

Structural basis for triplet repeat disorders: a computational analysis

Pierre Baldi^{1,*}, Søren Brunak², Yves Chauvin³ and Anders Gorm Pedersen²

¹Department of Information and Computer Science and Department of Biological Chemistry, College of Medicine, University of California, Irvine, Irvine, CA 92697-3425, USA, ²Center for Biological Sequence Analysis, The Technical University of Denmark, DK-2800 Lyngby, Denmark and ³Net-ID, Inc., San Francisco, CA 94114, USA

Abstract

Motivation: Over a dozen major degenerative disorders, including myotonic dystrophy, Huntington's disease and fragile X syndrome, result from unstable expansions of particular trinucleotides. Remarkably, only some of all the possible triplets, namely CAG/CTG, CGG/CCG and GAA/TTC, have been associated with the known pathological expansions. This raises some basic questions at the DNA level. Why do particular triplets seem to be singled out? What is the mechanism for their expansion and how does it depend on the triplet itself? Could other triplets or longer repeats be involved in other diseases?

Results: Using several different computational models of DNA structure, we show that the triplets involved in the pathological repeats generally fall into extreme classes. Thus, CAG/CTG repeats are particularly flexible, whereas GCC, CGG and GAA repeats appear to display both flexible and rigid (but curved) characteristics depending on the method of analysis. The fact that (1) trinucleotide repeats often become increasingly unstable when they exceed a length of approximately 50 repeats, and (2) repeated 12-mers display a similar increase in instability above 13 repeats, together suggest that approximately 150 bp is a general threshold length for repeat instability. Since this is about the length of DNA wrapped up in a single nucleosome core particle, we speculate that chromatin structure may play an important role in the expansion mechanism. We furthermore suggest that expansion of a dodecamer repeat, which we predict to have very high flexibility, may play a role in the pathogenesis of the neurodegenerative disorder multiple system atrophy (MSA).

Contact: pfbaldi@ics.uci.edu, yves@netid.com, brunak@cbs.dtu.dk, gorm@cbs.dtu.dk

major degenerative disorders have been elucidated in rapid-fire sequence (Ashley and Warren, 1995; Ross, 1995; Gusella and MacDonald, 1994; Hardy and Gwinn-Hardy, 1998; Rubinshtein and Hayden, 1998). These diseases include:

1. Huntington's Disease (HD) (Huntington's Disease Collaborative Research Group, 1993; Ross and Hayden, 1998).
2. Dentatorubral-pallidoluysian atrophy (DRPLA) (Koide *et al.*, 1994; Miwa, 1994; Nagafuchi *et al.*, 1994; Ikeuchi *et al.*, 1995; Tsuji, 1998).
3. Several forms of spinocerebellar ataxia (SCA-1, SCA-2, SCA-3, SCA-6, SCA-7) and Friedreich's ataxia (Orr *et al.*, 1993; Campuzano *et al.*, 1996; Junck and Fink, 1996; Paulson *et al.*, 1997; Koenig, 1998; Lee, 1998; Orr and Zoghbi, 1998; Paulson, 1998; Pulst, 1998; Stevanin *et al.*, 1998).
4. Spinobulbar muscular atrophy (SBMA) or Kennedy's Disease (Spada *et al.*, 1991; Brooks and Fischbeck, 1995; Beitel *et al.*, 1998).
5. Fragile X syndrome (FRAXA) and its variations with milder or no phenotypical effects (FRAXE, FRAXF) (Nelson, 1995; Gusella and MacDonald, 1994; Eichler and Nelson, 1998; Skinner *et al.*, 1998; Gecz and Mulley, 1999).
6. Myotonic dystrophy (MD) (Timchenko *et al.*, 1995; Gusella and MacDonald, 1994; Hamshere *et al.*, 1998).

All these diseases have a predominantly hereditary component and are characterized by the expansion of existing trinucleotide repeats upon transmission from parent to offspring (see Table 1 for a summary of important features of triplet repeat disorders). Here we briefly review characteristics of these diseases with an emphasis on features

Introduction

In the last few years, the genetic origins of over a dozen

*To whom all correspondence should be addressed.

Table 1. Summary of features associated with triplet repeat diseases. Sex bias refers to the sex in which the repeat is most often expanded. The number of triplet repeats in normal and affected individuals is noted. If known, the number of repeats above which the repeat becomes unstable is also noted

Disorder	Repeat	Location	Number of repeats	Sex bias
Kennedy's disease (spinobulbar muscular atrophy)	CAG	Coding	Normal 9–36 Affected 38–62 Unstable >47	Paternal
Huntington's disease	CAG	Coding	Normal 6–35 Affected 36–121 Unstable >35	Paternal
Spinocerebellar ataxia 1	CAG	Coding	Normal 6–35 Affected 40–81	Paternal
Spinocerebellar ataxia 2	CAG	Coding	Normal 14–32 Affected 33–77	Paternal
Spinocerebellar ataxia 3	CAG	Coding	Normal 12–40 Affected 67–82	Paternal
Spinocerebellar ataxia 6	CAG	Coding	Normal 4–17 Affected 20–30	
Spinocerebellar ataxia 7	CAG	Coding	Normal 7–17 Affected 38–130 Unstable >35 (?)	Paternal
Dentatorubropallidoluysian atrophy (DRPLA)	CAG	Coding	Normal 3–36 Affected 49–88	Paternal
Autosomal dominant pure spastic paraplegia (ADPSP)	CAG	(Coding ?)		
Myotonic dystrophy	CTG	3' UTR	Normal 5–30 Affected 50 to >700 Unstable 36 – 50	Maternal
Fragile X syndrome (FRAXA)	CGG	5' UTR	Normal 5–52 Affected 200 to >1000	Maternal
FRAXE mental retardation	GCC	5' UTR	Normal 6–25 Affected 130 to >700 Unstable > 60	
Friedreich's ataxia	GAA	Intron	Normal 7–22 Affected 200 to >900 Unstable 34–65	

connected with the expansion mechanism. Based on the assumption that this is closely connected to the unusual DNA-helical structural features of the repeat sequences, we then present a computational analysis of the structural features of all possible repeat classes. We also speculate about similar expansions of longer-base repeats.

Mechanism of action

The exact mechanism by which a triplet repeat mutation causes disease varies. This is clearly indicated by the fact that the currently known repeat expansions are found both

in 5' UTRs, in 3' UTRs, in introns and within coding sequences of various affected genes (Table 1) (Ashley and Warren, 1995; Gusella and MacDonald, 1994; Rubinsztein and Amos, 1998; Rubinsztein and Hayden, 1998). For instance, fragile X mental retardation is associated with an expanded CGG repeat in the 5' UTR of the *FMR1* gene (Nelson, 1995; Eichler and Nelson, 1998). This results in hypermethylation of CpGs in the repeat and in its immediate proximity, thereby leading to repression of the *FMR1* promoter and consequently a lack of expression of the gene. Several other disorders – including Huntington's

disease – are associated with CAG repeats that are present in coding sequence and are translated into polyglutamine tracts (Orr and Zoghbi, 1998; Rubinsztein and Hayden, 1998). Interestingly, the polyglutamine mutation results in a gain-of-function, possibly by causing toxic intracellular aggregates in neurons of affected patients (Perutz *et al.*, 1994; DiFiglia *et al.*, 1997; Scherzinger *et al.*, 1997; Orr and Zoghbi, 1998; Rubinsztein and Hayden, 1998). The exact length of repeat associated with disease varies between the different disorders, although there is a general tendency that the repeats associated with loss-of-function can be much longer than those expansions that are gain-of-function mutations (Rubinsztein and Hayden, 1998). As an example, the CGG expansion at the FRAXA fragile site (a typical loss-of-function mutation) can be more than 1000 repeats long in affected individuals (Eichler and Nelson, 1998), whereas the CAG repeat in patients with spinocerebellar ataxia-1 (SCA-1) typically lie in the range 40–81 repeats (Orr and Zoghbi, 1998). In this context it is also interesting that a large fraction of homeopeptide repeat-containing proteins in *Drosophila*, man and mouse have been found to play a role in the development of the central nervous system (Karlin and Burge, 1996).

Non-Mendelian inheritance

Non-Mendelian inheritance is a typical feature of triplet repeat diseases (Wells, 1996; Pearson and Sinden, 1998b; Rubinsztein and Hayden, 1998). Thus, while typical monogenic diseases display similar phenotypes in different family members with the same mutation, many triplet repeat diseases show more severe phenotypes and/or earlier age of onset in successive generations. The molecular basis of this phenomenon, which is known as anticipation, has to do with some characteristic peculiarities of the repeat disorders (Rubinsztein and Amos, 1998; Rubinsztein and Hayden, 1998). First, increased trinucleotide repeat length is generally associated with increased severity of disease phenotype, and normal, non-disease, chromosomes always have shorter repeats than chromosomes in affected individuals. Second, repeats on non-disease chromosomes are relatively stable when transmitted from one generation to the next, and are believed to only rarely expand in steps of one or a few repeats. However, with increasing size of the repeat the chance of further expansion rises dramatically, and disease alleles typically show large changes in size on transmission from parent to offspring. Expansion seems to occur during meiosis or later in gametogenesis. In some cases a threshold length has been found above which the instability suddenly rises. This length is not necessarily the same as the length required for manifestation of the disorder.

Another non-Mendelian feature of some triplet repeat disorders is that paternal and maternal transmission may be associated with different probabilities of expansion (Ta-

ble 1). Specifically, all the polyglutamine CAG-repeat diseases mentioned above show a (more or less pronounced) paternal bias for expansion, a phenomenon that is probably connected to specific events during spermatogenesis (Lee, 1998; Orr and Zoghbi, 1998; Paulson, 1998; Pulst, 1998; Ross and Hayden, 1998; Stevanin *et al.*, 1998; Tsuji, 1998). The CGG repeat at the fragile site FRAXA and the CTG repeat associated with myotonic dystrophy show a maternal bias for expansion (Eichler and Nelson, 1998; Hamshere *et al.*, 1998), while there is no apparent bias related to the expansion of the GAA repeat associated with Friedreich's ataxia (Koenig, 1998).

Trinucleotides involved in repeat disorders

A triplet repeat can be described in terms of different unit trinucleotides depending on what strand and triplet frame that is chosen. Thus, the repeat CGGCGGCGGCGG ... can be said to be a repeat of the triplet CGG, and also of its reverse complement CCG. Ignoring repeat boundaries, however, the sequence can also be described as a repeat of the shifted triplet pairs GGC/GCC and GCG/CGC. In this way, the 64 different trinucleotides can be divided into 12 possible repeat classes (see, for example, Table 2 for a listing of the classes; note that, strictly speaking, repeats of the triplet pairs AAA/TTT and CCC/GGG are more precisely described as mononucleotide repeats).

Currently, we know of triplet repeat disorders with trinucleotides from three of the 12 (or 10) classes: CAG, CGG and GAA. Huntington's disease, DRPLA, Kennedy's disease and spinocerebellar ataxias 1, 2, 3, 6 and 7 are all associated with expanded CAG triplets in coding regions (Spada *et al.*, 1991; Huntington's Disease Collaborative Research Group, 1993; Orr *et al.*, 1993; Koide *et al.*, 1994; Nagafuchi *et al.*, 1994; Brooks and Fischbeck, 1995; Ikeuchi *et al.*, 1995; Campuzano *et al.*, 1996; Junck and Fink, 1996; Ross, 1997; Lee, 1998; Orr and Zoghbi, 1998; Paulson, 1998; Pulst, 1998; Ross and Hayden, 1998; Stevanin *et al.*, 1998; Tsuji, 1998). As mentioned above, the triplets in all these disorders are translated, in the CAG frame, into polyglutamine tracts. Autosomal dominant pure spastic paraplegia (ADPSP) is a neurodegenerative disorder characterized by both inter- and intrafamilial variation and anticipation, and has also been reported to be linked with an expanded CAG repeat that appears to be translated into polyglutamine (Nielsen *et al.*, 1997; Benson *et al.*, 1998). Myotonic dystrophy is again associated with a CAG-type repeat, but in this case it is the reverse complement (CTG) that is expressed (in the 3' UTR of the mRNA) (Timchenko *et al.*, 1995; Gusella and MacDonald, 1994; Hamshere *et al.*, 1998). Fragile sites FRAXA, FRAXE and FRAXF are all associated with CGG-type repeats (Nelson, 1995; Gusella and MacDonald, 1994; Eichler and Nelson, 1998; Skinner *et al.*, 1998). This is also the case for the group of so-

called polyalanine diseases (synpolydactyly, cleidocranial dysplasia and oculopharangeal muscular dystrophy). In these disorders the repeat expansion is present within coding DNA and is translated, in the GCG frame, into polyalanine tracts (Goodman *et al.*, 1997; Mundlos *et al.*, 1997; Brais *et al.*, 1998). These diseases are apparently not associated with anticipation and dynamic mutations and will therefore not be treated in this work. Finally, the origin of the most common hereditary ataxia – Friedreich's ataxia – has been linked to very long GAA triplet repeats (Campuzano *et al.*, 1996; Koenig, 1998).

In addition to trinucleotides, repeats of other sequence elements with lengths ranging from two to at least 60 base pairs are also known to be unstable, and hence highly polymorphic in the human population (Jeffreys, 1997). (It is this fact which makes these so-called micro- and mini-satellites useful for forensic DNA typing.) The repeats may or may not be associated with disease. For instance, it has been found that a repeated 12-mer upstream of the *EPML* gene displays intergenerational instability and is associated with myoclonus epilepsy (Laloti *et al.*, 1997, 1998), while a similarly unstable, AT-rich, 42 bp repeat is involved in the fragile site FRA10B (Hewett *et al.*, 1998). However, it does seem that disease-causing trinucleotide repeats are special, at least among microsatellites. Thus, based on the distribution of microsatellite allele sizes in a set of human sequences, it has been estimated that dinucleotides have a higher mutation (i.e. expansion) rate than non-disease-causing trinucleotides, which on their part have a higher rate than tetranucleotides (Chakraborty *et al.*, 1997). The trinucleotide repeats associated with disease were, however, estimated to have a rate higher than all the other three classes.

Mechanism of expansion

It is currently not clear exactly how the expansion of triplet repeats occur, why expansion frequency depends on repeat length, and why some diseases display a sex bias. It is, however, generally assumed that unusual structural features of the repeats play a role, and several models for expansion have been proposed, involving alternative DNA structures in erroneous DNA replication, recombination, or DNA repair (Wells, 1996; Pearson and Sinden, 1998a,b; Moore *et al.*, 1999). There is mounting evidence that the formation of hairpins in Okazaki fragments (during replication of the lagging strand) is likely to be involved in the expansion process (Chen *et al.*, 1995; Gacy *et al.*, 1995; Wells, 1996; Mariappan *et al.*, 1998; Miret *et al.*, 1998; Pearson and Sinden, 1998b), but many features remain unclear and open to interpretation. We here present a novel approach to this problem, namely a computational analysis of all the 12 possible repeat classes using several different (and independent) quantitative models of sequence-dependent DNA structure.

DNA structural models and results

There is mounting evidence that DNA structural properties beyond the double helical pattern play an important functional and regulatory role. This is not too surprising if one realizes that meters of DNA must be compacted into a nucleus that is only a few microns in diameter. In the nucleus DNA is packed into chromatin fibers. The fundamental repetitive unit of chromatin fibers is the nucleosome core particle which consists of 146 bp of DNA wound around a histone protein octamer. The chromatin complex structure of DNA and the positioning of nucleosomes along the genome have been found to play an important (generally inhibitory) role in regulation of gene transcription (Baldi *et al.*, 1996; Pazin and Kadonaga, 1997; Tsukiyama and Wu, 1997; Werner and Burley, 1997; Pedersen *et al.*, 1998).

It has been shown that the exact DNA sequence influences the three-dimensional structure of DNA. Based on different experimental or theoretical approaches, several computational models have been constructed that relate the nucleotide sequence to DNA flexibility and curvature (Ornstein *et al.*, 1978; Satchwell *et al.*, 1986; Goodsell and Dickerson, 1994; Sinden, 1994; Brukner *et al.*, 1995; el Hassan and Calladine, 1996; Hunter, 1996; Baldi *et al.*, 1998). These models are typically in the form of dinucleotide or trinucleotide scales, and while they agree on some structural features they also display inconsistent interpretations of some sequence elements. While there is no final consensus regarding these models, it is likely that each one provides a slightly different and partially complementary view of DNA structure. In this work, we apply several different models to the analysis of the properties of long triplet repeats of the general form $(XYZ)_n = XYZXYZXYZ \dots$

Bendability

One bendability model has been derived using the cutting frequencies of the DNase I enzyme (Brukner *et al.*, 1995). DNase I is known to preferably bind and cut DNA that is bent, or bendable, towards the major groove (Lahm and Suck, 1991; Suck, 1994). Thus DNase I cutting frequencies on naked DNA can be interpreted as a quantitative measure of major groove compressibility or anisotropic bendability. Such data have been used to calculate bendability parameters $B[XYZ]$ for the 32 complementary trinucleotide pairs. Aside from an averaging window and from small boundary effects, the global bendability B of a long triplet repeats can be approximated by the expression

$$B[(XYZ)_n] \approx n[B[XYZ] + B[YZX] + B[ZXY]] \quad (1)$$

In Table 2, we have listed in the first column each one of the 32 possible triplets associated with the triplet on the

Table 2. DNase I derived bendability: large values correspond to flexibility ($\Sigma B[XYZ] = B[XYZ] + B[YZX] + B[ZXY]$)

Triplet pair	$B[XYZ]$	$\Sigma B[XYZ]$
AGC/GCT	0.017	0.268
GCA/TGC	0.076	0.268
CAG/CTG	0.175	0.268
ATC/GAT	-0.110	0.218
CAT/ATG	0.134	0.218
TCA/TGA	0.194	0.218
AGG/CCT	-0.057	-0.013
GGA/TCC	0.013	-0.013
GAG/CTC	0.031	-0.013
AAT/ATT	-0.280	-0.030
TAA/TTA	0.068	-0.030
ATA/TAT	0.182	-0.030
CCC/GGG	-0.012	-0.036
ACG/CGT	-0.033	-0.049
GAC/GTC	-0.013	-0.049
CGA/TCG	-0.003	-0.049
ACT/AGT	-0.183	-0.068
TAC/GTA	0.025	-0.068
CTA/TAG	0.090	-0.068
AAG/CTT	-0.081	-0.091
GAA/TTC	-0.037	-0.091
AGA/TCT	0.027	-0.091
CCG/CGG	-0.136	-0.106
CGC/GCG	-0.077	-0.106
GCC/GGC	0.107	-0.106
AAC/GTT	-0.205	-0.196
ACA/TGT	-0.006	-0.196
CAA/TTG	0.015	-0.196
CCA/TGG	-0.246	-0.238
ACC/GGT	-0.032	-0.238
CAC/GTG	0.040	-0.238
AAA/TTT	-0.274	-0.822

reverse complementary strand ($XYZ/\bar{Z}\bar{Y}\bar{X}$). The triplet pairs are divided into the 12 different possible repeat classes mentioned above. The second column contains the bendability value according to the bendability scale of Brukner *et al.* (1995), and in the third column the global bendability $\Sigma B[XYZ] = B[XYZ] + B[YZX] + B[ZXY]$ associated with a given repeat class is listed. All the entries are ranked in order of decreasing total bendability. Obviously, all the triplets in the same class have the same total bendability.

There are a number of observations that can be made from this table.

First, $(AAA)_n = (A)_{3n}$ is the stiffest of all possible

repeats and by a large margin. Such regions of DNA are unlikely to bend easily and probably are bad candidates for nucleosome positioning when n is large. In fact, a number of promoters in yeast contain homopolymeric dA:dT elements. Such homopolymeric tracts are known from X-ray crystallography to be straight and rigid (Nelson *et al.*, 1987). Studies in two different yeast species have shown that the homopolymeric elements destabilize nucleosomes and thereby facilitate the access of transcription factors bound nearby (Iyer and Struhl, 1995; Zhu and Thiele, 1996). Interestingly, the sequence of the IT15 gene involved in HD has a repeat containing 18 adenine nucleotides at its 3' end.

Second, only two groups of triplets have global positive bendability, and they are well separated from the rest.

Third, all groups of triplets but one, contain at least one shifted triplet with low (negative) bendability. Such a triplet occurs every three positions. Because the DNA double helical pitch is about 10.5 bp (or 10.3 in the nucleosome), long-range curvature can be obtained only if highly bendable triplets are positioned in phase with this pitch. These two facts together pose severe constraints on the phase of one of these triplet repeats with respect to the double helix in any curved region of significant length.

Fourth, and perhaps most importantly, there is a single group of triplets represented by $(CAG)_n$ which is characterized by the highest global bendability and the fact that all shifted triplets have high (positive) bendability. The $(CAG)_n$ repeat is likely to lead to highly flexible stretches of DNA which are flexible in all positions.

Fifth, cases are described in the literature (Orr *et al.*, 1993) where the G of a few isolated CAG triplets within a long CAG repeat regions are replaced by a T. It should be noticed that the CAT triplet belongs to the second highest bendability class in the table above, and that the flexibility properties of such stretches therefore probably are preserved.

Sixth, the GAA repeat class is predicted to be rather rigid according to this scale.

Finally, the triplet repeat class containing CGG or GCC is found to be relatively rigid according to this scale.

From this analysis, we expect that very long stretches of $(CAG)_n$ are likely to have special structural properties and correspond to highly flexible regions. Although to the best of our knowledge the structural scales have not been used before in this context, the flexibility of extended CAG repeats has been experimentally verified (Chastain and Sinden, 1998). In accordance with their high flexibility, CAG/CTG repeats have been found to have the highest affinity for histones among all possible triplet repeats (Wang and Griffith, 1994, 1995; Godde and Wolffe, 1996). CGG/CCG repeats, on the other hand, seem completely unable to form nucleosomes (Godde *et al.*, 1996; Wang *et al.*, 1996).

Position preference

From experimental investigations of the positioning of DNA in nucleosomes it has been found that certain trinucleotides have strong preference for being positioned in phase with the helical repeat. Depending on the exact rotational position, such triplets will have minor grooves facing either towards or away from the nucleosome core (Satchwell *et al.*, 1986). Based on the premise that flexible sequences can occupy any rotational position on nucleosomal DNA, these preference values can be used as measures of DNA flexibility. Hence, in this model, all triplets with close to zero preference are assumed to be flexible, while triplets with preference for facing either in or out are taken to be more rigid. Note that we do *not* use this scale as a measure of how well different repeats form nucleosomal DNA. Instead, the absolute value, or unsigned nucleosome positioning preference, is used here as a measure of DNA flexibility.

In Table 3, we have listed in the first column all the triplet pairs divided into repeat classes. Column 2 contains the unsigned position preference value modified from the original signed scale $N[XYZ]$ (Satchwell *et al.*, 1986). The third column contains the unsigned nucleosome positioning value $\Sigma|N[XYZ]| = |N[XYZ]| + |N[YZX]| + |N[ZXY]|$ associated with a triplet and its shifts. Entries are ranked in order of decreasing total nucleosome positioning value. The main observations here are that:

- The CAG triplet class is found to have relatively low total unsigned nucleosome positioning value – another sign of flexibility partially confirming the previous result obtained with the bendability scale.
- Most importantly, the GCC and CGG triplets, on the other hand, are found at the top with very high global positioning preference value corresponding to a rigid repeat. It is exceeded only by the stiffest of all triplets, AAA/TTT.
- The GAA repeat class is predicted to be very flexible (ranks as the second lowest value considering that the last two classes have the same value).

These results are essentially unchanged if one uses the signed version of the scale.

Propeller twist angle

The dinucleotide propeller twist angle scale of el Hassan and Calladine (1996; Table 4) is based on X-ray crystallography of DNA oligomers. Dinucleotides with a large propeller twist angle tend to be more rigid than dinucleotides with a low propeller twist angle.

- Except for the exceptional homopolymeric C or G tract, the CCG/CGG expansion has the highest

Table 3. Position preference: small values correspond to flexibility ($\Sigma|N[XYZ]| = |N[XYZ]| + |N[YZX]| + |N[ZXY]|$)

Triplet pair	$ N[XYZ] $	$\Sigma N[XYZ] $
AAA/TTT	36	108
CCG/CGG	2	72
CGC/GCG	25	72
GCC/GGC	45	72
AAT/ATT	30	63
ATA/TAT	13	63
TAA/TTA	20	63
ACG/CGT	8	47
CGA/TCG	31	47
GAC/GTC	8	47
AGC/GCT	25	40
CAG/CTG	2	40
GCA/TGC	13	40
CCC/GGG	13	39
ACT/AGT	11	35
CTA/TAG	18	35
TAC/GTA	6	35
ACC/GGT	8	33
CAC/GTG	17	33
CCA/TGG	8	33
ATC/GAT	7	33
CAT/ATG	18	33
TCA/TGA	8	33
AAG/CTT	6	27
AGA/TCT	9	27
GAA/TTT	12	27
AGG/CCT	8	21
GAG/CTC	8	21
GGA/TCC	5	21
AAC/GTT	6	21
ACA/TGT	6	21
CAA/TTG	9	21

cumulative propeller twist angle, corresponding to the most flexible of all the repeat classes. This is in contradiction to the two previously used structural tables.

- The propeller twist angle of the CAG/CTG triplet repeat has average rank.
- Except for the A or T homopolymeric tract, the GAA expansion has the lowest cumulative propeller twist angle, corresponding to the most rigid triplet repeat. This is also in contradiction to the position preference scale, but in agreement with the DNase I scale.

Table 4. Propeller twist: small negative numbers correspond to flexibility ($\Sigma PT[XYZ] = PT[XY] + PT[YZ] + PT[ZX]$)

Triplet pair	$\Sigma PT[XYZ]$
CCC/GGG	−24.33
CGC/GCG	−29.22
GCC/GGC	−29.22
CCG/CGG	−29.22
CCA/TGG	−30.66
CAC/GTG	−30.66
ACC/GGT	−30.66
GCA/TGC	−34.53
CAG/CTG	−34.53
AGC/GCT	−34.53
GGA/TCC	−35.59
GAG/CTC	−35.59
AGG/CCT	−35.59
CGA/TCG	−36.61
GAC/GTC	−36.61
ACG/CGT	−36.61
TCA/TGA	−37.94
CAT/ATG	−37.94
ATC/GAT	−37.94
CTA/TAG	−38.95
TAC/GTA	−38.95
ACT/AGT	−38.95
ACA/TGT	−41.21
CAA/TTG	−41.21
AAC/GTT	−41.21
ATA/TAT	−45.52
TAA/TTA	−45.52
AAT/ATT	−45.52
AGA/TCT	−46.14
GAA/TTC	−46.14
AAG/CTT	−46.14
AAA/TTT	−55.98

Table 5. Protein-induced DNA deformability: large values correspond to flexibility ($\Sigma PD[XYZ] = PD[XY] + PD[YZ] + PD[ZX]$)

Triplet pair	$\Sigma PD[XYZ]$
CGC/GCG	22.2
GCC/GGC	22.2
CCG/CGG	22.2
CGA/TCG	18.9
GAC/GTC	18.9
ACG/CGT	18.9
CCC/GGG	18.3
CCA/TGG	18.2
CAC/GTG	18.2
ACC/GGT	18.2
GCA/TGC	15.9
CAG/CTG	15.9
AGC/GCT	15.9
CAT/ATG	15.9
ATC/GAT	15.9
TCA/TGA	15.9
CAA/TTG	15.0
AAC/GTT	15.0
ACA/TGT	15.0
GGA/TCC	12.7
GAG/CTC	12.7
AGG/CCT	12.7
TAA/TTA	10.8
AAT/ATT	10.8
ATA/TAT	10.8
CTA/TAG	10.7
TAC/GTA	10.7
ACT/AGT	10.7
AGA/TCT	9.5
GAA/TTC	9.5
AAG/CTT	9.5
AAA/TTT	8.7

Protein induced deformability

The protein induced deformability scale of Olson *et al.* (1998) is a dinucleotide scale derived from empirical energy functions extracted from the fluctuations and correlations of structural parameters in DNA–protein crystal complexes. In our work, we have found that the two scales are highly, but not perfectly, correlated. (This is perhaps interesting in a different context: the fact that the scale based on crystallography of *naked* DNA is correlated to the scale based on crystallography of DNA *in complex with protein*, immediately suggests that DNA structures seen in protein–DNA complexes may to some degree be determined at the DNA-sequence level. Or at least that the structure of DNA in the complex has to be consistent with

the inherent structural features of the DNA itself.)

As expected, the results (Table 5) are similar to those obtained with the propeller twist scale. In particular,

- The CCG/CGG expansion has the highest protein deformability corresponding a highly flexible triplet repeat.
- The protein induced deformability of the CAG/CTG triplet repeat has average rank.
- Except for the A or T homopolymeric tract, the GAA expansion has the lowest cumulative protein induced deformability, again corresponding to the most rigid repeat.

Table 6. Summary of repeat class structural features according to different models

Triplet repeat class	DNase I	Position preference	Propeller twist/Deformability
CAG	Most flexible	Very flexible	Very flexible
CGG	Very rigid	Very rigid (second most)	Very flexible (second most)
GAA	Very rigid	Very flexible (second most)	Very rigid (second most)

Discussion

The existence of disorders associated with the expansion of only a very small number of triplets raises some basic questions at the DNA level. First, why do particular triplets seem to be singled out? Second, what is the mechanism for their expansion and how does it depend on the triplet itself? Third, could other triplets or longer-base repeats be involved in other diseases?

In our analysis, we have seen that each special triplet seems to occupy one of the extreme ranges of one or several structural models. Conversely, for all the structural models considered at least one of the extreme regions of its spectrum is occupied by one of the special triplet repeat classes. A simple probabilistic calculation assuming random uniform ranking of the 12 triplet classes shows that this is unlikely to happen as a result of chance alone. And the other triplet classes do not seem to share these properties. For instance, the ACT class has the second highest bendability, but is hardly noticeable according to any other models. Furthermore, the triplet considered appear to be extremal even when other non-directly-structural scales are considered. This is the case with the dinucleotide base stacking energy scale (Ornstein *et al.*, 1978, Appendix, Table A1), which is somewhat related to the propeller twist angle scale, and where the GCC class is again extremal.

It is essential to notice that these extremal properties pertain to the triplet repeat *class*, rather than the triplet alone. A triplet that is not extremal for a given scale, may become extremal once its two shifted versions are considered. For instance, AGC has relatively low bendability when taken alone, but corresponds to the most bendable class when GCA and CAG are taken into account.

The CAG class of repeats was consistently found to be highly flexible using the majority of the four models (Table 6). This is also in accordance with experimental results (Chastain and Sinden, 1998). It is interesting that CAG repeats have been found to be the most efficient at nucleosome formation of all the possible triplet repeats. Taken together with the observation that

in several triplet repeat disorders the instability threshold is close to 50 repeats (Table 1; corresponding to 150 bp which is about the same as the length of DNA in a nucleosome core particle) it is tempting to suggest that nucleosome formation (or alternatively, DNA–histone interaction during replication or recombination) might be at the heart of the unusual expansion properties of these repeats. The observation that the threshold for instability of the *EPHI* dodecamer repeat is about 13 repeats (again corresponding to around 150 bp) is consistent with this idea. Alternatively, the fact that 150 bp is similar to the length of a eukaryotic Okazaki fragment, might be important for the threshold effect (Richards and Sutherland, 1994). It is also interesting that most CAG-type repeats show paternal bias for expansion. This suggests that the expansion mechanism is related to the process of spermatogenesis. It might be relevant that during this process, normal histones are replaced by other DNA-binding proteins (Grimes, 1986).

The CGG repeat class is predicted to be very rigid according to the DNase I and positioning preference scales (ranks as second most according to the latter). However, by the propeller twist and deformability scales it is predicted to be very flexible (Table 6). It is perhaps relevant to look at the single triplet pairs in this case. According to the position preference scale, the triplet pair CGG/CCG has the most flexible value (2%) – a position it shares with the CTG/CAG pair. However, the two shifted triplet pairs (CGC/GCG and GCC/GGC) have very rigid values (25% and 45%, respectively – the latter is the most rigid value in the table) and this causes the class to be in the rigid end of the scale.

Similarly, the GAA repeat is hard to interpret unambiguously: according to the DNase I and propeller twist deformability scales it is very rigid (second most on the latter), while it is very flexible according to the positioning preference scale (Table 6).

Better structural models may be needed to shed light on these discrepancies. However, it is important to remember that the computational models used in this work are based on mutually different and also rather indirect investigations of DNA structure. It is therefore likely that any single model correctly captures only some structural features of some sequence elements. For instance, as mentioned previously, DNase I preferentially binds and cuts on sites where the DNA is bent *or* bendable towards the major groove. This means that a high DNase I value can be caused by either a very flexible piece of DNA (isotropically flexible, or anisotropically flexible in the right direction), or alternatively by a piece of DNA that is rigid but curved with a compressed major groove. Similarly, it can be imagined that a piece of DNA might be highly flexible, but that other features of its structure (helical twist for instance) nevertheless

makes it unsuitable for being positioned anywhere on nucleosomes. In this context it is perhaps relevant that GGC repeats have been found to be highly flexible based on cyclization experiments and electrophoretic studies (Bacolla *et al.*, 1997; Chastain and Sinden, 1998), but that they nevertheless assemble into nucleosomes with very, very low efficiency (Godde *et al.*, 1996; Wang *et al.*, 1996). Thus, we believe it is fair to conclude that although two classes of triplets (GAA and GGC) display ambiguous characteristics using different models, our results are still a strong indication that they possess peculiar structural features.

Finally, it is worth noticing that the CAG class of triplet repeats, which was consistently found to be flexible according to all the models used here, is in fact special among the triplet repeats. Thus, CAG-class repeats are responsible for the majority of the currently known diseases (10 of the 13 mentioned in Table 1). Furthermore, it has been found in a model study in *Escherichia coli* that the CAG triplet repeat was the predominant genetic expansion product, and was expanded at least nine times more frequently than any other triplet (Ohshima *et al.*, 1996).

The 12-base unit in the *EPMI* repeat mentioned above has the sequence CCC CGC CCC GCG (Lalioi *et al.*, 1997, 1998). Of the 12 overlapping triplets that occur in repeats of this dodecamer, four are CCC triplets while the remaining eight belong to the GCC class – the same class as fragile X. Thus, it is possible that this sequence also possess special structural features similar to those seen in the triplet repeat diseases. Pursuing this line of thought, we constructed the following 12-mer: CTG CAG CAG CAG. This sequence consists solely of triplets belonging to the CAG class. Specifically, each of the three different triplet pairs in the CAG/CTG class occurs four times in repeats of the 12-mer (again, counting all 12 overlapping triplets). Hence, based on the tables used in this work, we predict that the structural properties of the dodecamer repeat are identical to those of the CAG repeat. We then proceeded to perform a search of human nucleotide sequences using two repeats of the 12-mer as query sequence. One sequence found to contain a perfect match to the 24-nucleotide sequence, was the mRNA for ZNF231 – a putative transcription factor expressed in the cerebellum (Hashida *et al.*, 1998). Flanking these 24 nucleotides in the ZNF231 mRNA are two CAG triplets which might be argued to also be part of the same structural region. Combined, the 30 base pairs correspond to a stretch of 10 glutamines and leucines (amino acids 2456–2465). Intriguingly, enhanced expression of ZNF231 has been found to be correlated with the neurodegenerative disorder multiple system atrophy (MSA) (Hashida *et al.*, 1998; Klockgether *et al.*, 1998). We tentatively suggest that the pathogenesis of MSA may be similar to that of the

neurodegenerative disorders analyzed above, and that it may involve expansion of the dodecamer (possibly from one repeat to the two found in the neuronal mRNA mentioned above).

In addition to the triplet repeats analyzed in this paper, the eukaryotic genome contains many other repeat sequences (Jurka *et al.*, 1992; Jeffreys, 1997). Besides the group of tandemly repeated satellites (such as those found at telomeres), these also include a large diverse group of so-called interspersed elements. These mostly represent inactive copies of transposable elements and have, like the repeats described in this paper, been found to impact phenotypes – and even to sometimes play significant roles in genome evolution. An obvious step for future work is to check the DNA structural properties of repeats belonging to these other classes.

Acknowledgements

The work of P.B. and Y.C. has been in part supported by an NIH SBIR grant to Net-ID, Inc., and currently (P.B.) by a Laurel Wilkening Faculty Innovation award at UCI. The work of S.B. and A.G.P. is supported by a grant from the Danish National Research Foundation.

References

- Ashley, C.T. and Warren, S.T. (1995) Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.*, **29**, 703–728.
- Bacolla, A., Gellibolian, R., Shimizu, M., Amirhaeri, S., Kang, S., Ohshima, K., Larson, J.E., Harvey, S., Stollar, B.D. and Wellls, R.D. (1997) Flexible DNA: genetically unstable CTG-CAG and CGG-CCG from human hereditary neuromuscular disease genes. *J. Biol. Chem.*, **272**, 16783–16792.
- Baldi, P., Brunak, S., Chauvin, Y. and Krogh, A. (1996) Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.*, **263**, 503–510.
- Baldi, P., Chauvin, Y., Pedersen, A.G. and Brunak, S. (1998) Computational applications of DNA structural scales. In *Proceedings of the 1998 Conference on Intelligent Systems for Molecular Biology (ISMB98)* The AAI Press, Menlo Park, CA, pp. 35–42.
- Beitel, L.K., Trifiro, M.A. and Pinsky, L. (1998) Spinobulbar muscular atrophy. In Rubinstein, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 85–104.
- Benson, K.F., Horwitz, M., Wolff, J., Friend, K., Thompson, E., White, S., Richards, R.I., Raskind, W.H. and Bird, T.D. (1998) Cag repeat expansion in autosomal dominant familial spastic paraparesis: novel expansion in a subset of patients. *Hum. Mol. Genet.*, **7**, 1779–1786.
- Brais, B., Bouchard, J.-P., Xie, Y.-G., Rochefort, D.L., Chretien, N., Tome, F.M.S., Lafreniere, R.G., Rommens, J.M., Uyama, E., No-hira, O., Blumen, S., Korcyn, A.D., Heutink, P., Mathieu, J., Duranceau, A., Codere, F., Fardeau, M. and Rouleau, G. (1998) Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.*, **18**, 164–167.
- Brooks, B.P. and Fischbeck, K.H. (1995) Spinal and bulbar muscular atrophy: a trinucleotide-repeat expansion neurodegenerative disease. *Trends Neurosci.*, **18**, 459–461.

- Brukner, I., Sánchez, R., Suck, D. and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.*, **14**, 1812–1818.
- Campuzano, V., Montermini, L., Molto, M.D., Pianese, L., Cossee, M., Cavalcanti, F., Montos, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizares, J., Koutnikova, H., Bidichandani, S.I., Gellera, C., Brice, A., Trouillaas, P., Michele, G.D., Filla, A., Frutos, R.D., Palau, F., Patel, P.I., Donato, S.D., Mandel, J.L., Coccozza, S., Koenig, M. and Pandolfo, M. (1996) Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, **271**, 1423–1427.
- Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J. and Deka, R. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. USA*, **94**, 1041–1046.
- Chastain, P.D. and Sinden, R.R. (1998) CTG repeats associated with human genetic disease are inherently flexible. *J. Mol. Biol.*, **275**, 405–411.
- Chen, X., Mariappan, S.V.S., Catasti, P., Ratliff, R., Moyzis, R.K., Ali, L., Smith, S.S., Bradbury, E.M. and Gupta, G. (1995) Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. *Proc. Natl. Acad. Sci. USA*, **52**, 5199–5203.
- DiFiglia, M., Sapp, E., Chase, K.O., Davies, S.W., Bates, G.P., Vonsattel, J.P. and Aronin, N. (1997) Aggregation of Huntington's in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, **277**, 1990–1993.
- Eichler, E.E. and Nelson, D.L. (1998) The FRAXA fragile site and fragile X syndrome. In Rubinsztein, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 13–50.
- Gacy, A.M., Goellner, G., Juranic, N., Macura, S. and McMurray, C.T. (1995) Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell*, **8**, 533–540.
- Gecz, J. and Mulley, J.C. (1999) Characterisation and expression of a large, 13.7 kb *fmr2* isoform. *Eur. J. Hum. Genet.*, **7**, 157–162.
- Godde, J.S. and Wolffe, A.P. (1996) Nucleosome assembly on CTG triplet repeats. *J. Biol. Chem.*, **271**, 15222–15229.
- Godde, J.S., Kass, S.U., Hirst, M.C. and Wolffe, A.P. (1996) Nucleosome assembly on methylated CGG triplet repeats in the fragile X mental retardation gene 1 promoter. *J. Biol. Chem.*, **271**, 24325–24328.
- Goodman, F.R., Mundlos, S., Muragaki, Y., Donnai, D., Giovanulli-Uzielli, M.L., Lapi, E., Majewski, F., McGaughan, J., Mckown, C., Reardon, W., Upton, J., Winter, R.M., Olsen, B.R. and Scambler, P.J. (1997) Synpolydactyly phenotypes correlate with size of expansions in HOXD13 polyalanine tract. *Proc. Natl. Acad. Sci. USA*, **94**, 7458–7463.
- Goodsell, D.S. and Dickerson, R.E. (1994) Bending and curvature calculations in B-DNA. *Nucleic Acids Res.*, **22**, 5497–5503.
- Grimes, S.R., Jr (1986) Nuclear proteins in spermatogenesis. *Comp. Biochem. Physiol. [B]*, **83**, 495–500.
- Gusella, J.F. and MacDonald, M.E. (1986) Trinucleotide instability: a repeating theme in human inherited disorders. *Annu. Rev. Med.*, **47**, 201–209.
- Hamshire, M., Newman, E., Alwazzan, M. and Brook, J.D. (1998) Myotonic dystrophy. In Rubinsztein, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 61–84.
- Hardy, J. and Gwinn-Hardy, K. (1998) Genetic classification of primary neurodegenerative disease. *Science*, **282**, 1075–1079.
- Hashida, H., Goto, J., Zhao, N., Takahashi, N., Hirai, M., Kanazawa, I. and Sakaki, Y. (1998) Cloning and mapping of ZNF231, a novel brain-specific gene encoding neuronal double zinc finger protein whose expression is enhanced in a neurodegenerative disorder, multiple system atrophy (MSA). *Genomics*, **54**, 50–58.
- el Hassan, M.A. and Calladine, C.R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.*, **259**, 95–103.
- Hewett, D.R., Handt, O., Hobson, L., Mangelsdorf, M., Eyre, H.J., Baker, E., Sutherland, G.R., Schuffenhauer, S., Mao, J.I. and Richards, R.I. (1998) FRA10B structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Mol. Cell*, **1**, 773–781.
- Hunter, C.A. (1996) Sequence-dependent DNA structure. *Bioessays*, **18**, 157–162.
- Huntington'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.
- Ikeuchi, T., Koide, R., Tanaka, H., Onodera, O., Igarashi, S., Takahashi, H., Kondo, R., Ishikawa, A., Tomoda, A., Miike, T., Sato, K., Ihara, Y., Hayabara, T., Isa, F., Tanabe, H., Tokiguchi, S., Hayashi, M., Shimizu, N., Ikuta, F., Naito, H. and Tsuji, S. (1995) Dentatorubral-Pallidoluysian atrophy: clinical features are closely related to unstable expansion of trinucleotide Cag repeat. *Ann. Neurol.*, **37**, 769–775.
- Iyer, V. and Struhl, K. (1995) Poly (dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.*, **14**, 2570–2579.
- Jeffreys, A.J. (1997) Spontaneous and induced minisatellite instability in the human genome. *Clin. Sci.*, **93**, 383–390.
- Junck, L. and Fink, J.K. (1996) Machado-Joseph disease and SCA3: the genotype meets the phenotypes. *Neurology*, **46**, 4–8.
- Jurka, J., Walichewicz, J. and Milosavljevic, A. (1992) Prototypic sequences for human repetitive DNA. *J. Mol. Evol.*, **35**, 286–291.
- Karlin, S. and Burge, C. (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci. USA*, **93**, 1560–1565.
- Klockgether, T., Lüdtke, R., Kramer, B., Abele, M., Bürk, K., Scöls, L., Riess, O., Laccone, F., Boesch, S., Lopes-Cendes, I., Brice, A., Inzelberg, R., Zilber, N. and Dichgans, J. (1998) The natural history of degenerative ataxia: a retrospective study in 466 patients. *Brain*, **121**, 589–600.
- Koenig, M. (1998) Friedreich's ataxia. In Rubinsztein, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 219–238.
- Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F. and Tsuji, S. (1994) Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nat. Genet.*, **6**, 9–13.
- Lahm, A. and Suck, D. (1991) DNase I-induced DNA conformation: 2 Å structure of a DNase I-octamer complex. *J. Mol. Biol.*, **222**, 645–667.

- Lalioi, M.D., Buresi, H.S., Rossier, C., Bottani, A., Morris, M.A., Malafosse, A. and Antonarakis, S.E. (1997) Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, **386**, 847–851.
- Lalioi, M.D., Scott, H.S., Genton, P., Grid, D., Ouazzani, R., M'Rabet, A., Ibrahim, S., Gouider, R., Dravet, C., Chkili, T., Bottani, A., Buresi, C., Malafosse, A. and Antonarakis, S.E. (1998) A PCR amplification method reveals instability of the dodecamer repeat in progressive myoclonus epilepsy (EPM1) and no correlation between the size of the repeat and age at onset. *Am. J. Hum. Genet.*, **62**, 842–847.
- Lee, C.C. (1998) Spinocerebellar ataxia type 6 (SCA6). In Rubinsztajn, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 145–154.
- Mariappan, S.V.S., Silks, L.A., III, Bradbury, E.M. and Gupta, G. (1998) Fragile X DNA triplet repeats, $(GCC)_n$, form hairpins with single hydrogen-bonded cytosine. Cytosine mispairs at the CpG sites: isotope-edited nuclear magnetic resonance spectroscopy on $(GCC)_n$ with selective ^{15}N -labeled cytosine bases. *J. Mol. Biol.*, **283**, 111–120.
- Miret, J.J., Pessoa-Brandao, L. and Lahue, R.S. (1998) Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA*, **95**, 12438–12443.
- Miwa, S. (1994) Triplet repeats strike again. *Nat. Genet.*, **6**, 3–4.
- Moore, H., Greenwell, P.W., Liu, C.P., Arnheim, N. and Petes, T.D. (1999) Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc. Natl. Acad. Sci. USA*, **96**, 1504–1509.
- Mundlos, S., Otto, F., Mundlos, C., Mulliken, J.B., Aylsworth, A.S., Albright, S., Lindhout, D., Cole, W.G., Henn, W., Knoll, J.H.M., Owen, M.J., Mertelsmann, R., Zabel, B.U. and Olsen, B.R. (1997) Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. *Cell*, **89**, 773–779.
- Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M., Takeda, T., Tadokoro, K., Kondo, I., Murayama, N., Tanaka, Y., Kikushima, H., Umino, K., Kurosawa, H., Furukawa, T., Nihei, K., Inoue, T., Sano, A., Komure, O., Takahashi, M., Yoshizawa, T., Kanazawa, I. and Yamada, M. (1994) Dentatorubral and pallidolysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nat. Genet.*, **6**, 14–18.
- Nelson, D.L. (1995) The fragile X syndrome. *Semin. Cell Biol.*, **6**, 5–11.
- Nelson, H.C.M., Finch, J.T., Luisi, B.F. and Klug, A. (1987) The structure of an oligo(dA)-oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.
- Nielsen, J.E., Koefoed, P., Abell, K., Hasholt, L., Eiberg, H., Fenger, K., Niebuhr, E. and Sorensen, S.A. (1997) Cag repeat expansion in autosomal dominant pure spastic paraplegia linked to chromosome 2p21-p24. *Hum. Mol. Genet.*, **6**, 1811–1816.
- Ohshima, K., Kang, S. and Wells, R.D. (1996) CTG triplet repeats from human hereditary diseases are dominant genetic expansion products in *Escherichia coli*. *J. Biol. Chem.*, **271**, 1853–1856.
- Olson, W.K., Gorin, A.A., Lu, X., Hock, L.M. and Zhurkin, V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, **95**, 11163–11168.
- Ornstein, R.L., Rein, R., Breen, D.L. and MacElroy, R.D. (1978) An optimised potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers*, **17**, 2341–2360.
- Orr, H.T. and Zoghbi, H.Y. (1998) Polyglutamine tract vs. protein context in SCA1 pathogenesis. In Rubinsztajn, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 105–118.
- Orr, H.T., Chung, M., Banfi, S., Jr, Kwiatkowski, T.J., Servadio, A., Beaudet, A.L., McCall, A.E., Duveick, L.A., Ranum, L.P. W. and Zoghbi, H.Y. (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, **4**, 221–226.
- Paulson, H.L. (1998) Spinocerebellar ataxia type 3/Machado–Joseph disease. In Rubinsztajn, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 129–144.
- Paulson, H.L., Perez, M.K., Trottier, Y., Trojanowski, J.Q., Subramony, S.H., Das, S.S., Vig, P., Mandel, J.L., Fischbeck, K.H. and Pittman, R.N. (1997) Intranuclear inclusions of expanded polyglutamine protein in spinocerebellar ataxia type 3. *Neuron*, **19**, 333–344.
- Pazin, M.J. and Kadonaga, J.T. (1997) SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein–DNA interactions? *Cell*, **88**, 737–740.
- Pearson, C.E. and Sinden, R.R. (1998a) Slipped strand DNA, dynamic mutations and human disease. In Wells, R.D. and Warren, S.T. (eds). *Genetic Instabilities and Hereditary Neurological Diseases* Academic Press, New York, pp. 585–621.
- Pearson, C.E. and Sinden, R.R. (1998b) Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Curr. Opin. Struct. Biol.*, **8**, 321–330.
- Pedersen, A.G., Baldi, P., Brunak, S. and Chauvin, Y. (1998) DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.
- Perutz, M.F., Johnson, T., Suzuki, M. and Finch, J.T. (1994) Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci. USA*, **91**, 5355–5358.
- Pulst, S.-M. (1998) Spinocerebellar ataxia type 2. In Rubinsztajn, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 119–128.
- Richards, R.I. and Sutherland, G.R. (1994) Simple repeat DNA is not replicated simply. *Nat. Genet.*, **6**, 114–116.
- Ross, C.A. (1995) When more is less: pathogenesis of glutamine repeat neurodegenerative diseases. *Neuron*, **15**, 493–496.
- Ross, C.A. (1997) Intranuclear neuronal inclusions: a common pathogenic mechanism for glutamine-repeat neurodegenerative diseases? *Neuron*, **19**, 1147–1150.
- Ross, C.A. and Hayden, M.R. (1998) Huntington's disease. In Rubinsztajn, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 169–208.
- Rubinsztajn, D.C. and Amos, B. (1998) Trinucleotide repeat mutation processes. In Rubinsztajn, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 257–268.
- Rubinsztajn, D.C. and Hayden, M.R. (1998) Introduction. In Rubinsztajn, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 1–12.
- Satchwell, S.C., Drew, H.R. and Travers, A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**,

- 659–675.
- Scherzinger, E., Lurz, R., Turmaine, M., Magiarianin, L., Hollenbach, B., Hasenbank, R., Bates, G.P., Davies, S.W., Lehrach, H. and Wanker, E.E. (1997) Huntington-encoded polyglutamine expansions from amyloid-like protein aggregates in vitro and in vivo. *Cell*, **90**, 549–558.
- Sinden, R.R. (1994) *DNA Structure and Function*. Academic Press, San Diego, CA.
- Skinner, J.A., Foss, G.S., Miller, W.J. and Davies, K.E. (1998) Molecular studies of the fragile sites FRAXE and FRAXF. In Rubinsztein, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 51–60.
- Spada, A.R.L., Wilson, E.M., Lubahn, D.B., Harding, A.E. and Fischbeck, K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**, 77–79.
- Stevanin, G., Daviid, G., Abbas, N., Dürr, A., Holmberg, M., Duyckaerts, C., Giunti, P., Cancel, G., Ruberg, M., Mandel, J.-L. and Brice, A. (1998) Spinocerebellar ataxia type 7 (SCA7). In Rubinsztein, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 155–168.
- Suck, D. (1994) DNA recognition by Dnase I. *J. Mol. Recognition*, **7**, 65–70.
- Timchenko, L., Monckton, D.G. and Casey, C.T. (1995) Myotonic dystrophy: an unstable CTG repeat in a protein kinase gene. *Semin. Cell Biol.*, **6**, 13–19.
- Tsuji, S. (1998) Dentatorubral–pallidoluysian atrophy (DRLPA). In Rubinsztein, D.C. and Hayden, M.R. (eds). *Analysis of Triplet Repeat Disorders* BIOS, Oxford, UK, pp. 209–218.
- Tsukiyama, T. and Wu, C. (1997) Chromatin remodeling and transcription. *Curr. Opin. Genet. Dev.*, **7**, 182–191.
- Wang, Y.-H. and Griffith, J.D. (1994) Preferential nucleosome assembly at DNA triplet repeats from the myotonic dystrophy gene. *Science*, **265**, 669–671.
- Wang, Y.-H. and Griffith, J.D. (1995) Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. *Genomics*, **25**, 570–573.
- Wang, Y.-H., Gellibolian, R., Shimizu, M., Wells, R.D. and Griffith, J.D. (1996) Long CCG triplet repeat blocks exclude nucleosome – a possible mechanism for the nature of fragile sites in chromosomes. *J. Mol. Biol.*, **263**, 511–516.
- Wells, R.D. (1996) Molecular basis of triplet repeat diseases. *J. Biol. Chem.*, **271**, 2875–2878.
- Werner, M.H. and Burley, S.K. (1997) Architectural transcription factors: proteins that remodel DNA. *Cell*, **88**, 733–736.
- Zhu, Z. and Thiele, D.J. (1996) A specialized nucleosome modulates transcription factor access to a *C.glabrata* metal responsive promoter. *Cell*, **87**, 459–470.

Appendix: base stacking energy

Table A1. Base stacking energy in kilocalories per mole ($\Sigma BS[XYZ] = BS[XY] + BS[YZ] + BS[ZX]$)

Triplet pair	$\Sigma BS[XYZ]$
TAA/TTA	–15.76
AAT/ATT	–15.76
ATA/TAT	–15.76
AAA/TTT	–16.11
CTA/TAG	–21.11
TAC/GTA	–21.11
ACT/AGT	–21.11
GAA/TTC	–21.96
AAG/CTT	–21.96
AGA/TCT	–21.96
ACA/TGT	–22.45
CAA/TTG	–22.45
AAC/GTT	–22.45
TCA/TGA	–22.95
CAT/ATG	–22.95
ATC/GAT	–22.95
CCC/GGG	–24.78
GGA/TCC	–24.85
GAG/CTC	–24.85
AGG/CCT	–24.85
CCA/TGG	–25.34
CAC/GTG	–25.34
ACC/GGT	–25.34
GCA/TGC	–27.94
CAG/CTG	–27.94
AGC/GCT	–27.94
CGA/TCG	–30.01
GAC/GTC	–30.01
ACG/CGT	–30.01
GCC/GGC	–32.54
CCG/CGG	–32.54
CGC/GCG	–32.54