. Five to 54 CGGs are present in the normal population, but 55–200 CGGs occur in FXTAS patients, up to 1000 repeats are present in FXS patients (Verkerk et al., 1991; Hagerman and Hagerman, 2004). The 3′UTR of the **DMPK** gene implicated in DM1 contains five to 37 CTG repeats in the normal population and 50–3000 CTGs in DM1 patients (Brook et al., 1992). In contrast, the repeat expansions that occur in the ORF are smaller. In the **HTT** gene involved in HD, six to 35 CAG repeats are present in the normal population, and 36 to more than 100 in HD patients (Duyao et al., 1993). The general types of pathogenic mechanisms are as follows: toxic RNA gain-of-function; toxic protein gain-of-function; and mutant transcript and mutant protein loss-of-function (Orr and Zoghbi, 2007). The latter mechanism beside FXS (Sutcliffe et al., 1992) includes Friedreich’s ataxia, in which the expression of a product from the mutant gene is inhibited by the expanded GAA repeats located in the **FRDA** gene.
intron of the FXN gene (Campuzano et al., 1996). Mutant RNA toxicity was first shown to be directly involved in the pathogenesis of DM1 (Taneja et al., 1995) and then in FXTAS (Tassone et al., 2004).

The largest group of TREDs is caused by expanded CAG repeats present in the ORF of various individual genes. These repeats are translated into abnormally elongated polyglutamine (polyQ) tracts in proteins to which the toxicity is commonly attributed, and the diseases are often named polyglutamine or polyQ disorders. The mutant proteins, however, derive from mutant CAG repeat-containing transcripts that may also be toxic. This view of TREDs pathogenesis has recently been complicated by the discovery of repeat associated non-ATG translation (Zu et al., 2011), suggesting that toxic proteins may also be derived from repeats previously thought to be noncoding. Also the CAG repeats-driven -1 frameshift observed in Drosophila and mammalian models of SCA3 contributes to the increased complexity of pathogenic mechanisms in TREDs (Stochmanski et al., 2012).

In the next section, we will pay special attention to TREDs-related transcripts and specifically to the structures of their triplet repeat regions. In the following sections, we will present the toxic roles of mutant RNA repeats that have already been identified in triggering the TREDs.

### Structure of triplet repeat regions in TREDs transcripts

RNA structures formed by TNRs alone and structures formed by the repeats in their host transcripts may differ. It was therefore necessary to demonstrate the architecture of the TNR regions in the TREDs-related transcripts. The solution structures of the main normal variants and mutant variants of the transcripts from a number of TREDs genes (Figure 3) may also be derived from repeats previously thought to be noncoding. Also the CAG repeats-driven -1 frameshift observed in Drosophila and mammalian models of SCA3 contributes to the increased complexity of pathogenic mechanisms in TREDs (Stochmanski et al., 2012).

The CUG repeat region from the 3′UTR of the DMPK transcript implicated in DM1 was studied first. Five CUG repeats - the most frequent configuration in the normal population - formed no secondary structure, 11 repeats formed unstable hairpins, 21 repeats formed semi-stable hairpins, and the expanded 49 repeats formed fairly stable hairpins (Napierala and Krzyzosiak, 1997). The repeat hairpins were slippery, i.e., there was no significant ‘freezing’ of the repeat slippage imposed by the sequences flanking the repeats. The cleavage patterns obtained in the structure probing experiments gave the impression of an enlarged hairpin terminal loop. Several different transcript conformers resulting from the alternative alignments of the CUG repeats were present under the structure probing conditions, and gave that impression. Overall, the clear structural differences between the long normal CUG repeat (e.g., 21 CUG) and mutant CUG repeat were the length and rigidity of the repeat hairpin stem (Figure 3A).

The structural features of the CAG repeat regions from the transcripts of the ATN1, CACNA1A and ATXN3 genes implicated in dentatorubral-pallidoluysian atrophy, SCA6 and SCA3, respectively, were studied using a similar experimental scheme (Michlewski and Krzyzosiak, 2004). The repeats in ATN1 (Figure 3B) were shown to form several slipped hairpin variants, as demonstrated for the CUG repeats in the DMPK transcript. In contrast, the CACNA1A repeats formed hairpins stabilized by the base-pairing interactions of the repeat flanking sequences (Figure 3C).

The repeat region of the ATXN3 transcript has an unusual architecture as the 3′ terminal CAG repeats base pair with the pseudorepeats flanking the CAG repeat on its 5′ side (Figure 3D). The (CAG)C/G SNP located at the border of the repeated sequence, and its 3′ flanking specific sequence was shown to modulate the hairpin structure.

The most recently studied CAG repeat regions of the HTT and AR transcripts differ considerably from the repeat regions of other TREDs-related transcripts (de Mezer et al., 2011). The difference involves the presence of another type of triplet repeat in the close neighborhood of the expansion-prone CAG repeats. In the HTT RNA (Figure 3E), this neighbor is a polymorphic CCG repeat most frequently containing either 7CCG or 10CCG separated by 12 nucleotides of a specific sequence from the CAG repeat tract. One question was whether the different repeats form separate hairpins or participate in the formation of a single composite hairpin structure. In vitro transcripts containing a number of the most common CAG and CCG length combinations were analyzed using a battery of structure probing reagents, and in each case, a single hairpin structure was demonstrated to be formed by the compound repeats (de Mezer et al., 2011). In the fairly stable base portion of the hairpin, the 5′ part of the CAG repeats is engaged in base-pairing with all seven or 10 CCGs. The consecutive C-G and G-C base pairs are followed periodically by the nonstandard A-C pairs. The central portion of the tripartite hairpin stem is composed of the partially base-paired CAG repeats and 12 nucleotides specific HTT sequence. Only the upper portion of this hairpin is composed solely of CAG repeats. Mutant
HTT transcripts differ from the normal transcripts in the length of this upper portion of the hairpin, while the other elements of the hairpin architecture remain the same.

In the AR transcript (Figure 3F), not one but two other repeat tracts are present in close proximity to the expandable CAG repeats. Three CUG repeats directly precede the long CAG repeat tract, which is separated by 18 nucleotides of a specific sequence from the monomorphic tract of six CAG repeats. Structure probing experiments revealed that the CUG repeats pair with the last three CAG repeats of the long repeat tract and form a fully paired clamp at the base of the hairpin portion consisting of pure

---

**Figure 3** Structural organization of TNRs in the context of the flanking sequences of different TREDs transcripts (A–I). Simplified models were proposed based on experimental structure probing. Normal and mutant transcripts (with characteristic triplet repeat ranges) are shown on the left- and right-hand sides, respectively. Different colors represent different repeated sequences: red, CUG; green, CAG; blue, CGG; purple, CCG repeats; and gray, specific flanking sequence. The normal transcript variants contain either pure TNR tracts or interrupted repeats (interrupting bases are marked by orange circles). The predicted structures of the mutant transcripts that were not experimentally confirmed are indicated by an asterisk. **Alleles containing 39–44 interrupted repeats are normal, but those containing pure repeats are disease alleles.**
CAG repeats. In addition, this repeat hairpin is further stabilized by the base-pairing system of the nearest specific sequences flanking the repeats. The short tract of the six CAG repeats does not contribute to the formation of the long hairpin structure (de Mezer et al., 2011). The issue of possible interactions between different triplet repeat tracts present within a single transcript is not only relevant to the TREDs-related RNAs discussed above; in the human transcriptome, as many as 170 annotated mRNAs contain multiple TNR tracts. The predicted secondary structures of selected transcripts from this group suggest that even very distant TNR sequences may interact with each other, possibly making the final structure more compact (Kozlowski et al., 2010).

Most of the normal variants of several TREDs-related genes contain specific interruptions (base substitutions) within the repeat tracts that are absent in the expanded mutant alleles. The structural role played by the repeat interruptions in the transcripts was analyzed in the ATXN1, ATXN2 and FMR1 transcripts implicated in SCA1, SCA2 and FXS1, respectively. The CAU triplets interrupting the CAG repeats in the ATXN1 transcripts (Figure 3G) and CAA interspersions present within the CAG repeats in the ATXN2 transcripts (Figure 3H) both destabilize the CAG repeat hairpin structures (Sobczak and Krzyzosiak, 2004a, 2005). The hairpin destabilization strategy depends on the number and localization of the repeat interruptions within the CAG repeat tracts. The base substitutions enlarge the existing terminal loops, nucleate extra loops or destabilize the stem portions of the semi-stable CAG repeat hairpins. The AGG interruptions of the CGG repeats present in the FMR1 transcript (Figure 3I) destabilize the CGG hairpin in a similar manner (Napierala et al., 2005).

Taken together, the structural probing experiments revealed that the hairpin structures formed by the CUG, CGG and CAG repeat regions of the transcripts of nine genes implicated in TREDs show similarities and differences. Some of the common and distinct features of these hairpins had already become apparent from the biochemical and crystallographic studies of isolated repeats. After placing the repeats in the context of their natural flanking sequences and including the structural analysis of many naturally occurring variants of the interrupted repeats, new features of the repeat regions became evident. The nearest flanks were shown to provide some extra stabilization to all repeat hairpins investigated, with the exception of the CUG repeat hairpin from the DMPK transcript and CAG repeat hairpins from the ATXN1 and CACNA1A transcripts. The additional base pairs extend the repeat hairpin stem and cause greater exposure to protein interaction. Different repeat interruptions present within most normal ATXN1, ATXN2 and FMR1 transcripts were all shown to destabilize the hairpin structures. The most prominent difference between the non-interrupted mutant repeats and interrupted normal repeats is the formation of single long hairpin by the former and smaller split hairpins by the latter. This structural information has been gained not for full-length transcripts but for their fragments harboring the repeats. The experiments were performed in solution conditions typical for structure probing, i.e., only partly resembling the cellular environment. In the full length transcripts, however, the repeat region folding is the same or very similar, as demonstrated for the relatively short ATXN3 mRNA (Michlewski and Krzyzosiak, 2004). The question that remains to be addressed is whether the stability of different repeat hairpins is sufficiently high to be preserved in cellular conditions in the presence of various RNA binding proteins.

RNA toxicity in TREDs – aberrant alternative splicing

Missplicing caused by expanded CUG repeats

The pathogenesis of DM1 focusing on RNA toxicity has been the subject of detailed examination over the past decade. The commonly accepted pathomechanism implies that the expanded CTG repeat gives rise to a mutant transcript retained in the cell nucleus, where it sequesters the MBNL family proteins into nuclear foci (Fardaei et al., 2001; Mankodi et al., 2001; Cardani et al., 2006) and causes alternative splicing aberrations of multiple genes. The aberrant splicing of several developmentally-regulated genes has been linked to the clinical symptoms of DM1, e.g., the missplicing of insulin receptor (INSR) to insulin resistance (Savkur et al., 2001), chloride channel 1 (CLCN1) to myotonia (Charlet et al., 2002), sarcoplasmic/endoplasmic reticulum Ca\(^{2+}\) ATPase 1 (SERCA1) to muscle wasting (Kimura et al., 2005), troponin T type 2 (TNNT2) to defective cardiac functions (Philips et al., 1998), and microtubule-associated protein tau (MAPT) to cognitive deficits (Jiang et al., 2004). The splicing aberrations cause fetal isoforms of these proteins to persist in DM1 adults. Thus, the most prominent molecular hallmarks of DM1 are nuclear foci formation in which the mutant transcript colocalizes with the MBNL1 protein and alternative splicing deregulation of MBNL-regulated transcripts (Ranum and Cooper, 2006). The loss-of-function of the MBNL1 protein is accompanied...
in myotonic dystrophy tissue by the gain-of-function of antagonistic splicing factor CUGBP1. The elevated levels of CUGBP1 (Timchenko et al., 2001) are explained by the increased stability of the protein resulting from its phosphorylation (Kuyumcu-Martinez et al., 2007). The study that identified MBNL1 as the protein that directly binds to the CUG repeats demonstrated that this binding occurs in a length-dependent manner, with a strong preference for mutant repeats (Miller et al., 2000). The RNA baits that were instrumental in the MBNL1 identification had earlier been shown to form semi-stable hairpins (Napierala and Krzyzosiak, 1997). These results therefore suggested that the protein binds preferentially to the double-stranded CUG repeats. The binding of MBNL1 and CUGBP1 to double stranded CUG and single stranded CUG repeats, respectively, was also demonstrated by electron microscopy (Michalowski et al., 1999; Yuan et al., 2007). Further studies showed that MBNL1 binds similar RNA structures in its natural splicing substrates and CUG repeats (Yuan et al., 2007). The biochemical analyses demonstrated efficient MBNL1 binding to both the single-stranded and double-stranded repeats (Yuan et al., 2007; Mykowska et al., 2011). The crystal structure of the hexaribonucleotide CGCUGU bound to the zinc-finger domains of MBNL1, however, suggests that the protein may even more efficiently bind the single-stranded substrate (Teplova and Patel, 2008). Several studies have been performed to determine the nucleotide sequence preferences for MBNL1 binding, showing that MBNL1 binds to GC dinucleotides preferably embedded in the YGCY (Y=U,C) sequence motifs found in its natural splicing substrates (Warf and Berglund, 2007; Yuan et al., 2007; Warf et al., 2009; Goers et al., 2010; Fu et al., 2012). Recent studies with model oligoribonucleotides have suggested that MBNL1, which contains four CCCH zinc-fingers, has a very flexible structure and can bind GC dinucleotides in highly diverse sequence configurations (Cass et al., 2011).

Another question that has not yet been satisfactorily resolved is how many proteins other than the MBNL family are trapped in the CUG RNA foci (Wojciechowska and Krzyzosiak, 2011b). A recent study identified the p68 helicase as a protein that binds to CUG-RNA-bait and colocalizes with the mutant transcript in DM1 foci (Laurent et al., 2012). The p68 helicase is involved in many cellular processes, including the transcription, processing, export, translation and degradation of RNA. Notably, it acts as a modifier of MBNL1 activity in regulating the alternative splicing of exon 5 of troponin T (TNNT2) transcripts (Laurent et al., 2012). The p68 is postulated to modify hairpin structures formed by CUG repeats as it does on the stem-loop structure of TNNT2 RNA to increase its accessibility to MBNL1. It is therefore important to continue to search for other proteins likely to be present within the nuclear foci because the loss of these proteins may also contribute to RNA toxicity.

**Missplicing caused by expanded CAG and CGG repeats**

In recent years a number of *in vitro* and *in vivo* studies conducted in different model systems have demonstrated that some hallmarks of RNA toxicity in DM1 could also be triggered by expanded CAG repeats. The over-expression of CAG and CUG repeats of four different lengths – five, 30, 74 and 200 repeats – in HeLa and neuroblastoma SK-N-MC cells resulted in nuclear foci formation only in cells expressing 74 and 200 CUG or CAG repeats (Mykowska et al., 2011). The appearance of nuclear foci correlated with deregulation of the alternative splicing of several endogenous MBNL1-sensitive transcripts, e.g., *CLCN1* and *SERCA1*, which are misspliced in DM1 cells and expressed in HeLa and SK-N-MC cells. The strengths of the missplicing effects triggered by exogenous CAG or CUG repeats were similar. The splicing deregulation of several MBNL1-sensitive genes was also observed in the absence of RNA foci formation in HeLa cells transfected with short RNAs composed of CAG or CUG repeats. This finding suggests that a high concentration of short CAG or CUG repeat RNAs compensates for their reduced length and triggers the same missplicing events as longer mutant repeats. Earlier, the presence of RNA foci and MBNL1 colocalization with these foci was observed in COS-M6 cells overexpressing either 960 CAG repeats or 960 CUG repeats (Ho et al., 2005). The colocalization of endogenous mutant *HTT* and *ATXN3* transcripts with MBNL1 within nuclear foci was also observed in patient-derived HD and SCA3 fibroblast cells. However, the missplicing effects of endogenous *INSR* and *SERCA1* transcripts were only clearly observed in SCA3 and HD cells expressing the longer repeats of 69 CAG and 74 CAG, respectively (Mykowska et al., 2011). The colocalization of MBNL1 with the mutant transcript in the cell nucleus likely indicates the direct binding of the protein to repeat sequences, as earlier demonstrated for the mutant CUG repeat transcript (Miller et al., 2000) that was supported by the results of a filter binding assay. The affinity of recombinant MBNL1 for CAG repeat RNA was only slightly lower than for a CUG repeat of the same length. The similarity in MBNL1 binding was observed
for three repeat lengths: 54, 20 and seven CAGs and CUGs (Yuan et al., 2007; Mykowska et al., 2011).

The in vivo evidence for CAG repeat toxicity was first obtained in a Drosophila model of SCA3 (Li et al., 2008). The expression of untranslated CAG repeats of pathogenic length led to neurodegeneration in the absence of a mutant polyQ protein. The expression of translated CAA-interrupted CAG repeats resulted in a less severe phenotype than the expression of translated pure CAG repeats, which supported the importance of RNA structure for toxicity. The CAG repeat toxicity at the RNA level was also demonstrated in a worm system (Wang et al., 2011). Both CAG and CUG repeats of pathological length were shown to form nuclear foci, in which the mutant transcript colocalized with the nematode ortholog of MBNL1, CeML1. The disease phenotype was partially reversed by CeML1 over-expression. The expression of 200 untranslated CAG repeats was also shown to be deleterious in transgenic mice (Hsu et al., 2011). Very recently, the differences in the phenotypes of the YAC and BAC mouse models of HD expressing full-length human huntingtin were systematically analyzed, and the different content of CAA interruptions present in the HTT transcripts of these mice was considered responsible for the observed phenotypic differences (Pouladi et al., 2012).

The expanded CGG repeats also show the ability to trigger the misregulation of alternative splicing but differ from the CUG and CAG repeats in the pool of affected transcripts. The CGG repeats recruit a number of cellular proteins including Sam68, hnRNP-G and MBNL1 sequentially into nuclear foci (Sellier et al., 2010). The splicing factor Sam68 loses its regulatory function due to sequestration by the CGG repeats resulting in aberrant alternative splicing, as demonstrated for the ATP11B and SMN2 transcripts in FXTAS patients.

Although well established and commonly accepted, the mechanism of RNA toxicity that results in the aberrant alternative splicing of the MBNL-sensitive genes is unlikely to be the only pathomechanism triggered by the mutant CUG repeat RNA in DM1. Similar missplicing effects triggered by long CAG repeats – exogenous and endogenous, as well as translated and untranslated – are also unlikely to be the sole mechanism of RNA toxicity in the polyglutamine diseases. The same may apply to other TREDs in which the nuclear aggregation of mutant RNA results in the sequestration of important alternative splicing factors, and causes the downstream missplicing effects. The results of recent studies discussed in the following section and depicted in Figure 4 also support other mechanisms of RNA toxicity beyond spliceopathy.

Other mechanisms of RNA toxicity identified in TREDs and their models

Altered nuclear transport and export

The nuclear transport of pre-mRNA processing intermediates and export of mature mRNAs are among the steps of the mRNA lifecycle in which abnormal transcripts bearing lengthy repeat expansions may behave differently from their normal counterparts. For example, mutant DMPK transcripts accumulate in the cell nucleus within the ribonucleoprotein (RNP) foci and never reach the cytoplasm when the repeat expansion reaches 200 CUGs or more (Amack et al., 1999). Fluorescent in situ hybridization (FISH) experiments have demonstrated that the intracellular paths of normal and mutant transcripts split at the very early steps. The normal DMPK primary transcript is displaced from its transcription site to the closely located splicing speckle, enters the speckle and becomes spliced, undergoes further nuclear processing steps and is then exported to the cytoplasm. In contrast, the mutant transcript that contains the expanded CUG repeat does not enter the splicing speckle but is retained peripherally (Holt et al., 2007) and the signal from the mutant CUG repeat foci is very close to but does not overlap the signal from the splicing speckle marker, SC35 (Smith et al., 2007). Despite undergoing splicing, the mutant DMPK transcript ends its cellular journey at this early step and is trapped within the RNP foci. Protein hnRNP H has been shown to inhibit the export of the mutant transcript (Kim et al., 2005), and enhanced transcript export has been demonstrated for Staufen 1 (Ravel-Chapuis et al., 2012). The over-expression of Staufen 1 observed in DM1 patients can counteract the toxic effect of the expanded CUG repeats and play a protective role. The nuclear accumulation of transcripts containing expanded CAG repeats has also been reported (Tsoi et al., 2011). Experiments conducted using the Drosophila model system, transgenic mice and human cells have identified the protein responsible for the nuclear export of transcripts harboring expanded CAG repeats as the U2AF65 protein, which binds directly to the expanded CAG repeats and forms a complex with the nuclear export receptor NXF1. Although the initial data suggest that U2AF65 binding is specific to the CAG repeat, the issue of binding specificity must be clarified by further studies. The authors proposed that decreased levels of U2AF65 enhance mutant transcript retention in the nucleus and that the cell nucleus is the site where expanded CAG repeats exert their toxicity (Tsoi et al., 2011).
Expanded TNRs may also influence the nuclear retention of other mRNAs by different mechanisms. Two transport-related proteins, Pur-α and hnRNP A2/B1, are sequestered by and colocalize with expanded CGG repeats in the nuclear foci of FXTAS patients. Pur-α interacts directly with Rm62, the Drosophila ortholog of the p68 helicase, and purine-rich single strand RNA (Jin et al., 2007; Hokkanen et al., 2012). The CGG-induced phenotype can be rescued by Pur-α over-expression, but a specific role for this protein still has to be determined. A CGG repeat-dependent posttranscriptional decrease of Rm62 expression leads to the accumulation of Hsp70 transcripts in the nucleus (Qurashi et al., 2011). The absence of Hsp70 cytoplasmic localization interferes
with the cell’s response to stress through the inability to produce a sufficient amount of its protein product. Another RNA-binding protein that can interact directly with CGG repeats is hnRNP A2/B1, which in addition to splicing, plays a role as a trans-acting factor in neuronal RNA transport pathways (Sofola et al., 2007a; Muslimov et al., 2011). This protein has recently also been suggested to recruit heterochromatin protein 1 (HP1) to inhibit the expression of specific transposons (Tan et al., 2012). The hnRNP A2/B1 depletion by CGG sequestration results in insufficient recruitment of HP1 and increases the expression of transposons, promoting neurodegeneration.

**Altered microRNA and RNAi pathways**

The newly-discovered function of the MBNL1 protein as a cytoplasmic regulator of microRNA biogenesis implicates an alteration of the microRNA processing pathway in the RNA toxicity that occurs in DM1 (Rau et al., 2011). The study by Rau et al. demonstrated that the altered processing of the microRNA miR-1 is linked to heart defects in DM1 patients and also provided mechanistic explanation for this observation. In normal cells, MBNL1 wins the competition with known microRNA-processing regulator LIN28 for pre-miR-1 loop binding. As a result, the mature miR-1 normally produced in heart tissue down-regulates its target genes, including the calcium channel CACNA1C and gap-junction channel GJA1. In DM1 cells, the active pool of MBNL1 is strongly decreased, and thus LIN28 can bind to pre-miR-1 and enable its 3′ end uridylation, which inhibits precursor processing by Dicer. The decrease in miR-1 expression results in the up-regulation of its targets and contributes to cardiac dysfunction (Rau et al., 2011). As many other microRNA precursors contain sequence motifs recognized by MBNL1 in their hairpin loops, the search for further microRNAs deregulated by the same mechanism in myotonic dystrophy tissue may provide insight into the scale of microRNA deregulation in DM1. Another issue to be addressed is whether the nuclear step of microRNA processing could be affected by the sequestration of important RNA binding proteins by the expanded repeats. In addition to MBNL1, the multifunctional p68 and p72 RNA helicases were shown to colocalize, bind directly and be sequestered by expanded CUG and CAG repeats (Laurent et al., 2012). These helicases are well-known factors involved in the Drosha step of microRNA processing (Fukuda et al., 2007). In light of these findings, further research should address the possibility that both the nuclear and cytoplasmic processing steps of microRNA biogenesis contribute to RNA toxicity, not only in myotonic dystrophy but also in CAG and CGG repeat expansion diseases.

The involvement of the RNA interference (RNAi) pathway in the pathogenesis of TREDs has been proposed by several authors (Handa et al., 2003; Malinina, 2005; Krol et al., 2007). The RNAi mechanism implies that expanded repeats are cleaved by Dicer, enter the RNA-induced silencing complex and down-regulate transcripts containing complementary target sequences. Recently, the RNAi mechanism has been demonstrated to operate in several cellular and in vivo TREDs models. The bidirectional transcription that is frequent in human cells and also occurs through the triplet repeat regions of most TREDs genes has been modeled in *Drosophila* (Lawlor et al., 2011; Yu et al., 2011). The co-expression of expanded CAG and CUG repeats in fly cells from suitable vectors resulted in a neurodegenerative phenotype. The complementary repeats likely formed a double-stranded RNA cleaved by Dicer-2 into 21 nucleotides CAG/CUG repeat siRNAs active in silencing in an Argonaute-2-dependent manner. An earlier study using a *Drosophila* model of FXTAS showed that the co-expression of CGG- and CCG-containing transcripts resulted in an Ago2-mediated reversal of their independent toxicities (Sofola et al., 2007b). When the functional role of the naturally occurring HTT antisense transcript (HTTAS) was investigated, the regulation of the sense HTT transcript was demonstrated to be partly dependent on the Dicer and RNA-induced silencing complex pathways (Chung et al., 2011). More recently, the RNAi mechanism was demonstrated in human cell lines expressing mutant CAG repeats in the sequence context of HTT exon 1. Mutant CAG repeats, both translated and untranslated, gave rise to toxic small RNA (sCAG) in a Dicer-dependent manner and caused a downstream silencing effect in an Ago2-dependent manner (Banez-Coronel et al., 2012; commented on in Rudnicki et al., 2012). This observation is important from the perspective of RNA structure because the cytotoxic effects were only triggered by pure CAG repeats that formed hairpins, not by CAA-interrupted CAG repeats that did not form hairpins. From the functional point of view, the ability to detect the effects observed in cellular HD models in transgenic HD mice and human post-mortem HD brain tissue is important. In the context of that study, the composition of the active sCAG fraction and possible involvement of Drosha cleavage prior to Dicer cleavage in the process of the sCAG production remains to be clarified (Rudnicki et al., 2012). Taken together, the microRNA and RNAi pathways contribute to the RNA toxicity triggered by expanded CAG and CUG repeats in at least some TREDs, but the scale and importance of this contribution remain to be determined.
Global changes in transcriptomes

Several recent studies have used oligonucleotide microarrays to reveal all substantial changes in the cell transcriptomes caused by pathologically expanded CUG, CAG and CGG repeats in human tissue and animal models of selected TREDs. An important finding from one of these studies was the activation of genes responsive to double-stranded RNA and interferon-regulated genes by mutant CUG repeats in the cataracts of DM1 patients (Rhodes et al., 2012). The authors concluded that the aberrant activation of innate immune response genes, such as PKR, OAS, TLR3 and the Rig-I-like helicases, plays an important role in DM1. They also hypothesized that a similar activation of this cell defense system may occur in other TREDs in which mutant transcripts harbor long double-stranded RNA-like hairpins. Indeed, the cellular sensors of long double-stranded RNA and their protein responders (reviewed in Olejniczak et al., 2010) may participate in the induction of interferon response caused by RNAs containing expanded CNG repeats. Another study took advantage of mouse models of DM1 to demonstrate that in addition to the loss of Mbnl1 function and the consequent missplicing (to which only half of all observed changes could be attributed), the deregulation of numerous other mRNAs was triggered by the expression of mutant CUG repeat RNA (Du et al., 2010). These mRNAs include the messengers encoding extracellular matrix proteins, which bridge the molecular changes found in myotonic dystrophy with those observed in other muscular dystrophies.

Approximately 160 genes have been identified as being differentially expressed in Drosophila expressing 100 CAG repeats (Shieh and Bonini, 2011). One of these transcripts was the Hsp70 encoding chaperone protein, the up-regulation of which partially mitigated CAG repeat RNA toxicity. In Drosophila expressing the human disease-associated CAG or CUG repeats, other authors have observed the altered expression of components of the Akt/GSK3-β signaling pathway and suggested that the expression of long hairpin RNAs may also alter this pathway in the humans with the disease (van Eyk et al., 2011). The changes in mRNA and microRNA expression were analyzed in postmortem brain tissue of HD patients using microarrays and deep sequencing, respectively. Functionally distinct areas of the brain, especially the striatum and cortical regions, were shown to exhibit different profiles of altered gene expression, e.g., CNR1, CDRI, SLC14A1 and CD44, that parallel the known pattern of neurodegeneration (Hodges et al., 2006). The altered expression of many microRNAs, e.g., miR-100, miR-106b and miR-128 including changes in the levels and patterns of specific isomiRs, was also shown (Marti et al., 2010).

Taken together, the transcriptome-wide gene expression analyses provided long lists of genes with altered expression patterns in both human tissue and animal models of human disease. However, distinguishing the primary and secondary effects in the course of pathogenesis from such data and identifying the alterations triggered directly by the mutant transcript have been difficult.

Future perspectives

When considering any further studies that aim to identify the cellular processes and pathways altered by mutant transcripts in TREDs, one possible method is to more extensively exploit the modern tools of functional genomics and proteomics. To be more informative, such studies should include both human patient tissue and the appropriate animal models in which changes in gene expression could be analyzed in pre-symptomatic individuals and at all stages of development of the disease. In addition, carefully designed cellular models of the diseases expressing given repeat types of variable lengths at different levels in a broad range of cell types may shed new light on the process of pathogenesis triggered by mutant RNA repeats. Designing such models is more demanding in the case of polyglutamine diseases because the effect triggered by the mutant RNA must be separated from that induced by the mutant polyQ protein.

Another type of fishing expedition that should be considered is the isolation of all the proteins that bind to the mutant transcripts used as baits from cellular extracts or all of the proteins present in the nuclear inclusions formed by mutant transcripts. The alternative to fishing expeditions is more focused, hypothesis-driven research, which requires more conceptual work to generate attractive hypotheses and evaluate the likelihood of these hypotheses before designing validating experiments. A systematic approach to this issue would involve taking advantage of existing knowledge about the steps of the transcript lifecycle in human cells and making predictions as to which steps of the transcript’s cellular journey could be altered by the expanded CUG, CGG and CAG repeats that likely form long hairpin structures in cells. In Figure 4, the transcript nuclear processing steps are distinguished and separated from the cytoplasmic steps by the nuclear export event. This figure also includes previous observations concerning the cellular processes altered by toxic RNA repeats.
We anticipate that this cellular roadmap of the mutant transcripts will need to be presented at a much higher resolution several years from now, when many other alterations caused by the mutant repeats have been revealed.

Acknowledgements: This work was supported by the Ministry of Science and Higher Education (N N301 569340, N N401 572140, N N301 284837, N N302 633240) and the European Regional Development Fund within Innovative Economy Programme (POIG.01.03.01-30-098/08). The authors would like to thank Marzena Wojciechowska for her comments on the manuscript.

Received May 31, 2012; accepted July 6, 2012

References


Paulina Galka-Marciniak graduated in medical biotechnology in 2010 at the Poznan University of Medical Sciences, Poland. She is currently working on her PhD investigating the genetic approaches to RNAi-based experimental therapy of Triplet Repeat Expansion Diseases, the nonspecific immunological side effects triggered by RNAi and microRNA reagents and the mechanisms of microRNA biogenesis.

Martyna Urbanek is a MSc student in medical biotechnology at Poznan University of Medical Science and BCS student in algorithmics and programming engineering at Adam Mickiewicz University. Her research interests span nonspecific immunological effects of RNA interference reagent, RNA visualization in live cells and broadly defined bioinformatics including microarray analysis.
Professor Włodzimierz J. Krzyżosiak is Head of Laboratory of Molecular Biomedicine at the Institute of Bioorganic Chemistry, Polish Academy of Sciences in Poznan, Poland. He received his PhD degree from Adam Mickiewicz University in Poznan, and held fellowships at prestigious research centers in the UK, France, USA and Japan. Professor Krzyżosiak has a long-lasting interest in the biochemistry of RNA, and his research achievements include developing a method of RNA structure probing by metal ion-induced cleavages. He also has a long research record in studying the role of RNA repeats in molecular pathology of triplet repeat expansion diseases.