Biochemistry

© Copyright 1997 by the American Chemical Society Volume 36, Number 11 March 18, 1997

New Concepts in Biochemistry

Slipped Structures in DNA Triplet Repeat Sequences: Entropic Contributions to Genetic Instabilities

Stephen C. Harvey*

Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294

Received November 8, 1996; Revised Manuscript Received January 14, 1997[⊗]

ABSTRACT: Slipped DNA structures can occur in sequences with direct repeats. DNA triplet repeats, particularly $(CTG)_n$, $(CGC)_n$, and $(GAA)_n$, are known to be associated with several neurological diseases. Slippage is probably the cause of expansion of the number of repeats, a process called dynamic mutation, which is known to be the cause of the diseased state. Here it is shown that the conformational entropy associated with slippage is more destabilizing for long direct repeats (300−1000 base pairs) than shorter runs (10−30 base pairs), by about 2 kcal/mol. This contributes to the greater instability of longer sequences. Entropic considerations also favor the formation of simple bulges, rather than hairpin structures. A model is presented for dynamic mutations, and experimentally testable predictions are made that will allow the model to be tested.

Expansion of DNA triplet direct repeats, particularly $(CTG)_n$, $(CGC)_n$, and $(GAA)_n$, is associated with a variety of neurological diseases (Bates & Lehrach, 1994; Panzer et al., 1995; Sutherland & Richards, 1995; Campuzano et al., 1996). In normal individuals the number of repeats is preserved below some threshold value. Disease results from mutations that take *n* past that threshold, at which point the number of repeats becomes unstable and may rise with successive cell divisions. The probability of expansion increases with the number of repeats. These are called dynamic mutations. Slipped DNA structures are probably involved in these phenomena (Richards & Sutherland, 1992), resulting in the formation of simple bulges or hairpins (Wells, 1996).

Slippage during replication can cause the daughter strand to be longer than the parent strand (expansion), or shorter (contraction), depending on which strand bulges out, as shown schematically in Figure 1. For direct repeats containing one to four nucleotides, slippage of one repeat unit would produce a simple bulge with little or no internal secondary structure. Slippage of several repeat units might produce multiple bulges, multiple small hairpins, or a single large hairpin. Both bulges and hairpins could be associated with dynamic mutations, particularly in triplet repeat diseases (Wells, 1996), although direct evidence for the involvement of either of these structures in genetic expansion is still lacking.

During replication of a direct repeat sequence, slippage of the Watson−Crick duplex (Figure 2a) can create a bulge (Figure 2b). Such a structure is thermodynamically unfavorable, relative to the duplex. It is intuitively obvious that long direct repeats will be more prone to slippage than short ones. The purpose of this paper is to estimate the entropic advantage of slipping, as a function of *n*, providing a quantitative thermodynamic basis for examining dynamic mutations.

* Tel: 205-934-5028. FAX: 205-975-2547. E-mail: harvey@neptune.cmc.uab.edu.

[⊗] Abstract published in *Advance ACS Abstracts,* March 1, 1997.

S0006-2960(96)02771-7 CCC: $14.00 © 1997 American Chemical Society

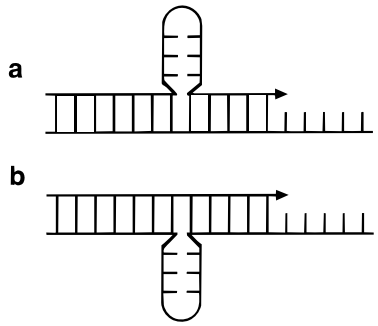

FIGURE 1: Slipped structures during replication will cause the daughter strand to have a different length than the strand being copied. (a) If slippage produces a bulge or hairpin in the daughter strand, the size of the genome will be expanded. (b) If slippage causes a bulge or hairpin in the parent strand, a deletion occurs. The slipped structure is represented here as a hairpin, but the results are the same for bulges.

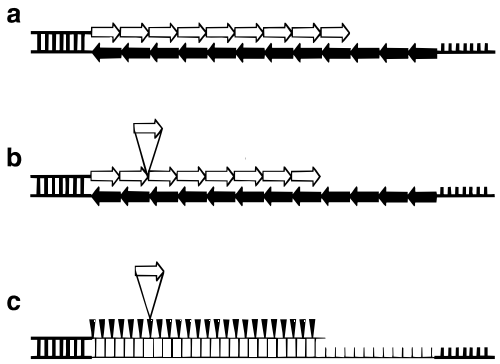

FIGURE 2: Slippage during replication. A region containing 12 direct triplet repeats (36 base pairs) is illustrated. Each black arrow represents a triplet sequence, e.g., CTG. Each white arrow represents the complementary triplet, e.g., CAG. (a) A duplex of 27 base pairs is formed when 9 repeats have been replicated without slippage. (b) One possible slipped structure. A CAG bulge is illustrated, and it is formed within a 24 base pair duplex. (c) Slippage could produce a bulge at any of 24 sites (black triangles); 8 of these are the isomers with CAG bulges, 8 have AGC bulges, and 8 have GCA bulges. The single- and double-stranded regions within the direct repeat are indicated by light lines, while the heavy lines represent sequences outside the direct repeat.

Let us denote the internal energy of the duplex state (Figure 2a) by E_0 and that of the slipped structure (Figure 2b) by E_1. If these two structures were the only possibilities, the probabilities of the two states would be

$$P_{\text{duplex}} = e^{-(E_0/RT)}/Z; \qquad P_{\text{slip}} = e^{-(E_1/RT)}/Z \tag{1}$$

where R is the universal gas constant and T the absolute temperature. Z is the partition function (Tinoco et al., 1985), determined by the normalization requirement that $P_{\text{duplex}} + P_{\text{slip}} = 1$, i.e.

$$Z = e^{-(E_0/RT)} + e^{-(E_1/RT)} \tag{2}$$

The energy cost of slippage is probably in the range 1−10 kcal/mol. The conclusions drawn in this study are independent of the exact value of the energy cost. It does, however, clarify the argument to examine a concrete example, so let us consider an intermediate value, using $E_1 - E_0 = 3$ kcal/mol. In this case, the probability of slippage is $P_{\text{slip}} = 0.007$.

The slipped state in Figure 2b is only one of a number of slipped states, all of which are isoenergetic, or nearly so. If the white arrows in Figure 2 represent a triplet repeat (such as CAG), if there are nine such triplets (Figure 2a), and if a three-nucleotide bulge is formed within a 24 base pair duplex (Figure 2b), then there are 24 possible slipped structures (Figure 2c). Correcting eq 2 to account for all 24 of these states gives

$$Z = e^{-(E_0/RT)} + \sum e^{-(Ei/RT)} \tag{3}$$

where the sum ranges from $i = 1$ to $i = 24$. Without loss of generality, we can set $E_0 = 0$. For illustrative purposes, let us assume that the energetic cost of slippage is 3 kcal/mol for all values of i. (The energetic differences between bulges with different sequences and the dependence of slip energy on the location of the bulge are second-order effects.) Since $RT = 0.6$ kcal/mol at 300 K, the partition function is

$$Z = 1 + 24\, e^{-5} = 1.16 \tag{4}$$

The probabilities of eq 1 are then

$$P_{\text{duplex}} = 1/Z = 0.86; \qquad P_{\text{slip}} = 24\, e^{-5}/Z = 0.14$$

An equivalent view is that this is a two-state system, with the slipped state having a 24-fold degeneracy. In this view, the partition function in eq 3 can be rewritten in terms of free energies

$$Z = e^{-(G_0/RT)} + e^{-(G_1/RT)} \tag{5}$$

The enthalpy H_i is essentially equal to the internal energy E_i (Tinoco et al., 1985). Equations 4 and 5 are then identical if we make the standard thermodynamic assignment, identifying the entropy of each state, S_i, with $R[\ln(m_i)]$, where m_i is the degeneracy of state i. The unslipped state has $m_0 = 1$ ($S = 0$). The 24-fold degeneracy of the slipped state gives $m_1 = 24$. The free energy change of slippage thus has components $\Delta H = 3.0$ kcal/mol and $T\Delta S = RT[\ln(24)] = 1.9$ kcal/mol, giving $\Delta G = G_1 - G_0 = \Delta H - T\Delta S = 1.1$ kcal/mol. The relative probability of slippage is $P_{\text{slip}}/P_{\text{duplex}} = e^{-\Delta G/RT} = 0.16$. The absolute probability is $P_{\text{slip}}/(P_{\text{duplex}} + P_{\text{slip}}) = e^{-\Delta G/RT}/Z = 0.14$, as required.

In short, slippage for this particular case is accompanied by an enthalpic cost of 3.0 kcal/mol and by an entropic advantage of 1.1 kcal/mol. The latter arises from the multiple ways in which slips can occur. It is important to note that the entropic advantage is independent of the value of ΔH.

These results can be generalized to the case of N base pairs of direct repeat, for which $T\Delta S = RT \ln(N)$. The enthalpic cost of slippage will be essentially independent of N, and it is offset by an entropic advantage that increases with N (Table 1). The most important result of this letter is that the free energy cost of the slipped state is 1−2 kcal/mol less for longer repeats than for shorter repeats. The greater instability of longer repeats has been directly demonstrated by altered patterns of gel electrophoretic

Table 1: Entropic Advantage of Slippage

length of direct repeat N (base pairs)	$T\Delta S = RT \ln(N)$ (kcal/mol)
10	1.4
30	2.0
100	2.8
300	3.4
1000	4.1

mobilities and by changes in sensitivity to digestion by nucleases (Pearson & Sinden, 1996).

This suggests a simple thermodynamic basis to the process of dynamic mutation. When there are few direct repeats, slips are rare, and the cellular repair systems are able to correct them. Direct repeat tracts of intermediate length are prone to more frequent slips that occasionally escape repair, leading to gradual expansion. Long tracts of direct repeat sequences have a high frequency of slippage, the repair systems are overwhelmed, and expansion is rapid. This is consistent with the observation that mutations in mismatch repair systems have been found to destabilize direct repeat sequences (Levinson & Guttman, 1987; Parsons et al., 1993; Strand et al., 1993).

Objections may be raised that the calculations ignore the complexities of DNA replication and repair *in vivo*, because they do not consider the effects of protein−DNA interactions and because they do not reflect variations in behavior due to differences in DNA sequence. Neither does this model address the question of different behaviors of trinucleotide repeats *vs* dinucleotide, tetranucleotide, and other repeat lengths. But whatever contributions these factors may make to ΔH, the entropic effects of multiple slipped conformers will still contribute to the total free energy changes. The length-dependent entropic contributions to the free energy changes (Table 1) are not altered by protein binding, nor do they depend on sequence.

Kinetic considerations do not alter the thermodynamic conclusions, either. It is known that slips only occur over short distances during replication (Schlötterer & Tautz, 1992), so it might be argued that, after replicating nine repeats (Figure 2a), kinetic barriers would prevent equilibration, and the 24 states in Figure 2c are not all kinetically accessible prior to the next step of replication. But it is not necessary for any of the 24 structures to be created directly by slippage of the structure in Figure 2a. Slippage can occur anytime during the replication of the first 27 nucleotides. The entropic advantage of slipping is exactly the same, regardless of when the three-nucleotide bulge is created.

It can also be shown that entropic considerations favor models in which the slipped structures form bulges rather than hairpins. For multiple slips, entropic effects are quite different for the two structures. If the first slippage nucleates a hairpin that grows with subsequent slippages, the second slip has only one place to go—into the existing hairpin—so the conformational entropy change accompanying the second slip is zero. However, if multiple slips create multiple bulges, then the second, third, and subsequent slips each receive nearly the same entropic advantage as the first, because each new bulge still has many possible locations. In Figure 2c, for example, 23 of the original 24 sites remain available to accommodate a second slip. For long direct repeats with multiple slips the entropic advantage for each slip is about the same.

It is more difficult to evaluate which structures enthalpic considerations would favor, bulges or hairpins. On the one hand, hairpins would appear to be favored, because the enthalpic cost of forming the junction and the unpaired bases in the loop is paid only once, and subsequent slips can be absorbed into the hairpin with no additional enthalpic penalty. On the other hand, the proposed hairpins are not based on perfect Watson−Crick duplexes, so an enthalpic price must be paid every time additional unpaired nucleotides are absorbed into the hairpin stem. It is unclear whether this price would be higher or lower than that paid for forming an additional bulge.

The entropic instability model can be tested experimentally *in vitro*. If this model is correct, the reactivity of direct repeat regions to chemical and enzymatic probes should be greater for longer direct repeats than for shorter ones. It can be shown that the reactivity should be directly proportional to the length of the direct repeat. [To first order, the probability of slippage, P, is the same for each position of a possible bulge (Figure 2c). If P is small, and if the direct repeat has a total length of N base pairs, then the number of expected slips is NP. Assuming that reactivity is proportional to the number of slips, then reactivity is directly proportional to N.] Thus, the model can be tested by careful measurements on the dependence of reactivity on length of the triplet repeat.

It would also be interesting to see if slipped structures are recognized by *in vitro* repair systems (Lahue et al., 1989). If so, one might use such systems to incorporate radiolabeled nucleotides into triplet repeat regions. The model presented here would predict that incorporation rates would increase in direct proportion with the length of the direct repeat. If the entropic instability model is sustained by *in vitro* studies, the next challenge would be to devise experiments testing it *in vivo* (Kang et al., 1995).

ACKNOWLEDGMENT

This research was initiated because of discussions with R. D. Wells. I thank J. Ordway, R. Gellibolian, A. Bacolla, P. J. Detloff, C. E. Pearson, and R. R. Sinden for critical discussions and R. D. Wells and R. R. Sinden for providing data prior to publication.

REFERENCES

Bates, G., & Lehrach, H. (1994) *BioEssays 16*, 277−284.
Campuzano, V., et al. (1996) *Science 271*, 1423−1427.
Kang, S., et al. (1995) *Nat. Genet. 10*, 213−218.
Lahue, R. S., et al. (1989) *Science 245*, 160−164.
Levinson, G., & Guttman, G. A. (1987) *Nucleic Acids Res. 15*, 5323−5338.
Panzer, S., et al. (1995) *Stem Cells 13*, 146−157.
Parsons, R., et al. (1993) *Cell 75*, 1227−1236.
Pearson, C. E., & Sinden, R. R. (1996) *Biochemistry 35*, 5041−5043.
Richards, R. I., & Sutherland, G. R. (1992) *Nat. Genet. 1*, 7−9.
Schlötterer, C., & Tautz, D. (1992) *Nucleic Acids Res. 20*, 211−215.
Strand, M., et al. (1993) *Nature 365*, 274−276.
Sutherland, G. R., & Richards, R. I. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 3636−3641.
Tinoco, I., Sauer, K., & Wang, J. C. (1985) *Physical Chemistry*, Prentice-Hall, Englewood Cliffs, NJ.
Wells, R. D. (1996) *J. Biol. Chem. 271*, 2875−2878.

BI962771E