6

Tandem DNA Repeats: Generation and Propagation in the Microgene Polymerization Reaction and *in vivo*

Mark Itsko[1,4], Eitan Ben-Dov[2], Avinoam Rabinovitch[3] and Arieh Zaritsky[1]
Departments of [1]Life Sciences and [3]Physics, Ben-Gurion University of the Negev, PO Box 653, Be'er-Sheva 84105, Israel,
[2]Achva Academic College, MP Shikmim, 79800, Israel
[4]Present address: Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA

1. Introduction

Short and long-motif periodicities are omnipresent in genomes of both eukaryotes (Britten & Kohne 1968) and prokaryotes (Hofnung & Shapiro, 1999). Repetitive DNA can amount to more than half of higher eukaryotic genomes, specifically in *Homo sapience* (International Human Genome Consortium, 2001) and *Zea mays* (Meyers et al., 2001). Even in prokaryotes it can account for about 6% of the total genome, specifically in *Mycoplasma pneumoniae* (Ruland et al., 1990) and *Neisseria meningitides* (Parkhill et al., 2000). Two major types of repetitive DNA exist (Dogget, 2000): tandem (head-to-tail contiguous) and interspersed (non-contiguous) with a specific pattern scattered all over the genome and lengths varying up to several hundred nucleotides (nt).

Emergence and propagation of interspersed repeats such as REP (Repetitive Extra-genic Palindrome) found in bacteria and *Alu* repeats that are abundant in the human genome are putatively attributed to reproduction processes mediated by transposons (Gilson et al., 1984) or retroviruses (Ullu & Tschudi, 1984). These are multi-step complex enzymatic processes and will not be considered here. This chapter deals with tandem repeats and their generation *in vitro* by the Microgene Polymerization Reaction (MPR) (Itsko et al., 2008a; 2008b; 2009), and proposes a mechanism for their *in vivo* generation as well.

Most tandem repetitive DNA sequences in higher eukaryotes are located near the chromosomal telomers or centromers where they play important roles in maintaining genome integrity (Blackburn, 1991) and segregation (Catasti et al., 1994). In addition to this un-transcribed but evidently functional repetitive DNA, the human genome contains many apparently non-functional repetitive DNA sequences in the forms of micro-satellites (a variety of di-, tri-, tetra-, and penta-nucleotide tandem repeats) and mini-satellites (30-35 bp long, with variable sequences and conserved cores of 10-15 bp) (Dogget, 2000). The number of repeats is prone to expand during replication because its constituent strands can slide over each other between the multiple complementary regions (Wells, 1996). Thermodynamically unfavorable structures bulging out from DNA duplex that accompany the strand sliding

process can be stabilized by the inner base-pairs which facilitate the expansion (Kang et al., 1995). Expansion of DNA repeats is associated with a variety of human hereditary diseases (Mirkin, 2007).

The conceptual views on repetitive DNA were changed from considering it as parasitic DNA (Orgel & Crick, 1980) up to necessary organizer of genomic information (Shapiro & von Sternberg, 2005). Furthermore, numerous periodicities encrypted in encoded proteins may reflect the evolution of modern coding sequences based on primordial oligomeric repeats (Ohno, 1987). Unused preexisted long repetitive sequences may yield coding frames expressing unique enzymes even currently (Ohno, 1984).

Different *in vitro* systems are used to study mechanisms underlying genomic repeat expansion and possible evolutionary aspects of this process. These systems include double- or single-stranded DNA with relatively short (4-8 nt) repetitive units and thermophilic DNA polymerases operating under isothermal conditions. Staggered re-annealing of constituent strands is putatively involved in expansion of repetitive DNA duplexes (Tuntiwechapikul & Salazar, 2002; Fig. 1A) whereas hairpin elongation was proposed for extension of single-stranded DNA with palindromes (Ogata & Miura, 2000; Ogata & Morino, 2000; Fig. 1B). Applying temperatures near the melting point of the hairpin-coil transition in the former or in the starting duplex in the latter profoundly facilitate such expansions. Total repetitive DNA synthesis is greatly accelerated if hairpin elongation is combined with endonucleolytic digestion of obtained long repetitive products (Liang et al., 2004; Fig. 1C).

The MPR, which was initially developed to produce artificial proteins containing repetitive motifs (Shiba et al., 1997), can be used to study not only DNA repeat propagation but their enigmatic generation as well. In this system, a medium-sized (40-50 nt) *non-repetitive* homo-*duplex* DNA (HD) evolves into multiple head-to-tail repetitive products during heat-cool cycles in the PCR.

In this chapter we consider this reaction, in strict terms of physical and polymer chemistry, to decipher the steps composing it and the mechanisms underlying their consecutive development. Several unorthodox views and consistent experimental results are described that obey basic thermodynamic rules, and analogous *in vivo* reactions are discussed in light of repeated motifs observed among subspecies of an entomopathogenic bacterial species.

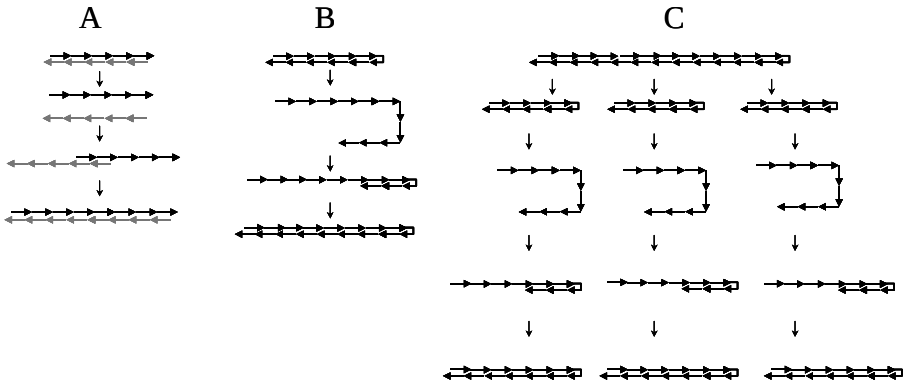

Fig. 1. *in vitro* systems for repeat expansion of a homo-duplex (**A**), single stranded (ss) DNA, (**B**) and ss-DNA with administered endonuclease (**C**). Black arrows, forward repeating units; grey arrows, reverse repeating units

Name	Sequence[a]	GC (%)	T_m[b] (°C)	ΔG_{OS}[c] (kcal mol^{-1})
NOMUL	5'-GG**AA**ATAGAAGAACTTAAATCTTTATTAG**AG**ATTAAC**ACA**GC-3' 3'-CC**TT**TATCTTCTTGAATTTAGAAATAATC**TC**TAATTG**TGT**CG-5'	28.6	67 (68)	-43
NOMU	5'-GGTGATAGAAG**AA**CTTAAATCTTTATTAGGAATT**AAC**CTGGC-3' 3'-CCACTATCTTC**TT**GAATTTAGAAATAATCCTTAA**TTG**GACCG-5'	33.3	68 (68)	-45
EVNA	5'-GGTGATAGAAG**TG**CTTAAATCTTTATTAGGAATT**GCT**CTGGC-3' 3'-CCACTATCTTC**AC**GAATTTAGAAATAATCCTTAA**CGA**GACCG-5'	38.1	71 (72)	-48
EVNAH	5'-GGTGATAG**TGC**TGCTT**TG**ATCTTTATTAGGAATTGCTCTGGC-3' 3'-CCACTATC**ACG**ACGAA**AC**TAGAAATAATCCTTAACGAGACCG-5'	42.9	73 (74)	-50

[a] Bold type letters indicate differences in composition between EVNA and NOMU. Underlined bold type letters indicate differences in composition between NOMUL and NOMU and between EVNA and EVNAH.

[b] Experimental and UNAFold-predicted (in brackets) melting temperature of the corresponding *OS* types.

[c] Association free energy (Visual OMP6-predicted) of complementary strands into *OS* at 37°C.

Reprinted from Itsko et al., 2008b with permission from Elsevier.

Table 1. Original homo-duplexes, called Original Singlets (*OS*) used in MPR research

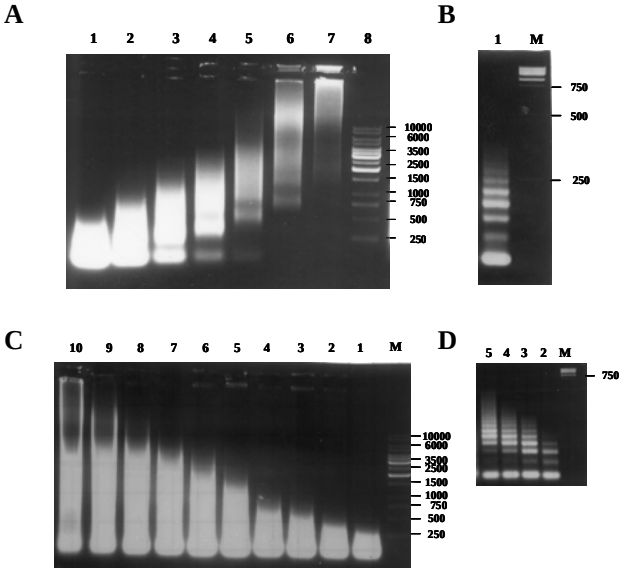

Fig. 2. Demonstration of MPR with EVNA *OS* (Table 1) and its concentration-dependence on the reactants. Change of product length with concentrations of EVNA with Vent (**A, B**) and of dNTP (**C** with Taq**, D** with Vent) on 0.8% (**A, C**) and 2.5% (**B, D**) agarose gels. (**A**, **B**) Lanes 1-7, 10.4, 5.2, 2.6, 1.3, 0.64, 0.32, 0.16 μM of EVNA, respectively. Lanes 8 or M, DNA ladder. (**C, D**) The following concentrations (in μM) of each dNTP were used: lane 1, 100; lane 2, 200; lane 3, 300; lane 4, 400; lane 5, 500; lane 6, 600; lane 7, 700; lane 8, 800; lane 9, 900; lane 10, 1,000. Concentration of the EVNA homo-duplex, 5.6 μM. Lanes M, DNA ladder. Reprinted from Itsko et al., 2008a with permission from Elsevier

2. Demonstration of the MPR

MPR can easily be demonstrated by applying long (up to 64 cycles) end-point-detection PCR heat/cool cycling conditions on reaction mixtures containing DNA in the form of homo-duplex (HD) of length 40-50 bp, the sequence of which does not include any repetitive motif (e.g., Table 1). The PCR product is composed of long DNA stretches of heterogeneous length that is visualized as a smear on loose agarose gels and resolvable into discrete bands of HD multiples on dense gels (Fig. 2).

3. Overall scheme of MPR

The MPR is kinetically divided into three stages: initiation, amplification and propagation (Fig. 3), each is subdivided into a number of steps that are considered later in the chapter. Here they are formulated in a simplified way with the following variables and parameters:

HD_i - homo-duplexes of DNA containing a variable number i repeats. Correspondingly, HD_1 is original non-repetitive homo-duplex called also OS (original singlet) with which the MPR starts, and HD_2 is doublet of homo-duplexes designated as D, including initial doublet (ID) that triggers propagation.

k_I is the constant rate of MPR initiation.

k_{Ampl} is the constant rate of ID amplification by OS.

k_{Pr} is the constant rate of MPR repeat propagation per PCR cycle, assumed to be independent of polymer length (measured in repeat units n).

The symbols embraced in square brackets designate concentrations of corresponding DNA species.

Initiation

The minimal repetitive unit that is prone to expand by staggered re-annealing and replication of overhangs (Fig. 3B) is D. Propagation is therefore initiated by generation of a so-called initial doublet (ID). The equation formulating the simplified process of initiation (Fig. 3 A) is:

$$2OS \xrightarrow{k_I} ID\,. \tag{1}$$

The mechanism of this reaction and its molecularity will be discussed in sections 5 and 6.

Amplification

The initiator (ID) can be amplified by the OS (Fig. 3 B) according to:

$$OS + ID \xrightarrow{k_{Ampl}} 2D\,. \tag{2}$$

Since ID and D are the same molecular species generated by different mechanisms, the reaction describes an autocatalytic process, in which the mass concentration of the initiator is rapidly brought to that of OS (see section **7** below). If $k_{Ampl} = k_{Pr}$, the amplification stage is kinetically included in the following propagation stage.

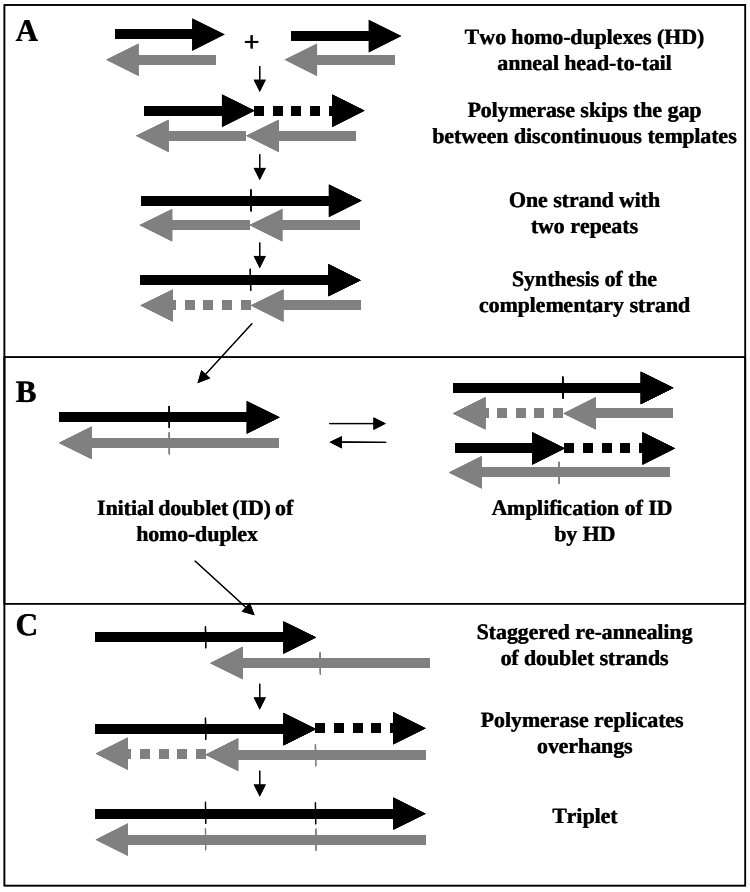

Fig. 3. A model for MPR. (**A**) Initiation. (**B**) The initiator (ID) and its amplification. (**C**) The propagation of ID in the first heat/cool cycle. Reprinted from Itsko et al., 2009 with permission from Elsevier

Propagation

After generation of doublets, the number of repeats in DNA is expanding according to

$$HD_i + HD_j \xrightarrow{k_{pr}} 2HD_n\,, \tag{3}$$

where $n \le i + j - 1$, since at least one repeat is always hidden in the overlap paired region (Fig. 3C).

Propagation is also an autocatalytic process resulting in an exponential growth of the number of repeats per one polymer molecule $\langle n \rangle_N$. This is justified as follows. In the extension reaction (eq. 3), two molecules of lengths n_1 and n_2 ($n_1 \le n_2$) yield two molecules with lengths in the range between n_1+1 and n_1+n_2-1. Assuming that the population of such repetitive products is uniformly distributed with a common difference of one repeat unit,

the average length of the product $\langle n\rangle = \big((n_1+1)+(n_1+n_2-1)\big)/2 = n_1+n_2/2$. Averaging over all possible reactions of this sort yields

$$\langle n\rangle_{N+1} = \langle n_1\rangle_N + \langle n_2\rangle_N/2\ . \tag{4}$$

Since $\langle n_1\rangle_N = \langle n_2\rangle_N = \langle n\rangle_N$ (belonging to the same distribution), the average length of the polymer increases by a factor of 1.5 at each cycle:

$$\langle n\rangle_{N+1} = 3\langle n\rangle_N\ /\ 2 \ \text{ or } \ \langle n\rangle_N = \langle n\rangle_0 \times 1.5^N\ . \tag{5}$$

Propagation is finished when nucleotides (dNTPs) are depleted.

4. Extent of polymerization in MPR

The final polymer length in chain-growth polymerization reactions is determined by the "kinetic chain length" defined as the number of monomer units consumed in the propagation stage per active center produced in the initiation stage "(Flory, 1953; Atkins, 1994) as visualized in Fig. 4. Accordingly, the mean MPR product size increases with decreasing *OS* and increasing dNTP concentrations (Itsko et al., 2008a). The nucleotide concentration [*dNTP*] presented in the reaction mixture determines the extent of propagation, whereas that of the [*ID*] (equal to $[OS]_0$ due to amplification (eq. 2)) represents active centers. Thus, in MPR-produced multiple repetitive DNA, the final length

$$\langle n\rangle\ =\ (1/m)\times[dNTP]/[OS]_0 \tag{6},$$

where *m* is the number of nucleotides composing one *OS*. Eq. 6 is concordant with the experimental results (Fig. 2).

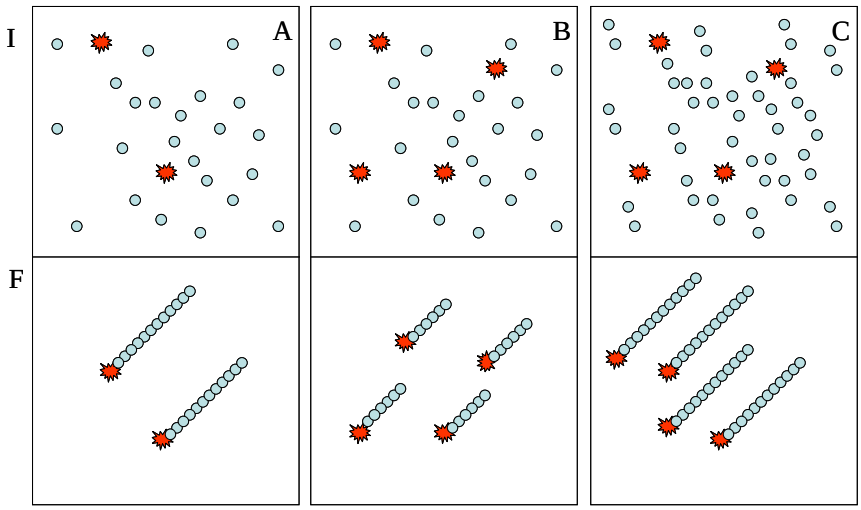

Fig. 4. Extent of polymerization in the MPR. Illustrative demonstration of its dependence on concentrations of dNTP (blue spots) and ID (red stars), at given concentrations of the reactants (**A**), at twice the concentration of ID (**B**) and at twice the concentrations of both (**C**). I, initial reactants; F, final products

5. Molecularity of the initiation stage

Progress of MPR can be followed by Real-Time PCR (Fig. 5). The initial concentration of *OS* ($[OS]_0$) determines the cycle number (the operative threshold cycle) *N*th (in cycle units *N*) at which the signal (the microgene expansion process) is first detected. A 2-fold decrease in $[OS]_0$ delayed *N*th by approximately 4 cycles (i.e., $\Delta N\text{th} = 4$; Fig. 5).
Following the initiation process, total MPR products (*Tot*) are propagated exponentially formulated by

$$[Tot]_N = [ID] \times (1+E)^N \qquad (7),$$

where *E* is the amplification efficiency per cycle. Average *E* determined experimentally is about 0.7 (between 0.8 and 0.57), larger than that calculated from general considerations (eq. 5). The generation-rate of the initial doublet *ID* is assumed to be proportional to the power *m* (molecularity) of the concentration of *OS* $d[ID]/dN \propto [OS]^m$ (Fig. 3, A, B). The rate of amplification that has exponential nature (eq. 2) is much higher than that of the initiation, and hence the consumption of *OS* molecules in the initiation process is negligible compared to that during the amplification stage. The initiation process would therefore not affect $[OS]_0$ hence [*ID*] would be $\propto N \times [OS]_0^m$. The MPR initiation is reasonably assumed (Itsko et al., 2008a) to occur in all $[OS]_0$ at the first cycle $N = 1$ (though the low sensitivity of RT-PCR detects the products at a later stage, i.e., at the *N*th). The experiments were performed with a series of 2-fold-$[OS]_0$ values (Fig. 5). The predicted ratio between [*ID*] of two successive dilutions is thus

$$\frac{[ID_2]}{[ID_1]} = \frac{\left(2 \times [OS]_0\right)^m}{[OS]_0^m} = 2^m \quad . \qquad (8)$$

The exponential propagation of *Tot* following generation of *ID* depends on the number of cycles and amplification efficiency (eq. 7). If *E* (derived from the slope in each line of Fig. 5) remains constant, the ratio between successive [*ID*] values in the RT-PCR experiments with two-fold different initial concentrations of *OS* may be derived at points with equal amount of $[Tot]_N$, as for example the threshold points (Fig. 5) according to

$$\frac{[ID_2]}{[ID_1]} = (1+E)^{(Nth_1 - Nth_2)}, \qquad (9)$$

and the molecularity m of the reaction can be derived by equating eq. 8 with eq. 9:

$$m = (Nth_1 - Nth_2) \times \frac{\ln(1+E)}{\ln 2} . \qquad (10)$$

The observed values of *E* however, do vary slightly (0.8–0.6) between successive lines (Fig. 5). A molecularity of 3.1 (ranging between 2.6 and 3.4) for the initiation process was estimated, using the average *E* value (0.71). It can thus be concluded that three *OS* must somehow interact to initiate doublet formation in the MPR.

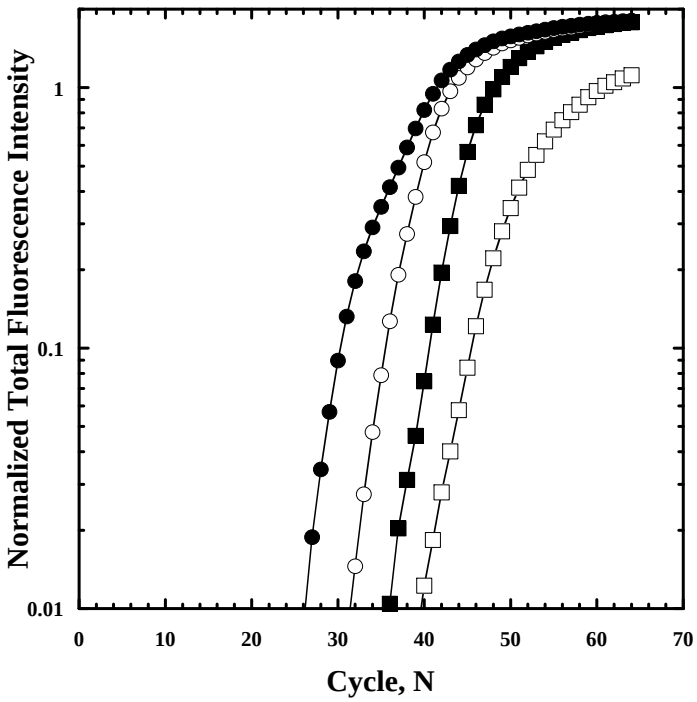

Fig. 5. Relative total MPR products generated by RT-PCR with the following initial EVNA homo-duplex concentrations (in μM): closed circles, 0.32; open circles, 0.16; closed squares, 0.08; open squares, 0.04. Reprinted from Itsko et al., 2008a with permission from Elsevier

6. Kinetics and thermodynamics of initiation

Third-order kinetics of initiation leads to a simple mechanism for the generation of *ID*: a rare and reversible association between three *OS* generates a nucleation complex (*NC*), (Fig. 6), which converts to *ID* according to:

$$3OS \overset{I}{\longleftrightarrow} NC \overset{II}{\longrightarrow} ID + 2OS \quad (11),$$

where *I* and *II* denote first (non-enzymatic fast equilibrium) and second (enzymatic rate-limiting) stages in the MPR initiation process.

One of these three *OS* (labeled in grey) aligns and bridges the other two, fixing them in the required proximity for the DNA polymerase to skip the inter-template gap while displacing the confronting non-template strand of the adjacent *OS*. This bridging occurs putatively through occasional Watson-Crick bonds between aligning and aligned homo-duplexes. Such putative bridging complexes are not covalently bonded, very unstable and have not been demonstrated directly. Experimental system for their revelation has still not been elaborated, but they can be predicted using the following software packages:

1. Visual OMP6 visualizes different 2nd-rank structures that can be potentially formed from constituent strands of *OS* (Fig. 7).

2. UNAfold is not so illustrative but much more comprehensive in providing schematic landscape of all hybridization patterns possible between complementary strands of *OS* and their temperature dependence (Fig. 8).

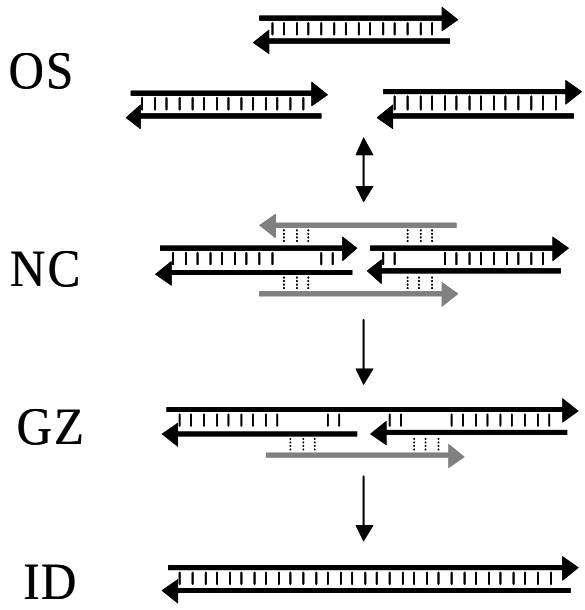

Fig. 6. A model for the MPR initiation. *OS*, original singlet *HD*s (a pair of complementary primers); *NC*, nucleation complex (an arrangement of three *OS*); grey strands correspond to the *OS* bridging the gap; *GZ*, gap zipper (conformation composed of half *NC* (*hNC*)) and a DNA polymerase-generated (by gap skipping) single strand of *ID*. Adapted from Itsko et al., 2008b with permission from Elsevier

UNAfold probability dot plots demonstrate that increasing temperature increases probabilities of generation of overall 2nd-rank structures and those that can be involved in *NC* (Fig. 8) among them. The van't Hoff equation that formulates this tendency is:

$$\frac{d\ln K_{1\to 2}}{d(1/T)} = -\frac{\Delta H_{1\to 2}}{R}, \tag{12}$$

where $K_{1\to 2}$ and $\Delta H_{1\to 2}$ are, respectively, the equilibrium constant and the standard enthalpy for the transition between more stable 1st-rank (*OS*) and less stable 2nd-rank (*NC*) structures in I part of eq. 11. $\Delta H_{1\to 2} > 0$ because the stability of 2nd-rank structures is lower. Therefore, as the temperature increases ($1/T$ decreases), $\ln K_{1\to 2}$ and hence $K_{1\to 2}$ must increase, reflecting a rise in the probability of the 2nd-rank conformations. Generation of *ID* from *NC* (Fig. 6, part II of eq. 11) requires keeping inside of *NC* 1st-rank

structure of *OS* for DNA polymerase to skip the inter-template gap by extending the pre-hybridized *OS* constituent strands. On the other hand, the equilibrium constant of the formation of this 1st-rank structure from separated complementary DNA strands will decrease with temperature according to the same van't Hoff equation (eq. 12) in which $\Delta H < 0$ due to exothermic property of the hybridization. Decrease in this constant results in impeding this polymerization step and overall amplification and propagation stages. Thus temperature rise increases the rate of initiation but decreases the rates of amplification/ propagation.

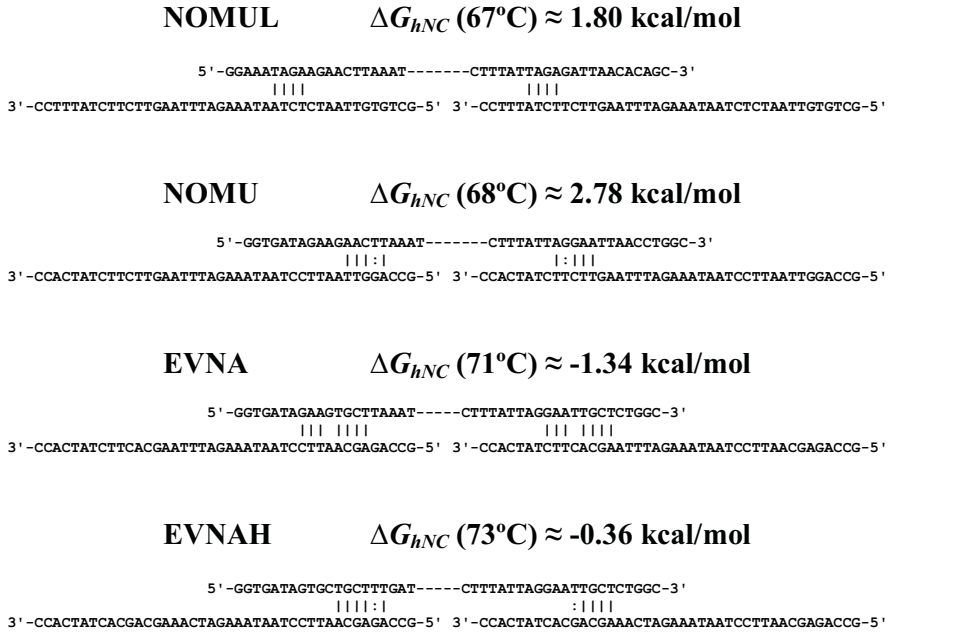

Fig. 7. Suggested structures of *hNC* (half *NC*) of *OS* types with approximate ΔG_{hNC} of their formation from single strands (at corresponding T_m), predicted by Visual OMP6. Conventional (|) and G:T pairings. Reprinted from Itsko et al., 2008b with permission from Elsevier

Following MPR kinetics by RT-PCR at different temperatures can test the above reasoning. Backward extrapolation of the amplification curves with different types of *OS* (next paragraph) to the first cycle retrieves the ratio $[ID]/[OS]_0$ and the value of $\ln(1+E)$ designated as amplification rate (A). The variability in changes of A (Fig. 9A) and $[ID]/[OS]_0$ (data not shown) with temperature is exceedingly high and not amenable for analysis. However these two parameters are strictly related in an exponential mode (Fig. 9B).

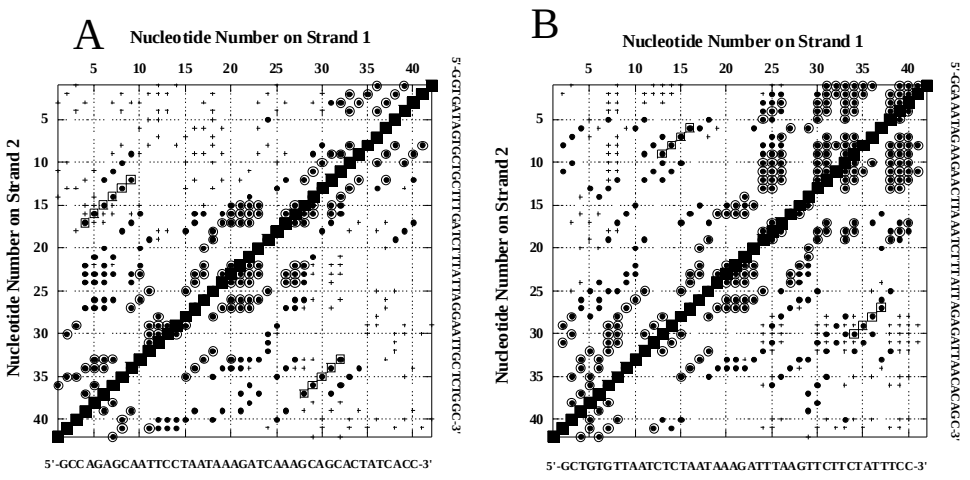

Fig. 8. Hybridization-probability dot-plot of EVNAH (**A**) and NOMUL (**B**). Each symbol designates pairing between strands' nucleotides. Filled-in squares designate pairs existing in *OS* at all tested temperatures with probability close to 1. Open circles (enclosing dots), 2nd-rank pairs at 68°C for NOMUL and 74°C for EVNAH. Dots, pairs existing at 71°C for NOMUL and 77°C for EVNAH, some of which (enclosed in open squares) form bridging structures (*hNC*) leading to *NC* (Fig. 7). Crosses, conformations existing at 74°C for NOMUL and at 80°C for EVNAH. Reprinted from Itsko et al., 2008b with permission from Elsevier

As *A* rose, the calculated ratios $[ID]/[OS]_0$ exponentially declined over seven decades. Increased *A* corresponds to a decrease in the applied temperature (Fig. 9A); hence, formation of *ID* is accelerated as temperature rises (Fig. 9B), consistent with the above reasoning. The high variability in the above parameters when they are plotted against temperature (Fig. 9A) seems to result from limited accuracy of the temperature maintenance by the heat block of the RT-PCR apparatus for the duration of each cycle (repeatedly launched by rapid cooling from 95°C to the desired temperature). The actual temperature that acts on each multiple sample may therefore differ from the registered one. The temperature-dependent amplification rate *A* (Fig. 9B) was used alternatively as a more stable indicator because it reflects the average temperature during the cycle. At each calculated *A*, the ratio $[ID]/[OS]_0$ for NOMUL was lower than that for EVNAH, consistent with lower stability of the putative bridging structure involved in NC formation of NOMUL than that for EVNAH (Fig. 7). In addition, at A=0.46 for NOMUL and 0.51 for EVNAH, the ratio lines were drastically bent (Fig. 9B), justified by lowering the fraction of *hNC* structures (Fig. 7) in overwhelmed number of emerging alternative pairings between constituent *OS* single strands at relatively high temperatures (Fig. 8, crosses) that correspond to lower amplification rates.

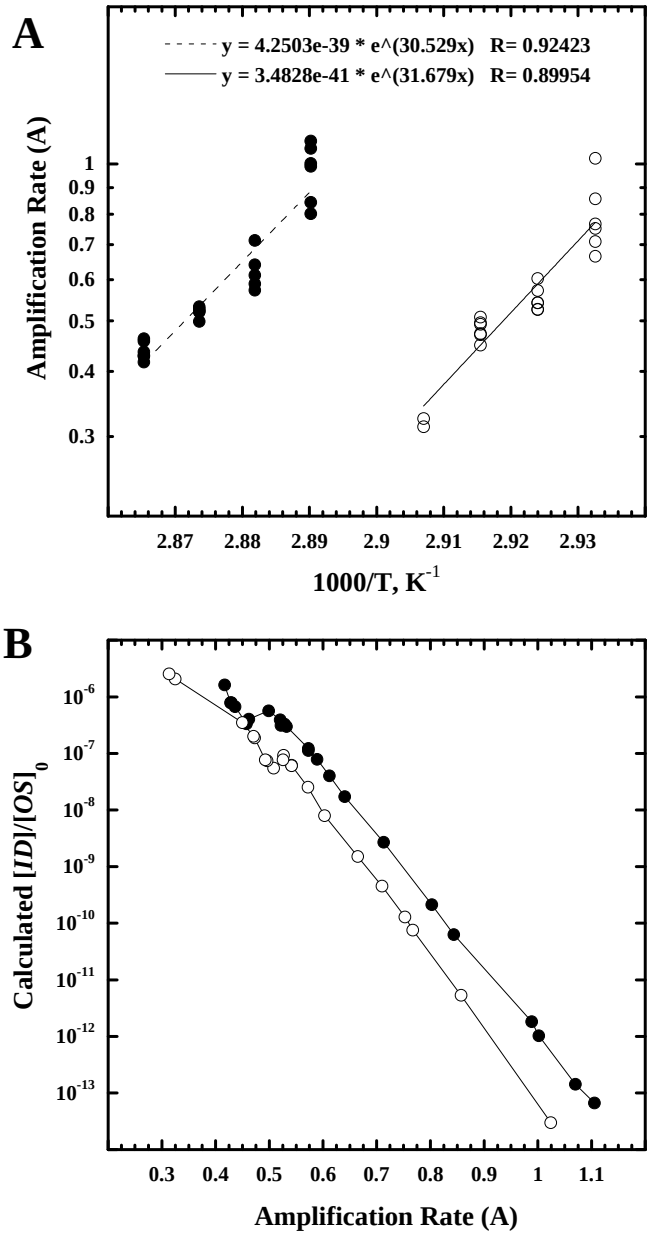

Fig. 9. **A**, Arrhenius plots of MPR amplification rate, with NOMUL (open circles) and EVNAH (filled-in circles). **B**, Dependence of calculated $[ID]/[OS]_0$ on MPR amplification rate, with NOMUL (open circles) and EVNAH (filled-in circles). Reprinted from Itsko et al., 2008b with permission from Elsevier

7. Kinetics of amplification

Kinetics of the MPR amplification stage demonstrates biphasic behavior for three of the four types of *OS* used (Fig. 10). Single-burst kinetics was seen with NOMU at all tested temperatures and with the rest of *OS*s at temperatures exceeding by 2-4°C the melting point of corresponding *OS*. The biphasic kinetics is explained by two parallel processes (with

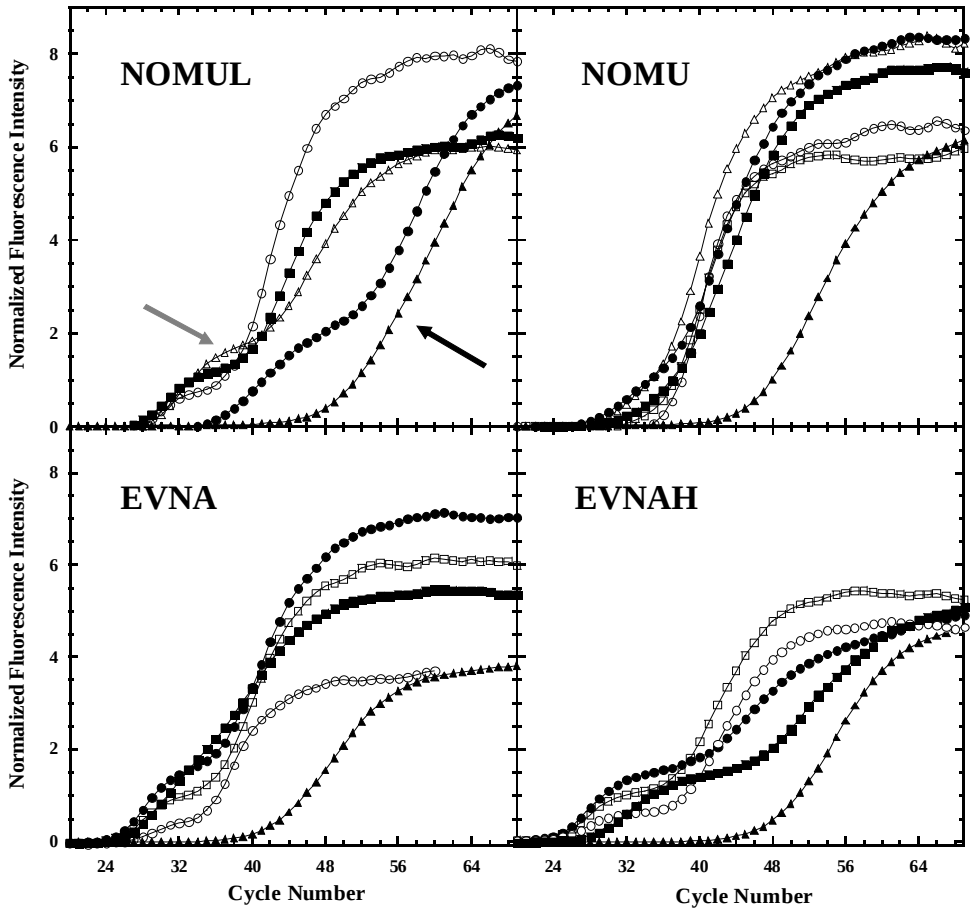

Fig. 10. MPR kinetics with four OS types: NOMUL (open circles, 68ºC; filled-in squares, 69ºC; open triangles, 70ºC; filled-in circles, 71ºC; filled-in triangles, 72ºC). NOMU (open circles, 68ºC; open squares, 69ºC; open triangles, 70ºC; filled-in circles, 71ºC; filled-in squares, 72ºC; filled-in triangles, 73ºC). EVNA (open circles, 71ºC; open squares, 72ºC; filled-in circles, 73ºC; open diamonds, 74ºC; filled-in triangles, 75ºC). EVNAH (open circles, 71ºC; open squares, 72ºC; filled-in circles, 73ºC; filled-in squares, 74ºC; filled-in triangles, 75ºC). Arrowheads point at biphasic (grey) and single-burst (black) kinetics. Reprinted from Itsko et al., 2008b with permission from Elsevier

distinct kinetic parameters) involved in the MPR expansion, amplification of the *ID* by *OS* and propagation of doublets (*D*) to triplets (*T*) and higher-level multiples through staggered conformations (*SD*) (Fig. 3C) Existence of an amplification stage was suggested for the first burst in the progressively increased fluorescence intensity during MPR (Fig. 10).
The amplification (Fig. 11) can be formulated by the chemical equation

$$OS + D \xleftrightarrow{I} 2H \xrightarrow{II} 2D\,. \tag{13}$$

where I and II denote first (fast equilibrium) and second (rate-limiting) steps of the process. The rate of amplification (II) can be expressed as $d[D]/dt = k[H]$, where k is the rate constant of the rate-limiting step and H is the hetero-duplex composed of hybridized *OS* and *D* (Fig. 11). At temperatures higher than T_m of *OS* (and obviously of *H*) and lower than that of *D*, step I involves melting of *OS*, fraying the ends of *D* and hybridization of the single strands of the former with the latter.
Assuming that this step occurs near its multi-state equilibrium with $K_H = [H]/[OS][D]$, the overall process is expressed by

$$d[D]/dt = kK_H[OS][D]\,. \tag{14}$$

k rises with temperature, but K_H decreases due to the dissociation of *H* into *D* and single stranded *OS*. In the reaction described by eq. 13, k rises less than the drop in K_H, so that the overall value kK_H decreases with temperature. That is reflected in the highly negative value (around -62 kcal mole^{-1}) of "$\Delta G^{\ddagger}_{Ampl}$ derived from the Arrhenius plots for NOMUL and EVNAH (Fig. 9A). In other chemical processes as well, an exothermic fast-equilibrium stage leads to negative $\Delta G^{\ddagger}$ of the overall process. Solving eq. 13 (for detailed derivation see Itsko et al., 2008b) yields an expression for total fluorescence (Flu_{Tot}):

$$Flu_{Tot} = \alpha[OS]_0\left(1 + (\gamma - 1)\frac{(e^{AN} - 1)}{R + e^{AN}}\right) \tag{15},$$

where $A = kK_H\varepsilon([OS]_0 + [D]_0)$, ($[OS]_0$ and $[D]_0$ are initial concentrations of *OS* and *D* correspondingly where $[D]_0$ is actually $[ID]$), N is the number of cycles and ε is the cycle period, $R = [OS]_0/[ID]$, α is the arbitrary coefficient expressing the fluorescence brightness of *OS*; γ is the ratio between the values of the fluorescence brightness of *D* and *OS*. Eq. 15 was used to approximate the first burst in fluorescence intensity of MPR curves (Fig. 10) and retrieve $[ID]/[OS]_0$ ratio and amplification rate A from them.

8. Thermodynamics of transition from amplification to propagation

Denatured single strands of doublet (*ssD*) can be hybridized in fully aligned manner generating doublet homo-duplex (*D*, right in Fig. 12A) or in staggered manner generating staggered doublet (*SD*, left). Transition from amplification to propagation is mediated by the appearance of *SD*. Obviously *SD* has lower stability than *D* due to half number of hydrogen bonds ($\Delta G_{SD\to D} < 0$). *SD* cannot readily return to *D* due to the energetic barrier ($\Delta G^{\ddagger}_{SD\to D}$) between these conformations, the magnitude of which is determined by the

stability (ΔG_{WSD}) of the additional 2nd-rank intermediate structure called weak staggered doublet (*WSD*) (Fig. 12B). The more stable it is, the more easily *SD* switches to *D* due to smaller $\Delta^{\ddagger}G_{SD\rightarrow D}$. *SD* switching to *D* prevents the propagation process and results in the biphasic mode of expansion (Fig. 10). Disappearance of the biphasic kinetics (Fig. 10) at temperatures higher than 71°C for NOMUL, 73°C for EVNA and 76°C for EVNAH (Table 1) is explained by increased probability of *SD* with temperature rise. UNAFold also predicts formation of *SD* at temperatures higher than: 73°C for NOMUL, 78°C for EVNA and 80°C for EVNAH (Fig. 13A). Formation of *SD* as a structure that is less stable than fully aligned hybridized *D* is stimulated by rising temperature according to the van't Hoff equation (eq. 12).

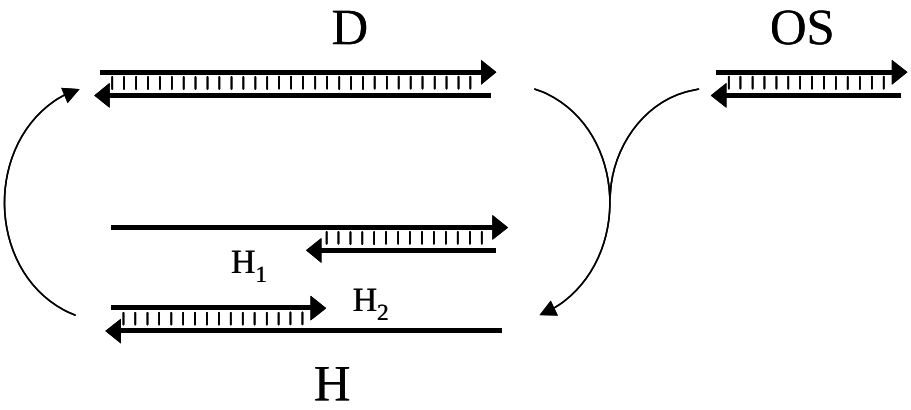

Fig. 11. The model for amplification: **OS**, Original Singlet *HD* (a pair of complementary primers), **D,** Doublet. **H_1** and **H_2**, hybridized denatured *D* and *OS*. Adapted from Itsko et al., 2008b with permission from Elsevier

Melting the doublets at high temperature and their quick-cooling afterwards can demonstrate *SD* conformations. Quick cooling prevents the *D* constituent strands from finding the final most stable fully-aligned conformation when they encounter each other but entrap them in first available conformation, mostly a staggered one. Such procedure was accomplished on MPR end products generated from all four *OS* types, with concentrations of dNTP and *OS* that limit the extent of expansion (Eq. 6 and Fig. 4), and revealed that out of the four, the *OS* of NOMU was the only one to diminish (bottom band, lane 5 in Fig. 13B), most likely due to its pairings with overhangs formed by staggered structures of multiple-repeated DNA. Thus, among all used *OS*s staggering of *D* of NOMU is the most facilitated process. It is also consistent with the observation that only this *OS* displayed single burst kinetics at all tested temperatures in RT-PCR (Fig. 10).

The facilitation of staggering and of transition from amplification to propagation stage in the NOMU case is reasoned by the least stability (highest ΔG_{hNC}; Fig. 7) of its bridging structure among all used *OS*s, that would reflect also the least stability of its *WSD* (Fig. 12B) because the latter is just a staggered tandem *OS*.

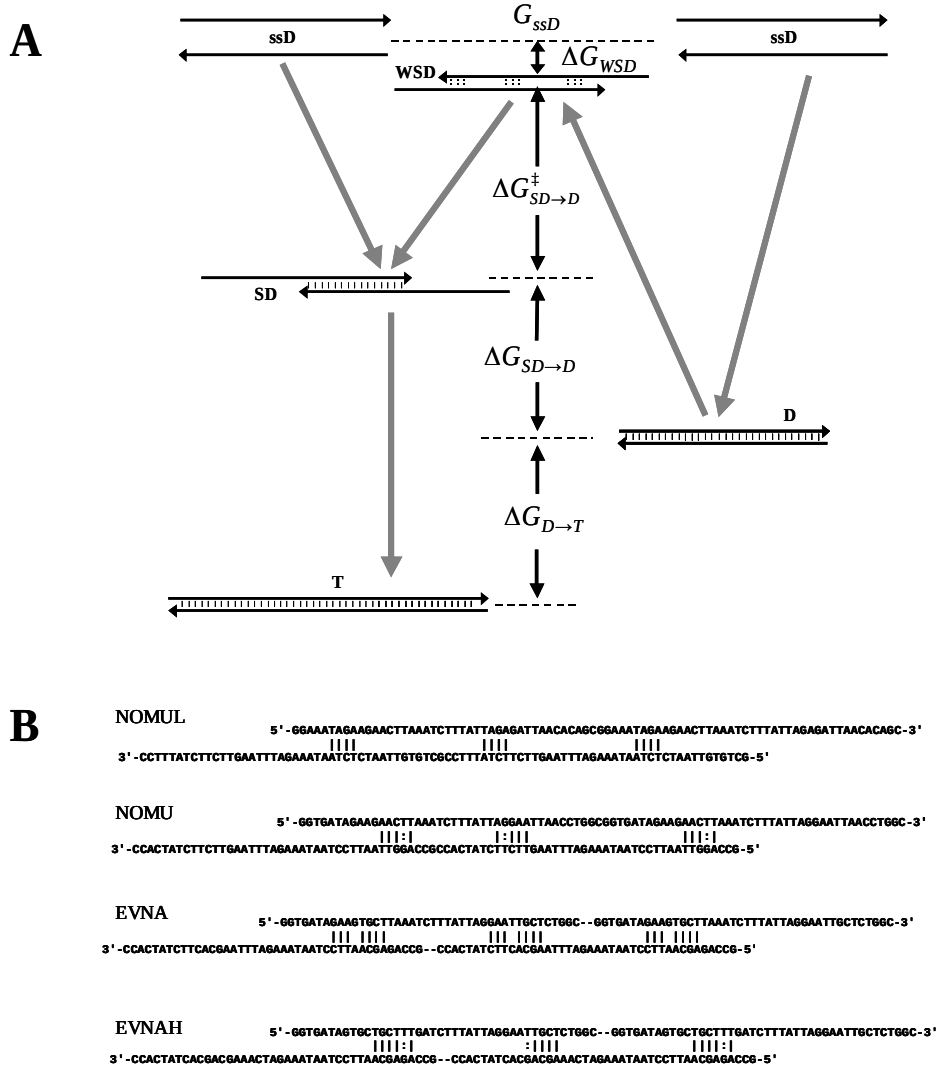

Fig. 12. Schematic free energy diagram of the various conformations that participate in the suggested transition from amplification to propagation. Dashed lines denote energy levels. G_{ssD} and $\Delta G_{D\rightarrow T}$ are, respectively, the free Gibbs energy of single stranded *D* and of the difference between *T* and *D*. **B**, *WSD* conformations (tandem staggered *OS*) of the four *D* types (related to Fig. 7). (Conventional (|) and G:T pairings). Adapted from Itsko et al., 2008b with permission from Elsevier

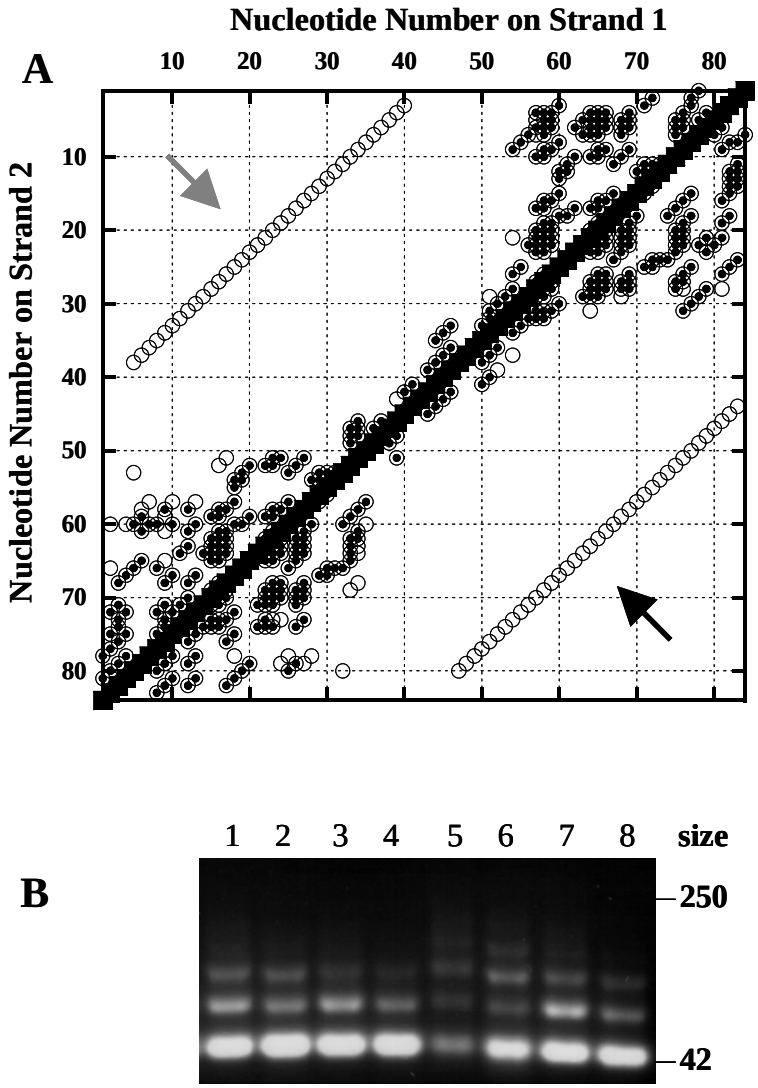

Fig. 13. Revealing SD structures *in silico* (A) and *in vitro* (B). **A**, Probability dot plot of hybridization of EVNAH's doublet. Each symbol designates pairing between strands' nucleotides. Filled-in squares designate pairs existing in *D* at all tested temperatures with probability close to 1. Open circles (enclosing dots or not, respectively) designate 2nd-rank pairs at 80°C or 81°C, with probabilities of approximately 10^{-6}. Arrows point at two lines corresponding to two *SD*, effective for propagation (black) and not (grey), **B**, MPR-generated products after 65 PCR cycles from 4 *OS* types, separated on 2.5% agarose gel. Lanes 1 and 2, with EVNAH; 3 and 4, EVNA; 5 and 6, NOMU; 7 and 8, NOMUL. Odd numbers designate samples denatured (10 min at 95°C) and then cooled rapidly. Reprinted from Itsko et al., 2008b with permission from Elsevier

9. Kinetics of propagation

The overall process of MPR propagation includes the following recurring set of three steps:

1. Aligning A_i forward and B_j reverse complementary DNA strands containing i and j repeats in staggered mode according to the equation:

$$A_i + B_j \xleftrightarrow{k_1, k_{-1}} A_i B_j \,, \tag{16}$$

   where k_1 and k_{-1} are rate constants of alignment and of melting, respectively.
2. Association of DNA polymerase with above staggered homo/hetero-duplex:

$$A_i B_j + Enz \xleftrightarrow{k_1, k_{-1}} A_i B_j Enz \,, \tag{17}$$

   where k_2 and k_{-2} are rate constants of association and of dissociation, respectively.
3. DNA-polymerase mediated filling-in overhangs according to:

$$A_i B_j Enz \xrightarrow{k_3} A_n + B_n + Enz \,, \tag{18}$$

   where k_3 is the rate constant of the enzyme-driven polymerization (turnover number) for filling in $(n - i)$ repeats.

The mathematical descriptions of these steps are:

1. $$\frac{d[A_i B_j]}{dt} = k_1[A_i][B_j] - k_{-1}[A_i B_j] - k_2[A_i B_j][Enz] + k_{-2}[A_i B_j Enz] \,, \tag{19}$$

2. $$\frac{d[A_i B_j Enz]}{dt} = k_2[A_i B_j][Enz] - k_{-2}[A_i B_j Enz] - k_3[A_i B_j Enz] \,, \tag{20}$$

3. $$\frac{d[A_{n(i,j)}]}{dt} = k_3[A_i B_j Enz] \,, \tag{21}$$

where square brackets designate concentrations of corresponding species and $A_{n(i,\ j)}$ is A_n generated from given i and j.

The rate of generation of intermediates such as $[A_i B_j]$ and $[A_i B_j Enz]$ is taken close to zero according to the assumption of steady state kinetics for them (Atkins, 1994). The final expression for the above chemical reaction is:

$$\frac{dA_{n(i,j)}}{dt} = \frac{k_3[Enz_{tot}][A_i][B_j]}{Enz_{tot}\, k_3/k_1 + K_D^{-1} K_M + [A_i][B_j]} \tag{22}$$

(for derivation see Itsko et al., 2009), where $K_M^{app} = K_M^{app*} + K_D^{-1} K_M$, $K_M^{app*} = Enz_{tot}\, k_3/k_1$, Enz_{tot} is the total concentration of enzyme, $K_D = k_1/k_{-1}$ is the equilibrium constant for duplex formation, $K_M = (k_{-2} + k_3)/k_2$ is the Michaelis-Menten constant.

Assuming $t = N\varepsilon$, where N is the number of cycles and ε is the cycle period, yields

$$\Delta A_{n(i,j)} = \frac{k_{Pol}[Enz_{tot}][A_i][B_j]}{Enz_{tot}\, k_3/k_1 + K_D^{-1} K_M + [A_i][B_j]} \,, \tag{23}$$

where $k_{Pol} = k_3\varepsilon$, and ε = 240 sec cycle^{-1} and ΔA_n [= $A_n(N+1) - A_n(N)$] is increments in $A_n(N)$ after one cycle.

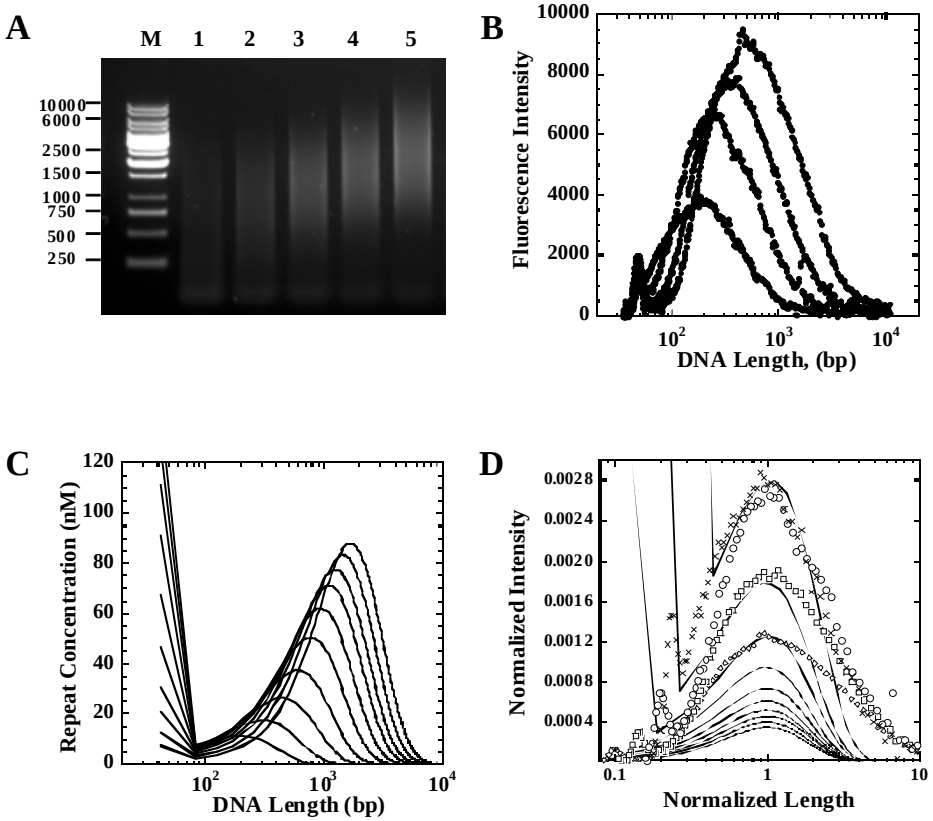

Fig. 14. Experimental versus model distributions. (**A**) Consecutive samples run on 0.8% agarose gel: M, DNA size markers; lane 1, cycle 16; lane 2, cycle 17; lane 3, cycle 18; lane 4, cycle 19; lane 5, cycle 20. (**B**) Scanned and digitized samples of cycles 17-20 (from **A**). (**C**) Matlab simulations. Final distributions of total MPR product at cycles 14-23. Model parameters: k_{Pol} = 0.09 repeat/cycle (k_3 = 3.75×10^{-4} repeat/sec), K_M^{app*} = 1 nM^2 (Eq. 22); k_I = 10^{-24} nM^{-3} sec^{-1} (Eq. 1); E = 0.735 (Eq. 7). (**D**) Juxtaposition of the model (full lines for cycles 14-23) with four (cycles 17-20) experimental distributions (crosses, open circles, open squares and open diamonds, respectively), each integrally normalized. Reprinted from Itsko et al., 2009 with permission from Elsevier

Calculation of the overall change A_n during a PCR cycle requires consideration of inflow to the A_n category from all the possible combinations of A_i and B_j that together allow the generation of such product and outflow of A_n to a group of longer lengths:

$$\Delta A_n = k_{\Pr}\left(\sum_{i=1}^{n-1}\sum_{j=n-i+1}^{\infty}\frac{\Delta A_{n(i,j)}}{i+j-1} - \sum_{j=2}^{\infty}\frac{(j-1)\Delta A_{n(n,j)}}{n+j-1}\right) \tag{24}$$

A similar equation may be formulated for ΔB_n. Since the initial concentrations of the complementary DNA strands dealt with here are identical, $A_1(0) = B_1(0)$, the propagation kinetics of A_n and B_n are identical, hence $B_i = A_i$.
The above set of difference equations (limited to 3,000) describing kinetics of MPR was solved numerically using Matlab 7 (MathWorks, Natrick, MA (http://www.mathworks.com)). The experimental results were compared with the model (Fig. 14) and satisfactory resemblance was found between them. A discrepancy was revealed between the skewness of the experimental and modeled distributions of final MPR length. The experimental distributions are skewed to longer lengths whereas the model predicts negative skewness. This means that for every (A_i, B_j) involved in the extension step (see eqs. 3 and 18) the generated product is much shorter than the expected maximum ((i+j-1)). This is consistent with lower processivity of the Vent polymerase (Kong et al., 1993).

10. MPR as an evolutionary process

MPR can be viewed as an evolutionary process to develop biological polymers such as DNA. The repeat generation aspect included in the MPR brings this process even closer to a common scenario of Darwinian evolution, namely, spontaneous appearance of the physico-chemical trait (*ID*) that is selected (amplified) since it directly contributes to its own reproduction in the form of repeats propagation. Theoretically, two variants of *ID* can be generated in the reaction mixture from a given *OS* with comparable probabilities, cis and trans (Fig. 15). Only the cis form is prone to propagation; the trans form is "infertile" since it cannot be aligned in a staggered manner. MPR polymers indeed contain only direct repeats and never alternating inverse repeats.

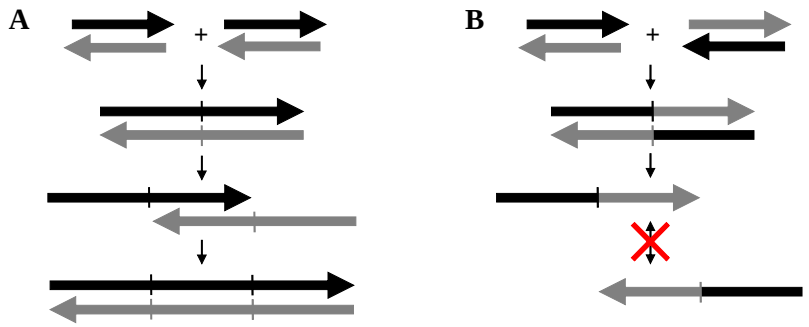

Fig. 15. Cis (**A**) and Trans (**B**) species of *ID* putatively generated in MPR initiation, and their potency in the subsequent propagation. Black and grey regions are complementary to each other. Arrowheads denote 3'

The thermodynamic stabilities of second-rank structures of different *OS*s involved in *NC* formation vary depending on sequence hence they have different propensities for propagation. In MPR conducted with a mixture of different *OS*s, the final polymers would contain predominantly the repetitive motif corresponding to *OS* with the most stable *NC* structure. This prediction is confirmed by the finding (Saito *et al.*, 2007) that the most abundant among 3 microgenic units in the combinatorial polymers generated from an equi-molar mixture is the one with the lowest free energy of formation of the *NC*-promoting 2nd-rank structure (analysis not shown), thus lending strong support to the model proposed and

described here. This feature can be exploited to design special types of proteins with desired repeated motifs *in vitro* and perhaps *in vivo* as well.

11. MPR and *in vivo* expansion

Repeating motifs in natural proteins are widely documented and described (http://www.ncl-india.org/trips/), and their possible activity mechanisms discussed (Katti et al., 2000). The fact that they derive of encoding genes is sometimes not sufficiently considered. During studies with the entomopathogenic *Bacillus thuringiensis* subspecies, we encountered several such cases, where DNA repeats exist, reflecting their encoded protein motifs. Specifically, helper Orf2 proteins of some *cry* operons include different numbers of repetitive motifs. For example, $Orf2_{cry2Aa}$, encoded by *orf*2 of the *cry2Aa* operon (Widner & Whiteley, 1989), contains 11 tandem repeats of a 15-amino-acids motif and seems to be involved in Cry2Aa crystallization (Ge et al., 1998). This motif is encoded by a 45-nt repeat in its gene $orf2_{cry2Aa}$ (Fig. 16A) that includes only 6 transitions and a single transversion among the 450 nt composing the first 10 repeats. All 5 additional changes are located at the 3' end of the last (11th) repeat. $orf2_{cry2Aa3}$ (Sasaki et al., 1997; Fig. 16B) includes 13 repeats with the same basic unit as $orf2_{cry2Aa}$. In this gene, the first 9.5 repeats are separated from the rest by a 12 nt long 'linker', i.e., a non-homologous insertion.
A second type of repeat unit among the known *orf*2 genes is found in $orf2_{cry2Ca}$ (Wu et al., 1991) and in $orf2_{cry2Ac}$ (Accession # AY007687). It differs from the repeat of $orf2_{cry2Aa}$ by two triplets that surround a conserved T(A/G)C found in both kinds of repeats and by an extra adjacent triplet. A string of 991 nt in $orf2_{cry2Ca}$ encoding almost 21 repeats of a 16-mer motif in Cry2Ca contains only 35 transitions (Fig. 16C), whereas the shorter (7 repeats) string in $orf2_{cry2Ac}$ is much less conserved and contains many deletions and insertions (Fig. 16D).

A. $orf2_{cry2Aa}$ (Access. # M23723)

```
ACG TAT AAT CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAT CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAT CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAT CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TAC ACA CAA GAT TTA ATT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC CCA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC CCA CAA GAT TTG AAT GTA
```

B. $orf2_{cry2Aa3}$ (Access. # D86064)

```
ACG TAT AAC CAA AGT CAG AAT GTA TGC GCG CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC ACA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC ACA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC ACA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC ACA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC GCG CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC ACA CAA GAT TTA GAA GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC ACA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC GCG CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AATGGTGTCTACAAG CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC GCG CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT GTA TGC ACA CAA GAT TTA GTT GAT
ACG TAT AAC CAA AGT CAG AAT ACT TAT ATA CAA GAT TTA ATT GAT
AAG TAT AAT CAA AGT CAG AAT
```

C. *orf2*$_{cry2Ca}$ Access. # X57252)

```
ACG TAT AAC CAA AGC CAA AATTGGCTG TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAC CAG AGC CAA AAT GGT GCC TGC GCG CAA GAT TTA ATG GAT
ACG TAT AAC CAA AGC CAA AAT GGT GCC TAC GCG CAA GAT TTA GTG AAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TGC GCG CAA GAT TTA ATG GAT
ACG TAT AAC CAA AGC CAA AAT GGT GCC TAC GCG CAA GAT TTA GTG AAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TGC GCG CAA GAT TTA ATG GAT
ACG TAT AAC CAA AGC CAA AAT GGT GCC TAC GCG CAA GAT TTA GTG AAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG AAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG AAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG AAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA GTG GAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TAC GCG CAA AAT TTA GTG GAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TGC GTG CAA GAT TTA GTG AAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TGC GTG CAA GAT TTA GTG GAT
ACG TAT AAC CAA AGT CAG AAT GGT GCC TAC G
```

D. *orf2*$_{cry2Ac}$(Access. # AY007687)

```
                               GT GCCTTAC GCG CAAAGAT TTA GTG GAAT
ACGGAAT AACCCAAAAGT CAG AAT GGT GCCNTAC GCG CAANGAT TTA AGT GAAT
ACG TATTAANCCAA AGT CAG AAT GGT GCC TAC GCG CAA GAT TTA AGT GNAT
ACGGTATAATC CAA AG  CAG AAT GGG CCC TAC GCG CAA AAT TTA GTG GAT
ACG TAT AAT CAA AGT CAG AAT GGT GCC TGC GTG CAA GAT TTA GTG AAT
ACG TAT AAT CAA AG  CAG AAT GGT GCC TGC GTG CAA GAT TTA GTG GAT
ACG TAT AAC CAA AGT CAG AAT GGT
```

E. *cry11Bb2* (Access. # HM068615)

```
            CAATAAT ACA AGC AGT GGG TAT GAG CAA GGA TAT AAC GAT
AAT TAT AAC CAA AAT ACA AGT AGT GGG TAT GAG CAA GGA TAT AAC GAT
AAT TAT AAC CAA AAT ACA AGT AGT GGG TAT GAG CAA GGA TAT AAC GAT
AAT TAT AAC CAA AAT ACA AGT AGT GGA TAT GAG CAA GGA TAT AAC GAT
AAT TAT AAC CAA AAT ACA AGC AGT GAG TAT GAG CAA GGA TAT AAC GAC
AAT TAT AAC CAA AAT ACA AGT AGT GGA TAT GAG CAA GGA TAT AAC GAT
AAT TAT AAC CAA AAT ACA AGT AGT GGA TAC GAG CAA GGA TAT ATT GAT
AAT TAT AGG CCA
```

Fig. 16. Individually aligned repeated motifs found in different *B. thuringiensis* (Bt) subspecies. **A**. *orf2*$_{cry2Aa}$ from Bt subsp. *kurstaki* HD-1; **B**. *orf2*$_{cry2Aa3}$ from Bt subsp. *sotto*. **C**. *orf2*$_{cry2Ca}$ from Bt S_1; **D.** *orf2*$_{cry2Ac}$ from Bt; **E**. *cry11Bb2* from Bt strain K34. Fonts: red, pyrimidine transitions; blue, purine transitions; underlined, transversions. Highlights: green, the conserved triplet in the middle of the motif; yellow, region of varied length (modulus 3); grey, sequence homology between *cry11Bb2* and *orf2*; red, non-homologous insertions

Repetitive motifs are encountered in certain *cry* genes themselves. The newly discovered Cry11Bb2 (Melnikov *et al.*, 2010) contains 7 tandem repeats of a 16-amino acids motif, the role of which is still to be determined. Seven out 16 triplets constituting the repetitive motif of Cry11Bb2 (highlighted grey in Fig. 16E) are identical/homologous in composition and location to those found in previously described *orf2* genes (Fig. 16, A-D). Repeated blocks have been found in other mosquito larvicidal δ-endotoxins (de Maagd et al., 2003): Cry11Ba1 (four repeats), Cry11Bb1 (5 repeats), Cry20Aa (eight), Cry27Aa (8 repeats).

The spectrum of replacements in the repeats of $orf2_{cry2Aa}$ and $orf2_{cry2Ca}$ is reciprocal on two sides of the conserved T(A/G)C triplet: in its 5′ part there is a bias to pyrimidine transitions whereas purine transitions are exclusively observed at its 3′ part. Intriguingly, such pattern of changes may point to occurrence of double strand break (DSB) in the middle of the repeat. Repair of DSB is mediated by trans-resection of opposite strands surrounding it in a way that transiently exposes generated ssDNA stretches to increased mutagenesis with characteristic pattern of transitions encompassing the DSB (Yang, 2008). Moreover, much evidences has accumulated in eukaryotic cells that point to blunt dsDNA ends as key intermediates in the process leading to gene amplification (Messer & Arndt, 2007; Pace et al., 2009; Mondello et al., 2010).

To survive DSB, cells exploit two major processes: homologous recombination (HR) and non-homologous end-joining (NHEJ). HR retrieves lost genetic information in error-free way from undamaged homologous template using Rad50/Mre11 complex in yeasts (Wyman & Kanaar, 2006) or RecBCD in *Escherichia coli* (Dillingham & Kowalczykowski, 2008). In contrast, NHEJ break repair is a ligation-like non-template-directed process that occurs between the non-homologous termini of a DSB and is therefore considered to be more error-prone. In the course of NHEJ, a DSB is repaired by attracting the Ku70/80 hetero-dimer, which recruits the ligase IV complex (comprised of ligase IV, XRCC4 and XLF) to seal the DNA ends (Mahaney et al., 2009). NHEJ, regarded as exclusive prerogative of eukaryotic cells, has recently been found in prokaryotes by identifying bacterial homologues of Ku protein (Doherti et al., 2001; Brissett & Doherty, 2009). Moreover, a functional NHEJ repair pathway is essential for spore viability in *Bacillus subtilis* under conditions that yield DSBs (Weller et al., 2002; Wang et al., 2006; Moeller et al., 2007). It is noteworthy that many of the bacteria that contain the Ku ligase system are capable of sporulation (*B. subtilis*, *Streptomyces coelicolor*) or spend long periods of their life cycle in the stationary phase (*Mycobacterium tuberculosis*, *Mesorhizobium loti*, *Sinorhizobium loti*) (Weller et al., 2002). The sporulating *B. thuringiensis* may contain NHEJ repair pathway as well.

Exogenous oligonucleotides complementary to the broken ends can efficiently target DSB for repair in yeasts (Storici et al., 2006). We propose that partial complementarity between sequences flanking the broken ends may assist their sealing and contribute to repeat generation via MPR-like slipped structures (Fig. 17, panels A-H). An example involving putative 2nd rank complementarity that can participate in the process to seal the broken ends (Fig. 17I) was revealed in Cry2Bb2 (Melnikov et al., 2010).

Once the initial doublet has been generated inside the chromosome, the repeats may propagate during DNA replication since the constituent strands may slide over each other between the multiple complementary regions. Second rank complementary structures similar to WSD in MPR (Fig. 12B) may stabilize such slipped structures. The alternative mechanism for the propagation is possible recurrence of DSBs in the middle of the generated repeats. Bridging the broken ends by occasional Watson-Crick bonds due to partial complementarity (Fig. 17D) can bring about nucleotides insertions/deletions during fill-in by DNA polymerase similar to those occurring during conversion of the *NC* to the *ID* in MPR (Fig. 6 and Itsko *et al*., 2008a). Selection for functionality of a given coding region can restore the lost frame by generating an additional triplet or dropping an existing one. This can result in the two/three triplets varied region (highlighted yellow in Fig. 16) surrounding the conserved T(A/G)C triplet in the 4 *orf2* versions and in *cry11Bb2* listed here (highlighted green in Fig. 16). The inserted linker in that region as in $orf2_{cry2Aa3}$ (highlighted red in Fig. 16B) is also consistent with this explanation.

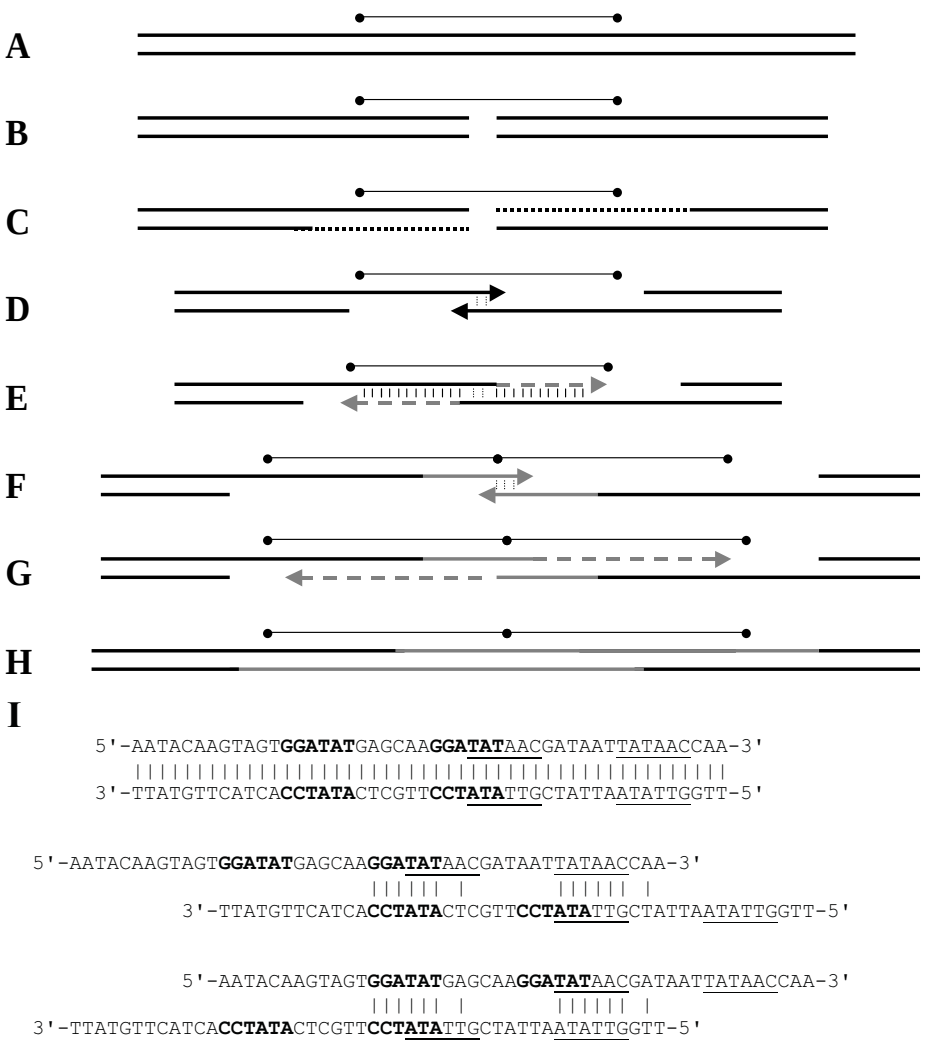

Fig. 17. Generation of sequence duplication due to DSB repair via NHEJ-like repair pathway. Chromosome region (**A**) suffers a DSB (**B**). **C**. Trans resection of DNA strands flanking the break by a dedicated exonuclease. **D**. Bridging the broken ends in staggered mode by occasional Watson-Crick bonds due to partial complementarity. **E**. Filling-in single stranded regions by DNA polymerase. **F**. Back slippage of the broken ends under repair that is mediated by a weak 2nd rank complementarity. **G**. Filling-in single stranded regions by DNA polymerase. **H**. Restored integrity of the chromosome with the generated duplicated region. Dumbbells designate schematic location of the sequence undergoing duplication. Newly synthesized DNA is designated in grey. **I**. Fully-aligned and two slipped structures existing between constituent strands of *cry2Bb2* repeat motif that may assist to bridge the DSB ends. Bold and underlined regions designate two sets of complementarity

It is intriguing that the AGT triplet in $orf2_{cry2Ca}$ (Fig. 16C) is always adjacent to CAG whereas its 5'- replaced AGC is adjacent to CAA, both combinations encode the same amino acids, Serine by AGT/C and Glutamine by CAG/A. This coincidence cannot be explained by selection for suppression of one transition by the other. Such replacements do affect stability of slipped structures presumably emerging during replication of this region. Without the replacements, almost triplet repetition AAT CAA AGT CAA AAT (Fig. 16C) would be explicitly observed that could putatively propagate in uncontrolled way through slipped structures, bringing about genetic instability. Strong selection for silent transitions that save triplet coding but prevent deleterious expansion by disrupting repetition pattern would thus be anticipated to stabilize the genome.

12. Concluding remarks

The MPR was dissected to sub-reactions and their thermodynamics and kinetics were analyzed. Different propensities of various HDs to expand into multiple repeating units were justified in terms of different stabilities of *NC*s engaged in MPR initiation. The proposed models, thermodynamic for initiation and kinetic for propagation, agree satisfactorily with experimental results.

The MPR with non-repetitive HD presents an optional chemical evolutionary system in which the thermodynamic advantage of very weak interactions results in biased proliferation of a certain reaction product. The learned approach to optimize MPR is necessary in protein engineering. The importance of studying this phenomenon lies far beyond applied interest; it may reflect primordial molecular evolution of primitive DNA sequences into complex genomes.

Molecular participants in the repeat expansion process and reaction conditions accompanying it *in vivo* obviously differ from those existing in MPR mixtures. The facilitating factor in the *in vitro* repeat expansion, temperatures manipulations above the T_m of expanding oligonucleotide *HD*, is impossible *in vivo*, where an ensemble of various enzymes is operating. DNA/protein complexes can select for generation of specific functionality-relevant hybridized species of DNA molecules out of competing structures. This principal difference affects only the rate of the processes and their efficiencies but not their thermodynamic feasibilities. Even strongly unfavorable energetic interactions (high ΔH values) between DNA species (putatively bringing to repeat generation and their propagation) can occur in both systems when overcome by entropic component of their multiplicity and redundancy ($T\Delta S$) (Harvey, 1997). Thus, the basic thermodynamic consideration about stabilities of 2nd rank structures underlying repeat propagation propensity can be common to the repeat expansion phenomenon in both arenas.

13. Acknowledgments

This investigation was partially supported by a grant from the United States-Israel Binational Science Foundation (BSF, number 2007-037), Jerusalem, Israel (to A.Z.), and a Levi Eshkol scholarship (to M.I.) from the Israeli Ministry of Science, Culture and Sports.

14. References

Atkins, P.W. (1994). *Physical Chemistry*. 5th ed., Oxford University Press, ISBN 0198557329, Oxford, UK, pp. 861–959.

Blackburn, E.H. (1991). Structure and function of telomeres. *Nature*, 350, 569-573.

Brissett, N.C. & Doherty, A.J. (2009). Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway. *Biochem Soc Trans*, 37, 539-545.

Britten, R.J. & Kohne, D.E. (1968). Repeated sequences in DNA. *Science*, 161, 529-540.

Catasti, P.; Gupta, G.; Garcia, A.E.; Ratliff, R.; Hong, L.; Yau, P.; Moyzis, R.K. & Bradbury, E.M. (1994). Unusual structures of the tandem repetitive DNA sequences located at human centromeres. *Biochemistry*, 33, 3819-3830.

de Maagd, R.A.; Bravo, A.; Berry, C.; Crickmore, N. & Schnepf, E. (2003). Structure, diversity and evolution of protein toxin from spore-forming entopathogenic bacteria. *Annu Rev Genet,* 37, 409–433.

Dillingham, M.S. & Kowalczykowski, S.C. (2008). RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol Mol Biol Rev*, 72, 642-671

Dogget, N.A. (2000). Overview of human repetitive DNA Sequences. *Current Protocols in Human Genetics*, APPENDIX 1B, DOI: 10.1002/0471142905.hga01bs08

Doherty, A.J.; Jackson, S.P. & Weller, G.R. (2001). Identification of bacterial homologues of the Ku DNA repair proteins. *FEBS Lett*, 500, 186-188.

Flory, P.J. (1953). *Principles of Polymer Chemistry*. Cornell University Press, ISBN 0801401348, Ithaca, New York.

Ge, B.; Bideshi, D.K.; Moar, W.J. & Federici B.A. (1998). Differential effects of helper proteins encoded by the *cry2A* and *cry11A* operons on the formation of Cry2A inclusions in *Bacillus thuringiensis*. *FEMS Microbiol Lett*, 165, 35-41.

Gilson, E.; Clément, J.M.; Brutlag, D. & Hofnung, M.A. (1984). Family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. *EMBO J*, 3,1417-1421.

Harvey, S.C. (1997). Slipped structures in DNA triplet repeat sequences: entropic contributions to genetic instabilities. *Biochemistry*, 36, 3047-3049.

Hofnung, M. & Shapiro, J. (1999). *Res Microbiol* 150 (special issue on bacterial repeats).

International Human Genome Consortium (2001). Initial sequencing and analysis of the human genom. *Nature*, 409, 860-921.

Itsko, M.; Zaritsky, A.; Rabinovitch, A. & Ben-Dov E. (2008a). Initiation of the microgene polymerization reaction with non-repetitive homo-duplexes. *Biochem Biophys Res Commun*, 368, 606-613.

Itsko, M.; Zaritsky, A. & Rabinovitch A. (2008b). Thermodynamics of unstable DNA structures from the kinetics of the microgene PCR. *J Phys Chem B*, 112, 13149-13156.

Itsko, M.; Rabinovitch. A. & Zaritsky, A. (2009). Kinetics of repeat propagation in the microgene polymerization reaction. *Biophys J*., 96, 1866-1874.

Kang, S.; Jaworski, A.; Ohshima, K. & Wells, R.D. (1995). Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli*. *Nat Genet*, 10, 213-218.

Katti, M.V.; Sami-Subbu, R.; Ranjekar, P. K. & Gupta, V.S. (2000). Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci,* 9, 1203-1209.

Kong, H.; Kucera, R.B. & Jack, W.E. (1993). Characterization of a DNA polymerase from the hyperthermophile archaea *Thermococcus litoralis*. Vent DNA polymerase, steady state kinetics, thermal stability, processivity, strand displacement, and exonuclease activities, *J Biol Chem*, 268, 1965–1975.

Liang, X.; Jensen, K. & Frank-Kamenetskii, M.D. (2004). Very efficient template/primer-independent DNA synthesis by thermophilic DNA polymerase in the presence of a thermophilic restriction endonuclease. *Biochemistry*, 43, 13459-13466.

Mahaney, B.L.; Meek. K. & Lees-Miller, S.P. (2009). Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. *Biochem J*, 417, 639-50.

Melnikov, O.; Baranes, N.; Einav, M.; Ben-Dov, E.; Manasherob, R.; Itsko, M. & Zaritsky, A. (2010). Tandem repeats in a new toxin gene from a field Isolate of *Bacillus thuringiensis* and in other *cry11*-like genes. submitted

Messer, P.W. & Arndt, P.F. (2007).The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol*, 24, 1190-1197.

Meyers, B.C.; Tingey, S.V. & Morgante, M. (2001). Abundance, distributionand transcriptional activity of repetitive elements in the maize genome. *Genome Res*, 11, 1660-1676.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease *Nature*, 447, 932-940.

Moeller, R.; Stackebrandt, E.; Reitz, G.; Berger, T.; Rettberg, P.; Doherty, A.J.; Horneck, G. & Nicholson, W.L. (2007). Role of DNA repair by non-homologous end joining (NHEJ) in *Bacillus subtilis* spore resistance to extreme dryness, mono- and polychromatic UV and ionizing radiation. *J Bacteriol*, 189, 3306–3311.

Mondello, C.; Smirnova, A. & Giulotto, E. (2010). Gene amplification, radiation sensitivity and DNA double-strand breaks. *Mutation Res*, 704, 29–37.

Ogata, N. & Miura, T. (2000). Elongation of tandem repetitive DNA by the DNA polymerase of the hyperthermophilic archaeon *Thermococcus litoralis* at a hairpin-coil transitional state: a model of amplification of a primordial simple DNA sequence. *Biochemistry*, 39, 13993-14001.

Ogata, N. & Morino, H. (2000). Elongation of repetitive DNA by DNA polymerase from a hyperthermophilic bacterium *Thermus thermophilus*. *Nucl Acids Res*, 28, 3999-4004.

Ohno, S. (1984). Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proc Natl Acad Sci USA*, 81, 2421-2425.

Ohno, S. (1987). Evolution from Primordial Oligomeric Repeats to Modern Coding Sequences. *J Mol Evol*, 25, 325-329.

Orgel, L.E. & Crick, F.H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284, 604-607.

Pace, J.K. II; Sen, S.K.; Batzer, M.A. & Feschotte, C. (2009). Repair-mediated duplication by capture of proximal chromosomal DNA has shaped vertebrate genome evolution. *PLoS Genet*, 5(5):e1000469.

Parkhill, J.; Achtman, M.; James, K.D.; Bentley, S.D.; Churcher, C.; Klee, SR.; Morelli, G; Basham, D.; Brown, D.; Chillingworth, T.; Davies, R.M.; Davis, P.; Devlin, K.; Feltwell, T., et al. (2000). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*, 404, 502-506.

Ruland, K.; Wenzel, R. & Herrmann R. (1990). Analysis of three different repeated DNA elements present in the P1 operon of *Micoplasma pneumonia*: size, number and distribution on the genome. *Nucl Acids Res*, 19, 637-647.

Saito, H.; Minamisawa, T. & Shiba, K. (2007). Motif programming: a microgene-based method for creating synthetic proteins containing multiple functional motifs. *Nucleic Acids Res*, 35(6):e38, doi:10.1093/nar/gkm017.

Sasaki, J.; Asano, S.; Hashimoto, N.; Lay, B.W.; Hastowo, S.; Bando, H. & Iizuka T. (1997). Characterization of a cry2A gene cloned from an isolate of *Bacillus thuringiensis* serovar *sotto*. *Curr Microbiol*, 35: 1-8

Shapiro, J.A. & von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biol Rev*, 80, 227-250.

Shiba, K.; Takahashi, Y. & Noda T. (1997). Creation of libraries with long ORFs by polymerization of a microgene. *Proc Natl Acad Sci USA*, 94, 3805-3810.

Storici, F.; Snipe, J.R.; Chan, G.K.; Gordenin, D.A. & Resnick, M.A. (2006). Conservative repair of a chromosomal double-strand break by single-strand DNA through two steps of annealing. *Mol Cell Biol*, 26, 7645-7657.

Tuntiwechapikul, W. & Salazar, M. (2002). Mechanism of in vitro expansion of long DNA repeats: effect of temperature, repeat length, repeat sequence, and DNA polymerases. *Biochemistry*, 41, 854-860.

Ullu, E. & Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature*, 312, 171-172.

Wang, S.T.; Setlow, B.; Conlon, E.M.; Lyon, J.L.; Imamura, D.; Sato, T.; Setlow, P.; Losick, R. & Eichenberger, P. (2006). The forespore line of gene expression in *Bacillus subtilis*. *J Mol Biol*, 358, 16–37.

Weller, G.R.; Kysela, B.; Roy, R.; Tonkin, L.M.; Scanlan, E.; Della, M.; Devine, S.K.; Day, J.P.; Wilkinson, A.; di Fagagna, F.D. *et al*. (2002). Identification of a DNA non homologous end-joining complex in bacteria. *Science*, 297, 1686–1689.

Wells, R.D. (1996). Molecular basis of genetic instability of triplet repeats. *J Biol Chem*, 271, 2875-2878.

Widner, W.R. & Whiteley. H.R. (1989). Two highly related insecticidal crystal proteins of *Bacillus thuringiensis* subsp. *kurstaki* possess different host range specificities. *J Bacteriol*, 171, 965-974.

Wu, X.; Cao, X.L.; Bai, Y.Y. & Aronson, A.I. (1991). Sequence of an operon containing a novel N-endotoxin gene from *Bacillus thuringiensis*. *FEMS Microbiol Lett*, 81, 31-36.

Wyman, C. & Kanaar, R. (2006). DNA double-strand break repair: all's well that ends well. *Annu Rev Genet*, 40, 363-83.

Yang, Y.; Sterling, J.; Storici, F.; Resnick, M.A. & Gordenin, D.A. (2008). Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast *Saccharomyces cerevisiae*. *PLoS Genet*, 4(11):e1000264.