CAG Repeats Containing CAA Interruptions Form Branched Hairpin Structures in Spinocerebellar Ataxia Type 2 Transcripts*

Received for publication, August 31, 2004, and in revised form, November 4, 2004
Published, JBC Papers in Press, November 8, 2004, DOI 10.1074/jbc.M409984200

Krzysztof Sobczak and Włodzimierz J. Krzyzosiak‡
From the Laboratory of Cancer Genetics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Nosalowskiego 12/14, 61-704 Poznan, Poland

Spinocerebellar ataxia type 2 (SCA2), one of the hereditary human neurodegenerative disorders, is caused by the expansion of the CAG tandem repeats in the translated sequence of the SCA2 gene. In a normal population the CAG repeat is polymorphic not only in length but also in the number and localization of its CAA interruptions. The aim of this study was to determine the structure of the repeat region in the normal and mutant SCA2 transcripts and to reveal the structural basis of its normal function and dysfunction. We show here that the properties of the CAA interruptions are major determinants of the CAG repeat folding in the normal SCA2 transcripts. We also show that the uninterrupted repeats in mutant transcripts form slippery hairpins, whose length is further reduced by the base pairing of the repeat portion with a specific flanking sequence. The structural organization of the repeat interruption systems present in other human transcripts, such as SCA1, TBP, FOXP2, and MAML2, are also discussed.

Numerous spinocerebellar ataxias (SCAs)1 make up the largest contribution to the variety of human hereditary neurological diseases that are caused by the expansions of short tandem repeats in single genes (1–3). More than 20 different autosomal dominant SCAs are clinically recognized. For 12 of them the respective genes and implicated mutations are known. In six SCA subtypes, SCA1, SCA2, SCA3, SCA6, SCA7, and SCA17, the underlying mutations are the expansions of trinucleotide CAG repeats present in the translated sequences of the functionally unrelated genes. The mutated CAG repeats give rise to expanded polyglutamine tracts in the encoded proteins, and the corresponding SCAs together with Huntington’s disease, dentatorubral-pallidoluysian atrophy, and spinal and bulbar muscular atrophy are known as polyglutamine expansion diseases.

SCA2 disease was also raised recently (10–16). The considered mechanism of transcript involvement in these diseases was similar to that demonstrated for the two types of myotonic dystrophy, DM1 and DM2 (11, 17, 18), and fragile X-associated tremor ataxia syndrome (14, 15, 19). In brief the expanded repeats in transcripts form stable hairpin structures that impair normal RNA metabolism and sequester some important cellular factors and alter the functions of other transcripts (11, 13, 16–18, 20).

The SCA2 gene (21–24) encodes protein ataxin-2 of an unknown function (Online Mendelian Inheritance in Man number 601517). Its normal alleles range in length from 14 to 32 CAG repeats, and the expanded mutant alleles range from 35 to 50 repeats (22–24). The intermediate length alleles belong to the so-called gray zone because they do not give rise to the disease in all carriers. The CAG repeat polymorphism in the SCA2 gene is not confined to its length only. Normal alleles usually contain one to three CAA interruptions (22–26). In this respect the SCA2 gene resembles SCA1, in which the repeats contain the CAA interruptions (25, 27, 28), and FMR1 gene involved in the fragile X syndrome and fragile X-associated tremor ataxia syndrome, which contains AGG interruptions within the CAG repeats (29).

The SCA2 alleles that harbor a single interruption are very common in various populations (25, 26). The CAA triplet is located asymmetrically within the repeated sequence. Typically it is positioned closer to the 3’-end of the CAG repeat and is followed by eight 3’ repeats. Because these SCA2 alleles range in length from 14 to 30 repeats, the pure CAG repeat tracts are most variable in their 5’-part (6–20 repeats). The alleles containing two interruptions occur most frequently in the 8-1-4-1-8 configuration (the interruptions are underlined). Their variants differ by one CAG triplet in one of the three pure repeat tracts (25). Only two SCA2 variants are known to contain three interruptions. They are composed of either 27 or 29 repeats in the 8-1-4-1-8 and 8-1-4-1-4-1-10 configurations, respectively. In all of the populations studied so far, the most prevalent are two SCA2 alleles containing 22 repeats in the 8-1-4-1-8 and 13-1-8 configurations. They account for about 90% of chromosomes. Uninterrupted normal SCA2 alleles are very rare, whereas the expanded mutant alleles usually lack interruptions (22–24, 26).

In this study we have determined the structures formed by the repeat regions of various normal and mutant SCA2 alleles in transcripts. We have asked intriguing questions related to the possible physiological functions of the SCA2 repeats in RNA and the possible involvement of the mutant transcript structures in SCA2 pathogenesis. More specifically, the questions we have asked were: What is the precise structural role of the CAA interruptions within the CAG repeats? What are the similarities and differences between the repeat interruption systems operating in the SCA1 (16) and SCA2 transcripts? To gain a still wider insight into the role of the repeat interruptions in RNA, we have extended these comparisons to tran-

* This work was supported by State Committee for Scientific Research Grants 2P06A-08828 and PBZ-KBN-040/P04/2001 and by Foundation for Polish Science Grant 8/2000. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
‡ To whom correspondence should be addressed. Tel.: 48-61-8528503 (ext. 136); Fax: 48-61-8520532; E-mail: wldokrzy@ibch.poznan.pl.
1 The abbreviations used are: SCA, spinocerebellar ataxia; nt, nucleotide(s); ODN, oligodeoxynucleotide.
scripts from several other human genes containing interrupted CAG repeats. We have also compared the structures of the repeat regions in various normal and mutant SCA2 transcripts and discussed the structural features of the mutant transcripts that could make them involved in RNA-mediated pathogenesis.

EXPERIMENTAL PROCEDURES

Preparation of DNA Templates and in Vitro Transcription—DNA templates for in vitro transcription were synthesized by PCR from gel-purified SCA2 amplification products (primer sequences SCA2-F, 5'-GGCTTGCGGACATTGGCAGC; GenBankTM NM_002973 positions from 564 to 755) and SCA2-R, 5'-GGCTTGCGGA-CATTGGCAGC; GenBankTM NM_002973 positions from 564 to 755) containing different sequence variants of the CAG repeat tracts as described earlier (25). In amplification reactions the forward primer (SCA2-FTT) containing a T7 RNA polymerase promoter (underlined) 5'-TAATACGACTCACTATAGGGCCCCTCACCATGT and reverse primer (SCA2-R) 5'-GGCTTGCGGACATTGGCAGC were used. The PCR was performed in a 50-μl reaction containing: a DNA template (gel-purified PCR product), 1 μM of each primer SCA2-FTT and SCA2-R, 10 mM Tris-HCl (pH 9.0), 1.5 mM MgCl2, 50 mM KCl, 0.1% Triton X-100, 200 μM of each of the dNTPs, and 1.5 units of Taq DNA polymerase (Promega) in the following conditions: 94°C for 1 min and then 35 cycles at 94°C for 1 s, 60°C for 1 s, and 72°C for 30 s.

A transcription reaction carried out in a 50-μl volume contained about 1 pmol of DNA template (10–20 μl of the PCR product described above), 400 units of T7 RNA polymerase (Epicenter), 3 μM guanosine, 0.5 μM each of ribo-NTPs, 10 mM dithiothreitol, 40 μM Tris-HCl (pH 7.9), 6 μM MgCl2, 2 μM spermidine, 10 mM NaCl, and 50 units of RNasin. Incubation was at 37°C for 1 h. The transcription products were purified in denaturing 6% polyacrylamide gels, visualized by Stains-all (Serva) staining, extracted from the gel slices by diffusion for 2 h at room temperature in 0.3 M NaOAc, 0.1% SDS, and 0.5 mM EDTA, and recovered by ethanol precipitation. 5'-End labeling was performed with 10 units of T4 polynucleotide kinase (Epicenter) and 50 μCi of [γ-32P]ATP (4500 Ci/mmol; ICN) at 37°C for 5 min, and RNA was repurified by electrophoresis on a 10% polyacrylamide gel in the presence of 7.5 mM urea and extracted as described above.

Nuclease Digestions and Lead Cleavages—Prior to structure probing reactions, the 32P-labeled RNAs were subjected to a denaturation and renaturation procedure in a solution containing 10 mM Tris-HCl (pH 7.2), 80 mM NaCl, 2 mM MgCl2, by heating the sample at 80°C for 1 min and then slowly cooling to 37°C. Limited RNA digestion was initiated by mixing 5 μl of the RNA sample (50,000 cpmp) with 5 μl of a probe solution containing either leads or nuclease S1 or ribonucleases T1, T2 in three different concentrations: Phosphate 0.25, 0.5, 0.7; 0.1, 0.2, 0.4 μl. RNase H digestion was performed by adding to the transcript solution, after denaturation and renaturation procedure (as above), one of the following three types of oligodeoxynucleotides: libraries of random 6-mers (100 pmol) or random 7-mers (100 pmol) or complementary to CAG repeats oligonucleotide (ctg-ODN) with sequence 5’-TGCTGCTTGCTGGA-3’ (5 pmol), and RNase H to the final concentration of 0.10 or 0.25 unit/μl. All of the reactions were performed at 37°C for 20 min and stopped by adding an equal volume of stop solution (7.5 μm urea and 20 μM EDTA with dyes) and sample freezing.

Analysis of the Reaction Products—To determine the cleavage sites, the products of lead-induced hydrolysis and nuclease digestion were separated in 10–15% polyacrylamide gels (acylamide/bisacrylamide, 37°C) containing 10% polyacrylamide and S1 ladder (not shown). The RNA fragments generated by nucleases S1 and H terminate with 3’-hydroxyls and migrate slower than the corresponding formamide and T1 fragments. Electrophoresis was performed at 1500 V and was followed by autoradiography at −80°C with an intensifying screen. The products of the structure-probing reactions were also visualized and analyzed by phosphorus imaging (Typhoon; Molecular Dynamics).

Electrophoresis in Nondenaturing Conditions—RNA structure homogeneity was analyzed for each transcript by the electrophoresis of radiolabeled samples in 12% nondenaturing polyacrylamide gel (dimension, 150/1401 mm; acrylamide/bisacrylamide, 29%) buffered with 45 mM Tris borate at a fixed temperature of 37°C. Prior to gel electrophoresis, the 32P-labeled transcripts were subjected to a denaturation and renaturation procedure in a solution containing 10 mM Tris-HCl (pH 7.2), 40 mM NaCl, and 1 mM MgCl2 by heating the sample at 80°C for 1 min and slowly cooling it to 37°C and mixed with an equal volume of 7% sucrose with dyes. Electrophoresis was conducted at a constant power of 15 W.

RNA Secondary Structure Modeling—RNA secondary structure prediction was performed using the mfold program version 3.1 (www.bioinfo.rpi.edu/applications/mfold) (30). This program is designed to determine the optimal and suboptimal secondary structures of RNA by free energy minimization using the nearest neighbor energy parameters, calculated for 1 s NaCl solution (pH 7.0) at 37°C.

RESULTS

Analysis of the Transcript Unique Sequence Configuration—Fourteen different SCA2 transcripts were analyzed in this study (Table I). They included 12 of the 24 normal variants of the SCA2 triplet repeat known to occur in a general population (25). These 12 variants represented all different types of the repeated sequence configuration found in the SCA2 gene. In all transcripts the repeat regions were surrounded by the same specific flanking sequences: 27 nt at the 5’-end and 32 nt at the 3’-end of the repeats. Seven transcripts contained a single CAA interruption

**Table I**

| Number of repeats | Sequence configuration | Transcript name |
|-------------------|-----------------------|-----------------|
| One CAA interruption | (CAG)14CAA(CAG)14 | 8-1-5 |
| 19                 | (CAG)14CAA(CAG)14 | 10-1-5 |
| 21                 | (CAG)14CAA(CAG)14 | 12-1-8 |
| 22                 | (CAG)14CAA(CAG)14 | 12-1-8 |
| 23                 | (CAG)14CAA(CAG)14 | 14-1-8 |
| 24                 | (CAG)14CAA(CAG)14 | 15-1-8 |
| Two CAA interruptions | (CAG)19CAA(CAG)19 | 8-1-4-1-8 |
| 25                 | (CAG)19CAA(CAG)19 | 8-1-4-1-9 |
| Three CAA interruptions | (CAG)23CAA(CAG)23 | 9-1-4-1-8 |
| 27                 | (CAG)23CAA(CAG)23 | 8-1-4-1-8 |
| 29                 | (CAG)23CAA(CAG)23 | 8-1-4-1-10 |
| No interruptions   | (CAG)36             | pure(CAG)36 |
| 37                 | (CAG)37             | pure(CAG)37 |

a The CAA interruptions are underlined.

b Mutant variants derived from SCA2 patients.
at various locations within the repeat tract, and the repeats of three other transcripts harbored two interruptions separated by four pure CAG repeats. In the remaining two normal transcripts containing one, two, and three CAA interruptions in the following repeat configurations: 15-1-8, 8-1-4-1-8, and 8-1-4-1-4-1-10 (interruptions are underlined in the figure legend and in addition shown in red in the figure panels). The transcripts were treated with: lead ions (panel Pb), RNase T1 (panel T1), RNase T2 (panel T2), and S1 nuclease (panel S1) at three different probe concentrations (increasing from left to right) specified under “Experimental Procedures.” Lane Ci, incubation control (no probe); lane F, formamide ladder; lane T, guanine-specific ladder obtained with T1 ribonuclease. The positions of nucleotides in the sequences flanking the repeats are italicized. Positions of selected G residues (G1, G2, etc.) of the CAG repeats (roman capital letters) and positions of CAA interruptions (red A letters and red arrowheads) are also shown. Blue arrowheads in the electropherograms show the positions of cleavages, which are similar in all of the analyzed transcripts and discussed in this paper. The S1 cuts in the repeat region, which are different in different transcripts, are indicated by vertical green lines. Fragments of autoradiograms corresponding to sequences forming the structure modules M1, M2, and M3 are also indicated. A, cleavages generated in three 5′-end-labeled transcripts containing one, two, and three CAA interruptions (red A letters and red arrowheads) are also shown. Blue arrowheads in the electropherograms show the positions of cleavages, which are similar in all of the analyzed transcripts and discussed in this paper. The S1 cuts in the repeat region, which are different in different transcripts, are indicated by vertical green lines. Fragments of autoradiograms corresponding to sequences forming the structure modules M1, M2, and M3 are also indicated. B, cleavages generated in the 3′-part of the 14-1-8 transcript, forming the M3 module, induced by lead ions, S1 nuclease, and T1 ribonuclease (other designations are as in A). The observed pattern of cleavages in this region is very similar in all of the analyzed transcripts. C, secondary structure model of the repeat region of SCA2 transcripts with invariable structures of the repeat flanking sequences shown in details. The cleavage sites, which are similar in all of the analyzed transcripts, are indicated for each probe used (see figure inset for probe designations and cleavage intensity classification). The RNA structure modules (M1, M2, and M3) are marked. The CAG repeat sequence is shown in red and spanned by arrows. The variable structure of the M2 module is shown schematically.
(37–39). All of the RNA structure probing reactions were carried out under very similar solution conditions, and important conclusions were based on comparisons of the results obtained with different transcripts and structure probes.

Overall Structures of All Analyzed Transcripts Show Several Similarities—Examples of structure probing results obtained for representative transcripts are shown in Fig. 1. It can be noticed from the inspection of the complete set of data that the same cleavage patterns are generated by the same probes in the repeat flanking sequences of all transcripts. For example, in those transcripts with one, two, and three interruptions (Fig. 1A), two regions of the repeat flanking sequences are highly susceptible to the S1, T1, T2, and lead cleavages. These regions are marked with blue arrowheads in Fig. 1A. The first susceptible region is located upstream the triplet repeat between nucleotides C12 and U16, and the second is located downstream of the repeat and includes the pentanucleotide CpGpCpCpCp, e.g. C101 to C105 in the 14-1-8 transcript (Fig. 1B). At both sides of these highly reactive regions, the RNA fragments highly resistant to cleavages are present, which suggests their involvement in the formation of two autonomous hairpin structures. The two highly reactive pentanucleotide regions men-
SCA2 CAG repeats containing one CAA interruption form different branched hairpin structures; variable module M2. A, patterns of nuclease S1 cleavages in three SCA2 transcripts 10-1-8, 12-1-8, and 15-1-8 containing one CAA interruption (marked by a red arrowhead) variably located in the repeat region. Vertical green lines show differences in cleavage patterns between transcripts. Transcripts are ordered according to the increasing total number of repeats. Other designations are as described in the Fig. 1A legend. B, nuclease S1 cleavage patterns generated in three similar SCA2 transcripts containing the following repeat configurations: 13-1-8, 13-1-9, and 14-1-8 (other designations as in Fig. 1A). C, proposed secondary structure of the SCA2 transcripts containing one CAA interruption and either an odd (14-1-8) or an even (13-1-8) total number of repeats. Positions of selected G residues of the CAG repeats are marked, and the localization of the A residues of the CAA interruption is shown by red ovals. For both transcripts two alternatively folded structural modules M2 are presented (conformer I and conformer II). Each is composed of two hairpins either M2a and M2b or M2a’ and M2b’. The designations of internal loops (a and a’ and hairpin terminal loops (b and c or b’ and c’) are also shown. Invariant structure modules M1 and M3 are not presented except for the uppermost structure. Other designations are as in the legend to Fig. 1C. D, left panel, patterns of ribonuclease T1 cleavages in two transcripts (CAG)17cl and (CAG)16cl containing 17 and 16 CAG repeats, respectively.
tioned above form terminal loops in the hairpins present in structure modules M1 and M3 (Fig. 1C).

Another common feature of all analyzed transcripts is the enhanced susceptibility to cleavages of both the 5’ and 3’ termini of the repeated sequence (also marked by blue arrowheads in Fig. 1A). The two 5’ CAG repeats show similar reactivity with all the three nucleases used. The ribonuclease T2 cleaves the C3pA5 and A3pG2 bonds (i.e. phosphodiester bonds present within the first CAG repeat), nuclease S1 cuts A5pG3, and ribonuclease T1 cleaves G1pC2. The next several repeats are resistant to cleavages by these nucleases (Fig. 1A). With regard to the 3’-terminal nucleotides of the repeated sequence, the last five CAG repeats show a great similarity in their cleavage patterns in all transcripts. This similarity is illustrated in more detail for the 14-1-8 transcript in Fig. 1B. The last four inter-nucleotide bonds of this sequence located between the G residues of CAG repeats 22 and 23 are strongly resistant to nucleases and lead-induced cleavages. This effect may be explained by the involvement of these nucleotides in the double helical structure of the M3 module (Fig. 1C). On the other hand, the strongly enhanced susceptibility to lead cleavages of the neighboring 11 nucleotides shown in Fig. 1A can be explained by their presence in between the M2 and M3 modules and in the irregular part of the module M3 (Fig. 1C). To sum up, our results suggest that as many as 10 nucleotides from the four 3’-terminal CAG repeats base pair with the 3’-flanking sequence (Fig. 1C), and the remaining repeats form the M2 module that has a variable structure depending on the number and localization of the repeat interruptions.

To make sure that the molecular architecture of the M3 module was properly assigned and that the specific repeat flanking sequences indeed do not base pair with each other, we have deconstructed this module in a way shown in Fig. 2C, so that most of the 3’-flanking sequence was excluded from the interaction postulated for the M3 in Fig. 1C. This was achieved by the hybridization of the 20nt-ODN oligodeoxynucleotide complementary to the terminal 20 nt of the 3’-flanking sequence (Fig. 2C). The highly efficient ODN binding as well as the structural homogeneity of the analyzed transcripts and their hybrids with the ODN was confirmed by the results of polyacrylamide gel electrophoresis carried out under nondenaturing conditions (Fig. 2A). The results of structure probing experiments performed before and after the 20nt-ODN hybridization are shown for the 12-1-8 and 15-1-8 transcripts in Fig. 2B. The cleavages generated in the transcript structure, which was rearranged by the 20nt-ODN, occurred as expected at different sites and pointed to the structure shown in Fig. 2C. In this structure all CAG repeats are involved in the formation of the rearranged M2 module composed of the repeated sequence and downstream pseudorepeats 5’-CGCCCGCCGCG-3’. Importantly, there are no structural differences in the 5’-flanking region (M1 module) between the 20nt-ODN hybridized and nonhybridized transcripts (the same reactive regions are shown by blue arrowheads in Fig. 2B). This indicates that the 5’- and 3’-flanking sequences indeed do not interact with each other in all of the transcripts analyzed in this study (Table 1).

**Interrupted CAG Repeats Form Branched RNA Structures**—Unlike the specific flanking sequences presented above, the repeats themselves form diverse branched structures (M2 module) in the analyzed transcripts. All seven transcripts harboring the single interruption are cleaved in two regions. One of these regions is located in the central portion of the 5’-pure CAG repeat tract, and the other is at the site of the interruption (Fig. 3, A and B). This cleavage pattern points to similar secondary structures of the repeat regions, which contain three-way junctions (loop a) from which two hairpins protrude (Fig. 3C). These form the M2b hairpin, which always contains a small 4 nt-terminal loop c harboring the interruption and M2a hairpin that has a more variable structure. Both the length of the M2a stem and its loop b size are different in different transcripts. The size of the terminal loop b depends on the total number of repeats in a transcript, consisting of 7 nt in transcripts with an odd number of repeats and 4 nt in those with an even number (upper and lower structures shown in Fig. 3C). This dependence is clearly demonstrated by the S1 and T1 nuclease cleavage patterns presented in Fig. 3 (A, B, and D). Because the length of the pure 5’-CAG repeat tract increases from 10 repeats in the 10-1-8 to 15 repeats in 15-1-8, the regions of enhanced S1 reactivity include repeats 4 and 5 and repeats 7 and 8, respectively (Fig. 3A). However, the structure of the M2 module is also dependent on the length of the 3’-pure CAG repeat tract. For example, in transcripts 13-1-8 (22 repeats) and 13-1-9 (23 repeats) that have the same length of the 5’-pure repeat tract, the nucleases S1 cleavage patterns are different. The regions of differences are marked with a vertical green line in Fig. 3B. On the other hand, the transcripts 13-1-9 and 14-1-8, which have the same total number of repeats, have similar nucleases S1 cleavage patterns. Furthermore, the number of cuts generated by nucleases in the central portion of the 5’-pure CAG repeat tract is higher than that which is needed to explain the terminal loop sizes. It is therefore likely that in each of the transcripts containing either odd or even numbers of repeats, two slightly different structures coexist (conformers I and II in Fig. 3C). The structures within each pair of conformers differ in the localization of the junction (loop a or a’) and terminal loop b or b’ motifs. For example in conformer I of the 14-1-8 transcript, terminal loop b is formed by repeats 6 and 7, whereas loop b’ is formed by repeats 7 and 8 in conformer II (Fig. 3C). Based on the intensities of the cleavages generated by the nucleases S1 (Fig. 3B) and T1 (Fig. 3D), it may be concluded that conformer I, which contains the hairpins M2a and M2b, predominates under the conditions of structure probing.

In transcripts containing three interruptions in their repeat tracts all three of these sites show enhanced susceptibility to cleavages (indicated by the vertical green lines in Fig. 1A). As many as four reactive regions exist within the repeated sequence in the transcript containing two interruptions in the 8-1-4-1-8 configuration. Two of these regions include interruptions, and the other two are in the pure CAG repeat tracts (Fig. 1A). The observed cleavage patterns are consistent with the more relaxed structures of the M2 module in transcripts containing two and three interruptions as compared with transcripts harboring one interruption (Fig. 4). The M2 modules in the former two classes of transcripts contain two junctions.

(described in Ref. 13) containing 17 and 16 pure CAG repeats surrounded at both sides with additional complementary sequences, which are marked as G-C clamp in the presented structure models. These transcripts form fixed secondary structures containing either a 7-nt or a 4-nt terminal loop (13). The ranges of cleavages generated by nucleases in these terminal loops are marked with a blue bracket in the electropherograms. The cleavage sites and intensities are shown for three enzymatic probes (as in C) in the terminal loops of the established hairpin structures. Right panel, RNase T1 cleavages generated in the loop b (b’) regions of two SCA2 transcripts 12-1-8 and 13-1-8 representing transcripts with an odd and even total number of repeats. The ribonuclease T1 cuts specific for the alternative terminal loops b and b’ are shown by green and blue lines, respectively. In the 12-1-8 transcript the 7-nt loop b comprises both G6 and G7 residues, and loop b’ harbors G5 and G7 residues. In 13-1-8 transcript the smaller 4-nt terminal loops b and b’ contain only one G residue, G6 or G7, in conformer I and conformer II, respectively.
(loops a and b) and three terminal loops (loops c–e), which are composed of the 4-nt sequence (either the 5'-AGCA or 5'-AACA) (Fig. 4).

The structural feature that all normal SCA2 transcripts have in common is the canonical W-C base pairing between the GpC sequence of the first two repeats and the complementary GpC sequence from the fourth and fifth repeat counting from their 3'-end, shown by an orange background in the structures presented in Fig. 4. Our experimental data suggest that this region is involved in coaxial stacking between the G-C base pairs located at the bases of both the M2 and M3 modules. The main evidence for this stacking interaction is the observed strong resistance of the sequence located between the M2 and M3 modules to the single strand specific nucleases and the strong resistance of the sequence located between the M2 and M3 modules. The region is involved in coaxial stacking between the G-C base pairs located at the bases of both the M2 and M3 modules. The main evidence for this stacking interaction is the observed strong resistance of the sequence located between the M2 and M3 modules to the single strand specific nucleases and the inability of this sequence to the short complementary ODN hybridization. The occurrence of the less energetically favorable 7-nt terminal loop b in the M2a hairpin of transcripts containing an odd number of repeats and a single interruption (Fig. 3C) is also in support of the postulated coaxial stacking.

Structure of CAG Repeat Region in Mutant Transcripts Is Different in Several Respects—The structures of two mutant transcripts, the pure(CAG)36 and pure(CAG)37, differ from those of all investigated normal transcripts in their M2 module only. In the approximately central portion of the CAG repeat tract of the pure(CAG)36 between the residues G13 and A17 (the guanyl residues of repeats 13 and 17), a number of cleavages are generated by all enzymatic probes used (Fig. 5, A–C). The strongest cuts occur between positions A16 and A17 and correspond exactly to the 4-nt terminal loop a of the hairpin structure formed by the 32 CAG repeats (Fig. 5D). This result also confirms the involvement of the 3'-terminal 10 nucleotides of the CAG repeat in the formation of the autonomous module M3. The remaining weaker cuts (G13–A16) can be assigned to the terminal loops of the alternatively aligned "slipped" hairpin structures formed by the decreased number of repeats: 29, 27, and 25, which have the terminal loops a', a", and harbor the residues G15, G14, and G13 (Fig. 5, C and D). The quantitative analysis of the decreasing intensities of the S1 cuts (Fig. 5B) as well as T1 and T2 cuts (Fig. 5C) allowed us to determine the shares of the alternative M2, M2', and M2" structures to be about 50, 30, and 20%, respectively. In Fig. 5D the structures of the M2 and M2' modules only are shown. Other evidence for the "slippery" nature of the M2 module structure comes from the analysis of cuts generated in the region separating the M2 and M3 modules (Fig. 5, B and C). In the case of the pure(CAG)36 variant containing the M2' module, this region is formed by as many as four CAG repeats (from A30 to A33 in Fig. 5D). In contrast, in the normal transcripts that contain interrupted repeats, no nuclease cuts were observed in this region. This difference is clearly demonstrated also by the RNase H cleavage patterns defined by the short ODN hybridization sites. Both probes of this type the random hexamers and cgt-ODN effectively hybridize to the spacer sequence between the M2 and M3 modules of the pure(CAG)36 and induce efficient RNase H cuts between A30 and A33 (Fig. 5A). It should be recalled here that in all analyzed normal transcripts that contain the CAA interruptions, the region between the modules M2 and M3 is completely inaccessible to the short ODN probes as shown for the 15-1-8 transcript in Fig. 5A (green vertical line). On the other hand, other sites are similarly accessible to these ODN probes in both normal and mutant SCA2 transcripts.

DISCUSSION
The RNA-mediated pathogenesis postulated for the increasing number of human diseases caused by the expansions of short tandem repeats in their responsible genes is one of the major recent developments in the field of dynamic mutations (40). The molecular basis of RNA toxicity is, however, poorly understood, partly because the related problems regarding the normal biological roles of various simple sequence repeats in transcripts remain unsolved (10, 12). In this study we have performed an in depth structure analysis of the CAG repeat region of the SCA2 transcript. This tran-
interrupted CAG repeats in SCA2 transcripts form several slipped hairpin structures. A, nuclease H cleavage products generated after the preincubation of the normal transcript 15-1-8 and mutant transcript pure(CAG)36 with short oligodeoxynucleotides, ctg-ODN. Blue arrowheads indicate regions of similar reactivity in both transcripts, whereas green vertical lines show regions of different reactivity. Other designations as described in the legend to Fig. 1A. B, cleavages induced by S1 nuclease in the mutant transcript pure(CAG)36 at two different probe concentrations (1 and 2 units/µl). Prior to nuclease S1 digestion this transcript was subjected to a denaturation/renaturation procedure in the structure probing buffer in the presence (+20nt-ODN) or absence (−20nt-ODN) of 20nt-ODN as described in the legend to Fig. 2A. The results of the experiment with 20-nt ODN-hybridizing mutant transcript confirm the structural conclusions drawn from similar experiments conducted with normal transcripts (see Fig. 2). C, phosphorus imaging peaks representing cleavage patterns in the repeat region (from G1 to G36) of the pure(CAG)36 transcript. The cleavage sites and intensities are shown for two probes: RNase T1 (dark blue) and RNase T2 (blue). The black line represents the guanosine specific T1 ladder. The peaks corresponding to the cuts in the terminal loops a, a', and a'' of alternative hairpin structures are indicated. D, secondary structure variants of the SCA2 mutant transcript pure(CAG)36. Cleavage sites in the alternative M2 module structures, the M2 containing terminal loop a, and M2' containing terminal loop a', are indicated for each probe used (see inset for probe designations). The putative binding sites of the ctg-ODN (orange lines) are also shown. Other designations are as in the legends to Figs. 1C and 3C. Invariant structure modules M1 and M3 are shown together with the structure of the M2 module only.

The approach we have used for the structure analysis of the repeat region of the SCA2 mRNA consisted of several steps. First, we applied the mfold program to predict the secondary structure of the entire mRNA and the structure of its repeat region composed of the repeats themselves and
specific flanking sequences of different lengths. By comparing these structures we selected the appropriate RNA module whose structure was expected to be preserved in the sequence context of the entire mRNA. This SCA2 mRNA fragment was subjected to detailed experimental structure analysis using chemical and enzymatic probes. We also assumed that the structure of the analyzed fragment well corresponds to that present within the endogenous SCA2 mRNA in its natural cellular environment. Both of these assumptions were strongly supported by our recent experimental results obtained for similar RNA structures containing the CAG repeats (16, 41). The experimentally determined structures of the SCA2 transcripts were then analyzed again with mfold to calculate their thermodynamic stabilities. This combination of RNA structure prediction and experimental structure probing proved highly informative also in studies of other RNAs performed in our laboratory (12, 13, 16, 20, 34, 41).

The analysis of numerous normal and mutant variants of the SCA2 transcript allowed us to establish the features of these structures, define the rules of their folding, and compare their

![Diagram](image-url)
folding patterns with those of other triplet repeat expansion disease-related transcripts. The interplay between the repeated and specific flanking sequence is an important element of this folding characteristics. It may also influence significantly the correlation between the repeat length, the stability of the hairpin structure it forms, and the ability of this structure to trigger RNA pathogenesis in cells. The repeat flanks may form base pairs with each other and stabilize the repeat structure considerably as was shown for the CAG repeats in the SCA1 (16) and SCA6 (41) transcripts (Fig. 6A). Alternatively, the flanking sequences may not interact with each other, allowing the repeats to form several slipped hairpins as demonstrated for the DMPK (20) and dentatorubral-pallidoluysian atrophy (41) transcripts. In the SCA2 transcript the nature of this interaction is different. The flanking sequences do not interact with each other, but one of them, the 3‘-flanking sequence, interacts with 3‘-terminal repeats. This kind of the interaction does not force the pure CAG repeats in the

**FIG. 7.** The predicted structures of the interrupted repeat regions of other human transcripts. Structure predictions were performed for the repeated sequences of three transcripts: FOXP2, MAML2, and TBP. The repeat configuration formulas defining the number of CAG and CAA triplets follow the system used for the SCA2 transcripts, and polymorphic CAG repeat tracts are shown in brackets. Red arrowheads indicate the positions of the G → A substitutions. For the TBP transcript three different repeat configurations were analyzed: 38Q represents the normal allele, and the remaining two configurations, 48Q and 55Q, represent mutant alleles found in SCA17 patients.
mutant transcript to a single alignment, but instead several slippery hairpin structures are observed (Fig. 6A). Also the way in which the CAG repeat structure fixation is achieved in the SCA2 transcripts containing the interrupted repeats is different from that described for the SCA1 and SCA6 transcripts (16, 41). It is most likely the coaxial stacking interaction between the M2 and M3 modules that is responsible for this effect. According to thermodynamic data the two G-C/G-C pairs involved in such stacking would have a significant destabilizing effect (3.8 kcal/mol contribution to the overall ΔG of the system).

An interesting picture also emerges from the analysis of structural effects caused by the CAA interruptions within the CAG repeats. The strong tendency of these G → A substitutions to localize in small 4-nt terminal loops of hairpins is evident. The CAG repeat shows enormous plasticity in the presence of even a single substitution of this type. It is noteworthy that none of the 12 analyzed SCA2 transcripts gives rise to the repeat region in the SCA2 transcript (m-1-4-1-4-1-n) is also in accord with the above rules, but it is entirely different in the SCA1 transcript (m-1-4-1-4-1-4-1-n) identified in healthy individuals. Note that thermodynamic stability of these structures decreases (ΔG value increases) as a function of the increasing number of CAA interruptions. In two other groups there are four examples of triplet repeats having a length characteristic for mutated tracts (36 repeats and 39 repeats), but two of them contain repeat interruptions (see text for more details).

Fig. 8. RNA structures of the repeat regions of various normal and mutant SCA2 transcripts. Simplified secondary structures of M2 modules of the SCA2 transcripts belonging to three groups differing in the repeat length. The first group harbors the longest normal tracts (29 repeats) containing different numbers of CAA interruptions (red arrowheads) identified in healthy individuals. Note that thermodynamic stability of these structures decreases (ΔG value increases) as a function of the increasing number of CAA interruptions. In two other groups there are four examples of triplet repeats having a length characteristic for mutated tracts (36 repeats and 39 repeats), but two of them contain repeat interruptions (see text for more details).
Interrupted CAG Repeats in SCA2 Transcripts

1-1-1-1-n). In the latter case this interruption system forces the CAG repeats to form two hairpin structures composed of pure repeats the (CAG)_m and (CAG)_n (16).

The results of the recent survey show that the CAG repeats are the most frequent triplet repeats in human mRNAs (12), and CAA interruptions (coding also for glutamine) disrupt the regularity of these repeat tracts most often. Moreover, the CAA interruptions occur in different numbers and at various positions within the CAG repeats in human transcripts. It was therefore tempting to make the first insight into the structural outcomes within the CAG repeats in human transcripts. It was regularity of these repeat tracts most often. Moreover, the CAA interruptions are the most frequent triplet repeats in human mRNAs (12), all SCA17 patients (44). Following are the examples of one population (29Q–42Q), and the interruptions are retained in the repeat region also shows length polymorphism in a normal transcript in the putative RNA-mediated pathogenesis of the disease is between 38 and 55 years of age (51). However, in one of the SCA2 families in which the 39Q tract was encoded by the genetically stable interrupted repeat 8-1-9-1-4-1-4-1-10, all five carriers developed the inherited Parkinsonism with motor complications but not cerebellar abnormalities at the age of 64–88 years (47). A number of other allelic variants of the SCA2 gene were also described. They contained the single CAA interruptions in the 3’ parts of their CAG repeat tracts, e.g. variant 27-1-8 in Fig. 8. These variants, however, were not associated with the SCA2 disease in most carriers (48, 49). In conclusion, the good correspondence of the RNA structures formed by the CAG repeats in SCA2 transcripts with the clinical data, together with the lack of such a correlation with the lengths of the poly(Q) tracts, may be considered supportive for the RNA contribution to SCA2 pathogenesis.

REFERENCES

1. Richards, R. L., and Sutherland, G. R. (1997) Trends Biochem. Sci. 22, 432–436
2. Cummings, C. J., and Zoghbi, H. Y. (2000) Hum. Mol. Genet. 9, 909–916
3. Bowser, R. P., and Wells, R. D. (2001) Prog. Nucleic Acids Res. Mol. Biol. 66, 159–202
4. Orr, H. T. (2001) Genes Dev. 15, 925–932
5. Schols, L., Bauer, P., Schmidt, T., Schulte, T., and Ries, O. (2004) Lancet Neurol. 3, 291–304
6. Koch, M. D., Moseley, M., Schut, L. J., Bovens, K. A., Bird, T. D., Day, J. W., and Ranum, L. P. (1999) Nat. Genet. 21, 379–384
7. Matsuura, T., Yamagata, T., Burgess, D. L., Rasmussen, A., Grewal, R. P., Watase, K., Kajhavi, M., McCullin, A. E., Davis, C. F., Zou, A., Chashi, M., Pulsit, S. M., Alkon, E. S., and Schmidt, J. L. D., Ranum, L. P., Zoghbi, H. Y., and Ashizuka, T. (2000) Nat. Genet. 26, 191–194
8. Holmes, S. E., O’Hearn, E. E., McNinis, M. G., Gorelick-Feldman, D. A., Kleiderlein, J. J., Callahan, C., Kwak, N. G., Inderesh-Ashworth, R. G., Sherr, M., Sumner, A. J., Sharp, A. H., Ananth, U., Seltzer, W. K., Boss, M. A., Viera-Saecker, A., Epplen, J. T., Riess, O. R., Boss, C., Atta, and Margolis, R. L. (1999) Nat. Genet. 23, 391–392
9. Ranum, L. P., and Day, J. W. (2002) Cur. Opin. Neurobiol. 12, 266–271
10. Jasinska, A., and Krzyzosiak, W. J. (2004) FEBS Lett. 567, 136–141
11. Broude, N. E., and Cantor, C. R. (2003) Expert. Rev. Mol. Diagn. 3, 269–274
12. Jochimski, M., Michlewski, G., Schobek, K., Kozudz, P., Napi- erala, M., and Krzyzosiak, W. J. (2003) Nucleic Acids Res. 31, 5463–5468
13. Schobek, K., de Mezer, M., Michlewski, G., Krol, J., and Krzyzosiak, W. J. (2003) Nucleic Acids Res. 31, 5468–5474
14. Hagerman, P. J., and Hagerman, R. J. (2004) Am. J. Hum. Genet. 74, 805–816
15. Orr, H. T. (2004) Trends Neurosci. 27, 233–234
16. Schobek, K., and Krzyzosiak, W. J. (2004) J. Biol. Chem. 279, 41563–41572
17. Tapscott, S. J., and Thornton, C. A. (2003) Science 293, 816–817
18. Miller, J. W., Urbinati, C. R., Teng-Umnuy, P., Stenberg, M. G., Byrne, B. J., Thornton, C. A., and Swanson, M. S. (2000) EMBO J. 19, 4439–4448
19. P. Zarnescu, D., Ziegelbaum, C., E. Li, Chami, L., McSweeney, M., Kedes, B., Boss, B. A., De- Jong, P., Rouleau, G. A., Auberger, G., Kerenberg, J. R., Figueroa, C., and Sabah, S. (1996) Nat. Genet. 14, 269–276
20. Innelt, A., Gaudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J. M., Weber, C., Mandel, J. L., Caneel, G., Abbas, N., Dwarz, A., Didierjean, O., Stevanin, G., and Brice, A. (1996) Nat. Genet. 14, 285–291
21. Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., Wakisaka, K., Ishida, Y., Kroech, T., Kunde, R., Sato, A., Tanaka, T., Shimo, T., Nishizawa, M., Shimizu, N., Nomo, Y., Segawa, M., Iwabuchi, K., Eguchi, I., Tanaka, H., Takahashi, H., and Tsujl, S. (1996) Nat. Genet. 14, 277–284
22. Schobek, K., and Krzyzosiak, W. J. (2004) Human Mutat. 24, 236–247
23. Choudhury, S., Mukerji, M., Srivastava, A. K., Jain, S., and Brahamchari, S. (2003) Hum. Mol. Genet. 10, 2437–2446
24. Orr, H. T., Chung, M. Y., Banfi, S., Kwiatkowski, T. J., Jr., Servadio, A., Beaudet, A. L., McCullin, A. E., Davick, L. A., Ranum, L. P., and Zoghbi, H. Y. (1993) Nat. Genet. 4, 221–226
25. Chung, M. Y., Ranum, L. P., Davick, A. L., Servadio, A., Zoghbi, H. Y., and Orr, H. T. (1993) Nat. Genet. 4, 254–258
26. Eichler, E. E., Holden, J. J., Popovich, B. W., Reiss, A. L., Snow, K., Tihubeau, S., Richards, C. S., Ward, P. A., and Nelson, D. L. (1994) Nat. Genet. 8, 85–94
27. Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) J. Mol. Biol. 288, 911–940
28. Krzyzosiak, W. J., Marciniie, T., Wierowski, M., Romby, P., Ebel, J. P., and Gege, R. (1988) Biochemistry 27, 5771–5777
29. Gornicki, P., Baudin, F., Romby, P., Wierowski, M., Krzyzosiak, W., Ebel, J. P., Ehresmann, C., and Ehresmann, B. (1989) J. Biomol. Struct. Dyn. 6, 971–984
CAG Repeats Containing CAA Interruptions Form Branched Hairpin Structures in Spinocerebellar Ataxia Type 2 Transcripts
Krzysztof Sobczak and Wlodzimierz J. Krzyzosiak

J. Biol. Chem. 2005, 280:3898-3910.
doi: 10.1074/jbc.M409984200 originally published online November 8, 2004

Access the most updated version of this article at doi: 10.1074/jbc.M409984200

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC’s e-mail alerts

This article cites 50 references, 12 of which can be accessed free at
http://www.jbc.org/content/280/5/3898.full.html#ref-list-1