Departamento de Bioquímica y Biología Molecular

Universidad del Pais Vasco

Structural study of CUG-repeating small RNAs complexed with silencing suppressor p19.

Doctoral Thesis

Elizaveta Katorcha

Leioa, 2012

Departamento de Bioquímica y Biología Molecular

Universidad del Pais Vasco

Structural study of CUG-repeating small RNAs complexed with silencing suppressor p19.

Doctoral Thesis

This Doctoral Thesis has been carried out in the Structural Biology Unit of CIC bioGUNE with the support of the PhD fellowship of CIC bioGUNE and grant BFU2007-61715 “Plan Nacional”.

Scientific supervisor: Dr Lucy Malinina

Elizaveta Katorcha

Leioa, 2012

Acknowledgements

I deeply thank Dr Lucy Malinina for giving me the opportunity to do my PhD thesis in her laboratory and for her enthusiasm about my project. I would like to thank her for always supporting me even when results did not come as fast as expected.

I would like to thank Dr Valeriya Samygina for inspiring me to grow as a person as well as a scientist.

I do not think this thesis would be possible without the marvelous people around me. I want to give a hug to Borja Ochoa for his true devotion and willingness to help, to Jevgenia Tamjar for supporting me at hard times, to Sandra Delgado for her enthusiasm and care.
I also thank ex-members of our laboratory: Dr Aintzane Cabo-Bilbao and Dr Felipe Goni. It was a pleasure to work with them.
I don't forget the people from CIC bioGUNE who welcomed me warmly and made my life easier and more fun. The coffees, excursions, fiestas and any other things we shared will stay in my memory for ever.

Other special thanks to the staff of the Department of Biochemistry and Molecular Biology at the University of the Basque Country. Without them, especially Dr Diego Guerin and Clemente Rodriguez, I would have been buried under a heap of papers. I am very grateful to Dr Alicia Alonso for accepting to be my “ponente” at the University and for her help and kindness.

I also would like to thank Dr Alexander Popov, Dr Garib Murshudov, and Dr Gleb Bourenkov for their help and suggestions about my work.

I am deeply grateful to my family and long-time friends who always supported and encouraged me during my life, whatever the decision I made. This work was possible thanks to your care, attention and love.

Table of Contents

1. Introduction

1.1.RNA interference

Double-stranded RNA (dsRNA) induces sequence-specific posttranscriptional gene silencing in many organisms by a process known as RNA interference (RNAi) [42]. In different organisms, the RNAi pathways comprise different proteins and mechanisms, but they all operate through processing a dsRNA into small RNAs, which determine the specificity of the response (**Figure 1**). During the past 13 years, following the discovery of RNAi [38], we have witnessed amazing developments in the study of small, noncoding RNA molecules in animals, plants, and fungi. First noticed as intermediates in an experimental silencing process that was at that time poorly understood mechanistically [39], small RNAs are now represented by many different species [40]: snoRNAs, miRNAs, siRNAs and piRNAs. One thing that all small RNAs have in common is that they bind to one of the members of the Argonaute (Ago) protein family, and subsequently act as guides to specifically target mRNA molecules.

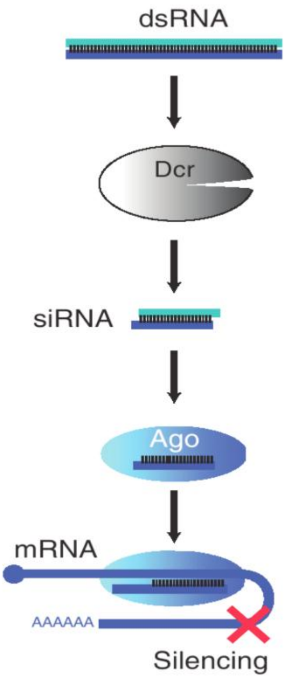


Figure 1. Schematic of RNA interference.

Long double-stranded RNA (dsRNA) becomes an origin for small interfering RNAs (siRNAs) cleaved from the dsRNA by the ribonuclease Dicer (Dcr). siRNAs are short dsRNA molecules (usually of 19 base pairs in length) provided with 2-nucleotide 3'-overhangs on either end. Molecules of siRNA bind to an Argonaute protein (Ago), which performs a selection of guide siRNA strand. Ago loaded by the guide strand forms a core of the RNA-induced silencing complex (RISC), which binds mRNA containing complementary sequence. RISC targets such mRNAs for silencing, commonly by degradation.

siRNAs (small-interfering RNAs), 20-25 nucleotides in length, are the central players in many RNAi pathways. They are usually produced from a long double-stranded RNA through cleavage by the endonuclease Dicer and then bound to an Ago protein. siRNAs recognize their targets by sequence complementarity and establish silencing of the recognized mRNAs (**Figure 2A**). The source of the initial dsRNA can be either exogenous (e.g. from viral replication) or endogenous. Expression of endogenous siRNAs is often tissue-specific or developmental-stage-specific, with a bias towards expression in reproductive tissues or stages [92]. While for some siRNAs the mode of action remains poorly understood, for others we have learned much about their biogenesis, the proteins they associate with, and the effects they can have on cells. siRNAs have a well-defined structure: a short dsRNA (usually of 19-bp) with 2-nt 3'-overhangs on either end.

1.1.1. Small RNA types

siRNAs. The first investigated cases of RNAi were triggered by exogenous dsRNA. Schematic of siRNA cellular pathways is presented in **Figure 2B**. Here, long exogenous dsRNA is cleaved into siRNAs by Dicer, a dsRNA-specific ribonuclease of RNase III family [41]. Produced siRNAs comprise two RNA strands (each bearing a 5′ phosphate and 3′ hydroxyl group), usually of 21 nt in length and paired in the manner that results in formation of a 19bp duplex provided with two-nucleotide overhangs at either 3’-end [42,43]. Strand that directs silencing is called the guide, whereas the other strand, ultimately destroyed, is called the passenger. Target regulation by siRNA is mediated by RISC (the RNA-induced silencing complex). siRNA is loaded into the RISC, the passenger strand is degraded and then the complex can proceed the silencing of cognate mRNAs [44]. In addition to an Ago protein and a small RNA guide strand, RISC may contain auxiliary proteins that extend or modify its function, in particular re-direct the target mRNA to a site of general mRNA degradation [44]. siRNAs are present in all three eukaryotic kingdoms: plants, animals, and fungi, - and provide (at least) plants and animals with anti-viral defense [45,46].

miRNAs. MicroRNA or miRNA is perhaps the most famous class of noncoding small RNAs. Main steps of miRNA cascades are shown in **Figure 2C**. miRNA genes are transcribed by RNA polymerases II and III into primary transcripts called the pri-miRNAs [47]. They are processed into pre-miRNAs in the nucleus by a Microprocessor complex, which contains the RNase III enzyme Drosha and the double-stranded RNA-binding protein DiGeorge syndrome critical region gene 8 (DGCR8; also known as Pasha) [48]. Pre-miRNAs are then transported by exportin 5 and RanGTP into cytoplasm, where they are further processed by the RNAse III enzyme Dicer. This results in miRNAs,

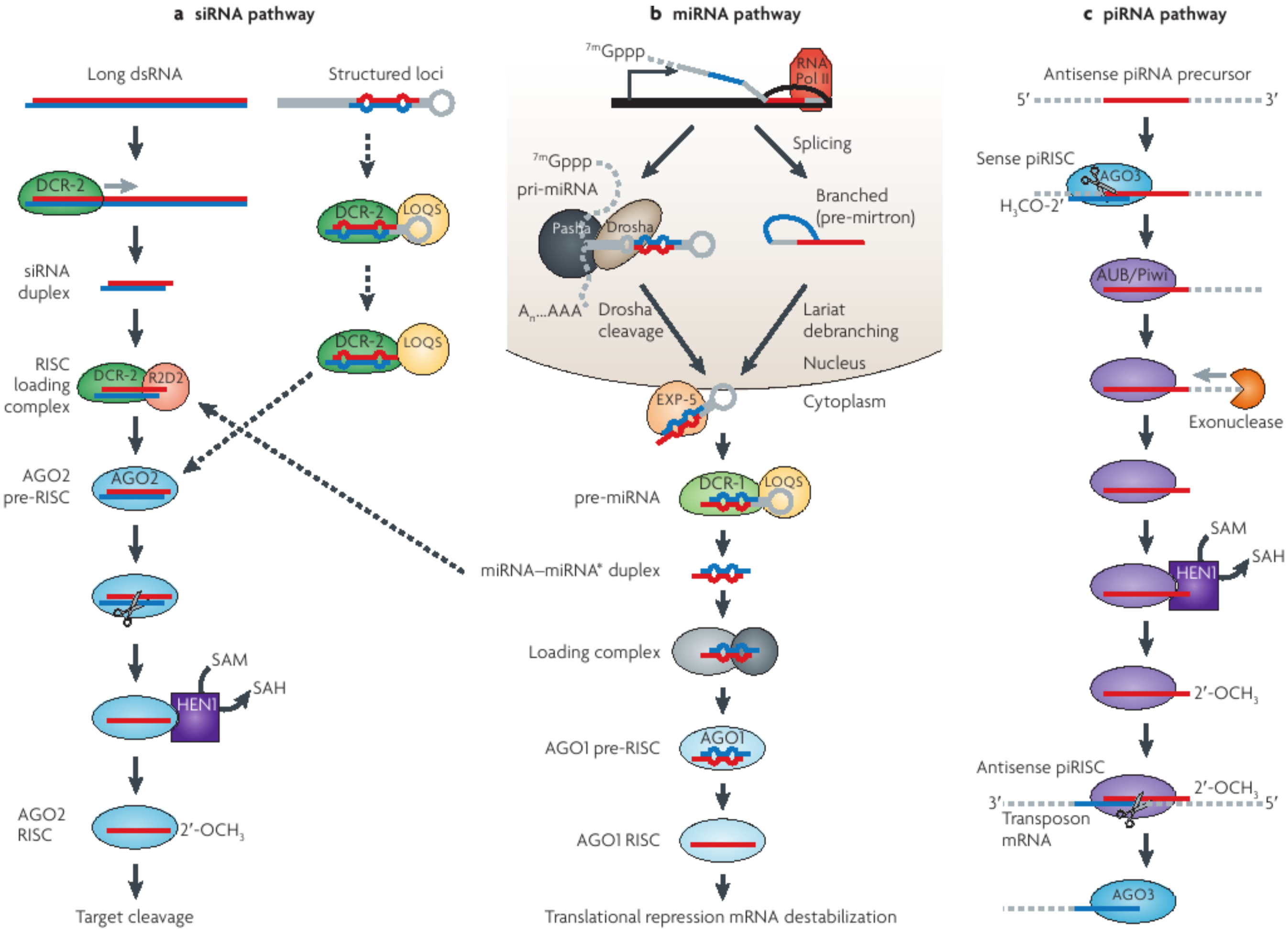

a siRNA pathway
Long dsRNA
Structured loci
DCR-2
DCR-2
LOQS
siRNA duplex
DCR-2
LOQS
RISC loading complex
DCR-2
R2D2
AGO2 pre-RISC
AGO2
SAM
SAH
HEN1
AGO2 RISC
2′-OCH3
Target cleavage
b miRNA pathway
7mGppp
RNA Pol II
Splicing
7mGppp
pri-miRNA
Pasha
Drosha
Branched (pre-mirtron)
An...AAA
Drosha cleavage
Lariat debranching
Nucleus
Cytoplasm
EXP-5
pre-miRNA
DCR-1
LOQS
miRNA–miRNA* duplex
Loading complex
AGO1 pre-RISC
AGO1
AGO1 RISC
Translational repression mRNA destabilization
c piRNA pathway
Antisense piRNA precursor
5′
3′
Sense piRISC
AGO3
H3CO-2′
AUB/Piwi
Exonuclease
SAM
SAH
HEN1
2′-OCH3
Antisense piRISC
2′-OCH3
3′
5′
Transposon mRNA
AGO3

Figure 2. Characteristics and biogenesis of small RNAs (cited from [40]).
(A) dsRNA precursors are processed by Dicer-2 (DCR-2) to generate siRNA duplexes containing guide and passenger strands. DCR-2 and the dsRNA-binding protein R2D2 (which together form the RISC-loading complex, RLC) load the duplex into Argonaute2 (AGO2). A subset of endogenous siRNAs (endo-siRNAs) exhibits dependence on dsRNA-binding protein Loquacious (LOQS), rather than on R2D2. The passenger strand is later destroyed and the guide strand directs AGO2 to the target RNA.
(B) miRNAs are encoded in the genome and are transcribed to yield a primary miRNA (pri-miRNA) transcript, which is cleaved by Drosha to yield a short precursor miRNA (pre-miRNA). Alternatively, miRNAs can be present in introns (termed mirtrons) that are liberated following splicing to yield authentic pre-miRNAs. Pre-miRNAs are exported from the nucleus to the cytoplasm, where they are further processed by DCR-1 to generate a duplex containing two strands, miRNA and miRNA*. Once loaded into AGO1, the miRNA strand guides translational repression of target RNAs.
(C) piRNAs are thought to derive from ssRNA precursors and are made without a dicing
step. piRNAs are mostly antisense, but a small fraction is in the sense orientation. Antisense piRNAs are preferentially loaded into Piwi or Aubergine (AUB), whereas sense piRNAs associate with AGO3. The methyltransferase HEN1 adds the 2′-O-methyl modification at the 3′ end. Piwi and AUB collaborate with AGO3 to mediate an interdependent amplification cycle that generates additional piRNAs, preserving the bias towards antisense. The antisense piRNAs probably direct cleavage of transposon mRNA or chromatin modification at transposon loci. SAH, S-adenosyl homocysteine; SAM, S-adenosyl methionine.

double-stranded 20–25 nucleotide (nt) intermediates with 2 nt overhangs on the 3' end [48]. One of the RNA strands is then loaded by Dicer into RISC, that then targets the 3' untranslated region of the target mRNAs by an imperfect match between the miRNA and the mRNA, to repress translation – although, it was found that the opposite effect of translation activation was also possible [49]. miRNAs have only been found in land plants, the unicellular green alga, *Chlamydomonas reinhardtii*, and metazoan animals, but not in unicellular choanoflagellates or fungi [50].

Piwi-interacting RNAs (piRNAs) are the most recently discovered class of small RNAs, and, as their name suggests, they bind to the Piwi clade of Ago proteins. Cellular interactions of Piwi-interacting RNAs are presented in **Figure 2D**. Animal Argonaute proteins can be subdivided by sequence relatedness into Ago and Piwi sub-families. piRNAs were first proposed to ensure germ line stability by repressing transposons when they were discovered in flies as a class of longer small RNAs (~25–30 nt) associated with silencing of repetitive elements [51]. Later, these "repeat associated small interfering RNAs"—subsequently renamed as piRNAs—were found to be distinct from siRNAs and miRNAs: they bind Piwi proteins and do not require Dicer for their production[52]. Moreover, they are 2′-*O*-methylated at their 3′ termini, unlike miRNAs, but like siRNAs in flies. piRNAs appear to be the youngest major small RNA family, having been found only in metazoan animals [50].

1.1.2. RNAi pathways

All RNA interference (RNAi)-related pathways involve an Ago protein (**Figure 2B-D**). The number of Ago genes varies widely, from one in the fission yeast *Schizosaccharomyces pombe* to 27 in the nematode *Caenorhabditis elegans*. Ago proteins identify the targets of an RNAi pathway through basepairing between the Ago-bound small RNA and the target RNA, to which they are capable of inducing a number of effects. Some Ago proteins have a catalyticaldomain that can cleave the targeted RNA molecule. Other Agos do not rely on target cleavage, either due to the absence of key catalytic residues in their active sites or because of slow enzyme kinetics. These Agos often involve the recruitment of additional factors to employ various mechanisms to affect the activity of their targets[53,54].

A number of RNAi pathways utilize double-stranded RNA (dsRNA) to generate small RNAs through the action of the enzyme Dicer. Sources of the dsRNA can vary between pathways [41]. For example, most miRNAs result from Dicer activity on intramolecular "fold-back" structures, or hairpins. Endogenous siRNAs (endo-siRNAs), another small RNA species, can be derived from more extended hairpin structures, as well as from dsRNA assembled through intermolecular basepairing between transcripts (described in mouse and in *Drosophila)* [47].

Another source of dsRNAs that has been found in *S. pombe*, plants, and *C. elegans*, is RNA-dependent RNA polymerases (RdRPs). The *S. pombe* RdRP enzyme Rdp1 synthesizes dsRNA at centromeric loci that is subsequently diced and loaded into the Ago1 protein to direct the formation of pericentromeric heterochromatin [55]. In the plant *Arabidopsis thaliana*, multiple RdRP enzymes are involved in intricate networks of different RNAi pathways. In each case, the RdRP enzyme appears to make dsRNA that is then used by one of the four Dicer-like enzymes as substrate [56]. In animals, RdRP activity has so far only been described in *C. elegans*. In this nematode, at least one RdRP enzyme, RRF-3, may be involved in producing dsRNA that is processed by Dicer (DCR-1). RRF-3, DCR-1 and a number of other factors, are involved in generating a subset of the endo-siRNAs in *C. elegans* [57].

Some RNAi pathways function in a Dicer-independent manner. For example, in *C. elegans*, a prominent population of endo-siRNAs is most likely derived directly from RNA-dependent RNA polymerase (RdRP) activity. In other words, the RdRP makes short RNA transcripts that directly bind to Ago proteins [58]. These small RNAs, known as 22G, are characterized by the presence of a triphosphate group at their 5′ end, possibly resulting from the first NTP residue used in their synthesis. Many different Ago proteins associate with these small RNAs [59]. Interestingly, most of these Ago proteins are not believed to trigger target cleavage, as they do not have all the residues required to be catalytically active, implying they employ different mechanisms to affect their targets.

Another Dicer-independent pathway is the so-called Piwi-pathway [60], specific to animals and most often specifically active in the germline. The Piwi-pathway is driven by a subclass of the Ago proteins called Piwi proteins, which bind a type of small RNA called piRNA [51]. The pathway seems to be particularly active in the germ line and functions in transposon silencing and epigenetic regulation. The exact mechanism of biogenesis is still unclear. Presumably, piRNAs originate from a long single-stranded RNA polymerase transcript (1–100 kilobases) that is often transcribed from a bidirectional piRNA cluster [48]. The precursor is subsequently processed into mature piRNAs of 23–32 nt in length by unknown mechanisms, which probably involve Piwi proteins but not Dicer [48].

The siRNA, miRNA and piRNA pathways were initially believed to be independent and distinct. However, the lines distinguishing them continue to fade. These pathways interact and rely on each other at several levels, competing for and sharing substrates, effector proteins and cross-regulating each other. Both the exogenous siRNA and miRNA pathways generate dsRNA duplexes containing usually a 19 bp double-stranded core flanked by 2 nt 3′-overhangs. A siRNA duplex is complementary throughout its core; a miRNA/miRNA* duplex may and usually does contain mismatches, bulges and G:U wobble pairs. In *Drosophila*, biogenesis of small RNAs is uncoupled from its loading into Ago proteins [61]. Instead, loading is governed by the small RNA structure: duplexes bearing bulges and mismatches are sorted into the miRNA pathway and subsequently loaded into Ago1 (primarily represses translation), while duplexes with greater double-stranded character partition go to load Ago2 (represses by target cleavage), the Argonaute protein associated with RNAi. Sorting creates competition between the two pathways for substrates [61]. In *Drosophila,* loading of a small RNA into one pathway decreases its association with other pathway.

The discovery of endogenous siRNAs derived from transposons of *D. melanogaster* further erases the distinctions between different RNAi pathways [62]. Endo-si-RNAs, are believed to be the main cause for silencing 'selfish' genetic elements in somatic cells, which lack the piRNA pathway [63]. Therefore, endo-siRNAs and piRNAs are fundamentally similar in that they defend organisms against nucleic-acid-based 'parasites'. Some endogenous siRNAs are processed from overlapping regions of functional genes and their cognate pseudogenes. This finding suggests that pseudogenes, which have been thought to be non-functional protein 'fossils', might regulate the expression of their founder genes [63].

This blurring of the boundaries between the different types of small RNAs has interesting evolutionary implications. The long stem–loop structures that are processed to form endo-siRNAs are reminiscent of the pre-miRNAs in plants. One hypothesis for the evolutionary origin of plant miRNAs is that new plant miRNA loci might evolve from the inverted duplication of founder loci, which when transcribed would result in hairpin RNAs [64]. These hairpin RNAs would be almost perfectly self-complementary. Since DCL1, the main miRNA-processing enzyme in plants, has a limited activity against such substrates, they should be processed by Dicer-like enzymes other than DCL1. . Subsequent acquisition of mutations, however, would result in a hairpin with imperfect complementarity, an appropriate substrate for DCL1. Thus, the stem–loop structures that originate endo-siRNAs might suggestively be gradually transformed into miRNA precursors. We can also imagine that similar evolutionary adaptation hasoccured with miRNA-encoding genes in *D. melanogaster,* in which DCR-1 could generate miRNAs rather than endo-siRNAs generated by DCR-2.

1.2.Repeats in the genome

Although considerable amounts of data are still lacking for full understanding of RNAi networks, it is evident that this system of interactions with multiple possibilities of inter-pathway cross-regulation is very complex. The presence of dsRNA is obligatory for the triggering of RNAi response and the mechanisms are seem highly dependent on sequence complementarity between the siRNA strands and, in the system of activated RISC, between siRNA and target mRNA. In the light of the new discoveries connected to RNAi pathways, the high and ubiquitous presence of non-coding genomic repeats, especially those found in UTRs (untranslated regions) and introns of mRNAs, can get an explanation.

Repetitive DNA sequences are interspersed throughout the human genome [1]. They account for almost 30% of the human genome [1-2]. It was a common belief that repetitive sequences are “junk DNA” without any biological functions and are merely passed on from one generation to the next through “selfish replication” [3]. However, in the past decade several important discoveries strongly indicate that at least some DNA repeats are biologically significant.

Repeats can be classified into two main categories based upon their locations in the genome: ones that are not associated with genes and the others that are present in the regulatory regions of genes. A pentanucleotide (AATGG) is tandemly repeated in the human centromere, which is the attachment point of two sister chromatids during mitosis [4]. The centromeric region stretches up to mega bases and shows a high degree of conservation of the pentanucleotide repeat. A hexanucleotide (TTAGGG) is repeated 30-100 times at the 3' DNA overhang of the human telomere, which is required for end-replication and chromosomal integrity [5]. This hexanucleotide sequence is highly conserved in vertebrates [6]. Even though they are not associated with any genes, the high level of their sequence conservation strongly argues for functional significance of the centromeric and telomeric DNA repeats.

Rapid progress in the sequencing of the human genome revealed many DNA repeats located in the regulatory (non-coding) regions of various genes, either at the 5' and 3' untranslated regions (5'/3'UTR) or inside introns. These repeats are referred to as microsatellites or minisatellites depending on their sizes. Microsatellites are tandem repeats of 1-7 nucleotides [7] whereas minisatellites are tandem repeats of 10-100 nucleotides [8]. These repeats have attracted special attention after the discovery that various human diseases are associated with either expansions or contractions of microsatellites and minisatellites [7-8].

1.2.1. Trinucleotide repeats in the genome

Trinucleotide repeats (TNRs) are a special class of microsatellites. These sequences have received special attention, primarily because of their pathogenic expansions that cause trinucleotide repeat expansion diseases (TREDs) [9]. More than 30 genetic disorders, mostly neurodegenerative and neuromuscular, are currently known to belong to this group [10,11]. In several TREDs, stable RNA structures formed by triplet repeats present in untranslated regions of the responsible genes are thought to be implicated in pathogenesis [12–15]; in some other TREDs, CAG repeats expressed as homo-Gln tracts in proteins suggestively give rise to pathogenesis [16–18].

The great majority of TNRs do not undergo pathogenic expansion [19] and a little is known about their normal function in human genes and transcripts. The features of TNRs that suggest their functionality include: (i) widespread occurrence in exons, (ii) formation of stable hairpin or quadruplex structures by some TNRs and (iii) coding for homo-amino acid (AA) tracts [19].

Previous studies [19] have shown that the occurrence of TNRs in the genome is strongly biased compared with their genomic frequency. Some TNR types are strongly overrepresented (CGG > CAG > GAC > AGG) while others are underrepresented (AAT > AAC > AAG) in exons. This is argued to be supporting the notion that TNRs are important functional genetic elements undergoing strong selective pressure and that the functionality of TNRs can be expressed at the protein, DNA (genetic) and RNA levels.

Length distribution also differs significantly between specific TNR types [19]. It is supposed that the length of a TNR presents equilibrium between its tendency to expand and the pressure against expansion (due to possible toxicity of long TNRs). Interestingly, average length does not differ significantly for a given TNR whether it is located inside or outside an exon. However, the probability of a TNR to be found in different mRNA regions is not the same. For example, AT-rich TNRs were found to be linked to the 3'UTRs, while CG-rich TNRs were linked to 5'UTRs and CAG repeats tended to collocate to ORFs (open reading frames). These observations have led to speculations about the functions and mechanisms of actions of TNRs in the genome and the transcriptome.

Functionality of TNRs can be expressed at the protein, DNA (genetic) and RNA levels. In proteins, the uninterrupted TNR sequences code for homo-amino-acid-tracts. The most frequent repeating amino acids are Gln, Ala, Glu and Leu [19]. Cys-, Arg-, Met- and Asn-tracts are very rare, while Tyr-, Trp-, Val-, Ile- and Phe-tracts do not occur at all. It was shown [93] earlier that the presence of such tracts may influence many protein properties and also cause protein toxicity. It was also shown that homo-amino-acid-tracks in some proteins are highly conserved across eukaryotes. A very few protein structures that contain homo-repeats were studied by X-ray crystallography, with homo-repeat tracts being usually reported to be disordered. This suggests that in most proteins, homo-repeats form unstable structures that can serve as flexible linkers or hinges. Such elements can suggestively facilitate interactions with other macromolecules. For instance, poly-Leucine tracts, frequently occurring in membrane proteins, are thought to collocalize these proteins to membrane.

The function of TNRs at the DNA (genetic) level is probably related mostly to their high mutability. A high mutation rate of microsatellites can increase plasticity and facilitate adaptation of certain classes of genes during evolution. For example, microsatellites located in rapidly evolving developmental genes were shown to differ significantly between morphologically different breeds of dogs [20] and may be considered a major source of phenotypic variation in evolution, facilitating a rapid response to selective pressure.

In RNA, TNRs can also modulate many different functions on the molecular level. It is shown that TNRs and other types of microsatellites in RNA can regulate gene expression [21,22], serve as protein binding sites and splicing enhancers [23], induce transcription slippage and influence RNA stability [24].

The above functions are related mainly to microsatellites localized in untranslated portions of transcripts [25]. The feature that may contribute most to the function of TNRs in RNA is their structure. The functional role of structures formed by TNRs is strongly supported by the correlation of the structure-forming potential of TNRs [26] with their overrepresentation in exons. The **Table 1** illustrates such a classification of TNRs, based on enzymatic and bilophysical analysis. The five TNR types most overrepresented in exons – CGG, GAC, CAG, CCG and CUG, - form stable hairpin structures in transcripts [26]. Notably, these TNRs include all four possible CNG repeats, where N can be A, U, C, or G, and GAC repeats, which areof low frequency in the genome. A common feature of CNG repeats in transcripts is their tendency to form double-stranded hairpin structures (‘dsRNA’) if repeat tracts are long enough. On the other hand, TNRs that remain single-stranded (see **Table 1**) are strongly underrepresented in exons, which suggests that hairpin-forming repeats might have functional roles in the regulation of gene expression.

Table 1. Structural classes of triplet repeat RNAs (cited from [66]). RNAs composed of triplet repeats fall into four structural classes according to enzymatic structure probing, migration in native gel and UV and CD spectra. Four different types of cleavage patterns could be distinguished for the 20 transcripts.

Class I	Class II	Class III	Class IV
Unstructured RNAs	Semi-stable hairpins	Stable hairpins	Stable tetraplexes
$(CAA)_{17}$	$(UAG)_{17}$	$(CAG)_{17}$	$(AGG)_{17}$
$(UUG)_{17}$	$(AUG)_{17}$	$(CUG)_{17}$	$(UGG)_{17}$
$(AAG)_{17}$	$(UUA)_{17}$	$(CCG)_{17}$	
$(CAA)_{17}$	$(CUA)_{17}$	$(CGG)_{17}$	
$(CUU)_{17}$	$(CAU)_{17}$	$(CGA)_{17}$	
$(CCU)_{17}$		$(CGU)_{17}$	
$(CUA)_{17}$			

1.2.2. Genomic CUG repeats and Trinucleotide Repeat Expansion Disorders

CUG repeats are among the most abundant TNRs in human transcripts, and their overrepresentation in coding regions implies a functional significance of these sequences. In mature mRNAs, the CUG repeat tracts occur most frequently in the protein-coding parts and, secondly, in5'- and 3'-UTRs [27]. The documented biological functions of CUG repeats in transcripts include modulation of efficiency and accuracy of pre-mRNA splicing [28], mRNA transport [29] and regulation of translation [30].

The CUG repeats are better known for the multiple system dysfunctions they cause in the mutated form

that occurs in myotonic dystrophy type 1 (DM1) patients [31]. The mutation leading to DM1 is the expansion of a CTG repeat, located in the 3' UTR of dystrophia myotonica protein kinase (DMPK) gene from normal 5–37 repeats to mutated 50–3000 repeats [32]. DM1 is an autosomal dominant disorder characterized by clinical anticipation, whereby mothers that barely manifest symptoms have children with severe forms of the disease including congenital DM1 [33]. The mapping of the gene to chromosome 19q13.3 and the identification of trinucleotide repeat expansion as the mutational basis provided a molecular explanation for the anticipation [34]. DM1 was the first disorder in which a trinucleotide repeat mutation (CTG) was found in the 3' UTR of a gene and a mechanism of “RNA toxicity” was proposed as an explanation of the pathology.

SCA8 is another disorder linked to CTG repeat expansion [35]. The increase of repeat number from 16-34 to 84 leads to ataxia (gross lack of coordination of muscle movements), slurred speech and nystagmus (involuntary eye movement). The CTG repeats in the DNA are untranslated but transcribed and are located in a non-coding region of the disease-associated gene, making an RNA-mediated mechanism worth considering. Molecular analysis failed to reveal an open reading frame in the CTG orientation, so it was proposed that SCA8 is caused by an RNA-mediated mechanism in which the SCA8 transcript was envisioned to function either as an antisense RNA to disrupt expression of the neighboring kelch-like 1 gene (KLHL1) that encodes an actin-binding protein or as a toxic CUG-containing RNA that alters the function of RNA-binding proteins analogous to DM1 [36].

Huntington disease-like 2 (HDL2) is an adult-onset, progressive, neurodegenerative autosomal dominant disorder. Abnormal movements, dementia, and psychiatric syndromes clinically characterize HDL2. Aspects of its neuropathology include prominent cortical and striatal atrophy and intranuclear inclusions. HDL2 is caused by a CTG/CAG expansion mutation on chromosome 16q24.3 in a variably spliced exon of junctophilin-3 in the CTG orientation [37]. This latter evidence indicates that HDL2 may not be a polyglutamine disorder, caused by a poly-Gln tract in the protein.

1.3. Possible RNA gain-of-function mechanisms for TREDs caused by CUG-repeats

A common feature of TREDs caused by CUG repeat expansions in DM1 is an observed misregulation of alternative splicing of numerous developmentally regulated transcripts [94]. This was also shown by observation of similar pathologic effects during expression of CUG repeats in normal cells [95]. Several possible mechanisms of pathology development were proposed to explain such misregulation.

Protein sequestration. Many data favor a mechanism caused by altered interactions of the implicated transcripts with two types of antagonistic splicing regulators: the CUG repeat binding protein (CUG-BP) and the muscleblind like (MBNL) protein [11]. In DM1 cells, the expanded CUG repeats cause a decrease in the cellular level of free MBNL because of its sequestration to nuclear foci and a simultaneous increase of the CUG-BP level by a yet unknown mechanism [11]. If the CUG repeat-containing transcripts and their binding proteins are strongly imbalanced this may result in severe disorders. DM1 is an example of a human disease in which the mechanism of RNA-mediated pathogenesis is generally accepted [65]. In this multi-system disease, the expansion of the CTG repeat located in the 3' UTR of the DMPK gene (myotonic dystrophy type 1, DM1) is the established source of pathogenesis. This expansion is thought to result in formation of long CUG-repeat containing hairpin structures in transcripts [66], which recruit proteins of the MBNL family with which they co-localize in nuclear foci observed in DM cells [67].

RNAi involvement in TREDs. Another possible explanation of DM1 pathogenesis [69] is schematically shown in **Figure 3**. It hypothesizes that long RNA hairpins formed by expanded CUG-repeat tracts can cause RNA silencing of genes that contain CNG-repeating regions. This hypothetic mechanism involves the RNAi pathway. In other words, it suggests the processing of long RNA-hairpins into small CUG-repeating RNAs, followed by their loading on RISC and targeting mRNAs containing CNG-repeating regions. Many indirect data gave rise to the conclusion that at RNA level, repeated CNG-triplet sequences form double helices (reviewed in [105, 66]). The special features of CUG repeats in RNA tracts were investigated previously. In the work by *Mooers et al.* [98], a 1.58 Å crystal structure of a 18-mer $(CUG)_6$ was described. The structure was originally described as statically disordered and the resulting model consisted of two superimposed duplexes. The double helices contained U*U pairs flanked by G-C pairs. The duplexes in

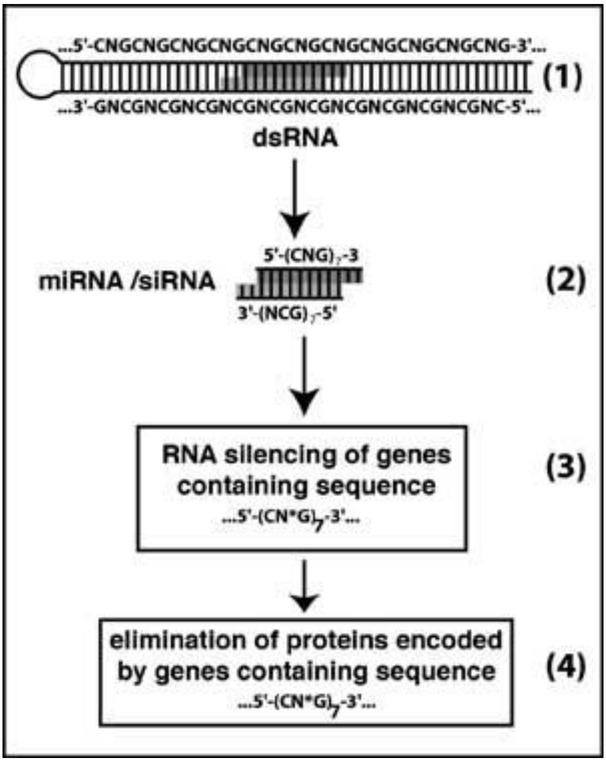

...5'-CNGCNGCNGCNGCNGCNGCNGCNGCNGCNGCNG-3'...
(1)
...3'-GNCGNCGNCGNCGNCGNCGNCGNCGNCGNCGNC-5'...
dsRNA
5'-(CNG)7-3
miRNA /siRNA
3'-(NCG)7-5'
(2)
RNA silencing of genes containing sequence ...5'-(CN*G)7-3'...
(3)
elimination of proteins encoded by genes containing sequence ...5'-(CN*G)7-3'...
(4)

Figure 3. A schematic illustrating the hypothetic molecular mechanism for TREDs (cited from [69]). N is a specific nucleotide of a triplet repeat (T, G, C, or A), while N* is complementary to N.

Stages of pathology development:

(1) transcription of d(CNG)n repeat expansions (where N = T, G, C, or A) promote the formation of the double-helical RNA hairpin, in which non-canonical N·N base pairs are flanked/stabilized at each side by two consecutive Watson-Crick C·G base pairs;

(2) such a 'double-stranded' RNA (dsRNA) becomes a source for microRNAs (miRNAs) or/and small interfering RNAs (siRNAs), which;

(3) negatively regulate the expression or cause the translational repression or mediate a post-transcriptional RNA silencing (siRNA guided mRNA degradation) of the genes containing the d(CN*G)n tracts with n ≥ 7 (which is the length of siRNA, 21-26 nucleotides, divided by a codon length 3; N* is complementary to N), or regions very similar to d(CN*G)n; and

(4) the linkage between triplet repeat expansion and the onset of genetic neuromuscular and neurodegenerative diseases is due to elimination of the proteins encoded by these genes.

the crystal lattice stacked end-to-end, forming long pseudo-continuous helices resembling stem structures of long CUG-repeat hairpins. The overall structure was similar to the A-form RNA, as expected, but the disambiguation of the electron density was difficult. It was determined that the distances between the C1' atoms of the paired uridines were ~10 Å but the U*U pairs appeared to lack hydrogen bonds.

Later in work [70], detwinning of the data from [98] was carried out with the purpose of gaining an unambiguous model. Moreover, in the same work, structure of the octomer $G(CUG)_2C$ of atomic resolution 1.23 Å was solved. It was found that in both cases (the 18- and the 8-mer duplexes) CUG repeats formed regular, well defined structural motifs with characteristic hydrogen-bonding pattern and interactions with the solvent. The higher resolution allowed for detailed structural description of U*U mismatch, flanked by two C-G Watson-Crick base pairs at either end [70]. It was supposed that the electrostatic charge distribution and surface features defined their properties and indicated the ligand binding potential of the CUG tracts. The U*U mismatches, homogeneous in their structural characteristics, were defined as ‘stretched U–U wobble’ pair.

Bearing in mind the quantities of CUG repeats in pathological cases (>50 for myotonic dystrophy type 1, [11]), it was of interest to structurally estimate longer CUG tracts. Such investigations are stated in literature before [66]. In that work, conclusions about $(CUG)_{17}$ transcripts are drawn upon biochemical and biophysical analysis. It was thus proved by enzymatic digestion and thermal melting curve measurements and UV and CD spectral analysis that $(CUG)_{17}$ repeats formed stable hairpin structure with 6 or 7 U*U mismatches flanked with two G-C pairs on each side [66].

In addition, it was shown that ribonuclease Dicer, which recognizes the dsRNAs and digests them into siRNAs, is also capable of processing CUG-repeating RNA sequences and producing small ‘siRNA’-like fragments [71]. It was observed that in DM1 cells, short CUG repeats, derived from the DMPK transcript, downregulate (via RNA interference) the HD and SCA1 transcripts, both containing CAG repeats. It is noteworthy that after silencing of Ago2 (main element of the RISC complex) by siRNA, the level of HD and SCA1 transcripts, measured by RT-PCR of the repeat regions, increased by about 45% and 40%, respectively. As anticipated, the level of DMPK transcript remained unchanged. These results demonstrate that in untreated DM1 cells, the CUG-repeating siRNAs derived from the mutant DMPK transcripts target complementary CAG-repeats in HD and SCA1 transcripts by using the RNAi mechanism.

In the same work, DM1 cells were transfected with $(CAG)_7$ repeat, which led to silencing of the pathological CUG-containing transcripts while leaving the normal allele mRNA level intact or only slightly decreased [71]. All these facts could be cited as possible indications for the presence of silencing mechanisms in TREDs pathology.

1.4. Overview of viral silencing suppressors, p19

RNA silencing is a potent surveillance system targeting parasitic RNA in a highly sequence-specific manner, manifesting as post-transcriptional gene silencing in plants or RNA interference in animals. These evolutionarily conserved processes are now known to be operative in most (if not all) eukaryotic organisms [72]. In plants, the RNA silencing pathway presents a formidable defense against viral pathogens. It is becoming increasingly evident that most, if not all plant viruses have adopted counter-defensive strategies to overcome the host silencing response on viral invasion.

The RNA silencing process in plants can be divided into two stages: initiation and maintenance. At the initiation stage, the presence of dsRNA in the host cell causes its digestion by Dicer (or Dicer-like) ribonuclease into siRNAs of 21–24 nucleotide length. Although it is often assumed that viral replicative RNA-forms provide the dsRNA substrates for DCLs, it is likely that highly structured regions of the genomic RNA are also primary targets for Dicer [73]. Furthermore, viral RNA may also be converted to a dsRNA by one of the RNA-dependent RNA polymerases (RdRPs) encoded by the plant host. siRNA produced by the action of DCLs is then recruited by RISC to mediate the sequence-specific digestion of homologous RNAs.

At the maintenance stage, silencing of homologous RNA persists in the absence of the dsRNA trigger. This is accomplished through a siRNA amplification process in which host RdRP synthesizes new dsRNA using siRNA as a primer, and the homologous cellular RNA as template. The unique feature of RNA silencing in plants is that its local induction generates sequence-specific signals that spread systemically throughout the plant [74].

. It is now well established that plant viruses encode RNA silencing suppressors (RSSs) to specifically counteract the RNA silencing-based defense and ensure successful invasion of the host plant. Interestingly, virologists have only recently begun to recognize the potentially important role of RSSs in modulating virus invasiveness in animal virus infections.

An interesting feature of plant viral suppressors characterized to-date is that none share any obvious sequence or structural similarity across viral families and groups. Many of them have been initially identified as pathogenicity determinants. It seems that evolutionary selection of a particular class of viral silencing suppressor has no relationship to any other primary protein function in the virus life cycle. RSS activity has been identified in proteins involved in many viral functions including ‘movement proteins’, viral replicases, replication enhancers, and transcriptional activators [75].

Perhaps the lack of similarities at either the nucleic acid or the protein level reflects differences at the mechanistic level as well. At present, two major classes of RSSs have been identified: suppressors that affect small RNA metabolism and those affecting systemic silencing [96]. The first type of suppressors reduces the accumulation of siRNAs, because these RSSs block Dicer activity to process dsRNAs. The second silencing-suppression strategy involves the recruitment of endogenous negative regulators of RNA-silencing. For instance, ERI-1 (Enhanced RNAi-1), one of the cellular inhibitors of RNA silencing that have been genetically identified in *C. elegans,* defines a novel subfamily of evolutionary conserved DEDDh nucleases that process siRNAs into shorter, inactive forms [80].

1.4.1. p19 - a ‘universal silencing suppressor’ from Tomato Bushy Stunt virus

The linear, positive sense ssRNA genome of tombusviruses encodes a 19 kDa protein (p19, ORF 5) that was first characterized as a symptom determinant [76] and later shown to suppress RNA silencing [77]. Initial experiments with different tombusviruses, including CNV, Cymbidium ringspot virus, and Tomato bushy stunt virus (TBSV), showed the p19 gene to assist systemic spread and symptom development in host plants [78]. Further studies showed that functional p19 was required for systemic invasion of TBSV in some hosts but not in others [79], suggesting that p19 might be important in antiviral defense of the host plants. Finally, p19was recognized as a suppressor of RNA silencing based on its ability to reactivate expression of a silenced GFP transgene in the systemic leaves of plants infected with TBSV carrying a p19 insert. Subsequently, several groups have independently demonstrated the potent RSS activity of p19 in different tombusviruses using the agro-infiltration assay [80].

In vitro, protein p19 was shown to bind natural and synthetic siRNAsunder stoichiometric conditions [80]. Therefore, it has been suggested [80] that in cell, the p19 function is to bind siRNAs, sequester them from RNAi pathway and block their incorporation into RISC. Impressive progress in studying structural and functional properties of the p19 [80-83,] resulted in recognizing it to become one of the best characterized viral silencing suppressor [97]. Notably, it was the first protein demonstrated to directly bind siRNAs, functioning presumably to prevent the siRNAs from entering the RISC complex [80]. Subsequently, the crystal structure of p19-siRNA complex was determined [81,82] and ‘elegantly explained’ the mechanism of p19-siRNA binding. Additional studies by several groups have now verified that the level of p19-siRNA binding in vivo correlates with the severity of viral infection [83,84].

The structural insights [81,82] indicate that RSS p19 acts as a molecular caliper to specifically select siRNA based on the length of the duplex region. The 19 base-pair duplex is cradled within the positively charged protein surface of a continuous eight-stranded β-sheet, formed by the p19 homodimer. Two α-helical “arms” project from opposite ends of the p19 dimer and position the “tryptophan-pair hands”, which stack over the terminal siRNA base pairs. Direct and water-mediated intermolecular contacts are restricted to the backbone phosphates and 2’-OH groups, consistent with sequence-independent siRNA recognition by the p19.

Phosphates and 2’OH groups mainly interact with basic and polar protein groups, while RNA-ends take part in stacking interaction with tryptophan residues. The RNA minor groove-facing surface of the central part of p19 beta-sheet is enriched in serine and threonine residues (totally, ten). These hydroxyl-carrying residues plus four trapped water molecules form a hydrogen-bond network with six sugar 2 '-OH groups. Interactions with phosphates and 2'-OH groups are mainly localized at the central portion of the RNA duplex and the duplex ends. Stacking between the tryptophan residues and terminal RNA base-pairs essentially provides a mechanism of RNA binding characterized by length specificity. Such bracketing only allows accommodating a 19 bp dsRNAs between two tryptophane “reading heads” [81,82]. Since p19 is known to recognize a range of siRNA lengths, the structural study of p19 complexed with small dsRNAs of other lengths is desirable.

1.5.Objectives of the thesis

This study aims on structural aspects of interaction between RNA silencing suppressor p19 and ‘CUG-repeating RNA sequences’ implicated in dystrophia myotonica and other human Trinucleotide Repeat Expansion Diseases. The research involves crystal structure determination of a protein-unbound CUG-repeating RNA and CUG-repeating RNA fragments of various lengths complexed with RNA silencing suppressor p19 from tombusvirus. The objectives of this work are as follows:

1. Additionally prove the tendency of CUG-repeating RNA sequences for the double-helical structure formation by means of crystallization and studying the crystal structure of RNA fragment pG$(CUG)_6$C; compare the pG$(CUG)_6$C-architecture observed in this work with atomic structure of G$(CUG)_2$C-fragment [70] and structure of $(CUG)_6$ sequence [98] refined against ‘untwined X-ray data’ in [70].
2. Prove the ability of complex formation between p19 and CUG-repeating RNA-sequences, and develop the experimental system for crystallographic study of p19-suppressor complexed with CUG-repeating RNAs: (a) select/make optimal protein constructs and (b) design RNA sequences for obtaining the high-resolution crystal structures.
3. Study the crystal structure of p19 with CUG-repeating RNA fragments of different design and length; explain how p19 interacts (if it does) with (a) CUG-repeating RNAs and (b) RNA fragments of different length; compare the CUG-repeat RNA structures in p19-bound form with those in free (protein-unbound) forms.
4. Analyze structural data, obtained in this work, in light of the hypothesis about possible involvement of RNA interference pathway in human

Trinucleotide Repeat Expansion Diseases [69].

2. Materials and Methods

2.1.Protein crystallography

2.1.1. Physical principles

X-ray crystallography is a method of determining the arrangement of atoms within a crystal, in which an X-ray beam of 'strikes' the crystal and diffracts into many specific directions. Given the angles and intensities of diffracted beams, we can reconstruct a three-dimensional picture of the electron density distribution within the crystal and therefore determine the mean positions of the atoms their mobility, chemical bonding and other characteristics.

Electromagnetic waves. X-rays are electromagnetic radiation with wavelengths of about 0.02 Å to 100 Å (1Å = 10^{-10} meters). They are part of the electromagnetic spectrum that includes wavelengths of electromagnetic radiation called light. Since X-ray wavelengths belong to the same order of magnitudes as atom sizes, they can diffract on atomic structures.

The energy of X-rays, like all electromagnetic radiation, is inversely proportional to their wavelength as given by the Einstein equation:

$$E = h\nu = hc/\lambda$$

where E = energy

h = Planck's constant, 6.62517 x 10^{-27} erg·sec

ν = frequency

c = velocity of light = 2.99793 x 10^{10} cm/sec

λ = wavelength

Thus, since X-rays have a smaller wavelength than visible light, they have higher energy. With their higher energy, X-rays can penetrate the substance easier than visible light. Their ability to penetrate depends on the substance density. Therefore, X-rays provide a powerful tool in medicine for mapping internal structures of the human body.

Bragg's law. Atoms in crystal interact with X-ray waves, as if they reflected the light. Since any crystal consists of periodically arranged atoms, the reflections occur from the sets of parallel planes of atoms characterized by a constant separation d, as illustrated in **Figure 4**. Incident and reflected X-rays always make the same angle with a set of parallel planes. However, only certain angle 'θ' dependent on plane-separation d (**Figure 4**) can 'display' noticeable intensity of reflected X-ray, since waves that come from different planes normally 'compensate' each other. Two X-rays shown in **Figure 4** are reflected from the atomic planes separated by distance d. Ray 1 reflects off the upper atomic plane at an angle θ. Similarly, Ray 2 reflects off the lower atomic plane at the same angle θ. While Ray 2 is in the crystal, however, it travels a distance of 2a longer than Ray 1. If this distance 2a is a multiple of wavelength (nλ),

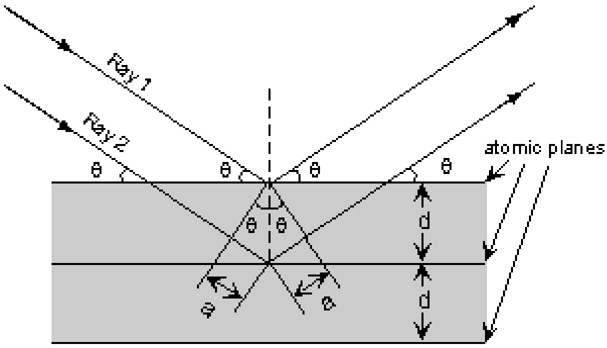


Figure 4. X-ray diffraction by crystal planes. Conditions that produce strong diffracted rays. If the additional distance traveled by the more deeply penetrating Ray 2 is an integer multiple of λ, then Bragg's law is met and Ray 1 and Ray 2 interfere constructively.

then Rays 1 and 2 will be ‘in phase’ on their exit from the crystal and intensity of reflected radiation will be a sum of intensities of reflected Ray 1 and Ray 2.

If distance 2a is not an integral number of wavelengths, then destructive interference takes place and the waves will not be as intensive as when they entered the crystal. Thus, the condition for constructive interference is

$n \lambda = 2a$

Since $2a = 2d \sin \theta$

the final equation can be expressed as

$n \lambda = 2d \sin \theta$

This equation is known as Bragg's Law for X-ray diffraction.

It says that any crystal diffracts X-ray ONLY in certain directions (at Bragg’s angles θ_i) that depend on crystal structure (*d-spacing* between the atomic planes) by condition:

$$2 \sin \theta_i = d_i / n \lambda$$

Laue equations, crystal planes and Miller indices. If the difference in path length between rays reflected from successive planes is equal to an integral number of wavelengths (that is, if Bragg’s Law is fulfilled), then the rays reflected from successive planes emerge from the crystal in phase with each other, interfering constructively to produce a strong diffracted beam. The planes are designated by a set of three numbers called lattice or Miller indices, *hkl*. For a crystal with cell parameters **a**,**b**,**c**, we have three Laue equations:

$$a(\cos\alpha_n - \cos\alpha_0) = h\lambda$$
$$b(\cos\beta_n - \cos\beta_0) = k\lambda$$
$$c(\cos\gamma_n - \cos\gamma_0) = l\lambda$$

where $\cos\alpha_0$, $\cos\beta_0$, $\cos\gamma_0$ are the direction cosines of the incident ray, and $\cos\alpha_n$, $\cos\beta_n$, $\cos\gamma_n$ are the direction cosines of the reflected ray in the crystal axis. Diffraction occurs when *h*,*k* and *l* are integers. The index *h* gives the number of parts into which the set of planes cut the edge a of each cell; the indexes *k* and *l* respectively give the number of parts into which the set of planes cut the edges b and c. Each set of parallel planes is treated as an independent diffractor and produces a single reflection.

The *hkl* planes can be described through a scattering vector **S** normal to the *hkl* plane and of length 1⁄d. The points at the end of these vectors form the reciprocal lattice. The reciprocal lattice is spatially linked to the crystal because of the way the lattice points are defined, so if the crystal is rotated, the reciprocal lattice rotates with it. Each reciprocal lattice point must be arranged with respect to the X-ray beam in order to satisfy the Bragg's law and produce a reflection from the crystal.

Ewald sphere and reciprocal lattice. The Ewald construction is of help to visualize, which Bragg planes are in the correct orientation to diffract. In 3D space, the Bragg law of X-ray diffraction is illustrated by the Ewald sphere shown in **Figure 5**, which is a convenient tool for constructing the X-ray diffraction pattern in the imaginary reciprocal lattice (unlike the real space, the reciprocal space is imaginary). Each point of reciprocal lattice corresponds to the atomic plane in real crystal, and it turns out to be convenient to consider a sphere of radius 1/λ to analyze the diffraction of X-ray radiation of wavelength λ. The crystal is in the center of the Ewald sphere, and the origin of the reciprocal lattice is at the crystal origin where the incident beam leaves the Ewald sphere (**Figure 5A**). If the reciprocal lattice point lies on the surface of the Ewald sphere, the length of the vector **S**, perpendicular to the reflecting plane is $2\ \sin\theta/\lambda=1/d$, that is the Bragg’s law (**Figure 5B**). In summary, the Ewald

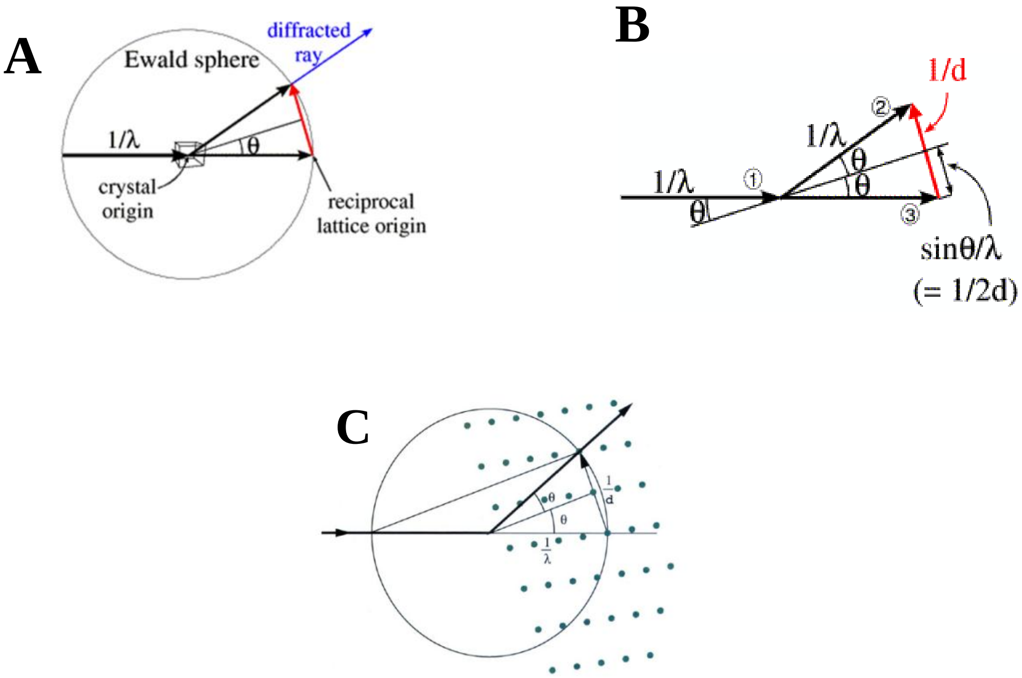


Figure 5. The Ewald construction. When the reciprocal-lattice point crosses the surface of the sphere (A), the trigonometric condition $1/d = (2/\lambda) \sin\theta$ is fulfilled (B). This is the three-dimensional illustration ofBragg's law $\lambda = 2d \sin\theta$ (C).

sphere covers all possible points of the reciprocal lattice, where reflecting planes (reflections) satisfy the Bragg equation. For simplicity, it is often drawn as a circle in two dimensions (**Figure 5C**).

Scattering factors. X-rays are basically scattered by electrons. The X-ray scattering on atom is characterized by the atomic scattering factor f(S):

$$f(S)=\int \rho(\mathbf{r})\exp[2\pi i\mathbf{r}\cdot\mathbf{S}]d\mathbf{r}$$

where $\rho(\mathbf{r})$ is the electron density distribution over the atom and **S** is the scattering vector.

The atomic scattering factor depends on the length |S| rather than the **S** vector direction. The bigger the angle θ (or the higher the resolution), the smaller the scattering factor.

Since atoms vibrate around their equilibrium position, the atomic scattering factors are affected by this vibration and therefore depend on the temperature. To account for atomic and molecular vibrations, the atomic scattering factor can be represented as follows:

$$f'=f\exp[-B\cdot\sin^2/\lambda^2]$$

where B is the Debaye-Waller temperature factor (also known as B-factor), a description of the uncertainty in the position of an atom within a crystal structure. This uncertainty may arise from thermal motion of the atom, leading to variations in the position of the atom between different copies of the unit cell; it may also arise from defects in the observed data. The temperature factor may also be represented by the symbol U, where $B=8\pi^2u^2$, u^2 is the mean-square amplitude of vibration of an atom or ion, and is directly related to the thermal energy.

Thermal motion and positional uncertainties may be isotropic (spheroidal) or anisotropic (ellipsoidal). While isotropic atomic motion is represented by one parameter (sphere radius), the anisotropic atomic displacement is represented by 6 parameters. Isotropic temperature factors should be used in crystallographic refinement at lower resolution. Anisotropic temperature factors may be used at high resolution refinement.

The X-ray radiation scattered by one unit cell - the smallest unit that can generate the entire crystal by translation operations alone - is known as the structure factor and symbolized by **F** or **F**(hkl). It is the Fourier transform of the scattering density (electrons in the molecules) sampled at the reciprocal lattice point hkl. The intensity of the scattered radiation is proportional to the square of the amplitude, $|F|^2$. The structure factor is represented by:

$$\mathbf{F}(hkl)=|\mathbf{F}(hkl)|\cdot\exp[-i\alpha(hkl)]$$

with $|\mathbf{F}(hkl)|$ representing the amplitude of the scattered wave, and α(hkl) its phase relative to the origin of the unit cell. **F**(hkl) can also be written as the sum of contributions from each volume element
of electron density (ρ) in the unit cell:

$$\mathbf{F}(hkl)=\int^{1}_{x=0}\int^{1}_{y=0}\int^{1}_{z=0}\rho(x,y,z)\exp\left[\pi i(hx+ky+lz)\right]dx\,dy\,dz$$

The Fourier transform equations show that the electron density is the Fourier transform of the structure factor and the structure factor is the Fourier transform of the electron density, therefore the electron density can be written as follows:

$$\rho(x,y,z)=1/V\ \Sigma_h\Sigma_k\Sigma_l\ |F(hkl)|\cdot\exp[-2\pi i(hx+ky+lz)+i\alpha(hkl)]$$

While structure amplitudes are directly obtained from measured reflection intensities, the phases are lost. This is known as the crystallographic phase problem. There are several ways of addressing the phase problem, which will be discussed later.

In summary, x-ray crystallography enables visualization of macromolecular atomic structures and facilitates the structure-based study of protein function. Specifically, one can study how proteins interact with other molecules, how they undergo conformational changes, and how they perform catalysis in the case of enzymes. Armed with this information we can design novel drugs that target a particular protein, or rationally engineer an enzyme for a specific industrial process.

2.1.2. Main stages of crystallographic study

Usually, crystallographic study includes several stages, each of which is important for the final result:

- protein expression, isolation and purification
- protein crystallization
- crystal data collection and processing
- model structure determination
- crystallographic refinement of the model
- model validation
- structural analysis of the final model

Often, the first bottleneck in the procedure is the protein purification stage. The problem is to obtain a concentrated (typically 5–15 mg/ml) and pure protein solution. Thus far, there are not obvious correlations between crystallization conditions and protein structure or family, neither sets of rules that could guarantee the production of good crystals. However, there is a variety of ‘factors to play’ to successfully arrange the crystallization of many proteins, including an appropriate choice of protein fragment (or protein-ligand complex) for crystallization, introduction of artificial modifications like ‘protein-surface mutagenesis’ [99], selection of protein-expression system etc.

Firstly, the choice of protein fragment to be crystallized is essential. Many proteins are composed of multiple functional domains with internal or terminal flexible regions. Commonly it is believed (and practiced the approach) that attempts to remove flexible and non-functional parts could be helpful for the crystal growth of the smallest functional protein core. In principle, this approach increases the probability of getting crystals because flexible parts might inhibit the packing of macromolecules in a crystalline array. However, the definition of “smallest functional domain” varies in practice.
Surface entropy reduction mutagenesis [99] is another option. In this method, linear clusters of amino acid side chains with high conformational entropy (e.g., Lys and Glu), which are presumed to lie on the surface of the protein, are replaced by methyl groups (Ala) in an effort to create new epitopes that will facilitate crystallization. A growing number of proteins have been crystallized in this manner, suggesting that the method may be of general utility. Yet, because is impossible to predict which cluster mutant(s) will crystallize, the probability of a successful outcome is proportional to the number of mutants that are screened. Consequently, surface entropy reduction mutagenesis is usually a time- and effort-consumable approach. The choice of expression system can also influence the protein properties. Variation of such features as proper folding, S-S bond formation, post-translational processing (proteolysis, N- and O-glycosylation, acylation, amidation, carboxymethylation, phosphorylation, and prenylation) in different cell type has been reported for many proteins [100].

Finally, isolation and purification steps can influence protein stability and, as a consequence, change important characteristics of a protein sample: purity and concentration. While choosing purification steps (like chromatography), the target molecule stability region (pH, salt concentration, temperature etc) should be taken into account. Purity is usually estimated with gel electrophoresis [101], and concentration can be assessed with spectroscopic methods, like UV-adsorption [102,103].

2.1.2.1. Protein purification

The development of methods for protein purification has been an essential pre-requisite for many advances made in structural biology. More than one purification step is often necessary to reach the desired protein purity. The appropriate choice of techniques, optimization of their performance and correct logical way of their application are main requirements for successful and efficient protein purification, with different chromatography techniques forming a powerful core combinations for purification.

The development of recombinant DNA techniques [104] has revolutionized the production of proteins in large quantities. Recombinant proteins are often produced in forms, which facilitate their subsequent chromatographic purification. However, this has not removed all purification problems: host contaminants, bad solubility, lack of structural integrity or biological activity. Although there may appear to be a great number of parameters to consider, with a few simple guidelines and application of the Three Phase Purification Strategy the process can be planned and performed, with only a basic knowledge of chromatography techniques.

The below guidelines for protein purification, cited from [89], can be applied to many proteins, because it is a sort of a systematic approach to the development of an effective purification strategy:

- Define objectives (requirements for purity, activity and quantity of final product)
- Define properties of target protein and critical impurities (to simplify technique selection and optimization)
- Develop analytical assays (for fast detection of protein activity/recovery and critical contaminants)
- Minimize sample handling at every stage (to avoid lengthy procedures which risk losing activity/reducing recovery)
- Minimize use of additives (additives may need to be removed in an extra purification step or may interfere with activity assays)
- Remove damaging contaminants early (for example, proteases)
- Use a different technique at each step (to take advantage of sample characteristics which can be used for separation - size, charge, hydrophobicity, ligand specificity)
- Minimize number of steps (extra steps reduce yield and increase time)

2.1.2.1.1. *Escherichia coli* strains and expression vectors

There are a number of E. coli strains in use in the laboratory practice. In this work, the most used bacterial expression host was *E.coli* BL21(DE3), which lacks the *lon* cytosolic protease and the *ompT* outer membrane protease [106]. BL21(DE3) contains integrated into its chromosome a copy of the T7 RNA polymerase under the control of the *lac*UV5 promoter. This means that the addition of a *lac* operon inducer such as lactose or IPTG will result in the expression of T7 RNA polymerase. DE3 is actually a bacteriophage λ that has been integrated into the genome (a lysogen) and cannot excise itself because the T7 RNA polymerase gene interrupts the *int* gene required for integration and excision. This makes it a stable source of T7 polymerase that does not need any antibiotic selection for its maintenance.

An expression vector, otherwise known as an expression construct, is generally a plasmid that is used to introduce a specific gene into a target cell. Once the expression vector is inside the cell, the cellular-transcription and translation machinery produces the protein encoded by the gene. The plasmid is frequently engineered to contain regulatory sequences that lead to efficient transcription of the gene carried on the expression vector. These include:

1. Antibiotic resistance gene. Kanamycin (one of the most common) was used in this work. This permits the selection of successfully transformated cells.
2. Origin of replication. The origin of replication (replicon) is the region of DNA that constitutes the binding site for DNA polymerase along with various cis-acting elements.
3. Cloning site. Traditionally, vectors have used a multiple cloning site, which is an area with multiple restriction enzyme sites. This allows the user to pick and choose restriction enzyme combinations for use in cloning the insert into the vector.
4. Promoter. Binding site for RNA polymerase. This is the genetic element that drives the expression of your protein. Most usefully, and most commonly, the promoter is inducible i.e. you can turn on protein expression when you want, usually by the addition of a chemical inducer to the growth media e.g. IPTG. The most commonly used promoters are the φ10 promoter from bacteriophage T7 (the 'T7 promoter') and the trp-lac hybrids trc and tac.
5. Terminator. Prevents the RNA polymerase from carrying on around the plasmid, and transcribing other genes downstream of your target. This can result in proteins being expressed in response to induction, which can disappointingly turn out to be the proteins of antibiotic resistance,

such as β-lactamase or chloramphenicol acetyl transferase. To guard against being deceived by this happening, it's always wise to use an empty vector as a negative control in your protein expression experiment.

6. RBS. The ribosome binding site (Shine-Dalgarno sequence) is where the ribosome attaches to the mRNA for translation.

2.1.2.1.2. Chromatography

Chromatography (from Greek χρωμα chroma "color" and γράφειν graphein "to write") is the collective term for a set of laboratory techniques for separation of mixtures. It involves passing a mixture dissolved in a "mobile phase" through a stationary phase, which separates an analyte from other molecules of the mixture based on differential partitioning between the mobile and stationary phases.

Affinity chromatography separates proteins on the basis of a reversible interaction between a protein (or group of proteins) and a specific ligand coupled to a chromatographic matrix. The technique is ideal for a capture or intermediate step in a purification protocol and can be used whenever a suitable ligand is available for the protein of interest. With high selectivity, hence high resolution, and high capacity for the protein of interest, purification levels in the order of several thousand-fold with high recovery of active material are achievable. Target protein is collected in a purified, concentrated form. In order to achieve higher sample purity several consecutive affinity purification steps can be employed.

Chelating Sepharose, when charged with Ni2+ ions, selectively binds proteins if complex-forming amino acid residues, in particular histidine, are exposed on the protein surface. The 6xHistidine-tag ((His)6-tag) Ni-NTA interaction is based on the selectivity and high affinity of Ni-NTA (nickel-nitrilotriacetic acid) resin for proteins containing an affinity tag of six consecutive Histidine residues at either the carboxyl or amino terminus. (His)6 fusion proteins can be easily bound and then eluted with buffers containing imidazole.

The (His)6 tag is one of the most common tags used to facilitate the purification and detection of recombinant proteins and a range of products for simple, one step purification of (His)6 fusion proteins are available. Polyhistidine tags such as (His)4 or (His)10 are also used. They may provide useful alternatives to (His)6 for improving purification results. For example, since (His)10 binds more strongly to the affinity medium, a higher concentration of eluent (imidazole) can be used during the washing step before elution. This can facilitate the removal of contaminants which may otherwise be co-purified with a (His)6 fusion protein.

Pseudo-affinity heparin chromatorgraphy is often used for nucleic acid binding proteins - an extremely diverse class of proteins sharing a single characteristic, their ability to bind to DNA/RNA. Heparin is a highly sulfated glycosaminoglycan with the ability to bind a very wide range of biomolecules. It has two modes of interaction with proteins and, in both cases, the interaction can be weakened by increases in ionic strength. In its interaction with DNA/RNA binding proteins heparin mimics the polyanionic structure of the nucleic acid. The elution is performed by increasing salt concentration.

Both Ni2+ affinity chromatography and heparin pseudo-affinity chromatography have been used in this work for p19 purification.

2.1.2.1.3. Polyacryamide gel-electrophoresis (PAGE)

SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis, is a technique widely used in biochemistry, forensics, genetics and molecular biology to separate proteins according to their electrophoretic mobility (a function of length of polypeptide chain or molecular weight).

SDS (sodium dodecyl sulfate) is a detergent capable of dissolving hydrophobic molecules and also carrying a negative charge (sulfate) attached to it. Therefore, proteins incubated with SDS, will be solubilized by the detergent, and covered with negative charges. The net charge will thus depend on the length (i.e. molecular weight) of the protein.

If the proteins are then put into electric field, they will all move towards the positive pole at the same rate, with no separation by size. So the proteins have to be moving in an environment that will allow size differentiation. The environment of choice is polyacrylamide, which is a polymer of acrylamide monomers. When this polymer is formed, it turns into a gel. Electricity can be used to pull the proteins through the gel; the entire process is thus called polyacrylamide gel electrophoresis (PAGE).

Following electrophoresis, the gel may be stained (most commonly with Coomassie Brilliant Blue R-250), allowing visualization of the separated proteins. After staining, different proteins will appear as distinct bands within the gel. It is common to run molecular weight size markers of known molecular weight in a separate lane in the gel, in order to calibrate the gel and determine the weight of unknown proteins by comparing the distance traveled relative to the marker.

2.1.2.2. Protein crystallization

Even when pure soluble protein is available, producing the high-quality protein crystals is another bottleneck for structure determination. Crystallization experiment is known as a complex, time-consumable, multi-parametric search. The crystal growth is a process, which is initiated by the formation of ordered nuclei. The process that causes molecules to arrange themselves in ordered nuclei rather than disordered precipitate is not well understood. Since many studies show that crystal formation never occurs prior the saturation stage, the conclusion can be easily made that protein samples should be free of aggregates to serve as an appropriate potential source for crystallization.

The solubility of any protein is limited. Once the limit is reached, the solution is no longer homogeneous, because a new ‘phase’ is appearing. Crystallization experiments are based on this phenomenon. Through variation of the solution conditions, the crystallographer tries to vary the solubility of the protein and cause crystalization by ‘slow pushing’ the experimental system to enter the supersaturation stage. The problems associated with producing protein crystals have stimulated fundamental research on protein crystallization. An important tool of such research is phase diagrams. A complete phase diagram (**Figure 6**) shows the state of a material as a function of all relevant variables of the system. For protein solution, the variables are the concentration of the protein, temperature,characteristics of the solvent (e.g., pH, ionic strength and the

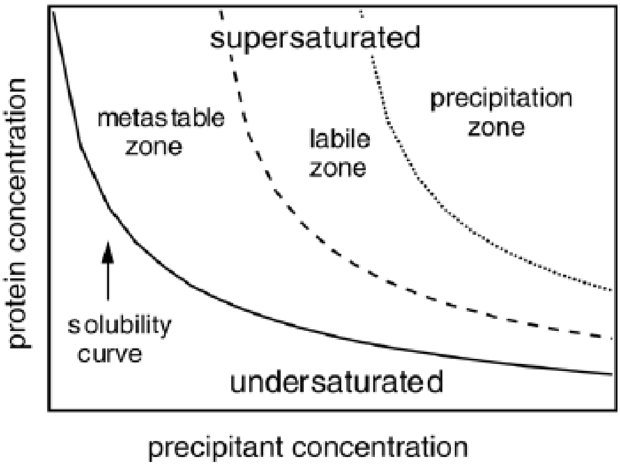


Figure 6. Protein crystallization phase diagram.
The line that separates *undersaturated* conditions from *supersaturated* is known as the *solubility curve.* A crystallization setup that is *undersaturated* or in the *metastable zone* will appear clear, however, the latter has the possibility of crystal growth if seeded. *Precipitation* is when the protein comes out of solution as an aggregate and therefore is not useful for crystallographic studies. The *labile zone* (or nucleation zone) is important since this is where crystal nucleation and initial growth occur. As the crystal forms the protein concentration will be depleted causing one to move from the labile to *metastable zone*.

concentration and identity of the buffer and any additives) etc. The most common form of the phase diagram for proteins is two-dimensional and usually displays the concentration of protein as a function of one parameter, with all other parameters held constant.

In principle, crystal formation could occur in any supersaturated protein solution, when protein concentration exceeds the solubility. In practice, crystals hardly ever form unless the concentration sufficiently exceeds the solubility. The supersaturation is required to overcome the activation energy barrier of nuclei formation and their consequent growth.

Since there is an energy barrier, the nucleation process takes time. If the supersaturation is too small, then the nucleation is slow and no crystal forms for the reasonable time. The corresponding area of the phase diagram is referred to as the “metastable zone” (**Figure 6**). In the “labile” or “crystallization” zone (**Figure 6**), the supersaturation is large enough to let a spontaneous nucleation to occur. If the supersaturation is too large, then disordered structures, such as aggregates or precipitates, may form. The “precipitation zone” is unfavorable for crystal formation, because the aggregates and precipitates form faster than the crystals. However, during the time, crystals are often observed to originate from precipitates, as well.

The three zones are shown in the **Figure 6**. Although the boundaries between them are not well-defined, the concept of different zones in phase diagram is useful in search for the appropriate conditions to produce crystals.

Various crystallization conditions often result in same difficulties of crystallization, namely : (i) the protein solution remains homogeneous and no new phase appears; (ii) a new phase appears, but it is not a crystal; (iii) crystals appear, but they are not suitable for protein structure determination because they produce a poor X-ray diffraction.

Nevertheless, very often the crystallization experiment can result in obtaining of high-quality crystals. It is important to note that crystal quality is mostly supported by the good choice of crystallization conditions and quality of the protein sample used for crystallization experiments.

2.1.2.2.1. Crystallization techniques

There are several common techniques for setting up crystallization experiments ("trials"), including sitting drop vapor diffusion, hanging drop vapor diffusion, batch, dialysis, and free interface diffusion.

The most common setup to grow protein crystals is the vapor diffusion approach, in which a small volume of protein solution are mixed in drops with small amounts of reservoir solutions containing precipitants. A drop of this mixture is put on a glass slide and sealed over the reservoir containing 500-1000ul of solution. Since the precipitant solution in the drop is less concentrated than the solution in reservoir, water is evaporating from the drop into the reservoir, therefore causing the slow elevation of both protein and precipitant concentration in the drop and consequent precipitation or crystallization of the protein.

The methods of hanging, sitting and sandwich drops are based on vapor diffusion. In the case of the sitting drops, larger drop volumes and/or additives lowering the surface tension (e.g. detergents) can be used. Sandwich drops – when the drop is sandwiched between two cover slides – allow for reducing/increasing the drop area exposed to the vapor chamber and therefore reducing/increasing the rate of the diffusion process, respectively.

In the dialysis technique, the sample is placed inside a dialysis cell, which is immersed in solution containing the crystallization agents. The agents diffuse through the cell membrane into the dialysis cell and reduce the macromolecule solubility.

In the batch method, concentrated protein solution is mixed with precipitant solution to produce a finally supersaturated protein mixture that therefore leads to crystallization of the protein. This can be done with batch volumes up to 1 ml or even higher [111], and typically results in larger crystals due to the larger volumes and the lower chance of impurities diffusing to the face of the crystal. This technique is by far the most expensive in terms of consumption of the solute macromolecule.

Liquid-liquid diffusion is performed when protein and precipitant solutions are layered on top of each other allowing a slow equilibration. Nucleation and crystal growth generally occurs at the interface between the two layers, at which both concentrations are at their highest values.

2.1.2.2.2. Crystallization conditions search

Crystallization of proteins is still the bottleneck in structure determination. In addition, there is no simple correlation between properties of the protein and crystallization conditions. Consequently, protein crystallization requires a broad screening of various crystallization conditions. An effective automation of this process was always desirable. Currently, many varieties of high-quality crystallization screens are available from Hampton Research and other companies. Using these screens is a common way of initiating crystallization trials. Although crystallization of different RNA-protein complexes could require different crystallization conditions, a large number of the crystals emerge from relatively limited range of conditions. For instance, the most used precipitants are PEGs, ammonium sulfate, citrate. The divalent ions (Mg2+, Ca2+, Mn2+) are usually employed, since they tend to bind to the RNA sugar-phosphate backbone and influence the protein/RNA binding. Various oligocations (like spermine or spermidine) can affect the crystal quality, as well. Thus, in every particular case, various conditions have to be tried in order to obtain the best crystal. However, the specific crystallization kits were developed for crystallization of RNA-protein (or DNA-protein) complexes.

Crystallization robots (e.g. Mosquito, available in our X-ray platform) make the initial search for crystallization conditions much easier and more effective. With the help of robots, each purified protein sample can be quickly tested for producing crystals in hundreds of different conditions. In addition, robots require very small quantities of protein for setting the complete experiment (since they have very high pipetting precision, the dropvolumes usually belong to the 25-1200nL range).

A CRYSTAL FARM (Bruker), available in our X-ray platform, is an automated protein crystallization and imaging system. It provides constant temperature maintaining for plates with drops during the crystallization time; it checks the test trays for presence of crystals and provides a simple interface for a user to monitor experiments. Tracking of the crystal growth is done by photographing the titer trays. Images are then made available via a web based interface. CRYSTAL FARM also has image recognition software that automatically finds crystals by means of an ingenious search algorithm that looks for solid particles in the digital images.

2.1.2.2.3. Specific features of RNA-protein complexes crystallization

RNA oligoribonucleotides (Dharmacon Research) for crystallization studies have to be chemically synthesized, subsequently deprotected and purified by denaturing polyacrylamide gel electrophoresis.

Also, for preparing an RNA sample for crystallization trials, the oligonucleotides have to be annealed in order to ensure correct secondary and ternary structure. Some authors recommend annealing by slow cooling in Mg2+ [90] or snap cooling on ice in the presence of monovalent (K+) and divalent (Mg2+) ions [91]. RNA can slowly convert to other structural form (e.g. form hairpins or dimers) at relatively low temperatures (10-30°C) so it is essential to set up the crystallization experiment as soon as possible after the annealing. RNA-protein complexes are normally prepared by mixing RNA and protein in appropriate stoichiometric relation. The initial stoichiometry is generally considered to be that of a theoretically expected ratio for complex formation or with a slight excess of RNA over protein. It can be useful to optimize the component ratios; this can be done by examining the complex over a series of ratios using the EMSA (electroforetic mobility shift analysis) to identify the ratio for better crystal quality.

It is also essential that neither the protein nor the RNA buffer contains any trace of RNAse. Overall, it is advised to use the minimal buffer conditions under which the complex is still soluble. Sometimes the solubility of the complex is higher than that of the pure protein, in which case it may be desirable to further concentrate the complex before crystallization.

2.1.3. X-ray data collection

Collecting X-ray diffraction data involves a number of choices and compromises, including choice of crystal, source, rotation range, exposure time and programs for integration and scaling.

2.1.3.1. Crystal harvesting: cryoprotection and cryocooling

To prevent the radiation damage of the protein crystal, data sets are practically always collected with the crystal placed in a stream of cold liquid nitrogen. Frozen crystals usually produce better data than those at room temperature, but the cryoprotection and freezing protocol must be optimized to avoid ice ring formation and minimize increases in mosaicity. One of the possible approaches for such optimization is to grow crystals in the cryoprotectant.

Cryopreservation of protein crystals has at least two advantages over room temperature methods. Firstly, it greatly reduces radiation damage of the crystallized protein, especially when irradiated with higher intensity radiation sources. Secondly, it provides for relatively simple storage and transportation of crystals for remote data collection. Typical cryopreservants include glycerol, sugars (glucose and sucrose), and polyethylene glycols. Cryopreserved crystals are usually stored at liquid nitrogen temperatures (77K). Glycerol (30%) or glucose (25%) is usually sufficient to cryoprotect most crystallized proteins. Lower concentrations of cryoprotectants are necessary in the presence of high concentrations of salts or polyethylene glycol. Generally, if a drop of well solutions vitrifies to a clear glass in a sample loop, the cryoprotectant concentration is sufficient for ice formation suppression. The simplest method of cryoprotection is to simply transfer crystals directly from their mother liquor to a drop of artificial mother liquor with the added cryoprotectant. Soaking in the cryoprotectant drop for as little as 30 seconds is usually sufficient to prevent ice formation or crystal cracking. When trying a cryoprotectant, the mother liquor composition should be first considered. For instance, if the crystal was grown using 2-methyl-2,4-pentanediol (MPD), PEG with molecular weight of less than 1000 as the precipitant, it is most likely ready for freezing without any additional cryo-protectant. Thus the original mother liquor of any kind should be the first to be tested. On the other hand, crystals that grow in salt should be washed in a solution containing a cryoprotectant, or may need to go through a complete mother liquor exchange with a cryoprotected solution in order to be cooled to cryo-temperatures. In general the best choice of cryoprotectant is the one that most closely resembles the composition of the mother liquor unless that is mostly salt. If the crystal grows in PEG, then ethylene glycol, the "monomer" of PEG, would be a good first choice. One may then try glycerol and even low molecular weight PEG. If the crystal grows in low MPD concentration, a higher MPD concentration would be the first option and so on. The nature of the search is trial and error.

2.1.3.2. X-ray radiation sources

X-rays of a suitable wavelength range for protein crystallography (~0.8 - 2.3 Å) are generated by three commonly used devices: X-ray tubes, rotating anodes and synchrotrons.

X-ray tubes consist of a filament that acts as a cathode. Electrons are emitted by the glowing cathode and accelerated across the vacuum towards the anode, which consists of a metal target made of a specific material, usually copper for protein crystallography. As the electron beam impacts the anode, the high kinetic energy of the electrons is converted during deceleration into X-rays producing a) a continuous spectrum consisting of "braking radiation" and b) emission lines, Kα and Kβ, characteristic for electronic transitions caused in the copper anode material. The Kα and Kβ emissions (1.54 and 1.39 Å, respectively) have an intensity that is several orders of magnitude higher than the "braking radiation".

The X-rays are filtered to a single wavelength of Kα (made monochromatic). The filtering out of the "braking radiation" and other emission lines (Kβ) is done by filters, monochromators or X-ray mirrors. This simplifies the data analysis, and also removes radiation that degrades the crystal. The wavelength of the filtered X-ray in copper anode sources is 1.54Å. The resulting monochromatic beam is collimated and focused onto the crystal. Collimation is done either with a collimator (basically, a long tube) or with an arrangement of curved mirrors.

When X-rays are produced by a rotating anode, the cathode and anode are housed under vacuum, in which the anode target rotates at high speed to efficiently distribute and dissipate heat. The wavelength of an in-house source such as a tube or rotating anode generator is fixed by the choice of anode target material and not tunable, as is the case at a synchrotron and the intensity of the source is less than that of a synchrotron.

For example, the MICROSTAR generator of the Bruker X8 PROTEUM X-ray system is equipped with a copper rotating anode, Montel multilayer optics to assure monochromatic X-rays. It has a 2.7 kW anode power loading on a 100 µm focal spot (27 kw/mm2) and ultra-high intensity: up to 8 x 10^{10} X-rays/mm^2 -sec.

At a synchrotron facility, bunches of electrons, several GeV in energy, move in a large, carefully steered, closed electron beam loop containing bending elements and linear segments, collectively called the storage ring. In each section, magnetic devices are inserted - bending magnets in the curved sections, insertion devices called wigglers and undulators in the straight sections - to bend, wiggle or undulate the path of the electrons while they pass around the ring. Due to the acceleration experienced in the bending magnets or insertion devices, the electrons emit a narrow fan of intense white (polychromatic) radiation ranging from soft UV to hard X-rays over a very tightly defined angle tangential to the ring. The radiation is 'tunable' by cutting out fine bands (few eV or 10-5 Å wide) of wavelengths appropriate for particular experiments with monochromator crystals that selectively pass the wavelength of choice. The intensity of X-rays generated by modern third generation synchrotron sources is so high that radiation damage of crystals has become a major concern, and this has given rise to the near-exclusive use of cryo-crystallographic techniques, in which crystals are kept at near-liquid nitrogen temperatures to minimize radiation damage. Synchrotron radiation has additional features that make it attractive for advanced applications. Because it is pulsed, it can be exploited for examining time-dependent phenomena, and because it is highly polarized, it can be used to examine polarization-dependent and angle-dependent effects.

The synchrotron beams emitted by the electrons are directed towards the "beamlines" which surround the storage ring in the experimental hall. Each beamline is designed for use with a specific technique or for a specific type of research. Experiments run throughout the day and night. Each beamline includes: an optics cabin, housing the optical systems used to "tailor" the X-ray beam to have the desired experimental characteristics; an experimental cabin which contains the support mechanism and sample environment for the sample to be studied. One or more detectors record the information produced as a result of the interactions between the X-ray beam and the sample; a control cabin which allows the researchers to control their experiments and to collect the data.

The European Synchrotron Radiation Facility (ESRF) is a joint research facility supported by 19 countries (18 European countries and Israel) situated in Grenoble, France. The ESRF operates the most powerful synchrotron radiation source in Europe, and is generally considered to be a world leading research facility. It has an annual budget of around 80 million euros, employs over 600 people and is host to more than 3500 visiting scientists each year.

Research in the ESRF focuses, in large part, on the use of X-ray radiation in fields as diverse as protein crystallography, earth science, materials science, chemistry and physics. Facilities such as the ESRF offer a flux, energy range and resolution unachievable with conventional (laboratory) radiation sources.

The ESRF physical plant consists of two main buildings: the experiment hall, containing the 844 meter circumference ring and forty tangential beamlines; and a block of laboratories, preparation suites, and offices connected to the ring by a pedestrian bridge. The linear accelerator electron gun and smaller booster ring used to bring the beam to an operating energy of 6 GeV are constructed within the main ring.

The ESRF site forms part of the "Polygone Scientifique", lying at the confluence of the Drac and Isère rivers about 1.5km from the centre of Grenoble. It is served by local bus and the Lyon airport coach, which stops at the Place de la Résistance just outside the site.

Currently, the most used X-ray sources in macromolecular crystallography are rotating anodes and synchrotrons because of their higher power output. High-intensity radiation is of value when collecting data from weakly diffracting and/or small crystals (and macromolecular crystals tend to have low diffraction compared to small inorganic crystals). Output form X-ray tubes is about 20kW limited by the amount of heat that can be dissipated from the anode by circulating water. Higher X-ray power can be obtained from rotating anode sources (about 100kW), where the powerful electron bombardment is spread over a much larger piece of metal. Rotating anode sources are more than ten times more powerful as tubes with fixed anodes.

Synchrotrons are the most powerful X-ray sources. Although synchrotron sources are available only at storage rings and require the crystallographer to collect data away from the usual site of work, there are advantages that compensate for the inconvenience. Firstly, X-ray data that require several hours of exposure to a rotating anode source can often be obtained in seconds or minutes at a powerful synchrotron source. Another advantage is that X-rays of selectable wavelength can be helpful in solving the phase problem with multi-wavelength anomalous diffraction (MAD) method.

2.1.3.3. X-ray detectors

X-ray experiments in macromolecular structural biology involve the measurement of intensities for separated reflections. That is why the single canal X-ray diffractometers are used in structural biology not very often. Much better results are arising from usage of area detectors. Area detectors are just any type of X-ray detectors that can collect diffraction information on array at once. There are two most common types of area detectors used: Charge Coupled Device or CCD detectors and Image Plate detectors.

CCD detectors are now used in a variety of ways for X-ray imaging. They are available with up to few thousand pixels, and this size is permanently grown up.

From the other hand the pixel size is permanently reduced and is about 10-50μm with readout times of less than 1 s. In most scientific applications, CCD detectors are cooled to below -30ºC to reduce background noise. In most systems, a thin phosphor screen converts the incident X-rays into optical photons, which the CCD detects. A commonly used phosphor is $Gd_2O_2S(Tb)$, which has a high efficiency and a light decay time of a few hundred microseconds. When used as a detector for macromolecular crystallography, a large phosphor screen (up to thousand mm^2) is usually coupled to the CCD with a tapered optical fiber. Under the light beam, the CCD chip is charged and this electric charge is read out and digitized. Furthermore, modern detectors are based on modular design, where CCD modules can be tightly stacked into 2 x 2 or larger arrays to increase the active area without adding additional readout time (all the modules are read out in parallel).

The heart of the Image Plate is a storage phosphor screen. When the storage phosphor is exposed to X-rays, secondary electrons are trapped in so-called metastable F-centers. The number of F-centers produced is proportional to the X-ray energy. The most common storage phosphors are the barium fluorohalides (mixtures of BaF and BaBr or BaI). After the exposure, these metastable centers can be excited by a read laser to release visible photons in a process known as photostimulation or bleaching. These photostimulated photons can then be detected by an appropriate detector (typically a photomultiplier tube with a multilayer filter to reject the scattered read laser light). By scanning the laser across the surface of the plate it is thus possible to determine the integrated number of X-rays incident at each location.

The biggest advantage of the Image Plate scheme is that it allows a relatively large active area (up to 345 mm diameter). Perhaps, their biggest disadvantage is their relatively long readout time since it typically takes several tens of microseconds to bleach each pixel on the image plate. The total readout time for the entire plate is typically on the order of 1 to 2 minutes. This long readout time is a serious disadvantage in experiments at synchrotron beamlines. The other principle disadvantage of the Image Plate is its relatively low sensitivity.

2.1.3.4. Data collection strategy and processing

The physical process of diffraction from a crystal involves the interference of X-rays scattered from the electron clouds around the atomic centers. The ordered repetition of atomic positions in all unit cells leads to discrete peaks in the diffraction pattern. The geometry of this process can alternatively be described as resulting from the reflection of X-rays from a set of hypothetical planes in the crystal; the visualization of this is the Ewald sphere. It is apparent, that for a stationary crystal in any particular orientation (the so-called “still” exposure, see **Figure 7**), only a fraction of the total number of Bragg reflections will satisfy the diffracting condition of $n\lambda = 2d \sin\theta$ □□The number depends on the density of the reciprocal lattice and hence on the unit-cell dimensions. A small-molecule crystal with short unit-cell dimensions and a sparsely populated reciprocal lattice may not give rise to any diffraction in some orientations. Crystals of macromolecules have unit-cell dimensions much larger than the wavelength of the radiation used, and several reciprocal lattice points (reflections) will lie on the surface of the Ewald sphere in any crystal

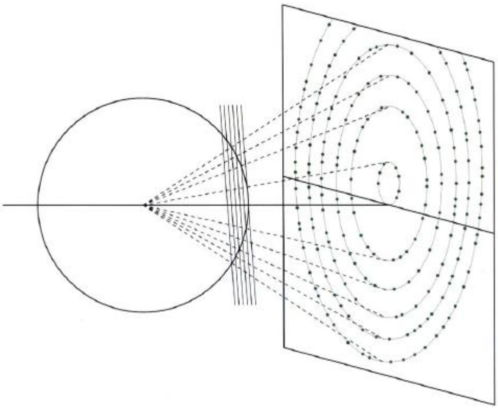


Figure 7. Stationary crystal in an X-ray beam.
A still exposure with a stationary crystal contains only a small number of reflections arranged in a set of narrow ellipses.

orientation.

In general, to observe the diffraction from a number of reflections, the reciprocal-lattice points have to be moved to the surface of the Ewald sphere or the sphere radius has to be changed so that different reflections will lie on its surface. The approach, using a constant Ewald sphere and therefore a selected wavelength (monochromatic radiation), requires that the crystal be rotated to bring successive reflections into diffraction. If the crystal is only rotated about a single axis, this is called the rotation method; this is the most common procedure used for recording diffraction data in macromolecular crystallography.

The main purpose of a crystallographic data collection is to extract the required structural information from a crystal, given finite available experiment time and the limited crystal lifetime in an X-ray beam. Incorrect choice of data collection strategy can lead to a failure of the experiment.

Both the software and the hardware provide the possibility of varying numerous parameters in order to optimize the quality of the data set. The influence of parameters, the most important of which are listed below, are of importance for the quality of the data.

Some redundancy produces more accurate data and allows for reliable rejection of outliers. It is an ancient principle of accurate measurement to measure something many times and take the average. With a fast read-out detector such as a CCD, collection of 180 or even 360 of data is reasonably fast and this also simplifies strategy.

Completeness, both in geometrical coverage of reciprocal space and the full intensity range is very important. Systematic omission of data will distort all maps. The geometric strategy may be complicated if the detector is not centrally placed on the beam; however, strategy simulations are available in a number of programs and should be used.

The maximum resolution of a data set may be reset after examination of data-reduction statistics. To collect the data, the detector may be positioned a little closer than the apparent maximum resolution, provided that the spots are resolved.

Exposure time needs to be set to long enough to give reasonable statistics at the highest resolution, but not so long as to overload the detector with the strong low-angle spots, nor to give too much radiation damage. More than one pass with different exposure times may be required to catch the full dynamic range of data.

Rotation width per image should be set to resolve the longest axis on rotation, taking into account the reflection width. Narrower image widths may improve data quality.

The selection of data-acquisition parameters in the case of protein crystals is always a compromise between many requirements. Each approach has its advantages and disadvantages in terms of experimental constraints and goals. Influence of data-acquisition parameters on the data quality and quantitative estimations of relationships between these parameters and data collection statistics can be analyzed with special data collection strategy programs (BEST, Bruker Strategy etc). It is possible to significantly minimize the time of X-ray data collection by correct prediction of strategy of data collection.

For rotating the crystal in X-ray beam a special loading mechanism called goniometer is used. Goniometers are very precise mechanics and by means of three rotation axes - allow crystal samples to be brought to any orientation in space, fulfilling Ewald's requirements to produce diffraction. All these movements can be programmed in an automatic mode, with minimal operator intervention. The Bruker X8 PROTEUM X-ray system possesses a *Kappa* goniometer head, which is designed for orienting crystals on the data collection axis, i.e. centring and alignment of the crystal. The crystals should be mounted on a sample holder with a constant flow of cooled nitrogen for keeping 100K temperature during data collection. The goniometer allows re-orienting crystals while keeping the crystal in the X-ray beam, gain better spot shapes from deformed crystals (e.g. bent crystals) and adjust for highest completeness of collected data.

2.1.3.4.1. Crystals and symmetry.

A mostly complete dataset can be collected on any crystal by rotating the crystal through 180 degrees solid angle. If the detector can be rotated around (in 2θ, like in PROTEUM system) to collect higher resolution data it can be necessary to collect more than 180 degrees of data to compensate for this. A small amount of data will be lost in the so-called blind region due to the curvature Ewald sphere: and lies along the rotation axis in a curve bi-conical shape. This region is often effectively collected elsewhere by virtue of crystallographic symmetry (except in the case of space group P1 where you need to re-orient the crystal to collect this data).

If the highest symmetry axis in the crystal is N-fold, then the minimum number of degrees that will have to be collected is 180/N. This is the minimum value - if the crystal is in a non-optimal orientation more data should be collected. Theoretically, the best orientation is with the highest symmetry axis almost aligned with the rotation axis of data collection. The worst orientation is with it aligned perpendicular to that axis. Consequently, for a successful data collection one must consider the symmetry of the crystal.

Crystals are made up of identical parallelepiped-shaped blocks called unit cells that constitute a three dimensional translation lattice (**Figure 8**).

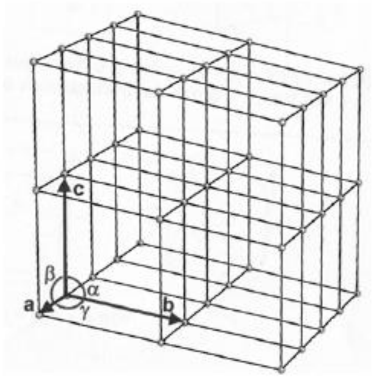


Figure 8. A three dimensional translation lattice.
The translations in a three-dimensional lattice may be described in terms of three linearly independent, i.e. non coplanar, vectors, **a**, **b** and **c**. The angles between the pairs of vectors **b** and **c**, **c** and **a**, and **a** and **b** are defined as α, β, and γ, respectively.

The cell is defined by the vectors ***a***, ***b*** and ***c***; they define the length |a|, |b|, |c|, and the angles α, β, γ which characterize the unit cell. The volume V of the unit cell can be calculated as follows:

$V=abc[1-\cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cdot\cos\beta\cdot\cos\gamma]^{1/2}$

The unit cell is the smallest unit that can generate the entire crystal by translation operations alone. The content of the unit cell is obtained by repetition of its part, referred to as the asymmetric unit (AU), through the symmetry elements. Within the cell there can be several symmetry related asymmetric units with identical contents, but in general in different orientations. All possible 230 symmetry groups with their operators are described in volume A of the *International Tables for Crystallography*. For biological macromolecules, there are 65 possible space groups because of their chirality (i.e. reflection symmetries cannot be observed in such molecules).

To define the planes in the crystal, the Miller indices (h, k, l) have been introduced. The h, k, l terms define parallel planes with intercepts a/h, b/k, c/l on the three a, b, c axes of the unit cell with h, k, l small integer numbers. For example, the (234) planes, shown in **Figure 9**, cut the unit cell edges: **a** into two parts, **b** into three parts and **c** into four parts.

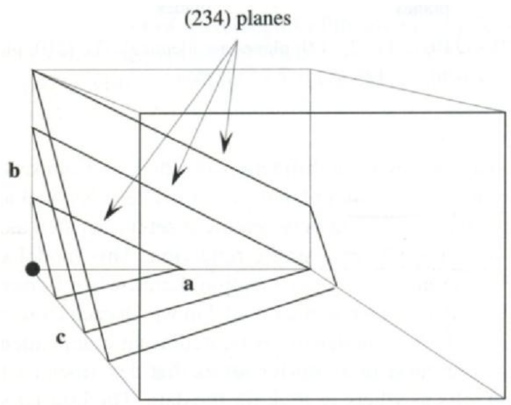


Figure 9. Geometrical model to interpret parallel planes with indices *hkl*

The intersection of three (234) planes with a unit cell. It can be observed that the (234) planes cut the unit cell edges a into two parts, b into three parts and c into four parts.

The symmetry of each finite object such as a molecule can be described by a self- consistent set of symmetry operations called a point group. The point group is thus the name given to the collection of symmetry elements of a finite object. There are 32 classes of point groups, given by the combination of the following symmetry elements:

- Mirror plane, which does not occur in crystals of proteins and DNA because they are chiral molecules.
- Rotation axis, characterized by a rotation about one axis of 360°/N where N can be 1, 2, 3, 4 or 6.
- Inversion point, which does not occur in crystals of proteins because they are chiral molecules.

By analysis of the rotational symmetry, crystals can be divided into seven crystal systems with defined characteristics and parameters.

There are fourteen Bravais Lattices which are categories of translation lattices: they can be primitive (P), body centred (I), face centred (F) and C lattices in the case of Monoclinic and orthorhombic systems having a position on the (001) face. The seven crystal systems and the fourteen Bravais Lattices are shown in **Table 2**.

Other symmetry elements are:

- Glide plane. Obtained by a combination of a mirror plane and a translation, it isnot possible for chiral crystals.
- Screw axis. A rotation is combined with a translation parallel to the rotation axis. The molecule is shifted by a fraction of unit cell and rotated.

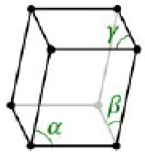
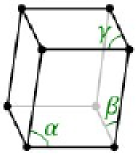
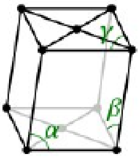
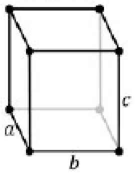
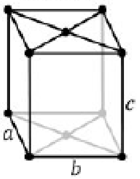
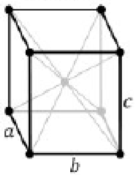
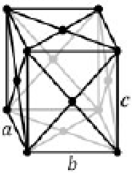
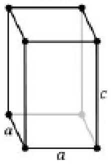
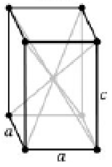
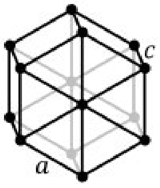
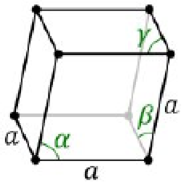
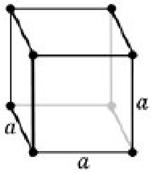
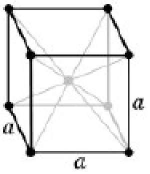
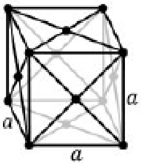
The combination of the 32 point groups with the Bravais Lattice and the screw axis and glide plane symmetry operations gives rise to 230 space groups of which only 65 are possible for chiral molecules.

The diffraction pattern of a crystal exhibits the same crystal symmetry but with an additional centre of symmetry, in the absence of anomalous scattering. The diffraction pattern symmetries are grouped in 11 Laue classes. The presence of symmetry elements like screw axes can be detected since they give rise to systematic absences in the diffraction pattern. The space group can often, but not always, be found

unambiguously considering the Bravais Lattice, the Laue symmetry and the systematic absences.

A special case of symmetry is non-crystallographic symmetry (NCS) through which the molecules within one asymmetric unit are related by appropriate operations.

Table 2. The Bravais Lattices.

The seven crystal systems	The fourteen Bravais Lattices
Triclinic $a \neq b \neq c$ $\alpha \neq \beta \neq \gamma$	$\alpha, \beta, \gamma \neq 90°$ 
Monoclinic $a \neq b \neq c$ $\alpha = \gamma = 90° \neq \beta$	P $\alpha \neq 90°$, $\beta, \gamma = 90°$  C $\alpha \neq 90°$, $\beta, \gamma = 90°$ 
Orthorhombic $a \neq b \neq c$ $\alpha = \beta = \gamma = 90°$	P $a \neq b \neq c$  C $a \neq b \neq c$  I $a \neq b \neq c$  F $a \neq b \neq c$ 
Tetragonal $a = b \neq c$ $\alpha = \beta = \gamma = 90°$	P $a \neq c$  I $a \neq c$ 
Hexagonal $a = b \neq c$ $\alpha = \beta = 90°$, $\gamma = 120°$ Rhombohedral $a = b = c$ $\alpha = \beta = \gamma$	R $a \neq c$  P $\alpha, \beta, \gamma \neq 90°$ 
Cubic $a = b = c$ $\alpha = \beta = \gamma = 90°$	P  I  F 

There are seven crystal lattice systems. In addition, the lattices can be primitive (only one lattice point per unit cell) or non-primitive (more than one lattice point per unit cell). Combining the seven crystal systems with the two lattice types yields the 14 Bravais Lattices (3 different cubic types, 2 different tetragonal types, 4 different orthorhombic types, 2 different monoclinic types, 1 rhombohedral, 1 hexagonal, 1 triclinic).

2.1.3.4.2. Parameters for estimation of collected data

To estimate the collected data quality, apart from completeness and redundancy, two more parameters are considered:

(i) R_{merge} - a measure of agreement among multiple measurements of the **same** (not symmetry-related) reflections, with the different measurements being in different frames of data or different data sets. R_{merge} is calculated as follows (I_i is the ith intensity measurement of reflection h, and <I> is the average intensity from multiple observations):

$$R_{merge} = \frac{\sum_h \sum_i \left| I_i - \langle I \rangle \right|}{\sum_h \sum_i I_i}$$

(ii) I/sigmaI – a signal-to-noise ratio.

2.1.3.4.3. Data processing

Data processing in macromolecular crystallography is the effort by which a user converts a set of raw diffraction data into a list of Bragg reflections with measured intensities. With modern crystallographic hardware the raw data consists of a set of two-dimensional detector images, each collected at a particular orientation of the crystal. Data processing may be broken down into four steps: calibration, determination of the unit cell, measurement of the integrated intensities, and merging and scaling the integrated data. Algorithms for each of these steps have been derived and validated, and are implemented in several commercially available software packages.

In recent years these packages have become more flexible and easier to use. For example, HKL2000-package is a set of programs designed for monocrystal X-ray diffraction data analysis. It consists of several subprograms coordinated by the graphical command center: XdisplayF for visualization of the diffraction pattern, Denzo for data reduction and integration, and Scalepack for merging and scaling of the intensities obtained by Denzo or other programs.

2.1.4. Structure determination

When waves are diffracted from a crystal, they give rise to diffraction spots. Each diffraction spot corresponds to a point in the reciprocal lattice and represents a wave with an amplitude and a relative phase. Photons are reflected from the crystal in different directions with a probability proportional to the square of the amplitude of this wave. The photons are counted, but the information about the relative phases of different diffraction is lost. **Figure 10** shows how the phase and amplitude of the overall scattered wave arise from the individual scattered waves.

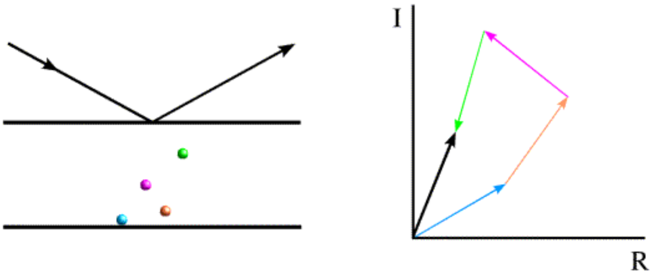


Figure 10. The structure factor.

Two Bragg planes are shown, together with four atoms. The relative phase (from 0 to 360 degrees) depends on the relative distance of the atoms between the planes that define a phase angle of zero. The atoms and their contributions to the scattering (represented as vectors) are shown in matching colors. The overall scattered wave is represented by a black vector, which is the sum of the other vectors. The overall scattering from a particular set of Bragg planes is termed the structure factor, and it is usually denoted **F**. Here, it is represented as a black vector.

The vector (amplitude and phase or, more properly, the complex number) representing the overall scattering from a particular set of Bragg planes is termed the structure factor, and it is usually denoted **F**. (The use of bold font indicates that it is a vector or complex number.) The structure factors for the various points on the reciprocal lattice correspond to the Fourier transform of the electron density distribution within the unit cell of the crystal. If an inverse Fourier transform is applied to the structure factors, the electron density can be calculated.

In the beginning, crystallographers worked on the structures of simple molecules and they could often make a good guess of the conformation of a molecule and even its packing in the crystal lattice. The guesses could be tested by calculating a diffraction pattern and comparing it to the observed one. If a hypothetical positions of the atoms are close to the real, then the calculated phases are approximately correct and a useful electron density map can be computed by combining the observed amplitudes with the calculated phases. If the model is reasonably accurate, such a map will show features missing from the model so that the model can be improved.

For proteins, a guess about how the structure will look like can only be made if a closely-related protein structure was solved before and even then efforts have to be made to find the correct orientation and location of the molecule in the unit cell. In principle, three techniques exist for solving the phase problem: the isomorphous replacement method, the multiple wavelength anomalous diffraction method, the molecular replacement method.

2.1.4.1. Multi-wavelength anomalous dispersion (MAD)

Multi-wavelength anomalous dispersion (sometimes Multi-wavelength anomalous diffraction; abbreviated MAD) is a technique used in X-ray crystallography that facilitates the determination of the structure of proteins or other biological macromolecules by allowing the solution of the phase problem. This is possible if the structure contains one or more atoms that cause significant anomalous scattering from incoming X-rays at the wavelength used for the diffraction experiment. Atoms in proteins which are suitable for this purpose are sulfur or heavier atoms, for example metal ions in metalloproteins. The most commonly used atom for phase determination via MAD, is selenium. The amino acid methionine can be substituted with selenomethionine by using selective media during protein expression. The use of the MAD technique in an experiment utilizing different wavelengths of X-rays generated at a synchrotron is an alternative to the original and still used method of phase determination via Multiple isomorphous replacement (MIR), which involves the preparation of heavy atom derivatives in a trial-and-error approach.

2.1.4.2. Multiple isomorphous replacement (MIR)

In isomorphous replacement, the idea is to make a change to the crystal that will perturb the structure factors and, by the way that they are perturbed, to make some deductions about possible phase values. It is necessary to be able to explain the change to the crystal with only a few parameters, which means that heavy atoms have to be used (atoms with large atomic number, *i.e.* many electrons). The introduction of a heavy atom will change the scattered intensity significantly. One reason for this is that "heavy" atoms contribute disproportionately to the overall intensity. On the other hand, all of the electrons in a heavy atom will scatter essentially in phase with one another. Because of this effect, different atoms contribute to the scattered intensity in proportion to the square of the number of electrons they contain. For example, a uranium atom contains 15 times as many electrons as a carbon atom, so its contribution to the intensity will be equivalent to that of 225 carbon atoms. As a result, the change in intensity from the addition of 1 uranium atom to a protein of 20kDa is easily measured.

In the case of two crystals, one containing just the protein (native crystal) and one containing in addition bound heavy atoms (derivative crystal), diffraction data from both can be measured. The differences in scattered intensities will then reflect the scattering contribution of the heavy atoms, and these differences can be used to calculate their location in the crystal. Their contribution to the structure factors can be computed. This allows to make deductions about possible values for the protein phase angles.

2.1.4.3. Molecular replacement (MR)

Molecular replacement can be used when there is a good model for a reasonably large fraction of the structure in the crystal. Usually, molecular replacement can be applied if the model is fairly complete and shares significant (according to some authors, at least 40%) sequence identity with the unknown structure. It becomes progressively more difficult as the model becomes less complete or shares less sequence identity.

To carry out molecular replacement, the model structure has to be placed in the correct orientation and position in the unit cell. To orient a molecule, three rotation angles are to be specified; to place it in the unit cell, three translational parameters are calculated. So if there is one molecule in the asymmetric unit of the crystal, the molecular replacement problem is a 6-dimensional problem which can be separated into two 3D problems. A rotation function can be computed to find the three rotation angles, and then the oriented model can be placed in the cell with a 3D translation function.

An understanding of the rotation and translation functions can be obtained most easily by considering the Patterson function. Even though the vectors are unresolved for a structure, the size of a protein, the way that they accumulate can provide a signature for a protein structure. The vectors in the Patterson map can be divided into two categories. Intramolecular vectors (from one atom in the molecule to another atom in the same molecule) depend only on the orientation of the molecule, and not on its position in the cell, so these can be exploited in the rotation function. Intermolecular vectors depend both on the orientation of the molecule and on its position so, once the orientation is known, these can be exploited in the translation function.

2.1.5. Crystallographic refinement

Models from a structure solution often give only a partial set of atoms in the unit cell. However this partial set of atoms can contain sufficient phase information to allow the user for location of the remaining atoms. From the atom types and relative positions in the initial model, a set of structure factors can be calculated.

The initial structural model contains errors that can be minimized through iterative model refinement. This is a process of adjustment of the atomic coordinates of the model in order to minimize the difference between experimentally observed structure factor amplitudes (Fobs) and those calculated from the model (Fcalc).

2.1.5.1. Minimization function. Other refinement parameters

Least-squares refinement

An optimization algorithm is used to minimize a target function by changing the parameters of the model. The function to minimize in least-squares refinement was given above in the general form

$S = \Sigma\ w(Yo\text{-}Yc)^2$

where *w* is an assigned weight reflecting the importance that this reflection makes to the sum. The weights usually represent an estimate of the precision of the measured quantity. The sum is taken over all measured reflections and the quantity *Y* is referred to as a measure of the strength of a reflection. In practice, *Y*, sometimes known as the structure-factor coefficient, may be either *I*, the intensity of the measured reflection, | F |, the magnitude of the structure factor, or F^2, the square of the structure factor.

Refinement against I, the measured intensities, has the merit of using the raw measurements directly, although it requires the incorporation in the refinement of the correction factors (scale factor, Lorentz–polarization and absorption) that are applied during standard data reduction. There are, however, problems of high statistical correlation when refining absorption parameters against anisotropic displacement parameters.

Refinement against $| F |$ involves mathematical problems with very weak reflections or reflections with negative measured intensities. There are also difficulties in estimating standard uncertainties $\sigma(F)$ from the $\sigma(F^2)$ values for weak or zero measured intensities.

Refinement against F^2 avoids these difficulties, and also reduces the probability of the refinement iterations settling into a local minimum. It also simplifies the treatment of twinned and non-centrosymmetric structures. For these reasons, it is probably currently the most frequently used technique, although it does rely heavily on the assignment of reasonable weights to individual reflections.

Maximum likelihood refinement

The principle of maximum likelihood formalizes the idea that the quality of a model is judged by its consistency with the observations. If a model is consistent with an observation, then -- if the model were correct -- there would be a high probability of making an observation with that value. For a set of relevant observations, the probability of generating such a set is an excellent measure of the quality of the model. For independent observations, the joint probability of making the set of observations is the product of the probabilities of making each independent observation.

In crystallography, let $P(| F_o | ; | F_c |)$ represent the probability of obtaining an observed structure factor F_o given a calculated value F_c. The joint probability is the likelihood function L:

$$L = \prod_{hkl} P(\,|F_o|\,;\,|F_c|\,)$$

Since it is more convenient to work with sums than products, one typically works with the negative logarithm of the likelihood function

$$L=-\Sigma_{\mathrm{hkl}} \log P(\,|F_o|\,;\,|F_c|\,)$$

The mathematical procedure for determining maximum likelihood then becomes that of minimizing *L*.

Restraints in refinement

The parameters being refined in a crystal structure determination are the *x, y,* and *z* positional parameters and the *U* isotropic or the six $U_{\mathrm{i,j}}$ anisotropic parameters for each atom. A typical refinement of *k isotropic* atoms would utilize 4 *k* atom parameters, 3 positional and 1 displacement parameter per atom. A typical refinement of *m anisotropic* atoms would require 9 *m* parameters, 3 positional parameters and 6 anisotropic parameters per atom. In addition to these atomic parameters, one overall scale factor K and B-factor are refined. This scale factor K accounts for a variety of items from the size of the crystal to the intensity of the radiation source, while B-factor accounts for the changes in K scaling depending on resolution.

The problem of how to perform least-squares structural refinement from X-ray diffraction measurements, taking into account the subsidiary structural information available (known bond lengths, bond angles *etc.*), was debated back in the 1960s. The question was whether to use *constraints* (precise specifications) or *restraints* (flexible specifications). Constrained refinement is used, for example, in some Molecular Reaplacement protocols (this is called rigid body refinement) to reposition the whole model in the unit cell without any further adjustment of individual atom coordinates. It will usually require further refinement with the use of restraints, where a degree of tolerance is built into the bond lengths and angles. Constraints have, in fact, been used sparingly in the past 40-50 years; restraints, on the other hand, have been used abundantly and are still widely employed.

In macromolecular crystallography, restraints are used to introduce *a priori* chemical knowledge in order to keep the model chemically correct while fitting it to the experimental data at lower resolution (the less is the resolution, the stronger becomes the weight W):

$$E_{\mathrm{TOTAL}} = W * E_{\mathrm{DATA}} + E_{\mathrm{RESTRAINTS}}$$

where

$$E_{\mathrm{RESTRAINTS}} = E_{\mathrm{BOND}} + E_{\mathrm{ANGLE}} + E_{\mathrm{DIHEDRAL}} + E_{\mathrm{PLANARITY}} + E_{\mathrm{NONBONDED}} + E_{\mathrm{CHIRALITY}} + E_{\mathrm{NCS}} + E_{\mathrm{RAMACHANDRAN}} + E_{\mathrm{REFERENCE}} + \ldots$$

With higher resolution the restrains contribution decreases – a molecule can be completely unrestrained for well ordered parts at subatomic resolution. Typically, each term in $E_{\mathrm{RESTRAINTS}}$ is a harmonic (quadratic) function:

$$E = \Sigma\, w * (X\mathrm{model} - X\mathrm{ideal})^2$$

and weight $w = 1/\sigma(X)^2$ is the inverse variance, in least-squares methods (e.g.

0.02Å for a bond length)
Making σ(X) too small is not equivalent to constraints, but will make weight infinitely large, which in turn will stall the refinement.

The weights used in least squares refinement are generally chosen to represent the relative influence an observation should have on the results. Weights typically include some term representing the statistical error of the measured data.
Usually, sources for restraints are (i) libraries created out of small molecules that are typically determined at much higher resolution, use of alternative physical methods (spectroscopies, etc); (ii) analysis of macromolecular structures solved at ultra-high resolution; (iii) pure conformational considerations (Ramachandran plot) or (iv) quantum-chemical calculations.

At low resolution the electron density map is not informative enough and a set of local restraints are insufficient to maintain known higher order structure (secondary structure), and the amount of data is too small compared to refinable model parameters. In this case it is useful to bring in more information in order to assure the overall correctness of the model:

- Reference model or point
- Secondary structure restraints (e.g. H-bond restraints for alpha helices, beta sheets, RNA/DNA base pairs)
- Ramachandran restraints (although they should never be used at higher resolution, since it is one of the few precious validation tools)
- NCS restraints/constraints (if there are multiple copies of a molecule/domain in the asymmetric unit that are assumed to have similar conformations).

Refinement Statistics

One way to judge how well the model fits the observed data is to calculate discrepancy factors. The progress of refinement is monitored using the conventional crystallographic indices R factor and R-free. The R factor is defined by the equation:

$$R_1 = \sum |(|F_o| - |F_c|)| / \sum |F_o|$$

R factor is a measure of agreement between the amplitudes of the structure factors calculated from a crystallographic model and those from the original X-ray diffraction data. The *R* factor is calculated during each cycle of least-squares structure refinement to assess progress. The final *R* factor is one measure of model quality.

Theoretical values of *R* range from zero (perfect agreement of calculated and observed intensities) to about 0.6 for a set of measured intensities compared against a set of random intensities. *R* factors greater than 0.5 (or 50%) indicate very poor agreement between observed and calculated intensities, and many models with $R \geq 0.5$ will not respond to attempts at improvement. An early model with $R \leq 0.4$ can usually be improved during refinement. A desirable target *R* factor for a protein model refined with data to 2.5 Å is considered to be ~0.2. Small organic molecules commonly refine to $R < 0.05$. However, the *R* factor must always be treated with caution, as an indicator of precision and not accuracy. Partially incorrect structures have been reported with *R* values below 0.1; many imprecise but essentially correct structures have been reported with higher *R* values.

A residual function calculated during structure refinement in the same way as the conventional *R* factor, but applied to a small subset of reflections that are not used in the refinement of the structural model is called free R factor (or R free) [115]. The purpose is to monitor the progress of refinement and to check that the *R* factor is not being artificially reduced by the introduction of too many parameters.

It is calculated in the same way as the conventional least-squares *R* factor, but uses a small subset of randomly selected reflections that are set aside from the beginning and *not* used in the refinement of the structural model. Thus R free tests how well the model predicts experimental observations that are not themselves used to fit the model. A fixed percentage of the total number of reflections is usually assigned to the free group.

Another statistic that is reported with crystal structure refinements is the goodness of fit, *S*. Technically, the goodness of fit is "the standard deviation of an observation of unit weight." In practice the goodness of fit shows how reliable the standard deviations of the positional and displacement parameters of the atoms really are. The standard deviations of the atomic parameters should be multiplied by the goodness of fit to give more realistic estimates of the standard deviations. These adjusted standard deviations can be compared with similar values from other structures. The goodness of fit is strongly influenced by the weighting scheme. Thus crystallographers will modify the weighting scheme to force the goodness of fit to have a value near to 1.0 and hence the standard deviations can be used directly as they are determined. For a refinement on F^2 the goodness of fit has the form:

$$GoF = S = \{\sum [w(F_o^2 - F_c^2)^2] / (n-p)\}^{1/2}$$

where n = number of measured data and p = number of parameters.

Correlations

When the shifts in pairs of parameters are not independent of each other the parameters are said to be *correlated*. Correlations can be either positive, shifts of the parameters in question have the same sign, or negative, shifts in the parameters have opposite signs. Correlations can assume any value from -1, complete negative correlation; to 0, no correlation; to +1, complete positive correlation. Large correlations, those with a magnitude between 0.5 and 1.0, are specifically noted by most refinement programs. To successfully refine parameters with large correlations, the starting model must be very close to its local minimum. Refining parameters with large correlations requires more cycles of refinement to achieve convergence.

Some large correlations are expected and quite reasonable. In nearly all structures with heavy atoms, the overall scale factor and the displacement parameters of the heavy atom(s) are correlated. Large correlations can also occur between the different anisotropic displacement parameters of any particular atom. If the unit cell angles are far from 90° then it is not uncommon to see large correlations between the corresponding x, y, and z parameters for a given atom. For example, in a monoclinic structure with $\beta > 100°$, the x and z parameters of a heavy atom are usually strongly correlated. In disordered structures, it is common to see large correlations between the positional and displacement parameters of atoms in close proximity with other atoms.

Some large correlations can also signal problems with the model. In particular, large correlations between the positional parameters of different atoms, e.g., the x parameter of one atom and the x parameter of another atom, when the atoms are not disordered, suggests that the space group may be wrong. A higher symmetry space group can usually be found that has symmetry operations that relate the two atoms being modeled separately in the lower symmetry space group.

2.1.5.2. Programs for refinement: CNS, REFMAC5

Several approaches are currently in most use for crystallographic refinement of macromolecules. The target function depends on several atomic parameters (coordinates, B-values, occupancies). The large number of adjustable parameters (at least 3 times the number of atoms) gives rise to a complicated target function. Depending on the refinement target optimization method, one can highlight:

-Gradient-driven minimization. Minimization function follows its local gradient.

-Simulated annealing. This method is good for escaping local minima. Annealing

-Grid search. This method is based on sampling parameter space within known range. It may be time inefficient for multiparameter systems, and not as accurate as gradient-driven refinement.

Among the most widely used programs for refinement are SHELXL [116], CNS [117], REFMAC5 [118] and some others. In this work, CNS and REFMAC5 programs were used to perform the refinement.

CNS., This program suite is used to address the multiple minima problem: there are many local minima in addition to the global minimum, which is the final goal of the refinement.

A solution is to use an optimization technique which is good at overcoming local minima. One of such techniques is simulated annealing. Annealing is a physical process wherein a solid is heated until all particles randomly arrange themselves in a liquid phase, which is then slowly cooled so that all particles arrange themselves in the lowest energy state. Simulated annealing is the computational simulation of the annealing process, which is available in CNS program and was used as a part of the refinement procedure in this work. Simulated annealing increases the probability of finding a more optimal solution than gradient-descent methods because motion against the gradient is allowed. The likelihood of this “uphill” motion is determined by the starting annealing temperature (which is recommended to be set to 5000K).

REFMAC5 is another program for the refinement of macromolecular structures, which has been used in this study. REFMAC5 is distributed as part of the CCP4 crystallographic suite [110].

The refinement by REFMAC is strongly based on Maximum Likelihood method and Bayesian statistics. The basic idea of maximum likelihood is quite simple: the best model is most consistent with the observations. Consistency is measured statistically, by the probability that the observations should have been made. If the model is changed to make the observations more probable, the likelihood goes up, indicating that the model is better. The probabilities have to include the effects of all sources of error, including not just measurement errors but also errors in the model itself. But as the model gets better, its errors clearly get smaller, which means the probabilities become sharper. The sharpening of probabilities also increases the likelihood, as long as they are no sharper than appropriate.

A useful feature of REFMAC5 is the TLS refinement. TLS (stands for Translation Libration Screw-motion) refinement allows to model anisotropic displacements of the model atoms at medium to low resolution. It does this by constraining the allowed displacements to a rigid body model, which requires 20 free parameters per rigid group or "TLS group".

Any displacement of a rigid body can be described as a rotation about an axis passing through a fixed point, together with a translation of that fixed point. The mean square displacement of a point in a rigid body is expressed in terms of three tensors **T**, **L** and **S**.

TLS parameterization allows to partly take into account anisotropic motions at modest resolutionand might improve refinement statistics. TLS parameters can be analyzed to extract physical significance.

2.1.5.3. Structure modeling programs

Thus, from a partial model of the structure, structure factor and sometimes refinement calculations are performed that are then followed by a difference electron density map calculation. New atoms are located from the map and included in the model. This process is repeated until all non-hydrogen atoms are located. The model improvement is, consequently, a combination of automatic and manual steps, first done by software and the latter by visualizing and modeling programs. Two programs were mostly used in the work to visualize the models and work with maps: TURBO-FRODO and COOT.

TURBO-FRODO is a general-purpose molecular modeling environment. It is designed for de novo modeling of macromolecules, polypeptides, and nucleic acids, by building up these macromolecules from experimental X-ray crystallographic and NMR data and displaying the resulting models in various forms. These forms include Van der Waals and Connolly's molecular dot surfaces as well as spline line surfaces and secondary-structure representation. Compact views are possible, including CPK, icosahedra, and ball-and-stick styles, either individually or in various combinations.

The program can be used to color molecules, either objectively or subjectively, in order to compare them with existing structures and to evaluate their geometry, fit and stack proteins, interactively mutate or chemically modify a protein, and assess the resulting conformational changes. Turbo-Frodo reads and displays the electron density maps. Skeletal maps are used as templates to start building up molecules with homologous protein fragments. A set of Crystallize options are available for various purposes, such as fitting molecules into electron density maps, taking the symmetry of molecules into account, and depicting molecular packing with water-accessible surfaces. These options can be used to evaluate the size of crystal channels and to decide whether diffusion or cocrystallization is the most feasible way to fit a ligand into the target protein.

With Turbo-Frodo, molecules are displayed full screen, and the functions take the form of pull-down menus, which are displayed only when they are needed. Turbo-Frodo enables the user to interact directly with a molecule via these pull-down menus and dials. Plot files can be generated on Postscript files, which can be printed out on any Postscript laser printer, such as that of a MacIntosh.

This software program is written in C language, making maximum use of the vast computing and graphics potential of the latest graphics workstations based on the Silicon Graphics Library. It is built onto a data base, the Heap File, in which molecules can be stored. The Heap File is a non-finite arborescent molecular data base that is limited in size only by users' disk capacity. Turbo-Frodo users can therefore have access to all the structures stored in the Protein Data Bank (PDB), providing they have enough space on their disk.

The program **COOT** (Crystallographic Object-Oriented Toolkit) is used to display and manipulate atomic models of macromolecules, typically of proteins or nucleic acids, using 3D computer graphics. It is primary focused on the building and validation of atomic models into 3-dimensional electron density maps obtained by X-ray crystallography methods, although it has also been applied to data from electron microscopy.

Coot displays electron density maps and atomic models and allows model manipulations such as idealization, real space refinement, manual rotation/translation, rigid-body fitting, ligand search, solvation, mutations, rotamers, and Ramachandran idealization. The software is designed to be easy-to-learn for novice users, achieved by ensuring that tools for common tasks are 'discoverable' through familiar user interface elements (menus and toolbars), or by intuitive behaviour (mouse controls). Recent developments have enhanced the usability of the software for expert users, with customisable key bindings, extensions, and an extensive scripting interface.

Coot is free software, distributed under the GNU GPL. It is available from the Coot web site at the University of York. Pre-compiled binaries are available for Linux and Windows from York, and for Mac OS X through Fink. Additional support is available through the Coot wiki.

Coot can be used to read files containing 3D atomic coordinate models of macromolecular structures in a number of formats, including .PDB, .MMCIF, and Shelx files. The model may then be rotated in 3D and viewed from any viewpoint. The atomic model is represented by default using a stick-model, with vectors representing chemical bonds. The two halves of each bond are colored according to the element of the atom at that end of the bond, allowing chemical structure and identity to be visualized in a manner familiar to most chemists.

Coot can also display electron density, which is the result of structure determination experiments such as X-ray crystallography and EM reconstruction. The density is contoured using a 3D-mesh. The contour level controlled using the mouse wheel for easy manipulation - this provides a simple way for the user to get an idea of the 3D electron density profile without the visual clutter of multiple contour levels. Electron density may be read into the program from CCP4 or CNS map formats, how it is more normal to calculate an electron density map directly from the X-ray diffraction data, read from an .MTZ, .HKL, .FCF or .MMCIF file.

Coot provides extensive features for model building and refinement - i.e. adjusting the model to better fit the electron density, and for validation - i.e. checking that the atomic model agrees with the experimentally derived electron density and makes chemical sense. The most important of these tools is the real space refinement engine, which will optimise the fit of a section of atomic model to the electron density in real time, with graphical feedback. The user may also intervene in this process, dragging the atoms into the right places if the initial model is too far away from the corresponding electron density.

2.1.5.4. Structure analysis and description

The conformational attributes, such as torsion angles and packing volumes, are not generally restrained during refinement, so their statistical distribution can justifiably be derived from data bases such as the PDB [112]. These attributes have been used in the development of a number of validation packages, including PROCHECK from CCP4 suite [114]. The purpose of such software is to (a) verify the syntax of the file, (b) check the consistency of an atomic model with the current library and identify outliers for further investigation, (c) detect gross errors in the structures, such as mistracing of the chain, (d) check for local abnormalities of stereochemistry and (e) produce global stereochemical quality criteria.

It is the checks made on the conformational properties, which are independent as far as possible of the restraints applied, that are of the greatest use in validation. For example, torsion angles, if not restrained during refinement, provide the basis for an excellent validation check. Parts which have unusual conformations warrant further investigation; they are possibly wrongly interpreted, or may be at the core of the structure's active site, where strained conformations could be extremely interesting.

Different checking programs address various aspects of structure validation and exploit different but to some extent complementary aspects of the structures, although there is a set of properties common to all. The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures. PROCHECK makes use of properties originally derived from a set of 119 non-homologous protein crystal structures at a resolution of 2.0Å or higher and having an R-factor no greater than 20% [113]. The standard uncertainties of several unrestrained parameters were shown to have a clear correlation with resolution. For example, the standard deviation in a protein's main-chain hydrogen bond energies decreases with improving resolution, as does the variation of some torsion angles [112].

However, there is a real danger of negative feedback; structures which have been erroneously forced into conformations to pass the validation checks then enter the data base and thereby artificially reinforce the expectations and keep the door closed to novel conformational features. To assess this, it is recommended to use empirical evidence as much as “common sense”. The following considerations can be kept in mind:

-The displacement parameters should be checked for signs of systematic error. For example, ellipsoids of several heavy atoms aligned in one direction may indicate the need for a better absorption correction. Nonspherical or large ellipsoids suggest that the model may need to include disorder.

-The final difference electron density map should have no abnormally high peaks or low valleys.

-The final *Rfactor* should be reasonably low for the quality of data. For large molecules, R-factor usually ranges between 0.6 (when comparing a random set of reflections with a given model) and 0.2 (for example for a well refined macro-molecular model at a resolution of 2.5Å).

For estimating these characteristics the programs with visualizing possibilities (Turbo, Coot) can be applied in combination with checking programs like

PROCHECK.

2.1.5.4.1. Model Visualization by Pymol

PyMOL is a cross-platform and open source enhanced molecular graphics program and is usually used on the final stages of refinement for visualization and generation of high-quality images. It excels at 3D visualization of proteins, small molecules, density, surfaces, and trajectories. It also includes features for molecular editing, ray tracing, and preparing movies.

Two unique and valuable features of this program over some other visualization program are the use of the powerful programming language (Python) and an emphasis on high-quality graphics.

PyMOL is an open-source molecular visualization system created by Warren Lyford DeLano and commercialized by DeLano Scientific LLC. It can produce high quality 3D images of small molecules and biological macromolecules, such as proteins. According to the author, almost a quarter of all published images of 3D protein structures in the scientific literature were made using PyMOL.

3. Results and Discussion

3.1.Protein expression and purification

3.1.1. Selection of protein fragment for co-crystallization experiments

In a crystallization experiment, a correctly selected protein fragment, which still conserves the protein function but bears no part that is strongly disordered and prevents the successful protein isolation and crystallization, is essential. In this work, the silencing suppressor p19 from the Tomato Bushy Stunt virus was used for co-crystallization with CUG-repeating RNA sequences. The protein had been shown to bind length-specifically to double-stranded RNAs [80]. The wild type p19 protein from Tomato Bushy Stunt Virus has 172 amino acids; the sequence is depicted in **Figure 11**.

```
1     MERAIQGNDA REQANSERWD GGSGGTTSPF KLPDESPSWT EWRLHNDETN
51    SNQDNPLGFK ESWGFGKVVF KRYLRYDRTE ASLHRVLGSW TGDSVNYAAS
101   RFFGFDQIGC TYSIRFRGVS ITVSGGSRTL QHLCEMAIRS KQELLQLAPI
151   EVESNVSRGC PEGTETFEKE SE
```

Figure 11. p19 fragments amino acid sequences.
p19 wild type sequence is shown from 1 to 172 residues; p19m sequence is marked in magenta; p19 cut sequence is further marked by underlining.

Previously, the entire p19 from CIRV (Carnation Italian Ringspot Virus) was crystallized by Vargason *et al.* [81]. Crystals diffracted up to 2.5Å resolution and belonged to the space group P6(1)22. This brought difficulties into siRNA refinement, as due to symmetry relations the RNA dyad was found to be in two different positions. As a result, the final model had two protein monomers and two RNA duplexes with occupancy of 0.4 and 0.6, making the RNA bases overlapping and thus less distinguishable.

In another work [82], the protein from TBSV was truncated at N- and C-termini. A p19 fragment containing 27–158 residues and double methionine mutations at L144 and L147 (introduced for facilitating phasing by MAD) referred to as p19m, was co-crystallized with siRNA and gave rise to a crystal, which was diffracting to 1.85 Å resolution and displayed substantially lowered disorder than crystals of entire p19 from [81]. The p19m crystal belonged to the space group R32, and contained one protein monomer and one strand of RNA duplex in asymmetric unit. In addition, it has been tested in [82] that the truncation and mutations had no effect on the ability of p19 to bind the siRNA duplex. Coloring in **Figure 11** highlights the positioning of the p19m fragment inside the entire sequence of p19.

In present work, we tried to co-crystallize the entire p19 with CUG-repeating RNAs that yielded a crystal diffracting to pretty low resolution (2.9Å). The crystal belonged to the space group P3(1) and had six protein monomers and six RNA strands in the asymmetric unit (AU). The electron density map was of acceptable quality and some novel structural details were derived (e.g. the first data were obtained about p19 structural variability). Although it was possible to refine the model to reasonable Rwork/Rfree parameters of 22.6/25.8, it was clear that the protein has to be better adapted to our needs of CUG-repeating RNA co-crystallization.

The choice of p19 fragment was essential for the crystal quality. In order to yield better crystals, we decided to additionally truncate 8 amino acid residues on the C-term that were disordered in crystal of p19m•siRNA in [82]. This

fragment is referred to as p19cut, and the sequence is shown in **Figure 11**. The approach worked well, and we managed to achieve comparable or even higher resolutions (1.73-1.96Å) than in previous work (1.85 Å) [82].
In this work, only structures acquired with shorter fragments of the protein – p19m and p19cut - are discussed, because of the higher crystal quality and better resolved structural details in co-crystallization experiments with truncated proteins.

3.1.2. Protein sample preparation

The p19m gene, designed with codon usage optimized for expression in Escherichia coli and cloned into pET28a (Novagen) vector containing a thrombin-cleavable His-tag at the protein N terminus, was kindly provided by Dr Keqiong Ye. A TAA stop-codon was inserted to truncate the last 8 C-terminal residues and thus generate the shorter p19cut fragment. Both p19 fragments (with codon usage optimized for expression in *E.coli*) were expressed within pET28a expression vector containing a thrombin-cleavable His-tag at the protein N terminus. As mentioned above, the p19m sequence contained double methionine mutations at L144 and L147, which were introduced for facilitating phasing by multiple-wavelength anomalous diffraction (MAD) in previous work [82]. *E.coli* BL21(DE3) cells were transformed with the vectors, grown on LB liquid medium and, after reaching optical density of 0.5-0.9, induced with IPTG. The proteins were overerexpressed at 25ºC for 25 hours in a shaker. Cells were harvested by centrifugation (25000 r.p.m., 60 min, 4°C).

To purify the p19 fragments, we have adapted a two-step affinity purification method used in [82]. Bacterial pellets were resuspended in HEPES buffer (25mM HEPES, 300 mM KCl, 25mM Imidazole, pH 7.0) and lysed either with cell disruptor (French press) or by sonication in an ice bath. After the cell lysis, fast protein liquid chromatography (FPLC) on AKTA purifier was used. The soluble fraction of the lysate was separated by ultracentrifugation (40000 r.p.m., 45 min, 4°C) and loaded into AKTA Purifier for subsequent Ni-chelate affinity purification (GE Healthcare 1ml HiTrap Crude column). The first step was a Ni2+ affinity column, the elution was performed by gradient of imidazole. The column was previously equilibrated in the 25mM HEPES, 300 mM KCl pH 7.0 buffer. The protein was eluted by gradient of imidazole; the elution buffer was 25mM HEPES, 300 mM KCl, 1M imidazole, pH 7.0. Fractions were collected at approximately 80% gradient. The typical chromatogram can be seen in **Figure 12A**; the flowthrough and the protein fractions are marked. After chromatography, some fractions were analyzed by SDS-PAGE (**Figure 12B**). The fractions enriched with the target protein were then pooled and incubated with thrombin protease in order to cut off the N-terminal His-tag (1U of thrombin for 1mg of His-tag fusion at room temperature for 10-12 hours). After that, the second chromatography step was performed. Here, the ability of nucleic acid-binding proteins to bind to heparin pseudo-affinity column was used. The resulting mixture was diluted three times (25mM MES pH 6.2) to decrease the salt concentration, which was now 100mM KCl, and used for additional purification by heparin chromatography (GE Healthcare 1ml Heparin column). The protein was eluted by gradient of KCl (elution buffer was 1M KCl, 25mM MES pH 6.2); typical chromatogram is shown in **Figure 12C**.

The fractions of p19 were pooled and concentrated on centrifugal filter units (Amicon Ultra, Millipore) to 0.8-2.0 mg/ml. Protein concentration was estimated during all purification and concentrating steps using the NanoDrop instrument (see **Figure 13** for example spectra).

For crystallization trials, p19cut (or p19m) was then mixed with the RNA solution at protein dimer:RNA duplex molar ratio of 1:1 or 1:2.

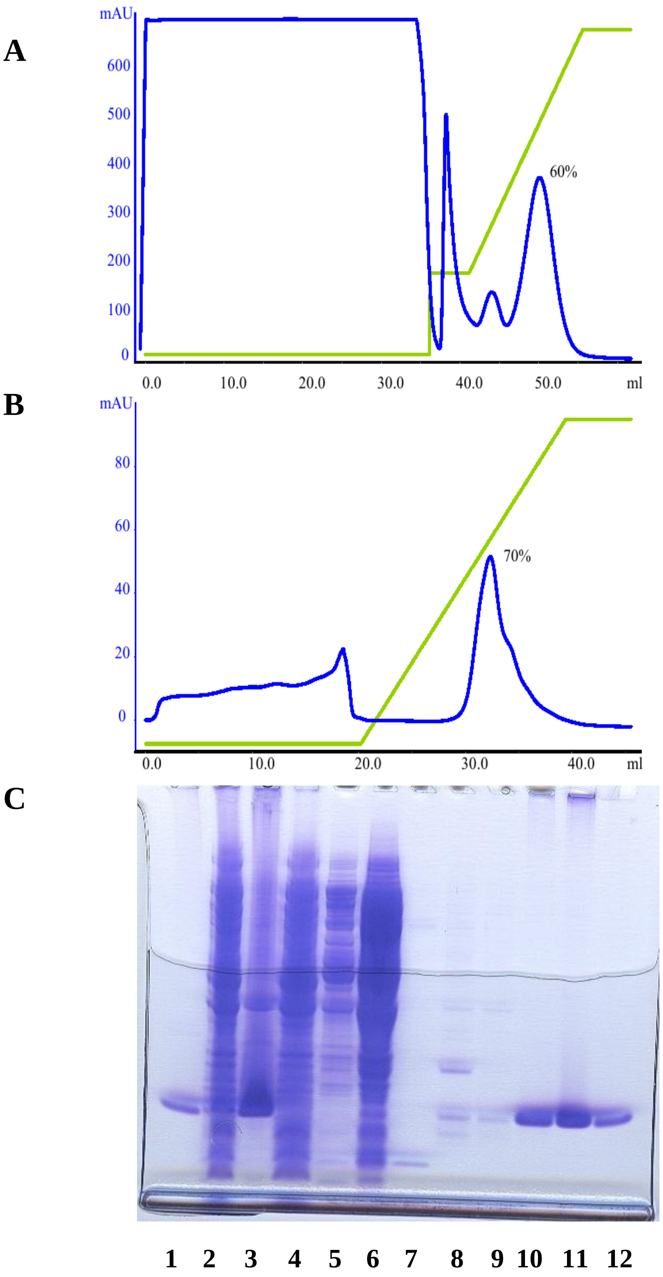


Figure 12. Purification of p19 fragments.
(A) Nickel affinity chromatogram. The adsorbance curve is shown in blue; the imidazole gradient is shown in green. After a large flowthrough fraction (0-35ml), a "washing" gradient step of 20% (35-40ml) and a small contamination peak (42ml) the target peak elutes at 60% elution buffer.
(B) Heparin chromatogram. The adsorbance curve is shown in blue, the gradient is shown in green. A flowthrough fraction (0-20ml), is followed by gradual ingrease of elution buffer; the target peak elutes at 70%.
(C) SDS gel, run after nickel chelate affinity chromatography. Lanes:
1 – pure p19cut as marker;
2 – cell lysate
3 – cell lysate: pellet
4 – cell lysate: supernatant
5 – flowthrough fraction
6 – "washing" peak
7 – contamination peak
8-12 – various fractions of the target p19cut peak

3.2.Design of RNA sequence for co-crystallization experiments

In our experiments, several **requirements to RNA sequences** were to be met:

- The sequences had to be constructed mostly (or completely) of uninterrupted CUG repeats. This was done in accordance to our aim to study special structural characteristics of CUG hairpins.

- The sequences had to form a duplex at least 19 base pairs long in order to be bound to and co-crystallized with p19.
- There had to be only one mode of 19 base pair duplex formation, so that in the crystal all complex molecules would be of the same kind. With this condition not fulfilled, problems in resolving and/or refinement of such structure could arise.
- The 19 base pair duplex had to be the most thermodynamically profitable in our crystallization conditions; any other secondary structure elements (e.g. hairpins) were to be avoided. In other words, it was essential to switch the free RNA state in solution towards duplex formation in order to facilitate its binding to p19 and subsequent crystal formation.
- In some cases, the duplex was to have 1 or 2 nt overhangs on the 5'-ends; this was done to improve crystal quality and obtain novel structural details of p19 interaction with small RNAs.
- The duplex had to be suitable in symmetry issues, i.e. had to possess a two-fold symmetry in order to avoid overlap of base pairs with occupancy of 0.5 as in [82]. This takes place because crystals of p19:RNA complex contain one protein monomer and half of the RNA duplex per asymmetric unit.

Our first choice was a 19 nucleotide uninterrupted CUG-repeating **sequence #1** in **Table 3**, which could form uninterrupted 19bp RNA duplex, as shown in the third column. Having 12 canonical G•C pairs and 7 U*U mismatches, it really formed a stable 2-fold symmetrical A-RNA duplex capable of binding to p19 protein and of yielding crystals of the complex. The crystal diffracted to 2.5 Å resolution, was isomorphous with the previous crystal of p19:siRNA complex

[82], and the structure from [82] was used as the initial model. However, the RNA chains possessed a certain level of disorder, especially on the duplex ends, which impeded high-quality structural result.

It was proposed then to stabilize the duplex ends with canonical G-C pairs instead of U*U mismatches. It was thus chosen a **sequence #2** (**Table 3**), also with a 2-fold symmetry. This replacement resulted in better crystals diffracting to 2.10Å resolution and reduced the disorder in the RNA duplex ends.

We then proved the unified RNA duplex formation mode by using a 5-bromouracil-containing sequence for the anomalous scattering experiment. The **sequence #3** (**Table 3**) was co-crystallized with p19m. The two anomalous peaks per chain were clearly observable at appropriate sides of the 4th and 16th nucleotides along the chain.

Although p19 was initially shown to bind RNA silencing-generated and synthetic 21 nucleotide siRNAs in vitro [80], it was proven capable of binding siRNAs containing duplex regions of longer than 19 base pairs (this type of double stranded RNA exists in living cells, for details see Introduction section “1.1.1. Small RNA types”). It was observed in the previously published structures [81,82] that the extra unpaired nucleotides on the 3'-term (the 2 nucleotide 3'-overhang, characteristic of natural siRNAs in cells), are disordered and do not establish any bonds with the protein, seemingly introducing no changes into the binding. Still, the mechanism of p19 binding duplexes of 20 or more base pairs was never studied structurally.

Table 3. RNA duplexes used in this work.

<table>
<tr><th>N</th><th>RNA strand sequence</th><th>Duplex-structure schematic</th><th>Aim</th></tr>
<tr><td>0</td><td>5´pG(CUG)6C</td><td>5’pGCUGCUGCUGCUGCUGCUGC
||*||*||*||*||*||*||
CGUCGUCGUCGUCGUCGUCGp5’</td><td>To crystallize as free (p19-unbound) CUG-repeating RNA tract</td></tr>
<tr><td>1</td><td>5´pUG(CUG)5CU</td><td>5’pUGCUGCUGCUGCUGCUGCU
||||*||*||*||*||*
UCGUCGUCGUCGUCGUCGUp5’</td><td>Uninterrupted 19nt CUG repeat with U*U mismatches and dyad symmetry</td></tr>
<tr><td>2</td><td>5´pGG(CUG)5CC</td><td>5’pGGCUGCUGCUGCUGCUGCC
|||*||*||*||*||*|||
CCGUCGUCGUCGUCGUCGGp5’</td><td>Reinforced duplex termini for better crystal quality</td></tr>
<tr><td>3</td><td>5´pGGCBrUG (CUG)3CBrUGCC</td><td>5’pGGCUGCUGCUGCUGCUGCC
|||*||*||*||*||*|||
CCGUCGUCGUCGUCGUCGGp5’</td><td>Confirmation of duplex accommodation with 5-bromouridine residues</td></tr>
<tr><td>4</td><td>5´pUUG(CUG)5CU</td><td>5’pUUGCUGCUGCUGCUGCUGCU
||||*||*||*||*||*
UCGUCGUCGUCGUCGUCGUUp5’</td><td>Duplex with 1 nucleotide 5’ overhang</td></tr>
<tr><td>5</td><td>5´pUUUG(CUG)5CU</td><td>5’pUUUGCUGCUGCUGCUGCUGCU
||||*||*||*||*||*
UCGUCGUCGUCGUCGUCGUUUp5’</td><td>Duplex with 2 nucleotide 5’ overhang</td></tr>
<tr><td>6</td><td>5´pG(CUG)6C</td><td>5’pGCUGCUGCUGCUGCUGCUG...C
||*||*||*||*||*||*|
CGUCGUCGUCGUCGUCGUC...Gp5’</td><td>20 base pair duplex: proof of p19 capability of unwinding "extra" base pairs</td></tr>
</table>

We tried adding one by one two 5'-overhanging nucleotides (see schematic in **Figure 14**). For studying the structural detail it was indispensable that the RNA duplex, as was mentioned before, formed uniformly and met the symmetry requirements of the R32 space group. Our aim was to see the importance, if any, of the 5'-overhang nucleotides which would form, as we hypothesized, upon unwinding of longer 20-23 bp RNA duplexes.

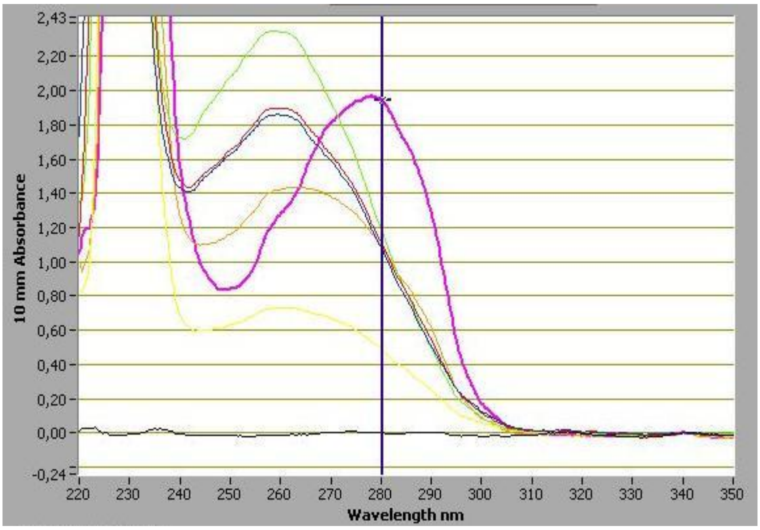


Figure 13. Example of spectra acquired with NanoDrop laboratory device.

Example spectra are shown: in black – a baseline (blank measurement); in purple, a typical protein spectrum (absorbance maximum close to 280nm); other five spectra are supposedly nucleic acid-contaminated (absorbance maximum close to 260nm). Using Lambert–Beer law, the protein concentration can be estimated with a formula C=Ads/Ext, where C is protein concentration, Ads is absorbance measured with NanoDrop, Ext is the extinction coefficient (for p19 it is approximately 1.8322).

The first 1 nucleotide 5'-overhang-bearing **sequence #4** (**Table 3**) had 12 C-G pairs, 7 U*U mismatches and an extra uridine on the 5'-terminus. It turned out that the 5'-phosphate of such overhang (the '-1' nucleotide, see **Figure 14**) was essential for creation of a new network of interactions of RNA with p19. This increased crystal quality and improved the resolution up to 1.85Å and brought a two amino acid protein loop (Asn28, Ser29) into ordered state through stacking interaction of Gln31 with the base moiety of the overhanging nucleotide.

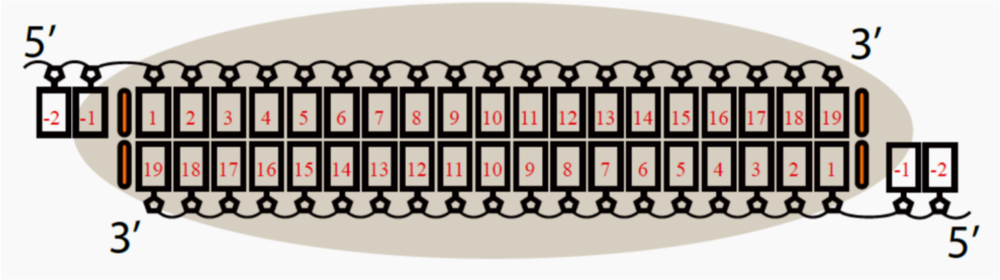


Figure 14. Schematic of 5’-overhang numbering.
The 19 base pair RNA duplex region is allocated inside the “caliper” of two tryptophan residues (shown as orange ellipses) from each side of the p19 dimer (shown as grey ellipse). The additional nucleotides (-1 and -2), if present, are found outside the “caliper”. The RNA 5’and 3’termini are marked.

Upon addition of a second extra base to the 5'-terminus (see **sequence #5** in **Table 3**), the crystal quality dropped – diffraction was only observed up to 2.5Å. Nevertheless, the allocation of the both overhanging nucleotides could be seen in the map: the system of interactions around the (-1) nucleotide phosphate was preserved and one new interaction of the (-2) phosphate was added.

Our suggestion of mechanism of p19 binding to longer duplexes (>19 base pairs) was that it involved unwinding of ¨extra¨ base pairs. To prove this hypothesis, we co-crystallized p19cut with a potential 20 base pair RNA **sequence #6** (see **Table 3**) that contained no overhangs. It was found that this RNA indeed forms a 19 base pair double-helical region while unwinding one terminal G-C base pair.

This CUG-repeating RNA was also crystallized without p19 (**sequence #0** in **Table 3**) and was shown to form a 20 base pair A-RNA duplex. Some structural differences in U*U mismatch stabilization were noted compared to CUG repeats in complex with p19 fragments.

3.3.Crystallization, Data Collection and Data Processing

Crystals were obtained of <u>p19 complex</u> with a set of siRNAs (see **sequences ##1-6** in **Table 3**) formed from uninterrupted CUG repeats: 19-21 nucleotides with either 0, 1 or 2 nucleotide 5'-overhang, and a 19 nucleotide siRNA with two incorporated 5-bromouracil nucleotides (5BrU) – for details, see **sequences ##1-6** in **Table 3**. All the sequences were used to form palindromic duplexes, resulting in 5-7 U*U mismatches in each case.

Crystals (example photos shown in **Figure 15**) were grown at 18°C by the method of vapor diffusion from hanging drops consisting of 1 µl each of complex (125 µM in 0.3 M KCl, 5 mM HEPES-KOH, pH 6.2) and reservoir solution (1.6 M ammonium sulphate, 0.1M Citric Acid pH 4.0). Sometimes RNA oligonucleotides (ordered from Dharmacon Research or synthesized in our laboratory) were previously annealed at 65°C for 1 min followed by a slow cooling on ice to select for duplex siRNA over hairpin RNA species. The siRNA was then mixed with p19 at 1:1 molar ratio and incubated on ice for 10-15 min. p19:siRNA crystals grew as thin rods with dimensions of 0.025 mm*0.025 mm*0.2 mm within 4-7 days.

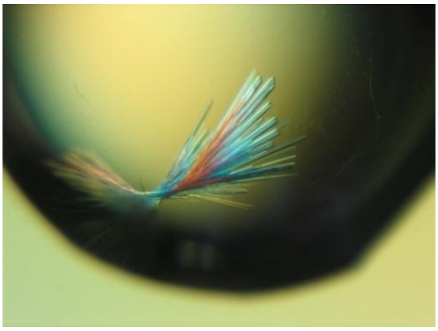


Figure 15. Microscopic photographs of p19:pUUG(CUG)$_5$CU complex crystals.

Crystals of <u>free RNA</u> (**sequence #0**, see **Table 3**) were obtained by the method of vapor diffusion from hanging drops. One µl of a 0.2 mM solution RNA (in DEPC-treated water) was mixed with 1 µl of reservoir solution (1.6 M ammonium sulphate, and either 0.1M Citric Acid pH 5.0, or 0.1M MES pH 6.0) and equilibrated over the reservoir solution at 18°C. One crystal with *trigonal prism* morphology, was obtained with dimensions of 0.1 mm*0.1 mm*0.25 mm within 15-20 days.

Prior to data collection, all crystals were transferred into 70% well solution:30% glycerol mixture and flash cooled to -180°C in liquid nitrogen. Datasets were collected for all p19:native siRNA complex crystals at beamline ID23-1 at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France to resolution 1.86-2.5Å. Diffraction data for the 5BrU p19:siRNA complex were collected to 1.96 Å resolution at beamline ID14-4 at the ESRF. Data set from the free RNA crystal (**#0** in **Table 3)** was collected using a rotating anode home X-ray source (Bruker PROTEUM8).
p19:siRNA crystals possessed diffraction anisotropy. Integration, scaling and merging of the diffraction data were performed by the HKL2000 suite of programs.

3.4.Structure Determination and Refinement

All the crystals of p19:siRNA complexes, of R32 space group (a = b = 89.36-91.25Å , c = 147.94-148.63Å), were isomorphous with the crystal from work [82]. Structures of p19 complexes with RNA sequences **##2** and **4** were determined using the structure [82] as initial model; the resulting coordinates were used in refinement of other structures (**## 1, 3, 5 and 6** in **Table 3**). Simulated annealing was performed and the initial 2Fo-Fc and Fo-Fc electron density maps were calculated by FFT program (CCP4 suite, Collaborative Computational Project 4, 1994) using phases derived from the initial model.

RNA sequence corrections were then introduced into the model with the help of the graphic program TURBO. The RNA duplex was constructed as “idealized” A-RNA model in TURBO-FRODO and was then inserted, pair by pair, into the electron density map calculated for the initial model. The refinement of the model was carried out by REFMAC5 program [8a] (CCP4 suite); solvent molecules were added to the model by program ARP/wARP [119] (CCP4 suite).

The model of p19cut or p19m includes residues 2-127 (corresponding to 24-149 in wild type) of each protein monomer: polypeptide chain of p19 cut terminates at residue 127 and the 15 additional residues of p19m, as was pointed out earlier, are disordered and therefore are not inserted in the model. Loop residues 28-29 are in most cases deleted due to poor, weak density for this region. The refinement with REFMAC5 was carried out for one protein monomer and one RNA strand (19, 20 or 21 nucleotides, depending on the RNA sequence) or, in some cases, for one protein monomer and a half of the 19 base pair RNA duplex that composes the asymmetric unit. Then the electron density map was calculated and, if the model displayed good RNA hydrogen bonding and stacking arrangement and overall good map quality, it was accepted in the further refinement which concluded in program REFMAC5. TLS parameters for two groups were refined: protein monomer and either one RNA strand or half of RNA duplex.

In case of difficulties due to high disorder in the mismatch areas of RNA duplex, molecular dynamics was carried out with the program CNS [107]. During CNS refinement the restrictions for Watson-Crick base pair distances were applied so that space group was changed for R3 and the non-crystallographic two-fold axis had to be generated and used. The final refinement stages, however, were concluded in REFMAC5 in R32 space group, and the model contained again one monomer of p19 and one strand of the RNA duplex. In case of 5-bromouracil-containing RNA, **#3** in **Table 3**, a single-wavelength anomalous dispersion (SAD) dataset was collected at the absorption peak of bromine atoms. (2Fo-Fc) anomalous map was calculated using the anomalous differences. Heavy atom sites were clearly visible in this map. The 5-bromouracils provide the two full-occupancy Br sites in the siRNA strand.

One RNA sequence, **#6** in **Table 3**, can form a 20 base pair duplex region (in contrast to all other RNA duplexes which form only 19 base pair duplex) and can thus bind to p19 protein in either of two positions, leaving one base pair unwinded alternatively from each side. Refinement was done in space group R3 by explicitly turning off van der Waals packing interactions amongst the two oppositely oriented RNA duplexes. For this, a special REFMAC5 script for handling such twinned data was used (the script was kindly provided by Dr Garib Murshudov). The sugar–phosphate backbones of oppositely oriented RNA duplexes were superimposable in the starting RNA model but separated during the structural refinement. The degree of separation correlates with the observed temperature factors, which are lowest for the protein-contacting segments, and highest for the solvent-exposed segments. It was anticipated [88] that the associated separation and temperature factors are both indicative of mobility within the bound RNA.

The free RNA crystal, sequence **#0** in **Table 3**, structure was solved by molecular replacement with the program AMoRe [108] using the data set processed at 2.5 Å resolution. A dataset was collected using a rotating anode home X-ray source ($I/\sigma = 2.3$ for the 2.59–2.50-Å resolution bin). The space group is P321 with unit cell dimensions of $a = b = 43.64$ Å, and $c = 158.56$Å. The duplex was constructed as "idealized" model in TURBO-FRODO, and was then used as a search model. In the solution, there are 1 and ½ duplexes in asymmetric unit; in other words, the second duplex was settled on crystallographic 2-fold axis. After that, each base pair was respectively treated

Table 4-I. Crystal data, processing and refinement statistics.

Protein RNA	p19cut pUG(CUG)5CU	p19m pGG(CUG)5CC	p19cut pUUG(CUG)5CU	p19m pGG(CUG)5CC-Br
Data collection				
Symmetry	Space group R32;　α=β=90°　γ=120°			
Cell dimensions				
a=b (Å)	89.83	90.798	89.36	90.97
c (Å)	148.345	148.489	147.94	147.84
Complexes per AU[a]	1 protein monomer; 1 RNA strand			
X-ray source	ESRF[b] ID23-	ESRF ID23-1	ESRF ID14-2	ESRF ID14-2
Wavelength	0.98035	0.98035	0.9330	0.98035
Maximal resolution (Å)	2.5	2.10	1.86	1.96
Rmerge(%)[c]	6.2 (32.9)	7.4 (55.6)	8.2 (64.4)	12.8 (99.5)
I/σIc	13.9 (3.1)	13.2 (1.78)	16.9 (2.5)	22.8 (5.04)
Completeness (%)	99.3	96.7 (98.0)	99.9 (99.9)	99.4 (98.9)
Redundancy	3.0 (3.0)	3.2 (3.3)	4.8(4.3)	22.2 (22.0)
Refinement				
Symmetry	R32			
Resolution range (Å)	15.0-2.5	15-2.10	15-1.86	15-1.96
Number of unique	7252	13862	19385	
Rwork/Rfree[d]	19.96/25.94	19.6/23.5	18.4/20.4	20.9/24.2
Model composition (AU)[e]				
Protein (amino	125	124	125	124
RNA (nucleotides)	19	19	20	19
Ions	2 SO42-	2 SO42-	2 SO42-; 1	2 SO42-; 1
Water	65	140	129	123
B-factors				
Protein (amino acids)	26.53	56.12	42.43	33.75
RNA	45.38	56.59	46.49	54.71

(nucleotides)				
Ions (SO42+)	56.20	68.00	48.68	49.54
Ion (Mg2+)	–	–	55.86	58.52
Water	48.44	63.29	51.93	53.95
Mean B-factor	32.50	56.90	44.07	40.755
RMSD bond length (Å)	0.015	0.013	0.012	0.010
RMSD bond angles (°)	1.791	1.53	1.43	1.45

Table 4-II. Crystal data, processing and refinement statistics.

Protein RNA	p19cut pG(CUG)6C	p19cut pUUUG(CUG)5CU	– pG(CUG)$_6$C
Data collection			
Symmetry	Space group R32; $\alpha=\beta=90^{\circ}$ $\gamma=120^{\circ}$		Space group P321
Cell dimensions			
a=b (Å)	89.78	90.39	43.64
c (Å)	148.22	148.91	158.56
Complexes per AU[a]	1 protein monomer; 1 RNA strand		3 RNA strands
X-ray source	ESRF[b] ID23-1	ESRF ID23-1	Bruker X8 PROTEUM
Wavelength	0.98035	0.9724	1.54
Maximal resolution (Å)	2.0	2.3	2.5
Rmerge(%)[c]	9.4 (99.7)	11.8 (63.0)	10.3 (47.6)
I/σI[c]	10.0 (1.7)	8.62 (1.86)	11.13 (2.28)
Completeness (%)	98.9 (98.0)	98.6 (98.5)	98.3 (92.4)
Redundancy	2.8(2.7)	2.7 (2.7)	6.6 (3.3)
Refinement			
Symmetry	R32		P321
Resolution range (Å)	15-2.0	15-2.3	15-2.5
Number of unique reflections	28241	10511	6470
Rfactor/Rfree[d]	18.2/22.1 18.6/22.8	19.8/25.7	20.3/27.9
Model composition (AU)[e]			
Protein (amino acids)	124x2	120	-
RNA (nucleotides)	20x2	21	60
Ions	4 $SO4^{2-}$; 2 Mg^{2+}	2 $SO4^{2-}$; 1 Mg^{2+}	5 $SO4^{2-}$; 1 K^{+}
Water	216	183	94
B-factors			

Protein	30.98	28.707	
RNA	52.92	47.023	24.65
Ions (SO4^{2+})	31.28	48.503	51.43
Ion (Mg^{2+})	38.69	47.030	35.39
Water	32.52	47.403	21.83
Mean B-factor	40.32	35.710	24.97
RMSD bond length (Å)	0.006	0.012	0.007
RMSD bond angles (^{o})	1.11	1.564	1.55

[a]AU, asymmetric unit; ESRF, European Synchrotron Research Facility; [b]ESRF, European Synchrotron Research Facility; [c]$R_{merge} = \Sigma_h \Sigma_i |I_{hi} - \langle I_h \rangle| / \Sigma \langle I_h \rangle$, where I_{hi} is the intensity of the ith observation of reflection h, and $\langle I_h \rangle$ is the average intensity of redundant measurements of the h reflections. Values in parentheses correspond to the last resolution shell; [d]$R_{work} = \Sigma ||F_o|-|F_c|| / \Sigma |F_o|$, where F_o and F_c are the observed and calculated structure-factor amplitudes; R_{free} is monitored with the 5% reflections excluded from refinement; [e]In case of refinement R3 space group NCS were used; the corresponding models contain 2 protein monomers and two RNA strands.

as a rigid body (the “fitting” operation in AMoRe program). The model was then refined using the REFMAC5 program. The refinement was carried out for one duplex and one 20 base pair strand that compose the asymmetric unit. At some stages of the refinement the restraints were applied to maintain the geometry of the Watson-Crick base pairs. Then the electron density map was calculated, the model displayed the good hydrogen bonding and stacking arrangement. It was accepted in the further refinement which concluded in program REFMAC5. TLS parameters for one group were refined. Crystal data, processing and refinement statistics for all structures are listed in the **Crystallographic Tables** (**Tables 4-I and 4-II**).

3.5.RNA parameter characterization

The conformational parameters of the RNA duplexes in structures presented in this work were calculated with program 3DNA [85] and can be compared to those for structures in [70].

The set of parameters with a common point of reference is needed to describe the three-dimensional arrangements of bases and base pairs in nucleic acid structures. There are two sets of parameters commonly in use in double helix conformational analysis: step parameters between neighbor base-pairs, and helical parameters which indicate the position and orientation of a base-pair relative to the helical axis. In each case, the axis taken as a starting point for the measurement is established in a different way: either a local axis for two base pairs or a global for the whole helix is calculated. Consequently, the values of local versus global helical rise and twist from these two sets of parameters can be quite different in nucleic acids which are bent or deviate significantly from homogenous B-DNA or A-RNA [87].

The parameters allow calculations to be carried out relative to the local helix axes (from one base pair to the next one), and relative to a long-range or global axis. The *x* direction of a local or base pair coordinate set should point along the short axis of the base pair, the y direction along the long axis and the z direction perpendicular to the plane of the pair, in a right-handed orthogonal axial set. (Directions of the *x*, *y* and *z* are considered below, following definitions of the parameters.) The long axis of a Watson-Crick base pair can be defined either by the line from the C6 of a pyrimidine to a C8 of a purine, or alternatively by the line from C6 of a pyrimidine to a hypothetical C8* atom on the purine, chosen so that the C6-C8* vector is parallel to the Cl '-Cl' vector. (The choice should be stated explicitly.) If desired, employment of axes along the three principal moments of inertia of a base pair may be incorporated as an extra user option, but should not replace the simpler definitions. Axes for calculating parameters of each base step should be chosen so that the same numerical values result (with only a possible change of sign) when going from base pair 1 to base pair 2, as from base pair 2 to 1. One way in which this can be accomplished is by choosing a local reference axis set intermediate between those of the base pairs themselves. See **Figure 16** for an illustration of base-pair parameters with definitions built upon qualitative guidelines established previously to specify the arrangements of bases and base-pairs in DNA and RNA structures [87].

Although useful for standard Watson-Crick hydrogen bonds, most of the aforementioned parameters become unusable (at least for comparison with standard values) for differently arranged bases in Non-Watson-Crick basepairs. It is important to note that the RNA structures described in this work contain mismatch base pairs and thus are no longer limited to the uniform staircase models first deduced by Watson and Crick. For structures that contain unusual bases, modified backbone atoms, mismatched residues, etc as well as drastically altered chain backbone conformations with single-stranded hairpins, inter-duplex pairs and looped-out bases a new set of parameters was proposed [86].

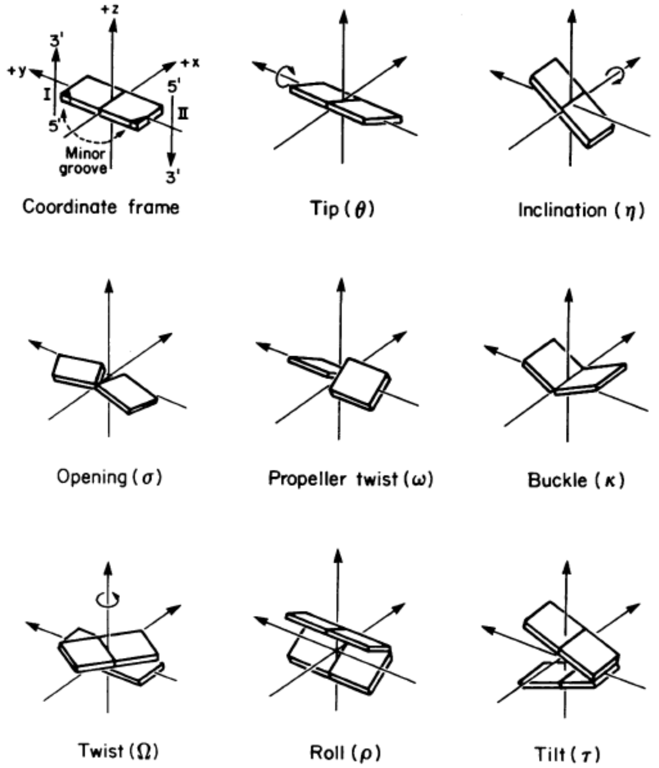


Figure 16. Definitions of various base-pair parameters involving two bases of a pair (upper two rows) or two successive base pairs (bottom row): **rotational parameters** (cited from [87]).
In the top row the motions of the two bases are coordinated, and in the middle row their motions are opposed. Columns at left, center and right describe rotations about the z, y and x axes respectively. The standard coordinate frame is defined at upper left.

The so-called sequence-independent measures were investigated, based on vectors connecting the C1' atoms of the paired residues, to avoid computational artifacts arising from non-canonical base pairing [86]. To calculate such parameters, a new coordinate frame was proposed: a right-handed coordinate frame attached to each base (**Figure 17**) follows established qualitative guidelines. The *x*-axis points in the direction of the major groove along what would be the pseudo-dyad axis of an ideal Watson-Crick base-pair, i.e. the perpendicular bisector of the C1′…C1′ vector spanning the base-pair. The *y*-axis runs along the long axis of the idealized base-pair in the direction of the sequence strand, parallel with the C1′…C1′ vector, and displaced so as to pass through the intersection on the (pseudo-dyad) *x*-axis of the

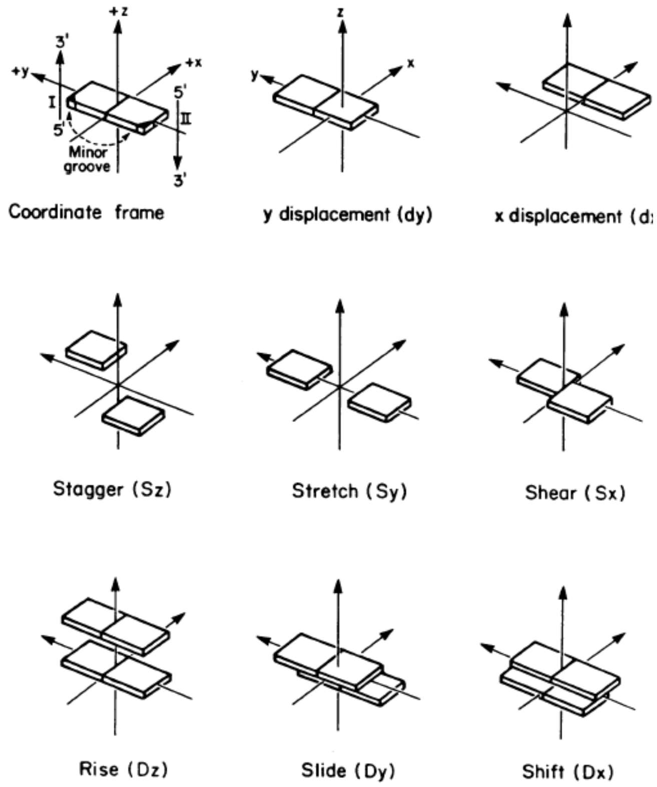


Figure 17. Definitions of various base-pair parameters involving two bases of a pair (upper two rows) or two successive base pairs (bottom row): translational parameters (cited from [87]).
In the top row the motions of the two bases are coordinated, and in the middle row their motions are opposed. Columns at left, center and right describe rotations about the z, y and x axes respectively. The standard coordinate frame is defined at upper left.

vector connecting the pyrimidine Y(C6) and purine R(C8) atoms. The *z*-axis is defined by the right-handed rule, i.e. $z = x \times y$. For right-handed *A*-RNA, the *z*-axis accordingly points along the 5′ to 3′ direction of the sequence strand. The location of the origin depends upon the width of the idealized base-pair, i.e. the C1′…C1′ spacing, $d_{C1'...C1'}$, and the pivoting of complementary bases, λ, in the base-pair plane (see **Figure 18**). The coordinates of the C1′ atoms establish the pseudo-dyad axis, i.e. the line in the base-pair plane where $y = 0$. The rotations of each base about a normal axis passing through the C1′ glycosyl atoms determine the Y(C6) and R(C8) positions used to define the line where $x = 0$ [86].

The parameters calculated in this work are: displacement of the middle C1'-C1' point from the helix (called displacement); inclination between C1'-C1' vector and helix axis, subtracted from 90 (angle); helical twist angle between consecutive C1'-C1' vectors (twist); helical rise by projection of the vector connecting consecutive C1'-C1' middle points onto the helical axis (rise).

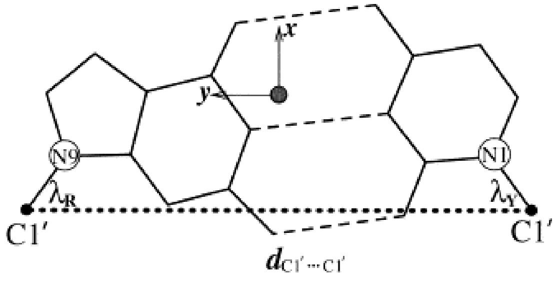


Figure 18. The right-handed coordinate frame attached to each base used in program 3DNA for sequence-independent measures calculations. Cited from [86].
Sequence-independent measures are based on vectors connecting the C1' atoms of the paired residues, to avoid computational artefacts arising from non-canonical base pairing. The schematic is an illustration of idealized base-pair

parameters, $d_{C1'...C1'}$ and λ, used respectively to displace and pivot complementary bases in the optimization of the standard reference frame for right-handed *A* and *B*-DNA, with the origin at • and the *x*- and *y*-axes pointing in the designated directions.

The stacking interactions were quantified in 3DNA by the shared overlap area, in Å^2, of closely associated base rings, i.e. the nine-membered ring of a purine R (A or G) and the six-membered ring of a pyrimidine Y (C, T or U), projected in the mean base pair plane.

3.6. Structural results

3.6.1. Free CUG-repeating RNA constitutes an A-form with no major perturbations to the double helix geometry

The 20-nucleotide sequence (**#0** in **Table 3**) was chosen as an example of uninterrupted CUG triplets with U*U mismatches, to see if this type of RNA would constitute an A-RNA and to estimate the structural details.

The studies [70,98] show that small CUG-repeating RNA fragments constitute an A-form. The structures formed by G(CUG)$_2$C and (CUG)$_6$ oligonucleotides have been described with diffraction data obtained with resolutions of 1.23 and 1.58 Å, respectively [70,98].

In the 2.5 Å X-ray structure (crystal data, processing and refinement statistics are listed in **Table 4-II),** the asymmetric unit contains three RNA pG(CUG)$_6$C strands forming one complete RNA duplex (strands A+B), while the second duplex is formed by strand D and its symmetry equivalent, D', related by the 2-fold crystallographic axis. The duplexes stack end-to-end, forming continuous columns parallel to the c axis . The model also contains ordered water molecules and five sulfate ions (**Table 4-II**). The electron density 2Fo-Fc map for entire RNA duplex is shown in **Figure 19**.

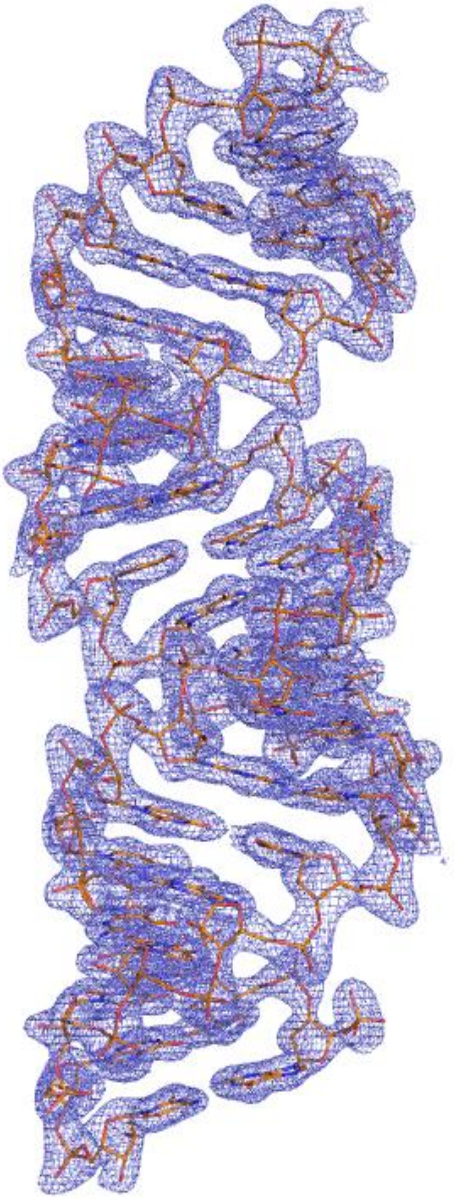


Figure 19. Free RNA duplex $[pG(CUG)_6C]_2$ crystal electron density map.
2Fo-Fc electron density map for $pG(CUG)_6C$ RNA crystal contoured at 1σ level and colored in blue. The asymmetric unit contains one 20 base pair duplex A+B and one strand E, which makes a duplex with its symmetry-related strand, E'. Here, only the duplex A+B is shown. The structure is presented with colored sticks, carbon atoms are colored yellow, nitrogen atoms – blue, oxygen – red, phosphate - orange.
The overall good quality of the map can be observed, in spite of the 2.5Å resolution.

As in works [98,70], it was similarly found that this 20-mer forms a A-RNA double-helix. All of the sugar residues are in the 3'-endo conformation. The volume per base pair in the crystal is about 1440 $Å^3$, which is a typical value for A-RNA or A-DNA crystals.

Displacement, angle (inclination between the inter-atomic C1'-C1' vector and the helix axis), rise and helical twist, calculated with program 3DNA [85] for A+B duplex structure, do not indicate any significant effects that can be attributed to the non-canonical base pairing. The average values are 6.9Å, 12.3°, 2.7Å, 32.1° respectively, which are typical of A-form.

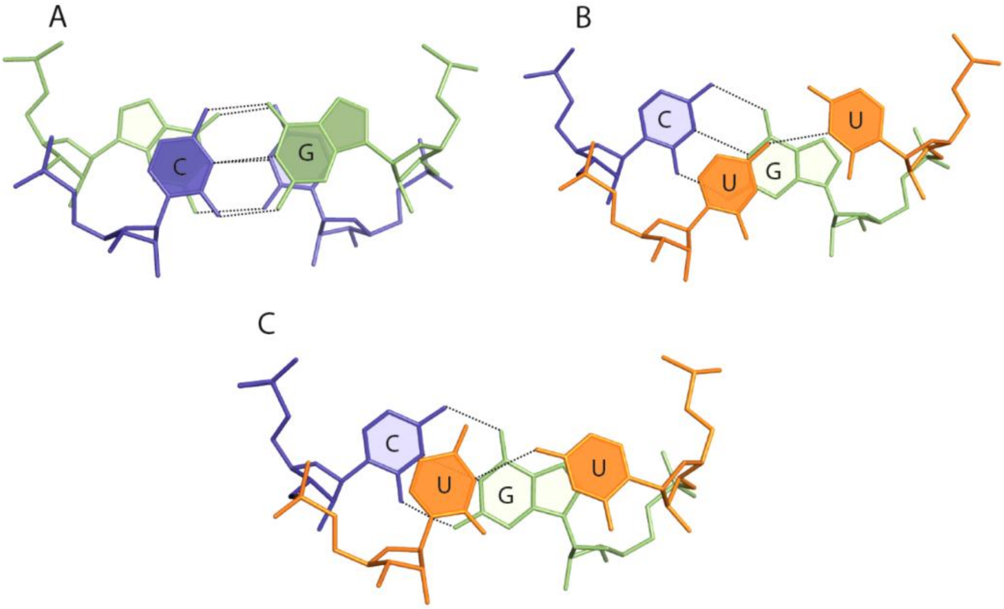


Figure 20. Stacking interactions for GC/GC step (A) and two kinds of CU/UG steps (B and C) depending on the conformation of the U*U pair. It can be noted that the GC/GC step possesses the highest overlap, while the steps involving the non-canonical pairing have pretty limited stacking interactions.

There is a significant difference in overlap area (i.e. area between polygons defined by atoms on successive bases, where polygons are projected in the mean plane of the designed base pair step) for the GC/GC steps (see **Table 5**). These steps were described in literature as having high stacking interactions [109]. The average value is 13.04Å^2 (standard deviation 0.65), whereas for two other steps it is 1.70Å^2 (standard deviation 0.55). This difference in stacking between canonical GpC-step versus non-canonical steps is depicted in **Figure 20**. There is also some regularity in the shift of the base pair steps: it is maximal for the UG/CU steps (1.12, SD=0.39), minimal for the CU/UG steps (-1.07, SD=0.43) and medium for the GC/GC steps (0.01, SD=0.22).

Every third base pair in the sequence is a mismatch U*U base pair. There are 9 mismatches (6 in A+B duplex and 3 in D+D' self-symmetrical duplex half) and 21 C-G pairs (14 in A+B duplex and 7 in D+D' self-symmetrical duplex half) per asymmetric unit. In this work, we will adopt different numbering schemes for mismatches (1st to 6th i.e. only counting mismatches, green in **Figure 21**) and C-G pairs (1st to 20th, i.e. counting each pair from the terminus). When speaking about individual residues, we will use their position in the appropriate strand if counted from the 5'-terminus (i.e. 1st to 20th in this structure).

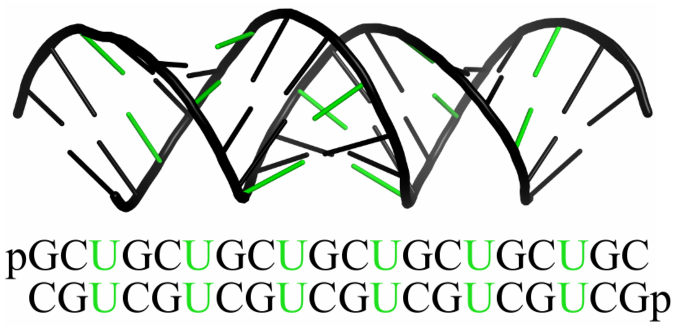


Figure 21. Overall packing scheme of CUG repeating RNA Every 3rd base pair is a U*U mismatch (highlighted in green), which is flanked from both sides by Watson-Crick C-G pairs.

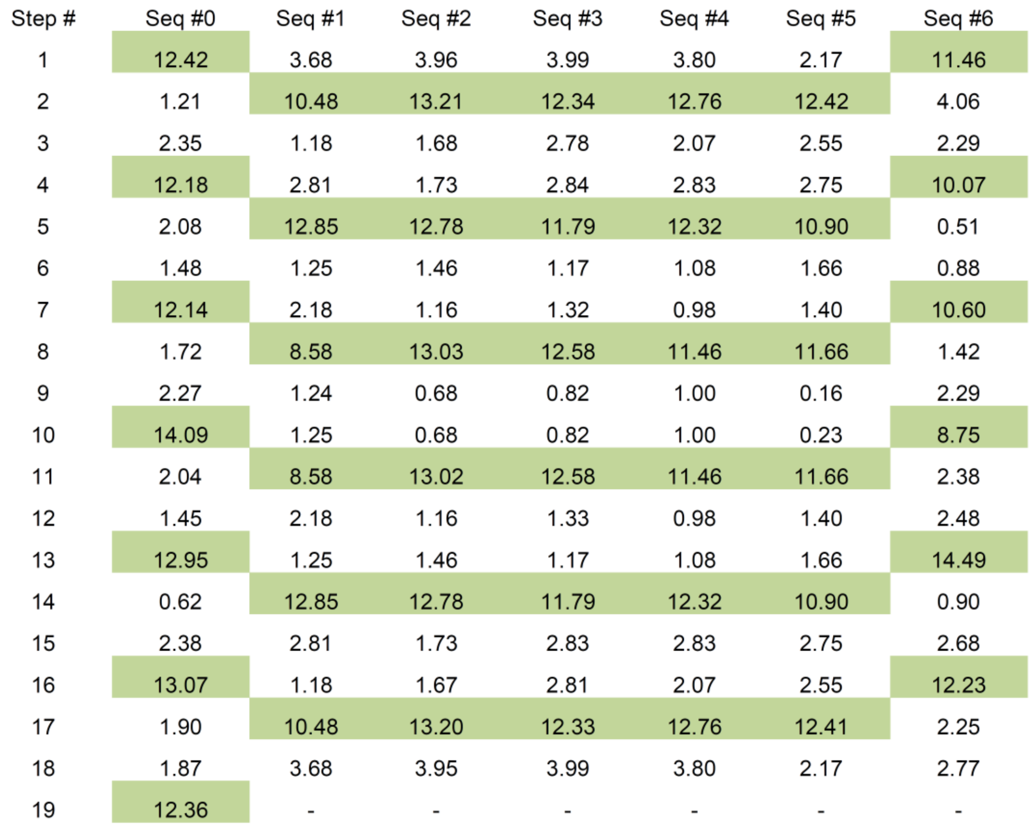
The mismatch U*U base pairs are flanked with two canonical C-G pairs on both sides, so that the overall packing scheme duplex A+B is the one depicted on **Figure 22** (for duplex D+D' it would be similar). 18 of the observed C–G base pairs form 3 hydrogen bonds (other 3 of the 21 C-G pairs have either one or two intra-basepair hydrogen bond lengths of more than 3.1Å), while 7 of the

9 U*U pairs interact via only one hydrogen bond between the carbonyl O4 atom of one base and the N3 amino group of the second U (the other two mismatches only form a repulsive O4-O4 contact). The residue accepting the H-bond is inclined towards the minor groove, as indicated by angle λ (between the glycosidic bond and the line joining the base-paired C1' atoms) (**Figure 22**). The value for the inclined bases is small, 34.3º (SD=7.4), compared to the average value for nucleotides of 55º. The inter-strand distance measured between the C1' atoms of the paired uridines remained typical for A-RNA—about 10.3Å, except for the first pair in duplex A+B where it was 9.3Å, which could be explained with the non-pseudo-infinite packing nature of duplexes in the crystal. The average for the analyzed duplex is 10.6Å, with standard deviation of 0.27Å. The average base pair opening for all U*U pairs is -20.5° (SD=5.2), irrespective of which U is inclined (**Table 2**). Mismatches are also characterized by stretch -1.26 (SD=0.11), with average of -0.49 and by shear of ±1.97 (average of ±1.02).

The above features are preserved in 8 of the observed U*U pairs. This pairing of uridines is described
in [70] and is called, according to the nomenclature introduced by Leontis and Westhof [120] as 'U*U cis (wobble) W+C+/W+C+', or 'stretched U*U wobble'. Overall, each CUG repeat assumes one of two distinct conformations depending on whether the uridine is inclined towards the minor groove (low λ) or not. In the A+B duplex, the 2nd, 4th and 6th uridines on strand A are inclined, thus the two strands are structurally different. Similarly, in the D+D' duplex 2nd, 4th and 6th uridines of strand D are inclined.

Ordered water molecules are associated with the U*U pairs, forming a characteristic pattern (**Figure 22**). In the minor groove one water molecule makes a hydrogen bond between the N3 amino group of the inclined uridine (low λ) and the O2 of the other uridine. This pattern is observed for four of the six U*U pairs in the duplex A+B but was not encountered in the duplex D+D'. The latter has two U*U mismatches that contain one water molecule makes a hydrogen bond between the O2 group of the one uridine and the O2 of the other uridine. All the remaining pairs were not observed to contain water molecules in the minor groove.

Table 5.
Comparison of stacking parameters (Å²) calculated with 3DNA program.
For RNA sequences in detail, see **Table 3**. Values for GC/GC steps are highlighted in green. The periodical nature of increasing and decreasing stacking values can be noted for all CUG repeats (sequences # 1 and #5 are not shown).



Step #	Seq #0	Seq #1	Seq #2	Seq #3	Seq #4	Seq #5	Seq #6
1	12.42	3.68	3.96	3.99	3.80	2.17	11.46
2	1.21	10.48	13.21	12.34	12.76	12.42	4.06
3	2.35	1.18	1.68	2.78	2.07	2.55	2.29
4	12.18	2.81	1.73	2.84	2.83	2.75	10.07
5	2.08	12.85	12.78	11.79	12.32	10.90	0.51
6	1.48	1.25	1.46	1.17	1.08	1.66	0.88
7	12.14	2.18	1.16	1.32	0.98	1.40	10.60
8	1.72	8.58	13.03	12.58	11.46	11.66	1.42
9	2.27	1.24	0.68	0.82	1.00	0.16	2.29
10	14.09	1.25	0.68	0.82	1.00	0.23	8.75
11	2.04	8.58	13.02	12.58	11.46	11.66	2.38
12	1.45	2.18	1.16	1.33	0.98	1.40	2.48
13	12.95	1.25	1.46	1.17	1.08	1.66	14.49
14	0.62	12.85	12.78	11.79	12.32	10.90	0.90
15	2.38	2.81	1.73	2.83	2.83	2.75	2.68
16	13.07	1.18	1.67	2.81	2.07	2.55	12.23
17	1.90	10.48	13.20	12.33	12.76	12.41	2.25
18	1.87	3.68	3.95	3.99	3.80	2.17	2.77
19	12.36	-	-	-	-	-	-

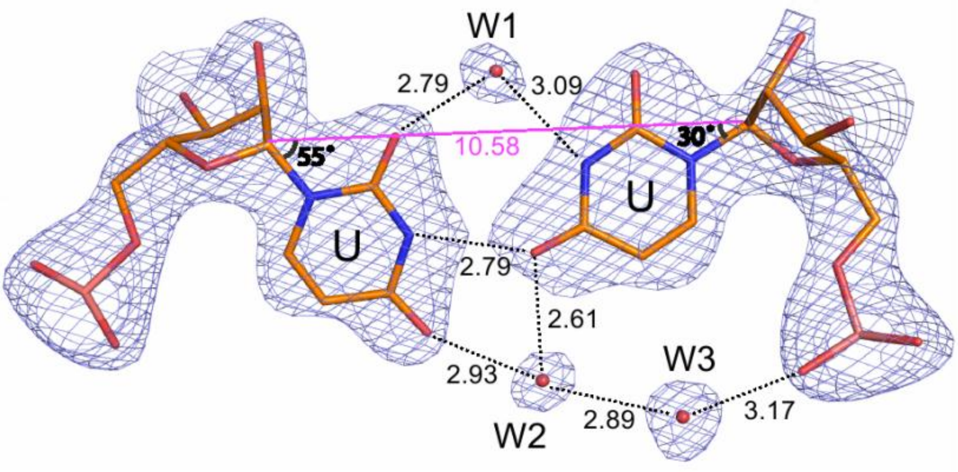


Figure 22. A representative observed ‘stretched U–U pair’ with a single direct H-bond N3-O4 and additional water-bridges supporting the pairing All the pairs in both analyzed crystal forms show the same conformation. One of the uridines is inclined towards the minor groove, and the λ angle, between the glycosidic bond and the line connecting C1' atoms (magenta line), is 30° or less, as opposed to the typical value of 55° . The distance C1' –C1' for the ‘stretched U–U pair’ is about 10.4Å, similar to the average value for an A-helix. Each type of U–U pair is solvated by three water molecules, one in the major groove (W1) and two in minor groove (W2, W3). The environment of the water molecules changes due to the inclination of one U.

Also, in two U*U pairs (4th and 6th) of the A+B duplex in the major groove, two water molecules can be found. One, W1 in **Figure 22**, is providing an H-bond between two O4 carbonyls of the uridines

while connecting them to the O6 carbonyl of the nearest guanosine adjacent to the inclined uridine. The

second water molecule, W2 in **Figure 22**, mediates an H-bond between W1 and the nearest PO43- group. Three more U*U pairs possess only W2. All other U*U pairs lack any water molecules in the major groove.

There are 4 sulfate ions present in the minor groove of the A+B duplex and 2 more in the minor groove of the D+D’ duplex (the latter two are symmetry-equivalent). The sulfates establish direct contacts with N2 groups of guanosines, O2' ribose atoms, and O2 groups of uridine and cytidines, and water-mediated contacts with O2' ribose atoms, O2 groups of uridine and cytidines, and phosphate groups.

The surface potential of the duplexes shows some regularity due to the periodicity of the CUG repeats. While the major groove is predominantly electronegative, with the cytidine N4 groups creating local electropositive points, the minor groove shows a regularity of electropositive areas created by guanine N2 amino groups, as was noted in [70] (**Figure 23**).

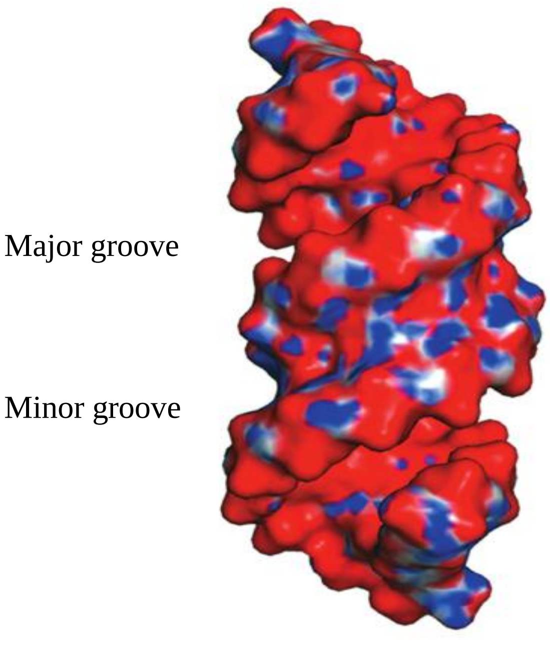


Figure 23. The pG(CUG)6C RNA duplex structure electrostatic potential surface (cited from [70])**.**
Red is negative, blue is positive.
The CUG repeats form regular, well defined structural motifs, whose characteristic hydrogen-bonding pattern, interactions with the solvent, the electrostatic charge distribution and surface features, define their properties and indicate the ligand binding potential of the CUG tracts. The major groove is predominantly electronegative with patches of positive potential due to amino groups of cytosines. The potential of the minor groove is complex and forms a pattern of alternating bands of positive and negative potential along the direction of the helix axis. The negative bands are formed by the electropositive atoms of stacking C, G and U, and the positive bands by the carbonyl oxygen atoms of U and C residues. The carbonyl groups of the inclined uridines protrude out of the minor groove and form bulges with high negative potential [70].

The electron density map is shown in **Figure 19**. It can be noted that, in spite of 2.5Å resolution, the quality of the map is good and there are no disordered regions. Because of this, we were able to observe and describe the same U*U mismatch stabilization modes as in [70], where the analyzed structure of a shorter duplex (octamer) is of atomic resolution of1.23Å.

3.6.2. CUG-repeating RNA duplexes can form complexes with silencing suppressor p19

Six complexes of CUG-repeating RNAs were tried in co-crystallization experiments with p19. It was established previously [82] that p19 from tombusvirus crystallized with siRNA in R32 space group, with the 2-fold symmetry axis coincident with the 19 base pair RNA pseudo-dyad.

All the p19:siRNA crystals were isomorphous with the crystal from [82]. Therefore, structures of complexes were determined using it as initial model.

3.6.2.1. p19cut:pUG(CUG)$_5$CU complex

The first fragment complexed with p19 was a 19 nt RNA strand constructed of consequent uninterrupted CUG repeats: sequence **#1** in **Table 3**. It can be noted that the sequence, when duplexed, possesses a 2-fold symmetry around the central U*U mismatch and that the two terminal pairs of the duplex are also U*U.

The crystals were checked for diffraction (~3.2-3.5 Å resolution at home X-ray source, Bruker Proteum) and then sent to ESRF, Grenoble, where a dataset was collected to resolution 2.5 Å on beamline 23-1 (*I*/σ = for the Å resolution bin).

The space group is R32; the unit cell dimensions are ***a*** = ***b*** = 89.827 Å, and ***c*** = 148.345 Å.

The 2.5 Å X-ray dataset was collected. Data were processed and scaled with HKL2000 in R32 space group (Table 4-I). The crystal was isomorphous to the one described in [82], so that the final p19-RNA structure from [82] (PDB entry 1R9F) has been taken as an initial model to calculate the original 2Fo-Fc electron density map. The idealized A-RNA duplex of an appropriate sequence was then built by using the TURBO-FRODO option ‘make-DNA’ and replaced the original RNA fragment in the initial model. The RNA model then was splitted into base-pairs, and each base-pair has been optimally fitted into 2Fo-Fc electron density map by using TURBO. Finally, the model was refined using the REFMAC5 program. The refinement was carried out in space group R32, with one protein monomer and one RNA strand composing the asymmetric unit. At some stages of the refinement the restraints were applied to maintain the geometry of the Watson-Crick base pairs combined with the use simulated annealing with the program CNS. The refined structure #1 was later used as an initial model in structure determination of a higher-resolution complex **#4 (Table 3**), which in turn brought the improved model for the structure #1. This improved model was finally re-refined by REFMAC against the proper dataset to values Rwork/Rfree = 19.96/25.94, indicated in Table 4-I. Overview structure of the complex is shown in **Figure 24**.

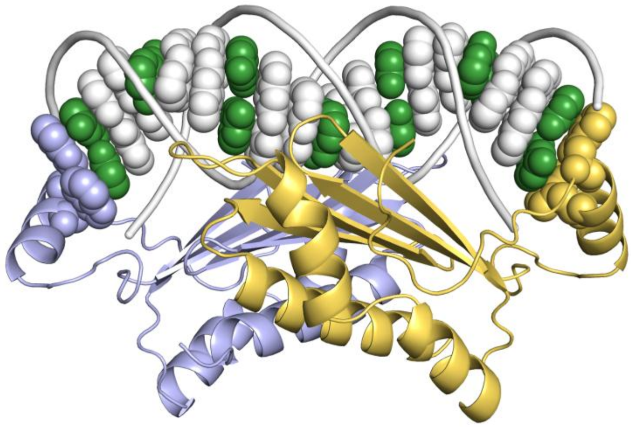


Figure 24. Overall structure of p19 bound to the RNA duplex.
The nucleotide sequences of duplexes are shown at the top. RNA bases in a sphere representation, and RNA backbones in a ribbon representation. Mismatches U-U, and Watson-Crick G-C base pairs are shown in green and white colors, respectively. Protein monomers are colored gold and blue, and are in a "ribbon" representation, whereas "caliper" tryptophan residues (colored gold and blue for different monomers) are shown in spheres.

polypeptide chain had good mean B-factor of 26.5 and was clearly visible in the map, the RNA moiety displayed higher B-factors of about 50,0. Still, the sugarphosphate backbone was well interpreted in the electron density map. However, RNA bases were partially disordered, especially on the duplex ends. This suggested further search for CUG-repeating RNA sequences in order to get high-quality crystals of p19/RNA complex.

3.6.2.2. p19m:pGG(CUG)$_5$CC complex

The next sequence tried for co-crystallization was sequence **#2** in **Table 3**. Here, it was supposed that the ends of the RNA duplex, when containing GC Watson-Crick base pairs instead of mismatches, would become more stable and less disordered. Indeed, better crystal quality and resolution were achieved.

X-ray data were collected on beamline ID23-1 to ~2.1 Å resolution (I/σ = 1.78 for the 2.17–2.10-Å resolution bin). The space group was R32 with unit cell dimensions of *a* = *b* = 90.798 Å, and *c* = 148.489 Å. Data were processed and scaled with HKL2000 in R32 space group. A higher-resolution structure, **#4** in **Table 3**, was used as the initial model for refinement (see **Section 3.6.2.4**). At some stages of the refinement the restraints were applied to maintain the geometry of the Watson-Crick base pairs and several trials of CNS molecular dynamics were conducted. With the changes in the sequence the level of disorder in the RNA duplex area was decreased significantly. The structure was refined to Rwork/Rfree 19.6/23.5. Crystal data, processing and refinement statistics are listed in **Table 4-I**.

Here, as in the free RNA, the increased overlap area difference for GC/GC steps compared to other steps was observed (see **Table 5**). Namely, the GC/GC steps have an average overlap area of 13.00Å^2 (SD=0.19), whereas the CU/UG or UG/CU steps 1.34Å^2 (SD=0.17). Notably, the terminal GG/CC step was characterized by a slightly increased stacking of 3.95Å^2 due to the overlap of two consecutive guanines. Then, a correlation of increased roll values was noted for the mismatch-including steps, i.e. UG/CU and CU/UG 10.63° (SD=3.82) as compared to GC/GC steps 3.3° (SD=2.1), and a sharply increased tilt of 8.07° for the first mismatch (all other tilt values lie in the range of -2.78-2.48°) which can implicate an increased double helix flexibility in the area of the U*U mismatches. Notably, in the Watson-Crick RNA in complex with p19, the increased roll values correlated with YR/YR steps, where Y – pyrimidine, R- purine.

There are 14 C-G pairs in this RNA sequence and all of them are involved in Watson-Crick interactions. Although, pairs number 3 and 6, characterized by an increased “opening” parameter, have one hydrogen bond weakened N4-O6 distance of 3.42 or 3.30Å whereas pair 5 has one hydrogen bond weakened and

one eliminated - N4-O6=3.61 and N1-N3=3.17Å. Average opening of those three pairs is 11.37° (SD=2.56), whereas for all other pairs it was less than 1° (in the idealized A-RNA the opening adopts negative values). Interestingly, an increased opening of 9.02° and 10.41° (there are two numbers cited as the two pairs are different although related via 2-fold non-crystallographic symmetry) was observed for the 5th and 15th pairs, C-G and U-A, in a complex of p19 with a Watson-Crick RNA [82], with N4-O6 distances of 3.45 and 3.61Å.
The 5th pair (and its symmetry-related 15th pair) in our complex is also characterized with the maximal deviation of buckle (-8.39), stagger (-0.47) and shear (-0.71) parameters, although the C1'-C1' distance is 10.3Å, which is normal for A-RNA. It also possesses the twist of 22.12°, while the average is 31.31° (SD=4.68). This all could be interpreted as local unwinding of the RNA double helix in complex with p19.

The 5 U*U mismatches are flanked by G-C pairs from both sides. There were no hydrogen bonds detected between the two uridines in a pair. On the contrary, a repulsive O4-O4 close contact of 2.85-3.29Å was observed in all U*U pairs. There was no significant deviation in the λ angles for the pairing uridines, as it was noted for the free CUG-repeating RNA. The C1'-C1' distances of 9.8-10.8Å, with mean value of 10.44 (SD=0.31) are inside the possible range for A-RNA. The U*U average opening in this structure was -8.8 (SD=1.97), whereas for the G-C pairs the average opening was 4.48° (SD=6.38). It seems that the U*U mismatches here do not have any H-bonding pattern, in contrast to the free CUG repeats (where a formation of one or two hydrogen bonds was observed), and are held only by the sugar phosphate backbone and the comparatively weak (see the overlap area) stacking interactions. They seem to bring the highest contribution into the bending of dsRNA, which is characterized by the increased roll and tilt parameters associated with them (described above).

3.6.2.3. p19m:pGGC[5-BrU]G(CUG)$_3$C[5-BrU]GCC complex

It was decided to use a bromine derivative of the sequence **#2** containing a 5-bromouracil in U positions #4 and #16 along the strand: sequence **#3** in **Table 3**. This was done to make sure that the RNA was packed inside the p19 “caliper” in only one position. Strong anomalous peaks in the electron density map were corresponding to the positions of the bromine atoms, and if the RNA arrangement was only one – that of the 19 base pair duplex – there would be no more than two anomalous peaks per asymmetric unit (which contained one protein monomer and half of the RNA duplex).

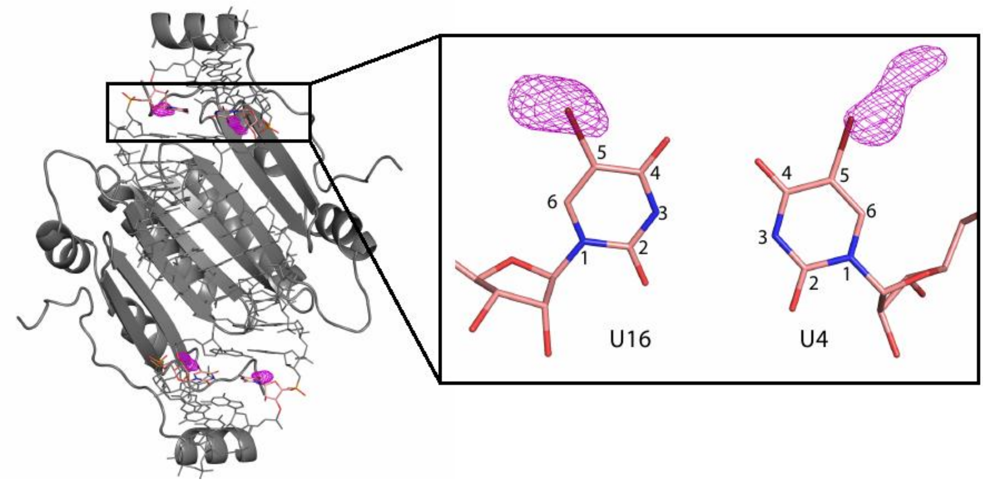


Figure 25. Confirmation of mismatch -containing RNA sequence accommodation in complex with p19 using 5-bromouridine substituted RNA.
The sequence and orientation of the dsRNA in the structure was confirmed using data collected from a crystal substituted with 5-bromouridine at positions 4 and 16 of the dsRNA. Crystals of p19 and 19 base pair dsRNA substituted with 5-bromouridine at position 12 were grown under identical conditions as the native complex. (2Fo-Fc) electron density maps at 1.96 Å resolution calculated using the protein alone contoured at 5σ is shown in magenta. While one of the peaks has two positions, both of them peaks to bromine atoms at the 5 positions of the 4 and 16 uridine bases. With the R32 symmetry, this gives 4 peaks per RNA duplex.

Data were collected at beamline ID14-4 (ESRF) up to resolution of 1.96 Å with the kind help of Dr. Gleb Bourenkov. The structure was refined with the use of phasing program OASIS (from the CCP4 suite). Anomalous electron density map was calculated separately and it was shown clearly that only two anomalous peaks existed in the asymmetric unit, corresponding to the bromine atoms of the 5-bromouracil residues in the (2Fo-Fc) density map contoured at 5σ cutoff level (**Figure 25**).

The 5-bromine modification changes the geometry of the double helix in the 5BrU*5BrU region due to the restricting effect of Br on the conformational freedom. The different mobility of the two bromine-bearing nucleotides can be noted in **Figure 25**: the map of 5-bromine suggests two preferred conformations of the U4 residue and less deviation for U16. The 5-BrU residues are have a decreased λ angle of 44.9-46.0° as compared to the respective 49.1-54.8° of the unmodified RNA, being thus shifted towards the minor groove. This also implicates a slightly different stacking pattern for the modified mismatch: the overlap areas for CBrU/BrUG and BrUG/CBrU are in the 2.78-2.83Å^2 range, which is increased compared to the 1.67-1.73Å^2 for the respective steps in the unmodified molecule (see **Table 3**). The extensive involvement of the bromine atom into the stacking interactions especially increases the overlap of 5-bromouracil with the subsequent guanine 1.43Å^2 (SD=0.09), whereas for the uracil-guanine it was 0.48Å^2 (SD=0.35). This can compensate the decrease of the O4-O4 repulsive contact from 2.85-3.29Å in the unmodified U*U mismatch to 2.50Å. The BrU*BrU has a slightly increased twist of 33.51° (as compared to 30.56 in the unmodified pair) and the preceding C-G pair has a sharply decreasing twist of 31.71° (37.91 in the U*U). Nevertheless, it is compensated elsewhere along the duplex, and the average twist of 31.47° (SD=3.25) does not deviate from the average for the non-brominated molecule, which is 31.31° (SD=4.68).

Apart from the above stated differences, there is little crystallographic evidence that bromination alters the structure compared with the native RNA. The native and brominated duplexes can be superposed with an r.m.s. deviation of less than 0.5Å.

3.6.2.4. p19cut:pUUG(CUG)$_5$CU complex

An attempt to stabilize the original U•U base-pairs at duplex termini in the p19/RNA crystal structure was done through addition of 5’- overhangs to the original sequence #1. Unlike 3’-overhangs that were facing outward the complex, the 5’termini of the RNA 19-mer faced the protein and made multiple contacts to it. To further prolong these interactions and stabilize 5’termini in the complex, we first introduced one 5’-nucleotide overhang. We wanted (i) to get more structural detail about the 5'-overhang interaction with p19 and (ii) improve crystal quality through additional stabilization of the RNA ends within the complex. Thus, the sequence was **#4** in **Table 3**and the crystal quality and resolution were increased.

Data were collected up to 1.85Å resolution (ESRF, beamline 23-1), showing that the crystal quality was indeed increased if compared to the p19:pUG(CUG)$_5$CUcomplex. The unit cell parameters were ***a*** = ***b*** = 89.362Å, and ***c*** = 147.936Å, space group was R32. After refinement (REFMAC5 from CCP4 suite and simulated annealing with CNS) an electron density map was calculated.

Here, the same stacking area patterns were noted (see **Table 5**), with average area for the GC/GC steps of 12.18 Å^2 (SD=0.59) and for other steps of 1.96 Å^2

(SD=1.11). In this complex, the tendency of decrease of overlap area from ends towards the middle of the duplex can also be observed, as for the pGG(CUG)$_5$CC sequence, with the first step having a clearly higher parameter than the other steps of the non-GC/GC group: UG/GU with overlap of 3.80Å^2 (as compared to 3.96Å^2 for GG/CC step in the pGG(CUG)$_5$CC duplex). The increased roll is associated with mismatch-including steps, and a highest deviation of tilt (-6.81°) was observed, again, for the 4th pair, a U*U mismatch.

The canonical G-C pairs, of which there are 12, are involved in Watson-Crick pairing. The 3rd, 5th and 6th pairs are characterized by increased opening (9.08°, 12.46° and 11.24°, respectively) and have a tendency to lose the hydrogen bonds adjacent to the major groove. The N4–O6 distances are 3.07, 3.46 and 3.60Å.

With the duplex bases well discernible, it could be observed that the U*U pairs acquired two different modes of stabilization, depending on the position of the mismatch. The terminal U*U (situated on the siRNA double-helical region edge), characterized by extensive stacking interactions with the p19 “caliper”, showed considerable deviation in C1'-C1' distance from the canonical A-form parameters - 8.96 Å as compared to approximately 10.5 Å, respectively - and established two inter strand hydrogen bonds: N3-O2 of 3.10 Å and O4-N3 of 2.89 Å (**Figure 26B**). The second U*U from the duplex terminus (i.e. the 4th base pair) also showed similar parameters: C1'-C1' of 8.8Å , N3-O2 of 3.11Å and O4-N3 of 3.34Å. Again, this shows that the base pairs (and especially mismatches) closer to the ends tend to deviate from the canonical A-RNA parameters.

The other “internal” U*U mismatches showed a C1'-C1' in the range 9.3-10.3

Å, closer to the A-form, and were not found to form any inter strand hydrogen bonds (**Figure 26A**). Moreover, the "internal" mismatches usually possessed a worse electron density map as compared to other nucleotides, suggesting the absence of one preferred position for such pairs. It is a major difference from the U*U pairs in the free RNA crystal, where, as was mentioned before, the mismatches can be interpreted unequivocally in the map.

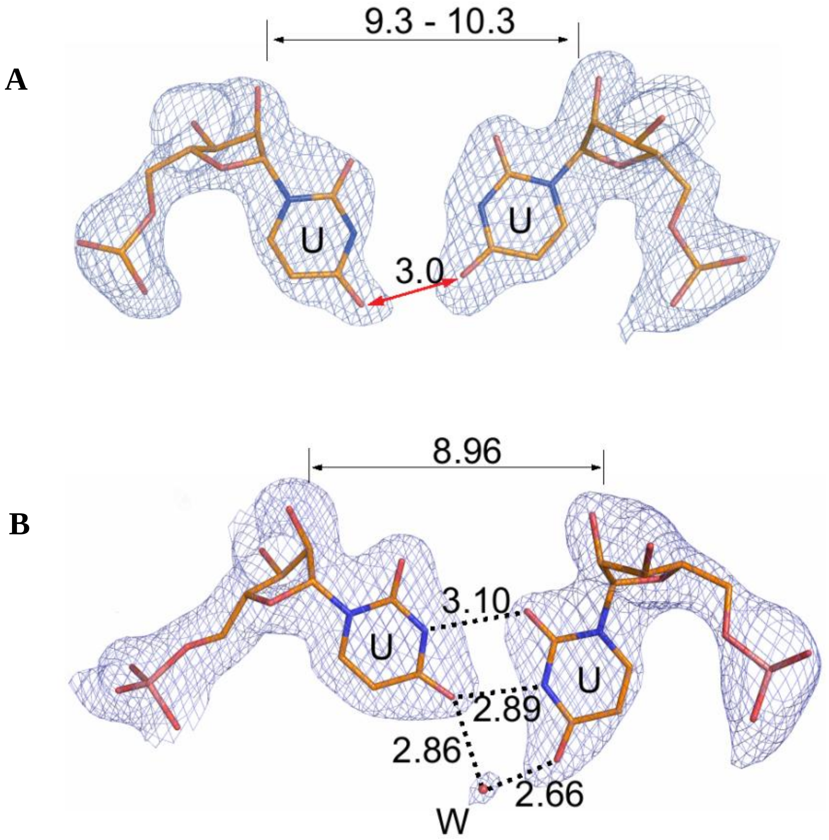


Figure 26. U*U mismatch pair stabilization inside p19:pUUG(CUG)5CU complex. Mismatched base pairs in the (2Fo-Fc) electron density maps contoured at 0.8σ are shown in blue, dashed lines represent hydrogen bonds, while red arrows illustrate close repulsive contacts. Black arrowed lines point to the C1' atoms, with values above the lines corresponding to the C1'-C1' distances in base pairs. Atom coloring code is "blue" for nitrogen, "red" for oxygen, "orange" for carbon and "pink" for phosphorous.
(A) Typical internal U*U base pair shows no stabilizing hydrogen bonds.
(B) Terminal U*U base pair stabilized by two direct hydrogen bonds and one water bridge in RNA/p19 complex

Nevertheless, at least one of the preferred U*U stabilization modes for these pairs can be characterized, as in the previous complexes, by repulsive O4-O4 contacts (values in the range of 2.69-3.35Å).

The overhang sugar and phosphate were well visible and their localization inside the protein could be traced. It was found that p19 protein had a special "pocket" for 5'-overhanging nucleotide, suggestively a binding site for the unwound ends of the siRNA. In **Figure 27**, the three possible variants of 5'-termini are presented - 0, 1 or 2 nucleotide 5'-overhangs (**Figure 27 A**, **B** and **C**, respectively). These three types of overhangs were studied in p19 complexes with sequences **#2**, **#4** and **#5** (the latter will be discussed further, see **Section 3.6.2.5**) from **Table 3**. Their accommodation in the protein cavity is depicted in the right column with spheres (RNA) and surface (protein), a stick representation highlighting the important contact with the protein residues in the middle column and accompanying schematics in the right column. The 5'-phosphate of the first duplex nucleotide (**Figure 27A**), was already established to be important for binding to p19 in previous work [81], and the phosphate of the overhang nucleotide brings new hydrogen bonds, which create a network of interactions (**Figure 27B**).

In this structure it was possible to investigate the 5'-overhang "pocket" properties. As can be seen from the **Figure 27B**, the three hydrogen bonds of the 5'-phosphate in non-overhang RNA (K38, Y51) as well as the stacking interactions of the bases with tryptophane residues of the p19 "caliper" are supplemented with four new contacts with the addition of one 5'-overhang nucleotide. Those are: one new stacking interaction of the overhang base moiety with Q31 side chain a hydrogen bond with the main polypeptide chain (the peptide nitrogen of G96), and two hydrogen bonds with R93 guanidine group connecting it to the phosphates of the 13th and 12th nucleotides of the opposite RNA strand of the duplex (**Figure 27B**, schematic). It could be suggested to be the mechanism for the viral suppressor to bind longer siRNAs *in vivo*.

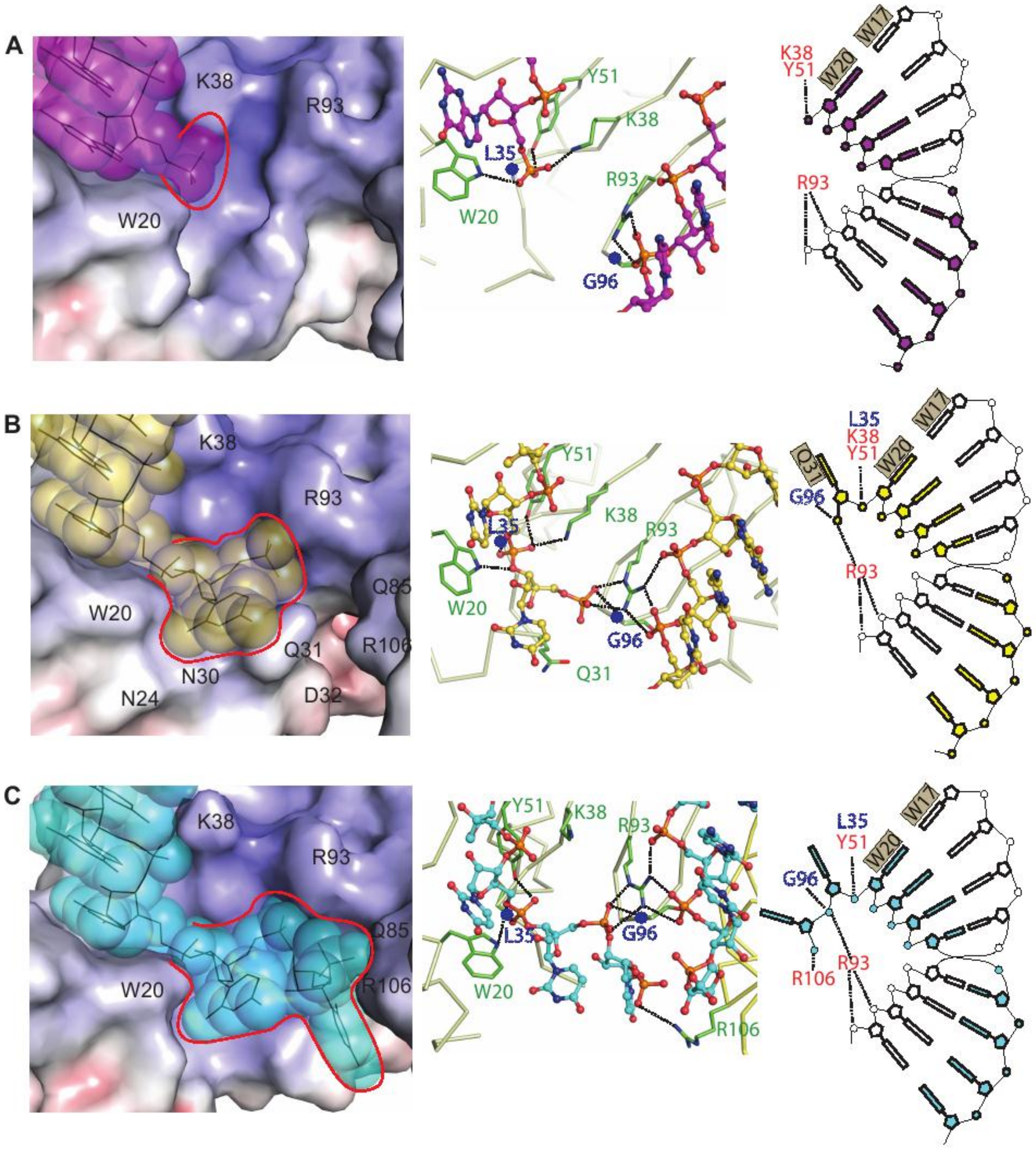


Figure 27. Accommodation of RNA 5'terminus with different overhang lengths in complexes with p19

The interactions between p19 amino acid residues and the RNA 5'-phosphate end; on the left panels p19 cavity is depicted by electrostatic surface view (blue, positive; red, negative; gray, neutral, some important amino acid residues are marked) with RNA atoms presented as transparent spheres with backbone in black; corresponding intermolecular interactions are shown in the corresponding middle panels; a schematic of interaction networks in the right panels.

On the schemes bonds are shown by dashed lines. Protein Cα-backbone is shown in pale-green, while important side chains are in green. Oxygen, nitrogen and phosphate atoms are shown in red, blue and orange respectively.

(A) pGG(CUG)5CC:p19 complex, RNA shown in magenta. There is no overhang present, the 5'-phosphate is highlighted with red line.

(B) pUUG(CUG)5CU:p19 complex, RNA shown in yellow. There is 1 nucleotide (uridine) 5'-overhang present, which is highlighted with red line.

(C) pUUUG(CUG)5CU:p19 complex, RNA shown in cyan. There is 2 nucleotides (uridines) 5'-overhang present, which is highlighted with red line.

3.6.2.5. p19cut:pUUUG(CUG)$_5$CU complex

To investigate the structural detail of p19 protein binding to longer siRNAs, a new sequence (**#5** in **Table 3**) was designed to introduce the two-nucleotide 5'-overhang into the duplex structure. The diffraction and overall crystal quality were decreased compared to one-nucleotide 5'-overhang-bearing duplex (**Table 4-II**). The details of p19 interactions with 2-nucleotide 5'-overhang were presented in previous paragraph (**Figure 27C**), where they were compared and contrasted to the binding patterns of 0- and 1-nucleotide overhangs. It was found that the 5'-phosphate of the "extra" nucleotide established an additional hydrogen bond with R109 residue of the other protein monomer (**Figure 27C**).

The observations concerning 5'-overhang binding show the principal ability of p19 to bind longer siRNAs by unwinding the terminal base pairs. The protein seems to have a special cavity for allocating the 5'-end unwound nucleotides, mostly by means of hydrogen contacts to the phosphates. The 3'-end nucleotides are accommodated outside of the cavity [81]. Although 2-nt 3'-overhang is present in Dicer products, it was found unimportant for p19:siRNA interaction in crystallization experiments: not clearly seen [82], yield good-quality crystals with no 3'-overhang.

The pair parameters are close to those of the pUUG(CUG)$_5$CU (sequence **#4** in **Table 3**) in complex with p19 (data not shown). Namely, the 5th G-C pair has a weak O6-N4 contact of 3.16Å and an increased opening of 8.45. Mismatches, of which there are 7, also resemble those of the previous complex. The terminal U*U is stabilized by two hydrogen bonds N3-O2 of 2.78Å and O4-N3 of 2.67Å with C1'-C1' distance of 8.8Å. The 4th pair (second mismatch from the terminus) has a C1'-C1' of 8.9Å and a hydrogen bond O4-N3 of 2.80Å. Other two U*U mismatches have repulsive contacts O4-O4 of 3.18 and 2.73Å and C1'-C1' distances of 9.9 and 10.8Å.

3.6.2.6. p19cut:pG(CUG)$_6$C complex

It was established by biochemical methods (EMSA, [81]) that p19 was capable to bind, albeit with lower efficiency, RNA duplexes longer than 19 base pairs. The details on 5’-overhang binding, discussed above (**Sections 3.6.2.4** and **3.6.2.5**), permitted us to propose a mechanism of RNA unwinding upon binding to p19. It was decided to test this suggestion from the structural point of view. A 20 base pair nucleotide was chosen correspondingly to our goals: possessing a 2-fold symmetry and constructed of consecutive uninterrupted CUG repeats: sequence **#6** in **Table 3**.

The complex was crystallized and crystals were harvested and flash-frozen as mentioned earlier. Data were collected up to 2.0Å resolution at ESRF beamline ID23-1 with unit cell parameters ***a*** = ***b*** = 89.781Å, and ***c*** = 148.223Å and space group was R32 (**Table 4-II**). The initial model for refinement was the structure **#4** (**Table 3**). An omit map was calculated and it was clearly visible that only 19, and not 20, base pairs were inside of the p19 “caliper”.

Three variants of the RNA accommodation could be proposed (**Figure 28A**): two of the base pairing shift generating a 19 base pair duplex part and either a 3'- or a 5'-overhang and the third one consisting in the unwinding of one terminal pair of the 20 base pair duplex. The first arrangement was rejected after the finding of the 5'-overhang (especially the 5'-phosphate and the ribose) in the difference map, collocated inside the protein cavity. The latter arrangement was chosen as the most probable as the unwinding of one base pair seemed energetically more profitable than losing 13 Watson-Crick GC pairs in favor of 12 less stable G*U pairs. Indeed, after refining the structure it was found that the crystal turned to be a twin, i.e. the RNA could adopt one of the two positions (**Figure 28B**,**C**) with 50% probability. It meant that the density map for every base pair position in this structure was in fact a superposition of signal for two base pairs: either G•C over C•G or a G•C over U•U (**Figure 28D**) and the unwound base pair was to be observed at both duplex edges with occupancy of 0.5. To resolve the superposition, a special script for REFMAC5 was used (including the new “TWIN” command for working with crystallographic twinning). Again, the RNA constitutes an A-form without major disturbances. Sequence-independent helical parameters were calculated using the C1' atoms in order to avoid possible calculation errors brought by non-canonical U*U mismatches. The average values of displacement, angle, rise and twist parameters are 7.33Å (SD=0.67), 13.77° (SD=2.67), 2.64Å (SD=0.36) and 31.52° (SD=4.19).

The overlap area difference pattern is analogous to the above stated complex (see **Table 5**). Namely, the GC/GC steps possess an average overlap area of 11.28Å² (SD=1.98), and the other steps 2.08Å² (SD=0.99). The increased roll parameter was also associated with the mismatch steps: 11.20° (SD=2.95), as compared to that of GC/GC steps which was 1.04° (SD=4.73).

The 13 canonical G-C pairs in this complex were all involved in Watson-Crick interactions. Here, the base pairs 4, 14 and 17 possessed the increased opening of 16.91°, 11.02° and 12.73°. The 4th pair appeared to lack two out of three of the canonical hydrogen bonds – the distances for N1-N3 (3.24Å) and O6-N4 (3.90Å) were increased. The 14th and 17th pairs had the distances N2-O2, N1-N3 and O6-N4 out of the range for normal hydrogen bonds (longer than 3.2Å). Although these perturbations could have been imposed by the symmetry of the RNA duplex – which, as was stated earlier, was refined resolving the two unequal superimposed base pairs in each position – the analysis of the self-

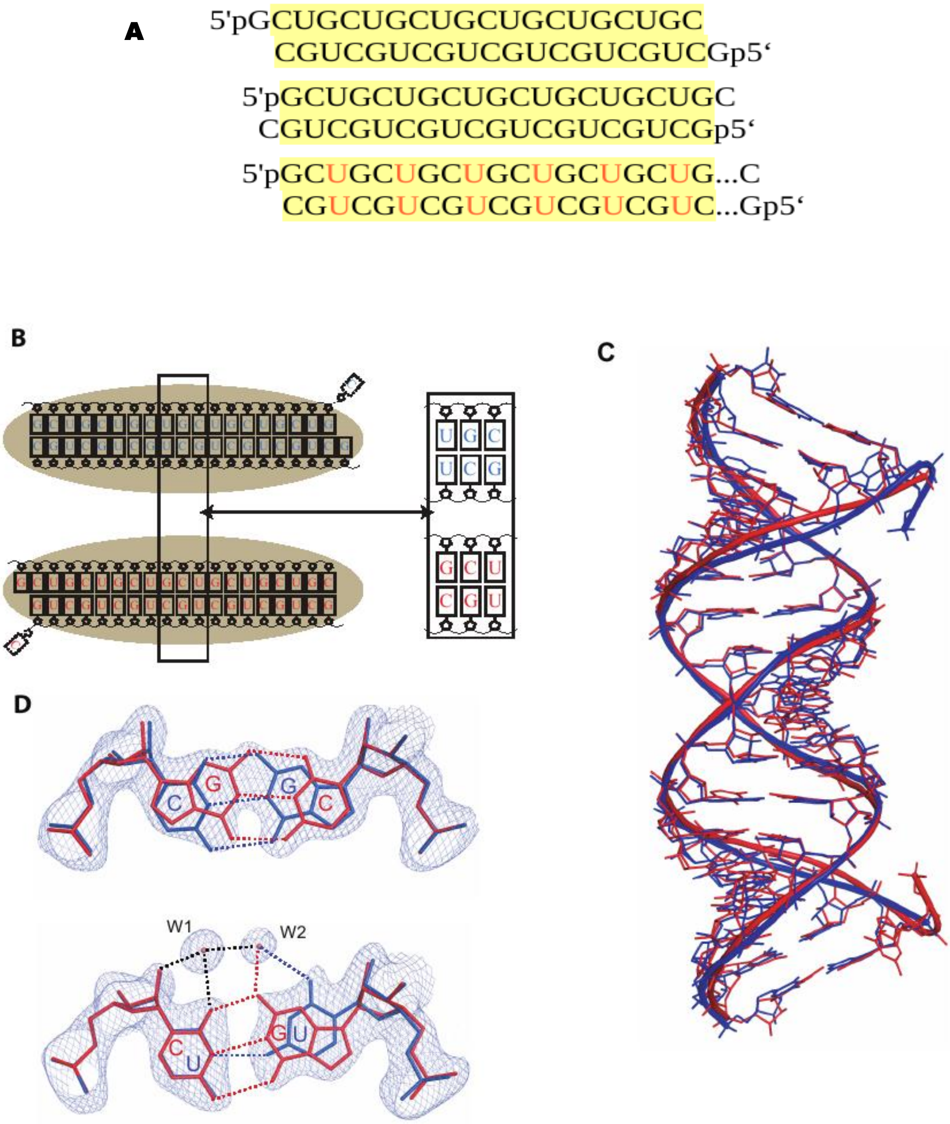


Figure 28. Binding of longer 20-basepair duplex by 19-basepair length specific p19

(A) Possible packing modes of the 20-mer pG(CUG)6C duplex in complex with p19: base pairing shift with generating a 1-nucleotide either (i) 5'- or (ii) 3'-overhang or (iii) one base pair unwinding (unwound base pair connected to the rest of the duplex via dotted lines). Uridine residues in mismatch pairs are depicted in orange font; double-helical regions of siRNAs are marked by yellow background.

(B) Schematic illustrating a superposition inside p19 cavity of two different pG(CUG)6C duplex positions (blue and red) with occupancy of 0.5 each.

(C) Overall view of the overloading duplexes with 0.5 occupancy each. Protein is not shown.

(D) Examples of basepair overload. Each basepair, depicted either in red or in blue, has occupancy of 0.5, (2Fo-Fc) electron density maps contoured at 0.8σ are shown in blue.

symmetrical sequence **#2** (see **Table 3**) may imply bending of the double helix in these regions in order for the RNA to adjust to the protein geometry.

The U*U mismatches do not form hydrogen bonds and mostly have repulsive O4-O4 contacts. However, mismatches of uridine (chain D, residue 12) – uridine (chain E, residue 9) does form an O4-N3 contact of 2.94Å, whereas U*U (chain D, residue 9 – chain E, residue 12) have O4-N3 of 3.40Å. This is reflected in the λ angles: 30.6°, 60.4°, and 24.6°, 64.5°, respectively. Notably, the C1'-C1' distance is 10.5Å and 10.9Å for these pairs, while it seems to decrease for the mismatches positioned closer to the RNA duplex terms: pairs 3D-18E, 15D-6E, 18D-3E have C1'-C1' of 8.6-9.2Å. This may again mean that the duplex parameters are closer to the canonical A-RNA near the center and tend to deviate more on the edges.

As was pointed out earlier, this complex has a one-nucleotide 5'-overhang with occupancy of 0.5.

Interestingly, whereas the 5'-nucleotide of the unwound base pair was visible in the map, the 3'-nucleotide could not be discerned in the electron density map. It is known that siRNAs *in vivo* possess a 2-nucleotide 3'-overhang – left after the processing of longer double-stranded RNA tracts by Dicer endonuclease.

3.7.Analysis of results

3.7.1. Free CUG repeats constitute A-RNA

In our work, we aimed to study CUG repeats in the light of their hypothetical involvement into cellular silencing pathways during TREDs pathology [69]. For this reason, we chose a longer 20-mer pG(CUG)$_6$C (sequence **#0** in **Table 3**) for crystallization studies in contrast to 18- and 8-mers used in previous works [70,98]. We employed a longer 20-mer duplex without having to overcome difficulties connected to crystal symmetry and detwinning [98]. Importantly, our CUG tracts possess a 5'-phosphate, which is observed in natural siRNAs [42,43].

It was additionally proved by means of X-ray crystallography, that free CUG-repeating RNA of sufficient length to constitute a siRNA (≥19 base pairs) forms A-RNA. The double helix in this case represents a special sequence pattern, where every U*U mismatch is flanked from both sides by two Watson-Crick G-C pairs. Each G-C pair has three direct hydrogen bonds whereas 7 from 9 mismatches present in the asymmetric unit (one duplex A+B and one half of the duplex D+D') are stabilized by one direct and one or two water-mediated hydrogen bonds (two other mismatches were observed to form only repulsive O4-O4 contacts).

3.7.2. Characteristic stacking patterns of CUG repeats

Apart from hydrogen bonding, the GC/GC step is known to have the highest stacking in right-hand double helices [109]. Therefore, a periodically repeating stacking motif of one high-overlapping step followed by two with low overlaps arises in CUG-repeating RNA duplexes. This characteristic pattern of CUG repeats is depicted in **Figure 29**, where an overlap area peak can be observed in every third position flanked by two low-stacking GU/UC steps (note that where the terminal mismatch base pairs were changed for G-C aiming to increase duplex stability, the periodic pattern is interrupted). The stacking diagrams in **Figure 29** illustrate the difference in stacking.

These two characteristics of G-C pairs add significantly to the stability and integrity of the double helix, as it was proved that G-C-enriched trinucleotide repeats show more tendency towards forming secondary structures as opposed to repeats lacking them. For example, biochemical and biophysical structure probing proved that (i) UUG, AAG, CUU, CCU, UUA and CCA tracts are unstructured and (ii) AUG, UAG, UUA, CAU and CUA only form unstable hairpins in solution [66]. The absence of any canonical base pairs – with the only possibility of forming wobble U*G pairs in every third position in

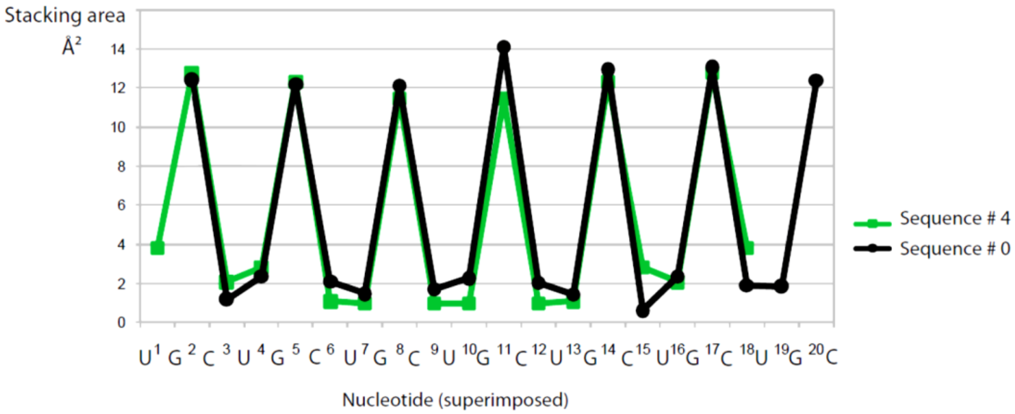


Figure 29. Stacking pattern in CUG-repeating sequences.
The periodic nature of stacking pattern is presented for a free CUG-repeating duplex (sequence #0, black) and a p19-bound duplex (sequence #4, green).The 20- and 19-nucleotide sequences are superimposed by CUG pattern. It can be noted the high stacking GC/GC step in every third position.

UUG - in the first group of tracts prevents the RNA repeat regions from pairing interactions. It is worth noting that the presence of Watson-Crick U-A pairs in the second group of tracts (forming semi-stable hairpins) and in the unstructured UUA tract seems not to be sufficient for yielding a stable A-form.

This can be contrasted to the fact that in the same investigation all CNG (N=U, A, C or G) and two GNC (GUC and GAC) tracts were found to form stable hairpin structures in solution [66]. This comes in agreement with intrinsic stacking energy calculations, which state -14.58 ±0.23 kkal/mol for GC/GC and -9.12± 0.09 for AU/AU steps [109], apart from the difference in hydrogen bond number: 3 per G-C pair and 2 per A-U pair.

3.7.3. Comparison with shorter CUG repeats from earlier studies

The structural characteristics of the CUG motifs observed in our protein-free RNA duplex are similar in general to the ones described in [70]. The atomic resolution of 1.23 Å in [70] permits to discern the details of the bases and backbone positions, particularly, more water molecules can be allocated in the U*U and, notably, in the G-C pair regions. Nevertheless, the overall mode of mismatch stabilization via one direct and one or two water-mediated hydrogen bonds combined with “canonical” G-C flanks and the increased overlap in GC/GC steps are the same. Moreover, the sequence-independent parameters of the RNA double helix, calculated with 3DNA program [85], remain in the standard deviation limits.

The essential novelty of our RNA structure compared to [70] is that it contains a noticeably longer 20 base pair duplex. In addition, in our structure there are 1½ duplexes per asymmetric unit, which permits us to observe a larger range of U*U mismatch stabilization modes – namely, 9 variants. Compared to the 18 base pair duplex described in [98], our crystal data did not require “detwinning”. The main limitation of refining against twinned data with twin fraction close to 0.5 was that meaningful omit maps cannot be calculated. Consequently, although the structure [98] made it possible to draw conclusions about the A-RNA nature of CUG repeats, the mismatches could not be interpreted in the map. In our work, we avoided the problems connected to twinning, which implicates that an unambiguous interpretation of U*U pairs could be achieved.

3.7.4. CUG repeats form complex with protein a silencing suppressor

In this work, we prove for the first time the principal ability of CUG repeats in RNA to form complex with a viral silencing suppressor – i.e. we demonstrate that CUG tracts are potentially able to enter the pathological RNAi pathway. Furthermore, our objective was to investigate the structural changes of the CUG repeats in complex with an RNA-binding protein specific for siRNAs. It is crucial for the dsRNA, in order to be involved in silencing cascade, to interact with different proteins, which form part of the cell silencing machinery [40]. For co-crystallization experiments we used a sequence-unspecific silencing suppressor p19 from TBSV.

For the purposes of optimizing X-ray diffraction, we have selected protein constructs. It was shown earlier, that the wild type p19 yielded lower quality crystals than the truncated p19m with 26 N-terminal and 14 C-terminal amino acids omitted [82]. Here, we designed another, shorter fragment, p19cut, 8 C-terminal amino acids shorter than p19m, which made it possible to achieve better resolution in co-crystallization experiments with CUG repeats. We tried different CUG-repeating RNA sequences for co-crystallization with p19 and managed to obtain high-resolution crystal structures. We had to take into account crystal symmetry, RNA duplex length and stability issues (for details, see **Section 3.2** “Design of RNA sequence for co-crystallization experiments”).

3.7.5. Structural “adaptation” of CUG repeats in complex with a silencing suppressor

The structures of p19 in complex with Watson-Crick RNA were published previously [81,82]. In this study, the nature of structural changes that the U*U mismatches bring into the geometry of A-RNA were analyzed. It was known from previous research [81,82] that the helix axis of the A-RNA double helix in complex with p19 was bent about 40° towards the protein. We compared the overall helix geometry of CUG repeats as free RNA or in a complex with the silencing suppressor. The bending can be observed in (**Figure 30**), where the same RNA sequence, pG(CUG)$_6$C, is superposed in free and p19-bound state (sequences **##0** and **6** in **Table 3**, respectively).

It was found that the U*U mismatch base pairs can adapt to the overall helix geometry, in our case, to helix bending. One of the characteristics of U*U mismatches is their conformational flexibility, because two pyrimidines are narrow as a pair compared with Watson-Crick base pairs and are therefore likely to have more conformational freedom, if placed between two G-C pairs with a C1’C1’ distance characteristic of A-RNA. It was observed that the stabilization modes of mismatches in the free RNA and in p19-bound

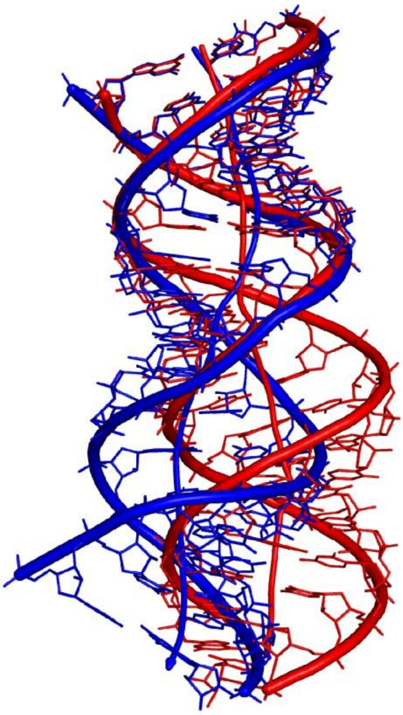


Figure 30. Helical bending of pG(CUG)$_6$C duplex upon binding to p19.
While in complex with p19, the A-RNA duplex shows helix axis bending towards the protein. Here, the duplex in bound (blue) and unbound (red) forms are superposed; superposition was carried out for the terminal part of the duplex. The global helical axis is shown as calculated with program 3DNA [85]. Bending and untwisting of the bound duplex can be observed.

RNA differ from each other. We suggest the reason to be the local compensation of the helix bending in the mismatch areas. U*U mismatches (except for the terminal mismatches) in our complexes are flanked from both sides by 2 C-G base pairs, and the GC/GC steps are characterized by extensive stacking. Therefore, it seems thermodynamically more profitable to shift the geometry of mismatches in favor of preserving the stacking of GC/GC. Local decreases of C1'-C1' distances, roll and tilt deviations (which are characteristic of bending in double-stranded nucleic acids) and repulsive O4-O4 contacts were noted in U*U mismatches and mismatch-including GU/UC steps and are supposedly compensated by the stacking of consecutive CG pairs, their hydrogen bonds and the interactions with the protein. The graphs in **Figure 31** illustrate the fluctuations of C1’-C1’ distance along the free duplex as opposed to p19-bound CUG repeats.

The central part of the bound helix deviates much less from the canonical A-form than its termini. The central part (base pairs from 7 to 13) bears the majority of contacts to p19 amino acids (namely, all hydrogen bond interactions) and seems to be crucial for binding with the protein. This part is supposed to be crucial for making the molecule distinguishable to the protein as its ligand – a double-stranded RNA, as opposed to dsDNA, which would lack the 2'-hydroxyl groups and therefore would not establish the ten water-mediated contacts with p19. This can be noticed in **Figure 31**.

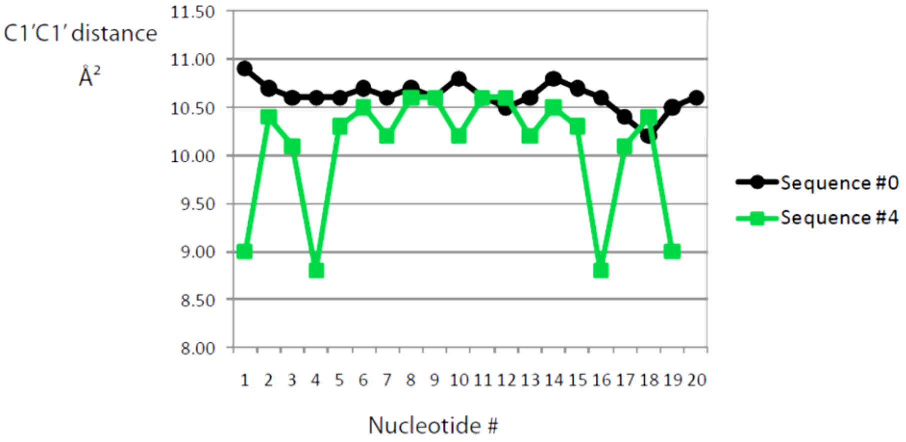


Figure 31. Comparison of C1’C1’ distances in free (sequence #0) and p19-bound (sequence #4) CUG-repeating RNA duplexes.

This brings a difference into the geometry of U*U mismatches. In the central part, the three mismatches seem to be constrained and held by the more strict A-RNA-like sugar phosphate backbone, preferring this position to the possibility of hydrogen bonds. In contrast, in the terminal parts of the helix (base pairs from 1 to 6 and from 14 to 19) U*U pairs have more deviations in roll, tilt and C1'-C1' distances, and tend to establish one or two weak hydrogen bonds. The terminal U*U, if present, has the strongest N3-O2 and O4 – N3 contacts (as it is less held in place by the double helix and depends more on the stacking interaction with Trp18 and Trp20); the the first 5' nucleotide in the double helical region also has a C2'-exo ribose conformation instead of C3'-endo, characteristic of A-RNA.

3.7.6. CUG duplex termini can unwind upon binding to the silencing suppressor

Another part of our structural analysis deals with the CUG-repeating RNA of >19 base pairs binding to p19. The ability of p19, which possesses length specificity towards siRNA duplexes of 19 base pairs, to bind molecules of ≥20 base pairs was reported before [81], but cannot get any explanation in terms of p19/RNA structure. In this work, we have shown the structural details of this interaction and analyzed the changes in CUG repeats upon binding to a protein.

As became clear from our study, silencing suppressor p19 has a specific center to bind the 5' overhang nucleotide and its 5' end phosphate. As it was shown before [81], the first 5' phosphate of the 19-mer siRNA with no overhangs increases the siRNA binding to p19; we observed at least three hydrogen bonds between this phosphate and the protein – W20, Y51 and the main chain nitrogen of L35 (**Figure 27A**). In case of the 1 nucleotide 5' overhang we observed a special “binding pocket” for the sugar and the phosphate of this “extra” nucleotide (**Figure 27B**). While the O2' atom of the sugar contracts the NE1 atom of W20, the phosphate establishes an extensive network of interactions. Being stabilized with a direct hydrogen bond to main chain (G96 nitrogen), it contacts the phosphate backbone of the siRNA (in the region of 13th and 14th nucleotides of the duplex) via R93 guanidium group (**Figure 27B**). In case of 21-mer with two “extra” nucleotides we observe a possible “tunnel” for accommodating the 5' overhangs while unwinding of longer sequences: it seems that the loop 49-52 is flexible and capable of adjusting for siRNAs of different length (**Figure 27C**). This could be the result of a viral adaptation to the variation in length of siRNAs naturally produced by Dicer; also it could be a result of virus adjusting to the pathways of secondary RNAi signals present in plants [64].

We then tried a 20 base pair RNA pG(CUG)6C in co-crystallization with p19. It was shown a possibility to unwind the “extra” base pair in order to bring the siRNA to a 19-base pair double helical region (**Figure 28B**). In this complex only location of 5’ends overhangs could be observed in the density map hence leaving us with two conclusions (i) 3’end overhang does not seem to play special role in binding of p19 to RNA and (ii) 5’end overhang on the contrary might have certain peculiarities.

This illustrates the structural adaptation of CUG repeats to the suppressor, with a possibility of regulation of the A-RNA duplex length by unwinding one or two terminal base pairs

In our study we show the principal ability of the viral silencing suppressor p19 from TBSV to bind siRNAs with CUG repeats that are characteristic of such Trinucleotide Repeat Expansion Disorders as Myotonic dystrophy 1. CUG-repeating siRNAs were shown earlier to be capable of triggering a silencing signal in vivo [71], which could mean their central role in pathology development in TREDs. The possibility of neutralization and elimination of this siRNAs may lead to promising therapeutic prospects for TREDs.

3.8.Overview of results in the light of current TREDs paradigm

TREDs, or trinucleotide repeat expansion disorders, are characterized by a naturally present trinucleotide repeat tract in a particular gene losing stability and thus starting to expand. The diseases are mostly muscle- or neurodegenerative and have a clear genetic nature. Most of them share a common (CNG)n genetic pattern of the trinucleotide repeats (where N is any nucleotide, which vary depending of a particular disorder [11]). These tracts are believed to be sometimes expanding during DNA replication due to a “slippage” of the replication fork during interaction with special structures formed by (CNG)n-repeating DNA [4] and thus lead to pathology. The trigger mechanism of the disorders is, however, still not clear.

It was reported, that there is a strong correlation between the repeat number and the development of pathology. The increasing of CTG repeats number from 5-37 to >50 in the gene coding DMPK protein and from 16-34 to >74 in SCA8 leads, respectively, to pathological phenotypes of myotonic dystrophy 1 and spinocerebral ataxia, both being severe muscle- and neurodegenerative conditions. Such a switch from healthy to pathological phenotype depending on the trinucleotide tract length could suggest a mechanism of disease development, connected closely to the properties of the repeat-bearing RNA. In this issue, it was especially interesting to investigate the peculiarities of CUG tracts.

A hypothesis explaining pathology developing in TREDs, formulated in [69], states that transcribed CNG-repeating RNAs can become involved into silencing processes. After transcription, such molecules are supposed to form hairpin structures and to be processed by cell endonuclease Dicer [71]. Afterwards, slices of such RNA can become the silencing guides built into the silencing complex (RISC) and thus lead to translational arrest or degradation of cognate mRNAs.

Clearly, an in-depth study of CUG-repeating RNA possibility to form hairpin structures capable of undergoing processing by cell silencing machinery and structural detail of mismatch-containing dsRNA was needed. CUG-repeats were already crystallized and discussed previously [70]. In that work, structures of $(CUG)_6$ (1.58 Å) and $G(CUG)_2C$ (1.23 Å) were studied. The first crystal had been yielded previously [98] and in [70] the authors applied a different refinement approach. Nevertheless, the crystal was a twin and the base pairs were resolved with difficulty. The latter crystal with atomic resolution had only two CUG repeats, whereas the natural siRNAs usually consist of 19-21 nucleotides.

In our work, we tried to avoid possible drawbacks caused by space group characteristics while using dsRNAs close to natural size - 19-20 base pairs. It was shown that an RNA consisting of six uninterrupted CUG repeats constitutes an A-form. Our crystal structure of $pG(CUG)_6C$ duplex demonstrates that the mismatch U*U pairs do not disturb significantly the A-form geometry. This implicates that such RNA can be processed by Dicer and thus be involved into the RNAi cascade. Indeed, this reaction has already been studied yielding biochemical evidence *in vitro* and *in vivo* [71].

It is crucial for the dsRNA, in order to be involved in silencing cascade, to interact with different proteins, which form part of the cell silencing machinery. Consequently, our aim was to investigate the structural changes of the CUG repeats in complex with an RNA-binding protein. The protein binding properties of CUG sequences have been mostly studied in relevance to the toxic features of mutant transcripts rather than in the context of the putative normal functions of the tracts [121]. These studies took advantage of various methods to identify proteins that bind these repeats and to characterize such interactions structurally.

First, the CUGBP1 protein was identified on the basis of its specific binding to single-stranded (CUG)8 incubated in HeLa cell nuclear extract. CUGBP1 is a member of the CELF (CUGBP and ETR-3-like factors) protein family, which regulates a number of post-transcriptional RNA processing steps including alternative splicing [122]. It was revealed that the protein only bound to the CUG hairpin base part, not proportionally to the (CUG) repeat number. Further investigation has shown that CUGBP1 does not co-localize with mutant transcripts in DM1 cells [123].

Swanson and colleagues succeeded in identifying an RNA-binding protein, which binds to CUG repeats in a length-dependent manner and regulates alternative splicing [122]. This protein, MBNL1 (muscleblind-like 1 protein) was shown to co-localize with mutant DMPK transcripts in a variety of DM1 patient cells and model organisms [124]. The structures of very short oligomers containing CUG motifs in complex with MBNL1 were determined by crystallography, and the results suggested that MBNL1 may efficiently bind single-stranded CUG repeats [125]. MBNL1 was also shown to bind (CUG)17 and (CAG)17 oligoribonucleotides with similar affinity, whereas non-hairpin forming repeats of the same lengths composed of AUG or UUA repeats did not bind MBNL1 under the same assay conditions [95]. However, structural details of a CUG duplex in complex with a protein have to our knowledge so far remained uncharacterized.

CUG tracts were proven to form hairpin structures [66] and to be processed by Dicer enzyme into repeats of approximately 21 nucleotides [71]. Our study shows that the CUG-repeating RNA of 19 or more base pairs displays a tendency to form a duplex. In this type of molecule, CG Watson-Crick pairs make the double-helix formation thermodynamically profitable (by providing 3 hydrogen bonds each and extensive stacking interactions between two consecutive GC pairs) while the U*U mismatches could bring additional flexibility to such a structure. Indeed, the mismatches are capable of adjusting to the dsRNA geometry dictated by the environment. For example, in free CUG-RNA they form one or two water-mediated hydrogen bonds and stabilize the unbent A-RNA. In bound state they promote the changes to the double helix by (i) losing the hydrogen bonds but keeping the major A-RNA parameters (C1'-C1'), like the central U*U pairs in our complexes, or by (ii) losing the canonical C1'-C1', roll and tilt but providing one or two direct hydrogen bonds, like the terminal mismatches. The CUG repeats in the double-helix form seem stable enough for forming a hairpin that would be then recognized by cell machinery and, at the same time, flexible enough to adapt to RNA-binding proteins that require slight changes in geometry.

4. Main Results and Conclusions

1. The 2.5 Å X-ray crystal structure of RNA oligonucleotide pG(CUG)$_6$C was determined and refined to Rwork/R-free=20.3/27.9. The structure comprises three 20 nucleotide strands that form RNA duplexes, in which all uridine residues are arranged as mismatched U*U base pairs. In total, there are nine U*U mismatches per asymmetric unit, with seven of them clearly displaying a common pattern of pairing supported by one direct and one (or two) water-mediated H-bonding.
2. The observed pattern of U*U mismatch coincides with that of recently found in the atomic resolution structure of short RNA fragment G(CUG)$_2$C [70] and results in U*U size similar to a Watson-Crick base pair. Overall double-helical structure of a "free RNA" CUG-repeat in pG(CUG)$_6$G sequence is similar to that of the 18 nucleotide crystal structure (CUG)$_6$ [98] derived in work [70] from untwined X-ray data
3. The newly determined crystal structure of pG(CUG)$_6$C and its general consistency with the atomic resolution structure of short RNA fragment pG(CUG)$_2$C [70] and 18 nucleotide RNA structure [98] after data untwining [70] prove the tendency of CUG-repeating RNA sequences for the double-helical structure formation. In addition, the structure clearly displays the A-RNA helix geometry, with U*U mismatches adapting this geometry, which indicates the potential of CUG-repeating RNA to appear involved in RNAi related pathways.
4. To prove the ability of complex formation between p19 and CUG-repeating RNA-sequences, the experimental system for

crystallographic study of p19-suppressor complexed with CUG-repeating RNAs were developed. The system involved (i) a truncated construct p19cut used in parallel with previously selected p19m-construct [82] and (ii) six RNA CUG-repeating sequences, stepwise-designed for obtaining the high-resolution crystal structures.

5. X-ray crystal structures of six RNA CUG-repeating sequences in complex with the viral silencing suppressor p19 were studied at resolution of 1.86 Å to 2.5 Å. A property of binding to a protein specifically aimed to eliminate RNAi intermediates testifies in favor of the hypothesis of silencing processes in TREDs (Trinucleotide Repeat Expansion Disorders) pathology suggested in [69].
6. Differences in U*U mismatch stabilization mode between "free" and protein-bound CUG repeat structures were observed. In the "free" form, a U*U pair fits without major disturbances into the A-RNA helix through additional stabilization of intra-base contacts via solvent molecules. In contrast, the mismatches in p19-bound CUG tracts are characterized to either form a repulsive contact or significantly deviate from the canonical A-RNA parameters (i.e. have a shorter C1'-C1' distance); it is clear that in this case the U*U is mainly stabilized by stacking interaction. Thus, U*U mismatches provide for the flexibility and geometric adaptation of the potentially "toxic" repeats to "free" and protein-bound environment achieved via at least three different modes of pairing described in this work.
7. Structural study of silencing suppressor p19 interaction with longer siRNAs was performed: for that, specific CUG-repeating RNA sequences of 20 and 21 nucleotides were designed. A new 5'-overhang binding center was found and described in p19 protein,

and a mechanism of unwinding “extra” base pairs was proposed. This draws attention to the potential importance of the 5’-termini in siRNA, while the focus was predominantly on 3’-overhangs as a characteristic trait of Dicer endonuclease products. From our work, it can be observed that 5’-phosphate and 1- or 2-nucleotide 5’-overhangs can significantly contribute to p19:siRNA interaction.

5. References

1. Moyzis RK, Torney DC, Meyne J, Buckingham JM, Wu J-R, Burks C, Sirotkin KM, and Goad WB (1989). The Distribution of interspersed repetitive DNA sequences in the human genome. *Genomics*, 4:273-289.
2. Stallings RL, Torney DC, Hildebrand CE, Longmire JL, Deaven LL, Jett JH, Doggett NA, and Moyzis R (1990). Physical mapping of human chromosomes by repetitive sequence finger printing. *Proc Natl Acad Sci USA*, **87**:6218-6222.
3. Orgel LE, Crick F. and Sapienza C. (1980). Selfish DNA. *Nature*, **288**:645-646.
4. Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, and Lin CC (1997). Human centromeric DNAs. *Human Genetics*, **100**:291-304.
5. Wright WE, Tesmer VM, Huffman KE, Levene SD, and Shay JW (1997). Normal human chromosomes have long G-rich telomeric overhangs at one end. *Genes and Development*, 11:2801-2809.
6. Moyzis RK (1990). The Human Telomere. *Structure & Methods*, 1:61-67.
7. Caskey CT, Pizzuti A, Fu Y-H, Fenwick RG, and Nelson DL (1992). Triplet repeat mutations in human disease. *Science*, **256**:784-789.
8. Krontiris TG (1995). Minisatellites and human disease. *Science*, **269**:1682-1683.
9. Smith CU (2010). Chapter 24: the coming of molecular biology and its impact on clinical neurology. *Handb Clin Neurol*, **95**:361-72.
10. Pearson CE, Nichol Edamura K, Cleary JD (2005). Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet*,.**6**:729-742.
11. Orr HT, Zoghbi HY (2007). Trinucleotide repeat disorders. *Annu Rev Neurosci*, **30**:575-621
12. Miller JW, Urbinati CR, Teng-Umnuay P, Stenberg MG, Byrne BJ, Thornton CA, Swanson MS (2000). Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J*, **19**:4439-4448.
13. Hagerman RJ, Leavitt BR, Farzin F, Jacquemont S, Greco CM, Brunberg JA, Tassone F, Hessl D, Harris SW, Zhang L, et al. (2004). Fragile-X-associated tremor/ataxia syndrome (FXTAS) in females with the FMR1 premutation. *Am J Hum Genet*, **74**:1051-1056.
14. Napierala M, Krzyzosiak WJ (1997). CUG repeats present in myotonin kinase RNA form metastable “slippery” hairpins. *J Biol Chem*, **272**:31079-31085.
15. Napierala M, Michalowski D, de Mezer M, Krzyzosiak WJ (2005). Facile FMR1 mRNA structure regulation by interruptions in CGG

repeats. *Nucleic Acids Res*, **33**:451-463.

16. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**:77-79.
17. The Huntigton's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**:971-983.
18. Gatchel JR, Zoghbi HY (2005). Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*, **6**:743-755.
19. Kozlowski P, de Mezer M, Krzyzosiak WJ (2010). Trinucleotide repeats in human genome and exome. *Nucleic Acids Res*, **38**(12):4027-39.
20. Fondon JW 3rd, Garner HR (2004). Molecular origins of rapid and continuous morphological evolution. *Proc Natl Sci USA*, **101**:18058-18063.
21. Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y (1999). Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett*, **455**:70-74.
22. Toutenhoofd SL, Garcia F, Zacharias DA, Wilson RA, Strehler EE (1998). Minimum CAG repeat in the human calmodulin-1 gene 5′ untranslated region is required for full expression. *Biochim Biophys Acta*, **1398**:315-320.
23. Richards RI, Holman K, Yu S, Sutherland GR (1993). Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. *Hum Mol Genet*, **2**:1429-1435.
24. Fabre E, Dujon B, Richard GF (2002). Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. *Nucleic Acids Res*, **30**:3540-3547.
25. Li YC, Korol AB, Fahima T, Nevo E (2004). Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol*, **21**:991-1007.
26. Sobczak K, de Mezer M, Michlewski G, Krol J, Krzyzosiak WJ (2003). RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res*, **31**:5469-5482.
27. Jasinska A, Michlewski G, de Mezer M, Sobczak K, Kozlowski P, Napierala M, Krzyzosiak WJ (2003). Structures of trinucleotide repeats in human transcripts and their functional implications. *Nucleic Acids Res*, **31**:5463–5468
28. Philips AV, Timchenko LT, Cooper TA (1998). Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science*, **280**:737–741
29. Taneja KL, McCurrach M, Schalling M, Housman D, Singer RH (1995).

Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues. *J Cell Biol*, **128**:995–1002.

30. Raca G, Siyanova EY, McMurray CT, Mirkin SM (2000). Expansion of the (CTG)(n) repeat in the 5′-UTR of a reporter gene impedes translation. *Nucleic Acids Res*, **28**:3943–3949.
31. Groenen P, Wieringa B. Expanding complexity in myotonic dystrophy (1998). *Bioessays*, **20**:901–912.
32. Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion JP, Hudson T, et al (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell*, **68**:799–808.
33. Koch MC, Grimm T, Harley HG, Harper PS (1991). Genetic risks for children of women with myotonic dystrophy. *Am J Hum Genet*, **48**:1084–91
34. Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, et al. (1992). Myotonic dystro- phy mutation: an unstable CTG repeat in the 3 untranslated region of the gene. *Science*, **255**:1253–55
35. Koob MD, Moseley ML, Schut LJ, Benzow KA, Bird TD, et al. (1999). An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nat Genet*, **21**:379–84
36. Nemes JP, Benzow KA, Moseley ML, Ranum LP, Koob MD (2000). The SCA8 transcript is an antisense RNA to a brain-specific transcript encoding a novel actin-binding protein (KLHL1). *Hum Mol Genet*, **9**:1543–51
37. Holmes SE, O’Hearn E, Rosenblatt A, Callahan C, Hwang HS, et al. (2001a). A repeat expansion in the gene encoding junctophilin-3 is associated with Huntington disease-like 2. *Nat Genet*, **29**:377–78
38. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, and Mello CC (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**:806–811.
39. Hamilton AJ, and Baulcombe DC (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, **286**:950–952.
40. Ghildiyal M, and Zamore PD (2009). Small silencing RNAs: an expanding universe. *Nat Rev Genet*, **10**:94–108.
41. Bernstein E, Caudy AA, Hammond SM, and Hannon GJ (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**:363–366.
42. Elbashir SM, Lendeckel W, and Tuschl T (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev*, **15**:188–200.
43. Elbashir SM, Martinez J, Patkaniowska A, Lendeckel W, and Tuschl T

(2001). Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J*, **20**:6877–6888.

44. Pham JW, Sontheimer EJ (2005). Molecular requirements for RNA-induced silencing complex assembly in the Drosophila RNA interference pathway. *J Biol Chem*, **280**(47):39278-83.
45. Ding SW, Voinnet O (2007). Antiviral immunity directed by small RNAs. *Cell*, **130**(3):413-26.
46. Aliyari R, Ding SW (2009). RNA-based viral immunity initiated by the Dicer family of host immune receptors. *Immunol Rev*, **227**(1):176-88.
47. Okamura K, and Lai EC (2008). Endogenous small interfering RNAs in animals. *Nature Rev Mol Cell Biol*, **9**:673–678.
48. Gangaraju VK, Lin H. MicroRNAs: key regulators of stem cells (2009). *Nat Rev Mol Cell Biol*, **10**(2):116-25.
49. Vasudevan S, Tong Y and Steitz JA (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science*, **318**:1931–1934.
50. Grimson A, et al (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, **455**:1193–1197.
51. Aravin AA, et al (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr Biol*, **11**:1017–1027.
52. Vagin VV, et al (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*, **313**:320–324.
53. Ender C, and Meister G (2010). Argonaute proteins at a glance. *J Cell Sci*, **123**:1819–1823.
54. Hutvagner G, and Simard MJ (2008). Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol*, **9**:22–32.
55. Grewal SI (2010). RNAi-dependent formation of heterochromatin and its diverse functions. *Curr Opin Genet Dev*, **20**:134–141.
56. Voinnet O. (2008). Use, tolerance and avoidance of amplified RNA silencing by plants. *Trends Plant Sci*, **13**:317–328.
57. Gent JI, Lamm AT, Pavelec DM, Maniar JM, Parameswaran P, Tao L, Kennedy S, and Fire AZ (2010). Distinct phases of siRNA synthesis in an endogenous RNAi pathway in C. elegans soma. *Mol Cell*, **37**:679–689.
58. Aoki K, Moriguchi H, Yoshioka T, Okawa K, and Tabara H (2007). In vitro analyses of the production and activity of secondary small interfering RNAs inC. elegans. *EMBO J*, **26**:5007–5019.
59. Gu W, Shirayama M, Conte D Jr, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ, et al. (2009). Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the C. elegans germline. *Mol Cell*, **36**:231–244.

60. Malone CD, and Hannon GJ (2009). Small RNAs as guardians of the genome. *Cell*, **136**:656–668.
61. Tomari Y, Du T, and Zamore PD (2007). Sorting of Drosophila small silencing RNAs. *Cell*, **130**:299–308.
62. Nilsen TW (2008). Endo-siRNAs: yet another layer of complexity in RNA silencing. *Nat Struct Mol Biol*, **15**(6):546-8.
63. Siomi H, Siomi MC (2009). On the road to reading the RNA-interference code. *Nature*, **457**(7228):396-404.
64. Chapman EJ, and Carrington C (2007). Specialization and evolution of endogenous small RNA pathways. *Nature Rev Genet*, **8**:884–896.
65. Ranum LP, Day JW (2004). Pathogenic RNA repeats: an expanding role in genetic disease. *Trends Genet*, **20**:506–512.
66. Sobczak K, Michlewski G, de Mezer M, Kierzek E, Krol J, Olejniczak M, Kierzek R, Krzyzosiak WJ (2010). Structural diversity of triplet repeat RNAs. *J Biol Chem*, **285**(17):12755-64.
67. Fardaei M, Rogers MT, Thorpe HM, Larkin K, HamshereMG, Harper PS, Brook JD (2002). Three proteins, MBNL,MBLL and MBXL, co-localize in vivo with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells. *Hum Mol Genet*, **11**:805–814.
68. Timchenko NA, Patel R, Iakova P, Cai ZJ, Quan L, Timchenko LT (2004). Overexpression of CUG triplet repeat-binding protein, CUGBP1, in mice inhibits myogenesis. *J Biol Chem*, **279**:13129–13139.
69. Malinina L (2005). Possible involvement of the RNAi pathway in trinucleotide repeat expansion diseases. *J Biomol Struct Dyn*, **23**(3):233-5.
70. Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Structural insights into CUG repeats containing the 'stretched U-U wobble': implications for myotonic dystrophy (2009). *Nucleic Acids Res*, **37**(12):4149-56.
71. Krol J, Fiszer A, Mykowska A, Sobczak K, de Mezer M, Krzyzosiak WJ （2007). Ribonuclease dicer cleaves triplet repeat hairpins into shorter repeats that silence specific targets. *Mol Cell*, **25**(4):575-86.
72. Tomari Y, and Zamore PD (2005). Perspective: machines for RNAi, *Genes Dev*, **19**:517–529.
73. Molnar A, Csorba T, Lakatos L, Varallyay E, Christophe Lacomme C, and Burgyan J (2005). Plant virus-derived small interfering RNAs originate predominantly from highly structured single-stranded viral RNAs. *J Virol*, **79**:7812–7818.
74. Voinnet O (2005) Non-cell autonomous RNA silencing. *FEBS Lett*, **579**(26):5858-71.
75. Voinnet O (2005). Induction and suppression of RNA silencing: Insights from viral infections, *Nature Rev Genet*, **6**:206–220.

76. Scholthof HB, Scholthof K-BG, and Jackson AO (1995). Identification of tomato bushy stunt virus host-specific symptom determinants by expression of individual genes from a potato virus X vector. *Plant Cell,* **7**:1157–1172.
77. Qu F, and Morris TJ (2002). Efficient infection of Nicotiana benthamiana by Tomato bushy stunt virus is facilitated by the coat protein and maintained by p19 through suppression of gene silencing. *Mol Plant Microbe Interact*, **15**:193–202.
78. Russo M, Burgyan J, and Martelli GP (1994). Molecular biology of Tombusviridae. *Adv Virus Res*, **44**:381–428.
79. Scholthof HB, Scholthof K-BG, Kikkert M, and Jackson AO (1995) . Tomato bushy stunt virus spread is regulated by two nested genes that function in cell-to-cell movement and host-dependent systemic invasion. *Virology*, **213**:425–438.
80. Silhavy D, Molnar A, Lucioli A, Szittya G, Hornyik C, Tavazza M, and Burgyan J (2002). A viral protein suppresses RNA silencing and binds silencing-generated, 21-to 25-nucleotide double-stranded RNAs. *EMBO J*, **21**:3070–3080.
81. Vargason JM, Szittya G, Burgyán J, Hall TM (2003).Size selective recognition of siRNA by an RNA silencing suppressor. *Cell*, **115**(7):799-811.
82. Ye K, Malinina L, Patel DJ (2003). Recognition of small interfering RNA by a viral suppressor of RNA silencing. *Nature*, **426**(6968):874-8.
83. Lakatos L, Szittya G, Silhavy D, and Burgyán J (2004). Molecular mechanism of RNA silencing suppression mediated by p19 protein of tombusviruses, *EMBO J*, **23**:876–884.
84. Chapman EJ, Prokhnevsky AI, Gopinath K, Dojia VV, and Carrington JC (2004). Viral RNA silencing suppressors inhibit the microRNA pathway at an intermediate step, *Genes Dev*, **18**:1179–1186.
85. Lu X-J, and Olson WK (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, **31**:5108–5121.
86. Olson WK, Bansal M, Burley SK, Dickerson RE, Gerstein M, Harvey SC, Heinemann U, Lu XJ, Neidle S, Shakked Z, et al (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *J Mol Biol*, **313**:229–237.
87. Dickerson RE, Bansal M, Calladine CR, Diekmann S, Hunter WN, Kennard O, et al (1989). Definitions and nomenclature of nucleic acid structure parameters. *EMBO J*, **8**:1–4.
88. Dauter Z. Data-collection strategies (1999). *Acta Crystallogr D Biol Crystallogr*, **55**(10):170317.
89. Protein Purification - Handbook. [Brochure] (2001). Amersham

Biosciences.

90. Golden BL and Kundrot CE (2003). RNA crystallization. *J Struct Biol,* **142**:98-107.
91. Obayashi E, Oubridge C, Krummel DP, Nagai K (2007). Crystallization of RNA-Protein Complexes. In: Doublie, S. (Ed.), Macromolecular Crystallography Protocols. Volume 1: Preparation and Crystallization of Macromolecules. 259-276. Totowa, New Jersey: Humana Press.
92. Landgraf, P. et al (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**(7):1401-14.
93. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, Whisstock JC (2005). Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res*, **15**:537-551.
94. Ranum LP, Cooper TA (2006). RNA-mediated neuromuscular disorders. *Annu Rev Neurosci*, **29**:259-77.
95. Mykowska A, Sobczak K, Wojciechowska M, Kozlowski P, Krzyzosiak WJ (2011). CAG repeats mimic CUG repeats in the misregulation of alternative splicing. *Nucleic Acids Res*, **39**(20):8938-51.
96. Roth BM, Pruss GJ, Vance VB (2004). Plant viral suppressors of RNA silencing. *Virus Res*, **102**(1):97-108.
97. Qu F, Morris TJ (2005). Suppressors of RNA silencing encoded by plant viruses and their role in viral infections. *FEBS Lett*, **579**(26):5958-64.
98. Mooers BH, Logue JS, and Berglund JA (2005). The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc Natl Acad Sci USA*, **102**:16626–16631.
99. Derewenda ZS, Vekilov PG (2006). Entropy and surface engineering in protein crystallization. *Acta Crystallogr D Biol Crystallogr*, **62**(Pt 1):116-24.
100. Khoury GA, Baliban RC, and Floudas C A (2011). Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep*, **1**:90.
101. Laemmli UK (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**(5259):680–685.
102. Layne E (1957). Spectrophotometric and turbidometric methods for measuring proteins. *Meth Enzymol*, **3**:447
103. Stoscheck, CM (1990). Quantitation of Protein. *Methods in Enzymology*, **182**:50-69.
104. Russell DW, Sambrook J (2001). Molecular cloning: a laboratory manual. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory.
105. Korade-Mirnics Z, Babitzke P, Hoffman E (1998). Myotonic dystrophy: molecular windows on a complex etiology. *Nucleic Acids Res*, **26**(6):1363-8.

106. Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol*, **185**:60-89.

107. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, **54**(Pt 5):905-21.

108. Navaza J (1994). *Acta Cryst*, **A50**:157–163.

109. Svozil D, Hobza P, Sponer J (2010). Comparison of intrinsic stacking energies of ten unique dinucleotide steps in A-RNA and B-DNA duplexes. Can we determine correct order of stability by quantum-chemical calculations? *J Phys Chem B*, **114**(2):1191-203.

110. Collaborative Computational Project, Number 4 (1994). Collaborative computational project number 4. The CCP4 suite: programmes for protein crystallography. *Acta Crystallogr D Biol Crystallogr* **50**:760–763.

111. Rayment I (2002). Small-scale batch crystallization of proteins revisited: an underutilized way to grow large protein crystals. *Structure*, **10**(2):147-51.

112. EU 3-D Validation Network (1998). Who Checks the Checkers? Four Validation Tools Applied to Eight Atomic Resolution Structures. *J Mol Biol*, **276**:417-436

113. Morris AL, MacArthur MW, Hutchinson EG and Thornton JM (1992). Stereochemical quality of protein structure coordinates. *Proteins*, **12**:345-364.

114. Laskowski RA, MacArthur MW, Moss DS and Thornton JM (1993a). PROCHECK: a program tocheck the stereochemical quality of proteinstructures. *J Appl Crystallog*, **26**:283-291.

115. Bruenger AT (1993). Assessment of phase accuracy by cross validation: the free R value. Methods andapplications. *Acta Crystallog D*, **49**:24-36.

116. Sheldrick GM (2008). A short history of SHELX. *Acta Cryst*, **A64**:112-122.

117. Bruenger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998). Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr D Biol Crystallogr*, **54**(Pt 5):905-21.

118. Murshudov GN, Vagin AA, and Dodson EJ (1997). Refinement of Macromolecular Structures by the Maximum-Likelihood method. *Acta Cryst*, **D53**, 240-255.

119. Lamzin VS, and Wilson KS (1993). Automated refinement of protein models. *Acta Cryst*, **D49**:129-147

120. Leontis NB, and Westhof E (1998). Conserved geometricalbase-pairing patterns in RNA. *Q Rev Biophys*, **31**:399–455.

121. Krzyzosiak WJ, Sobczak K, Wojciechowska M, Fiszer A, Mykowska A, Kozlowski P. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target (2011). *Nucleic Acids Res*, **40**(1):11-26.

122. Timchenko LT, Miller JW, Timchenko NA, DeVore DR, Datar KV, Lin L, Roberts R, Caskey CT, Swanson MS (1996). Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. *Nucleic Acids Res*, **24**:4407-4414.

123. Lin X, Miller JW, Mankodi A, Kanadia RN, Yuan Y, Moxley RT, Swanson MS, Thornton CA (2006). Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy. *Hum Mol Genet*, **15**:2087-2097.

124. Fardaei M, Larkin K, Brook JD, Hamshere MG (2001). In vivo co-localisation of MBNL protein with DMPK expanded-repeat transcripts. *Nucleic Acids Res*, **29**:2766-2771.

125. Teplova M, Patel DJ (2008). Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat Struct Mol Biol*, **15**:1343-1351.