February 2005

A macroscopic kinetic model for DNA polymerase elongation and high-fidelity nucleotide election

Steve Viljoen
Department of Biochemistry, Nebraska Wesleyan University, Lincoln, NE 68540, USA

Mark A. Griep
University of Nebraska-Lincoln, mgriep1@unl.edu

Michael Nelson
Megabase Research Products, Lincoln, NE 68540, USA

Hendrik J. Viljoen
Department of Chemical Engineering, *University of Nebraska-Lincoln*, hviljoen1@unlnotes.unl.edu

Computational Biology and Chemistry 29 (2005) 101–110, © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.compbiolchem.2005.02.003.

A macroscopic kinetic model for DNA polymerase elongation and high-fidelity nucleotide election

Steve Viljoen , Mark A. Griep , Michael Nelson , Hendrik Viljoen

Department of Biochemistry, Nebraska Wesleyan University, Lincoln, NE 68540, USA
Department of Chemistry, University of Nebraska, Lincoln, NE 68588, USA
Megabase Research Products, Lincoln, NE 68540, USA
Department of Chemical Engineering, University of Nebraska, Lincoln, NE 68588, USA

Abstract

The enzymatically catalyzed template-directed extension of ssDNA/primer complex is an important reaction of extraordinary complexity. The DNA polymerase does not merely facilitate the insertion of dNMP, but it also performs rapid screening of substrates to ensure a high degree of fidelity. Several kinetic studies have determined rate constants and equilibrium constants for the elementary steps that make up the overall pathway. The information is used to develop a macroscopic kinetic model, using an approach described by Ninio [Ninio J., 1987. Alternative to the steady-state method: derivation of reaction rates from first-passage times and pathway probabilities. Proc. Natl. Acad. Sci. U.S.A. 84, 663–667]. The principle idea of the Ninio approach is to track a single template/primer complex over time and to identify the expected
behavior. The average time to insert a single nucleotide is a weighted sum of several terms, including the actual time to insert a nucleotide plus delays due to polymerase detachment from either the ternary (template-primer-polymerase) or quaternary (+nucleotide) complexes and time

delays associated with the identification and ultimate rejection of an incorrect nucleotide from the binding site. The passage times of all events and their probability of occurrence are expressed in terms of the rate constants of the elementary steps of the reaction pathway. The model accounts for variations in the average insertion time with different nucleotides as well as the influence of G+C content of the sequence in the vicinity of the insertion site. Furthermore the model provides estimates of error frequencies. If nucleotide extension is recognized as a competition between successful insertions and time delaying events, it can be described as a binomial process with a probability distribution. The distribution gives the probability to extend a primer/template complex with a certain number of base pairs and in general it maps annealed complexes into extension products.

Keywords: Biochemical engineering; Molecular biology; Polymerase chain reaction; Mathematical model

1. Introduction

Template-directed nucleic acid synthesis is one of the greatest biochemical discoveries. Since the activity of the first RNA polymerase and DNA polymerase were studied in 1955 (Grundberg-Manago et al., 1955; Lehman, 2003), template-directed polymerases have contributed to our understanding of many subcellular processes, such DNA replication and repair, transcription and telomere homeostasis. The accuracy, relative ability to bypass lesions and ability to elongate polymers processively makes these enzymes crucial to understanding such biological processes as infection, cancer, and aging and also identifies them as important drug targets (Ahmed and Tollefsbol, 2003; Miura and Izuta, 2004). Finally, template-directed nucleic acid polymerases are used in many procedures basic to modern biotechnology including PCR and DNA sequencing (Smith, 1980; Saiki et al., 1985). The microscopic kinetic mechanisms and structure/ function relationships of *Escherichia coli* DNA polymerase I, T7 DNA polymerase, KlenTaq DNA polymerase and a *Bacillus* DNA polymerase have been worked out in some detail (Johnson, 1993; Kiefer et al., 1998; Li and Waksman, 2001). Comparative analysis suggests

that all DNA polymerases will have the same broadly conserved molecular properties and mechanisms, but that each polymerase will have its unique set of microscopic rates and equilibria (Jager and Pata, 1999; Joyce and Benkovic, 2004; Kunkel, 2004). The mechanisms of nucleotide insertion and incorporation are probably highly conserved although the details of the nucleotide selection process are just eginning to be understood. Despite these advances on the molecular level, we still lack an understanding of the relationship of these parameters to macroscopic properties such as chromosomal elongation rates. This lack of theoretical understanding has practical consequences. For instance, it is one of the obstacles that prevent users from selecting the best DNA polymerase for a given PCR application. Currently researchers choose a polymerase, work out the PCR conditions semi-empirically and then adjust if necessary. Our poor theoretical understanding is also an obstacle for those researchers trying to solve the many PCR problem areas such as the amplification of GCrich sequences, amplification of long sequences, maintaining low error frequencies, etc. There is a commercial impact in that many thermophilic DNA polymerases are available (Vieille and Zeikus, 2001) with various known biochemical properties (elongation rate, processivity, nucleotide selectivity, thermal half-life at 95 ◦C etc.) but it is not clear which properties are best for which PCR protocol. To bridge the gap between the microscopic and macroscopic in the realm of DNA polymerases, we have developed a model for DNA synthesis that accounts for the template nucleotide sequence and dNTP pool to predict DNA synthesis rates and error frequencies.

2. Mathematical model

The objective is to derive an expression for the extension rate of the polymerase. An outcome of the model is that it becomes possible to predict extension rate, fidelity, processivity and product yield in polymerase chain reactions. Whereas a microscopic model would focus on a complex sequence of steps and the details of the mechanism, the macroscopic model condenses the information into a few key steps. Although a reduction in complexity always ccompanies a reduction in detail knowledge, it is also true that if the rate-limiting step can be identified, then further data become largely superfluous. In an effort to formulate a macro-

scopic model that describes the average dynamics of a polymerase/template complex, it is necessary to conjecture some sequence of steps, even though the exact mechanism of the polymerase is an ongoing study. Johnson (1993) proposed a kinetic scheme for DNA propagation by T7 polymerase and it is schematically presented in Fig. 1. The binary complex primer/template is denoted as *A*1. This complex binds a polymerase molecule to form a ternary complex *A*0. The dNTP pool comprises of dATP, dCTP, dGTP, dTTP and dUTP.

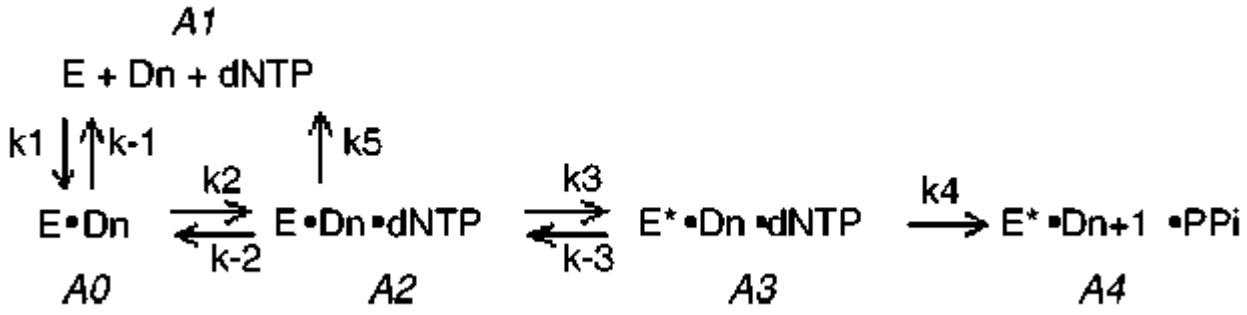

Fig. 1. Schematic of polymerase catalysed DNA elongation. *A*1 presents the primer/template complex, *A*0 is the ternary complex that includes the polymerase, *A*2 is the quaternary complex that forms after addition of a dNTP. The model accounts for polymerase detechment from states *A*0 and *A*2. The quaternary complex adopts an activated state *A*3 followed by phosphodiester bond synthesis, the release of pyrophosphate, and translocation of the enzyme. The schematic also identifies the notation to label the different states and the respective rate constants.

Uracil is formed by thermal deamination of dCTP, thus it will be present. Another component that is not usually considered part of the dNTP pool, but it competes for binding to the ternary complex is PP*i*. Pyrophosphate is a product of the extension reaction and it is responsible for the pyrophosphorolysis reaction that reverses extension and produces dNTP. After one of these components have bound to the ternary complex to form the quaternary complex *A*2, the polymerase may undergo a conformational change denoted as *A*3. The final step to *A*4 involves the release of PP*i*. Patel et al. (1991) noted that the rate limiting step is the formation of the activated complex *A*3 and the reaction following that is in equilibrium, because it is so much faster. The state *A*4 is a ternary complex similar to *A*0, but the number of base pairs has increased by

one. The scheme differs from the one proposed by Johnson because a reaction has been added to account for the detachment of the polymerase from the quaternary state, i.e. $A2 \rightarrow A1$. Due to the paucity of kinetic data, it is reasonable to assume that $k5$ and $k-1$ would be similar. The derivation of the kinetic rate expression that is presented in the following section differs from the conventional approach. There is a striking similarity between these two approaches and the two approaches that are used to derive the equilibrium energy distribution of an ensemble by statistical thermodynamics. It is well known that the same energy distribution is obtained either by sampling a large number of molecules at the same time or tracking a single molecule along a timeline to determine occupation periods of different energy levels. In kinetics, the classical approach to determine the rate of an elementary reaction is to sample the ensemble of reacting molecules over a short interval $\_t$ and find the number that has reacted. In the spirit of the ergodic principle, the alternative approach is to track a molecule over time and average the conversion times.

2.1. Reaction rates based on pathway probabilities

Ninio (1987) proposed an alternative approach to derive kinetic expressions for enzymatic reactions. For conventional reaction schemes, there is no benefit in this alternative approach. However, it is much better suited to handle the complexity of the nucleotide selection and editing.
The principle idea of the Ninio approach is to track a single enzyme/template complex over time and then determine its average behavior. Before we continue with the formal derivation, we demonstrate the approach for the classical Michaelis- Menten reaction scheme.

2.1.1. Example: Michaelis-Menten kinetics for enzyme-catalyzed reactions

The Michaelis-Menten scheme for enzymatic catalysts (Michaelis and Menten, 1913; Briggs and Haldane, 1925) is as follows:

$$E + S \underset{k_{-1}}{\overset{k_1}{\Leftrightarrow}} E \cdot S \xrightarrow{k_2} E + P,$$

but we prefer to use a more succinct notation that describes the enzyme's state:

$$A1 \underset{k_{-1}}{\overset{k_1}{\Leftrightarrow}} A0 \overset{k_2}{\rightarrow} A2.$$

Following the reaction pathway from left to right, the enzyme changes from free to bound and back to the free. The average time to go from state $A1$ to state $A0$ is

$$t_1 = \frac{1}{k_1[S]}$$

and the average time to go from state $A0$ to either state $A1$ or state $A2$ is

$$t_2 = \frac{1}{k_{-1} + k_2}.$$

The passage time $t = t1 + t2$ is the average time that elapses from the moment an enzyme leaves the free state until it enters the free state again. The rate of reaction is the probability to go from $A0$ to $A2$ (not to $A1$), divided by the passage time. Since the probability for the successful forward reaction ($A0 \rightarrow A2$) is

$$P_S = \frac{k_2}{k_{-1} + k_2},$$

the production rate of one enzyme molecule is $v1$ =

$$v_1 = \frac{P_S}{\langle t \rangle} = \frac{k_1 k_2 [S]}{k_1[S] + k_2 + k_{-1}}. \qquad (1)$$

If the total enzyme concentration is $[E]$, then the macroscopic rate expression is

$$v = \frac{k_1 k_2 [S][E]}{k_1[S] + k_2 + k_{-1}}. \qquad (2)$$

Still within the framework of Ninio's approach, there is an alternative interpretation that sheds more light on the dynamics of the enzyme. If an enzyme were to go from $A1$ to $A0$ and then back to $A1$, it would have completed a loop without producing any product. Let D denote this loop. Alternatively, the enzyme could follow the path $A1$ to $A0$ to $A2$ that also constitutes a loop, S. Both loops take the same time to complete and it corresponds to the passage time;

$$t = t1 + t2.$$

The probabilities for the events to occur are

$$P_{\mathrm{S}} = 1 - P_{\mathrm{D}} = \frac{k_2}{k_{-1} + k_2}.$$

Therefore, a fortuitous situation may arise where an enzyme could produce a product on its first attempt, i.e S and it will take time t. In a slightly less fortuitous case, the enzyme may experience one failed event, followed by a successful one, i.e. DS and it will take 2 t. The pattern is now clear. The set of all possible outcomes that end with a successful event is:

$$\{\mathbf{S, DS, D^2S, \ldots D^nS}\}$$

where $n \to \infty$. The expected time (expectancy value) for an enzyme to produce product P (it cannot be lucky all the time) is

$$\begin{aligned}\tau &= P_S(t) + P_D P_S(2t) + \ldots P_D^n P_S((n+1)t) \\ &= P_S(t)\{1 + P_D + P_D^2 + \ldots P_D^n\} + P_S P_D(t)\{1 + P_D(t) \\ &\quad + P_D^2(2t) + \ldots P_D^n(nt)\} = \langle t\rangle + \frac{P_D}{P_S}\langle t\rangle \\ &= \frac{k_1[S] + k_2 + k_{-1}}{k_1 k_2[S]} = \frac{1}{v_1}. \qquad \text{(3a–c)}\end{aligned}$$

Thus, τ is the average time it takes an enzyme to produce a product. The inverse of τ is its corresponding production rate $v1$. Since every enzyme may be expected to perform in this fashion, the macroscopic rate is $v = [E]v1$ and the result corresponds to Eq. (2). This interpretation helps us to understand the time-wasting loops that an enzyme may perform before it produces a reaction product.

2.1.2. Derivation of reaction rate of multi-substrate enzyme-catalyzed DNA extension

Consider the reaction scheme in Fig. 1. Starting at the left of the reaction scheme, the binary complex (formed during the annealing step) binds the polymerase with rate $k1[E]$ to form a ternary complex, identified as state $A0$. The polymerase concentration is denoted as $[E]$. The polymerase may detach from the ternary complex at a rate $k-1$, or it may bind a nucleotide with rate $k2[X]$. *Remark*: The term $[X]$ presents the sum of all species that may attach to the ternary complex. Apart from the nucleotides, dUTP (which forms by thermal deamination of dCTP) and PP*i* (a product of extension) may compete for attachment. Note that in the absence of competitive binding by dUTP or PP*i*, $[X]$ = [dNTP]. Competitive binding by dUTP or PP*i* is reduced by an excess [dNTP], as well as the use of enzymes such as dUTPase and pyrophosphatase.

No selection of dNTP precedes this step. Therefore, the random nature of thermal motion in the solution assures that the probability of the nucleotide being of type *i* (where *i*=A, C, T and G) is determined by the molar fraction of each of those species. For example, the molar fraction of dATP is xA = [dATP]/$[X]$. Thus, if the correct nucleotide is A, the probability that a dATP has

bound to the ternary complex is xA and the probability that it is not dATP is $1-x$A. If the correct nucleotide is bound to the ternary complex, the kinetics favors insertion. If the incorrect nucleotide has been selected, insertion may still occur, but it is not kinetically favored. At this point we make a very important distinction. The forward rate constants $k3$ and $k4$ change to lower values $k3w$ and $k4w$ if a wrong nucleotide has been selected. Note however that the nucleotides still bind to the ternary complex with the same rate constant $k2$. To continue, if the ternary complex progresses from state $A0$ to state $A4$ it means that one nucleotide has been added, but several events may delay the progress: (1) the polymerase may detach when the complex is in state $A0$ or in state $A2$ (reactions $A0 \rightarrow A1$ and $A2 \rightarrow A1$ in Fig. 1), (2) the polymerase may reject the nucleotide in state $A2$ (even if it is correct) and the reverse reaction occurs $A2 \rightarrow A0$ or if it proceeds to the activated complex$A3$ (3) the nucleotide fails to insert and the activated complex returns to the quaternary complex $A3 \rightarrow A2$. Thus, the advancement of a single enzyme/primer/template complex must be viewed as a probabilistic affair that consists of a series of events. A successful event S starts at $A0$ and ends at $A4$, but it may idle between states $A2$ and $A3$. Therefore, S can be presented as follows:

$$A0 \rightarrow (A2 \Leftrightarrow A3)^n \rightarrow A4.$$

The probabilities to proceed from $A0$ to $A2$, $A2$ to $A3$, $A3$ to $A2$ and $A3$ to $A4$ are:

$$P_{02} = \frac{k_2[X]}{k_2[X] + k_{-1}}; P_{23} = \frac{k_3}{k_3 + k_{-2} + k_{-1}};$$

$$P_{32} = \frac{k_{-3}}{k_4 + k_{-3}}; P_{34} = 1 - P_{32} = \frac{k_4}{k_4 + k_{-3}}.$$

If an incorrect nucleotide has attached to the ternary complex, P_{23}, P_{32} and P_{34} are replaced by:

$$P_{23w} = \frac{k_{3w}}{k_{3w} + k_{-2} + k_{-1}}; P_{32w} = \frac{k_{-3}}{k_{4w} + k_{-3}};$$

$$P_{34w} = 1 - P_{32w} = \frac{k_{4w}}{k_{4w} + k_{-3}}.$$

The probability to insert a correct nucleotide accounts for the idling:

$$P_S = P_{02} P_{23} P_{34} + P_{02}(P_{23} P_{32}) P_{23} P_{34}$$
$$+ \ldots P_{02}(P_{23} P_{32})^n P_{23} P_{34}, n \to \infty = \frac{P_{02} P_{23} P_{34}}{1 - P_{23} P_{32}} \qquad (4)$$

If $P23$, $P32$ and $P34$ in Eq. (4) are replaced by $P23w$, $P32w$ and $P34w$ the probability to insert an incorrect nucleotide is:

If P_{23}, P_{32} and P_{34} in Eq. (4) are replaced by P_{23w}, P_{32w} and P_{34w} the probability to insert an incorrect nucleotide is:

$$P_{\mathrm{S}w} = \frac{P_{02}\,P_{23w}\,P_{34w}}{1 - P_{23w}\,P_{32w}} \tag{5}$$

The second event is detachment (denoted as **D1**) from state $A2$. The same idling may occur as before:

$$A0 \rightarrow (A2 \Leftrightarrow A3)^n \rightarrow A1 \rightarrow A0.$$

The probability for event **D1** is

$$P_{\mathrm{D1}} = \frac{P_{02}\,P_{21}}{1 - P_{23}\,P_{32}}. \tag{6}$$

If the nucleotide is incorrect, the probability is

$$P_{\mathrm{D1}w} = \frac{P_{02}\,P_{21w}}{1 - P_{23w}\,P_{32w}} \text{ and } P_{21w} = \frac{k_{-1}}{k_{3w} + k_{-2} + k_{-1}}. \tag{7}$$

The third event (**F**) is rejection of the nucleotide in state $A2$ and return to state $A0$. Idling must be considered and the probabilities for correct and incorrect nucleotides become:

$$P_{\mathrm{F}} = \frac{P_{02}\,P_{20}}{1 - P_{23}\,P_{32}},\ P_{\mathrm{F}w} = \frac{P_{02}\,P_{20w}}{1 - P_{23w}\,P_{32w}}. \tag{8,9}$$

The last event (**D2**) is detachment from state $A0$ and the loop is $A0 \rightarrow A1 \rightarrow A0$. The probability for this event to occur is:

$$P_{\mathrm{D2}} = P_{01} = \frac{k_{-1}}{k_2[X] + k_{-1}}. \tag{10}$$

Associated with each probability is its characteristic time. Again we need to distinguish between correct and incorrect nucleotides. The times associated with P_{01}, P_{02}, P_{20}, P_{21}, P_{23}, P_{32}, P_{34} are

$$t_{01} = t_{02} = \frac{1}{k_2[X] + k_{-1}},$$

$$t_{20} = t_{21} = t_{23} = \frac{1}{k_3 + k_{-2} + k_{-1}},\ t_{32} = t_{34} = \frac{1}{k_4 + k_{-3}}.$$

Similarly, we define t_{20w}, t_{21w}, t_{23w}, t_{32w}, t_{34w}, by replacing $k_{3,4}$ with $k_{3w,4w}$ in the above equations.

The time for a successful event is τ_{S}

$$\begin{aligned}\tau_{\mathrm{s}} &= P_{02}\,P_{23}\,P_{34}[t_{02} + t_{23} + t_{34}] + P_{02}(P_{23}\,P_{32})\,P_{23}\,P_{34} \\ &\quad \times [t_{02} + (t_{23} + t_{32})t_{23} + t_{34}] + \ldots\, P_{02}(P_{23}\,P_{32})^n \\ &\quad \times P_{23}\,P_{34}[t_{02} + n(t_{23} + t_{32})t_{23} + t_{34}] \\ &= P_S[t_{02} + t_{23} + t_{34}] + \frac{P_S\,P_{23}\,P_{32}[t_{23} + t_{32}]}{1 - P_{23}\,P_{32}}.\end{aligned} \tag{12}$$

The times for the other events are:

$$\tau_{\mathrm{D1}} = P_{\mathrm{D1}}[t_{02} + t_{21} + t_{10}] + \frac{P_{\mathrm{D1}}\,P_{23}\,P_{32}[t_{23} + t_{32}]}{1 - P_{23}\,P_{32}} \tag{13a}$$

$$\tau_{\mathrm{F}} = P_{\mathrm{F}}[t_{02} + t_{20}] + \frac{P_{\mathrm{F}}\,P_{23}\,P_{32}[t_{23} + t_{32}]}{1 - P_{23}\,P_{32}} \tag{13b}$$

$$\tau_{D2} = P_{01}[t_{01} + t_{10}]. \qquad (13c)$$

In Eq. (13a–c), $t10 = 1/k_1[E]$ is the average time it takes the binary complex (template/primer) to bind a polymerase. The average passage time τ ($A0 \rightarrow A0, A4$) is the expectancy value of all possible events

$$\tau = \tau_S + \tau_{D1} + \tau_F + \tau_{D2} = t_{02} + P_{01}t_{10} + \frac{P_{02}[t_{20} + P_{23}t_{34} + P_{21}t_{10}]}{1 - P_{23}P_{32}}. \qquad (14)$$

The average passage time of an incorrect nucleotide is τw. The template sequence determines the correct nucleotide and it is realistic to expect that the rate constants of different nucleotide types differ. Theoretically speaking we should distinguish between average passage times τA, τC, τT, τG and τwA, τwC, τwT, τwG, but in practice the paucity of kinetic data rules out such a distinction. If we track a primer/template complex over a large number of passage times and observe that the template extends with NA +NT +NC +NG, each nucleotide type requires on the average $Ni/[xiP\mathrm{S} + (1 - xi)P\mathrm{S}w]$ passage times. The number of passage times associated with correct/incorrect nucleotide is $xiNi/[xiP\mathrm{S} + (1 - xi)P\mathrm{S}w]$ and $[1 - xi]Ni/[xiP\mathrm{S} + (1 - xi)P\mathrm{S}w]$, respectively. Therefore, the time to insert Ni nucleotides is $(xiNi\tau + [1 - xi]Ni\tau i)/[xiP\mathrm{S} + (1 - xi)P\mathrm{S}w]$. Thus, the extension of the template at site i occurs at the rate

$$v_i = \frac{x_i P_S + (1 - x_i)P_{Sw}}{x_i \tau + (1 - x_i)\tau_w}. \qquad (15)$$

This is a local rate at a point in the sequence and $1/vi$ is the average time to insert nucleotide i. The average rate for a template with composition N=NA +NT +NC +NG is

$$\nu_{\text{ave}} = \frac{N}{\sum_{i=\text{A,C,T,G}} N_i[x_i\tau + (1-x_i)\tau_w]/[x_i P_{\text{S}} + (1-x_i)P_{\text{S}w}]}. \qquad (16)$$

The error frequency is:

$$E = \left\{ \sum_{i=\text{A,C,T,G}} [N_i(1-x_i)P_{\text{S}w}]/[x_i P_{\text{S}} + (1-x_i)P_{\text{S}w}] \right\} /N. \qquad (17)$$

3. Elongation rates and template sequence It has been experimentally observed that G+C rich sequences copy slower, in other words, we expect that τG > τA,τT,τC. To investigate this point further, we propose that the rate constant $k3$ is nucleotide dependent. Note that this is only a phenomenological relation $k3(i)$. The model accounts for the dependence of rate on the four base pairs that immediately precede the insertion site. At position j in the sequence nucleotide type i must be inserted. Check the four base pairs before position j and set their number of G–C pairs equal to NG. The value of NG can vary between 0 and 4. The model states that the rate constant is lowered by 10% if a G must be inserted and by a further10% for each additional G+C pair in the previous four base pairs. When a G+C rich region is encountered, the rate could be reduced by as much as 50%.

$k3(i, j)$ = _ $k3$, i_= G

0.8$k3$, i= G_− 0.1$k3$ × NG,

where i is nucleotide type and j is sequence position. (18) The model is tested for two computer-generated sequences of a template of 1020 base pairs. A random generator assigns G, C, T or A values to the template based on probabilities. The first sequence (S1) has equal probability for any nucleotide type. A second sequence (S2) is generated, but the G+C content is 70%. The av-

erage time to insert a nucleotide of type i is calculated from Eq. (15) and Eq. (18) has been used to evaluate $k3$. The parameter values are listed in Table 1. It is assumed that the primers are 20 base pairs long, hence copying starts at site 21. In Fig. 2 the extension rate for template S1 is shown for the region $301 \leq j \leq 400$. As expected, the movement of the polymerase along the template is stochastic. The average extension rate is 90 nucleotides per second. When the calculation is repeated for template S2, the extension is slower (Fig. 3). In this case the average rate drops to 86 nucleotides per second. The increase in G+C content has led to only a slight decrease in the overall elongation rate. Keep in mind that this is a phenomenological study and the

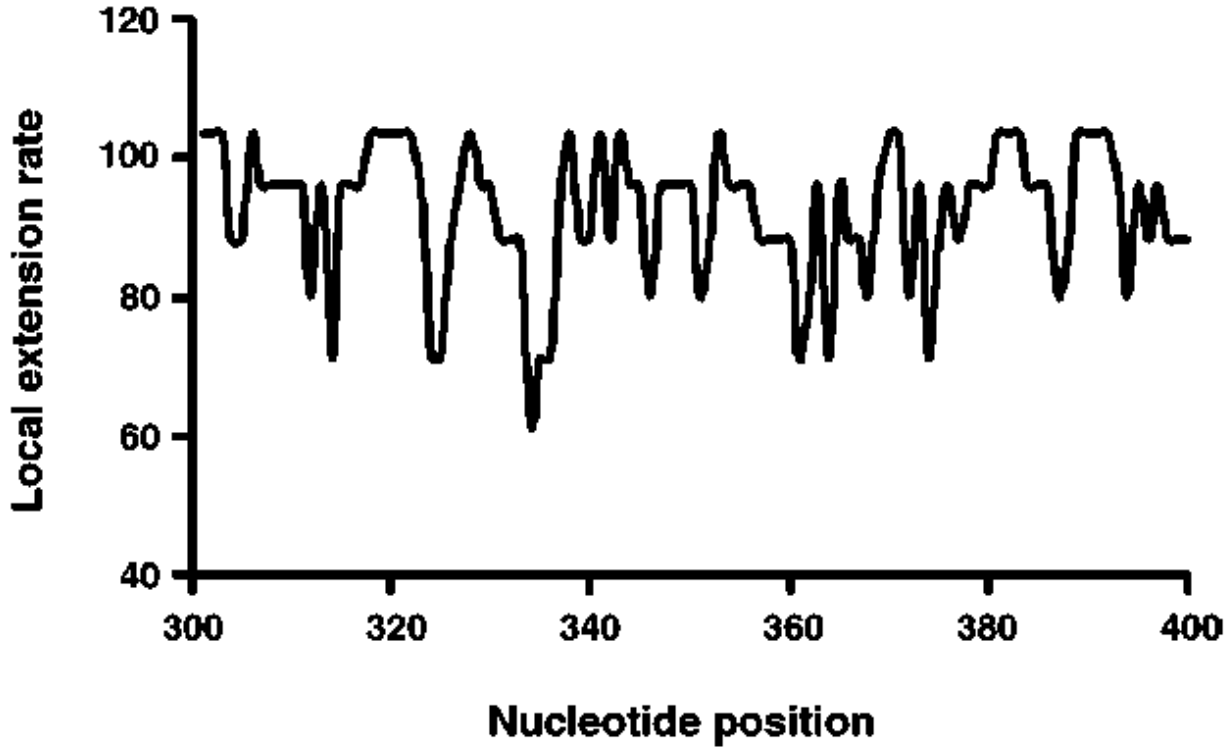

Fig. 2. Local extension rate vs. sequence S1 (50% G+C). The part of S1 between 300–400 is shown.

G+C effect was only included in $k3$. A comparison of Figs. 2 and 3 shows that the local dynamics have changed. The polymerase rates drop to 51 nucleotides per second in sequences that are G+C rich. The question that remains unanswered at this stage is whether these periods of very slow processing (it could also be interpreted as pausing) have any secondary effects, such as ed-

iting efficiency. Besides $k3$, there may be other parameters that are also affected by sequence composition, but at this stage the relations are not known. It is also known that polymerases of the *Pyrococcus* family have the ability to assess the sequence ahead of the insertion site and they detach if a U is detected (Greagg et al., 1999; Fogg et al., 2002; Connolly et al., 2003; Shuttleworth et al., 2004). It would be straightforward to modify the mathematical model to describe this behavior.

Table 1
Parameter values

Parameter	k_4	k_{4w}	k_3	k_{3w}	k_{-3}	k_1	k_{-1}	k_2	k_{-2}	k_5	[dNTP] (μM)	[E] (μM)
Value	3000	0.1	1000	0.01	1	40	0.2	4	$K_2[X]/1.6$	0.2	80	1

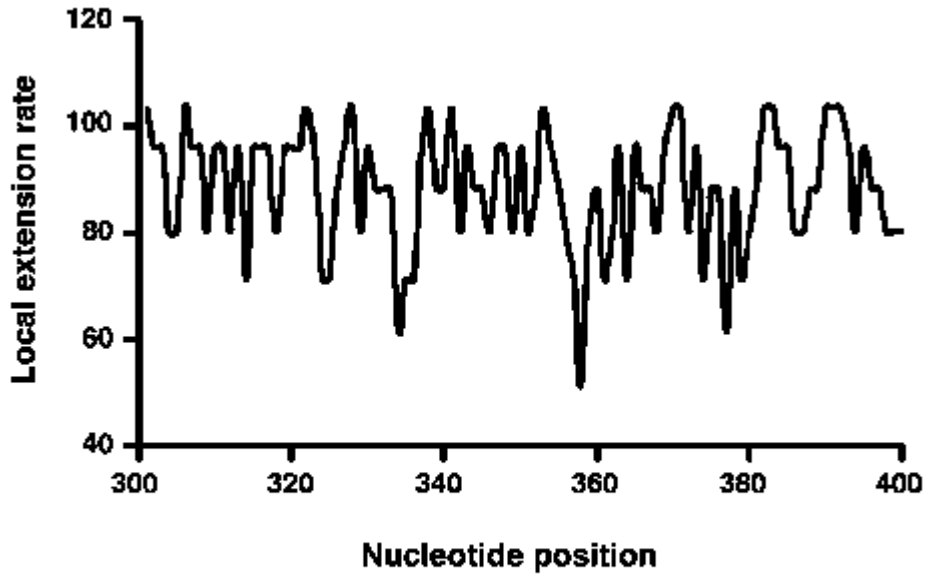

Fig. 3. Local extension rate vs. sequence S2 (70% G − C). The part of S2 between 300–400 is shown.

4. Estimates of elongation rates and error frequencies with varying [dNTP] pools Anotherway to bias the experiment is by adjusting the molar composition of dNTP. Changes in [*X*] will lead to changes in the probabilities and passage times. Template S1 is composed of equal amounts of all four nucleotides and it is expected that an equimolar dNTP pool will perform best. To investigate the effect of varying dNTP pools, the total dNTP concentration is fixed at 800 _M and [dGTP] is varied from 20 to 740_M with the balance equally distributed amongst the other three dNTPs. The elongation rate is calculated as a function of *X*G, the molar fraction of dGTP in the pool. As expected, the elongation rate is a maximum when *X*G is 0.25

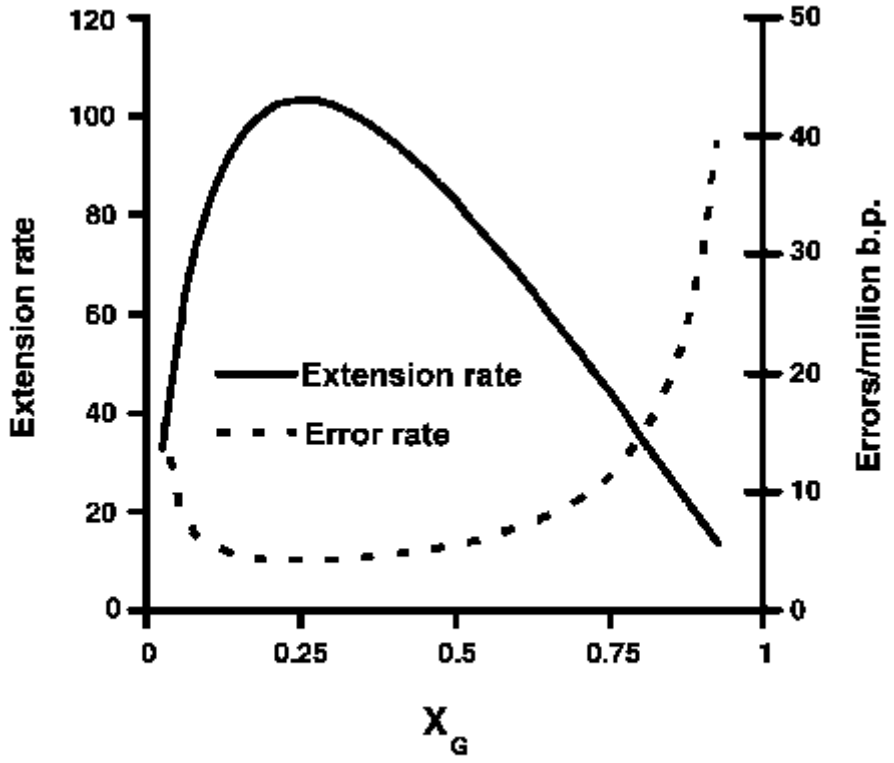

Fig. 4. Elongation rate and error frequency vs. XG for sequence S1.

(Fig. 4). The rate decreases sharply when the molar fraction is lowered, but the decrease in rate is less pronounced when *X*G is increased. This finding is logical. If *X*G is (0.25−0.24), there is a 1% chance that dGTP presents at the binding site. If *X*G is (0.25 + 0.24), the molar fraction of each one of the rest is (1−0.49)/3 = 0.17, which means a reduction from 25 Fig. 4. Elongation rate and error frequency vs. XG for sequence S1. to 17% to be present at the binding site. Clearly an increase in XG does not exhibit the same sensitivity as a reduction in its value. Of course, the reaction is inhibited when *X*G approaches one. In Fig. 4 the error frequency is shown as a function of *X*G. The parameters are listed in Table 1. The parameter *k*3 has been kept constant in this case. The lowest error frequency of 4.09/106 corresponds to equimolar conditions, *X*G = 0.25. The extreme values are 15.70/106 at *X*G = 0.025 and 39.60/106 at *X*G = 0.925. At equimolar conditions, the rate is a maximum and the complete extension of the template requires the least number of passage times. The results show that the error frequency is a minimum when the number of passage times is a minimum. The results further show that error frequencies are relatively resistant to changes in the molar fraction over a wide range. A plateau is observed from about 0.1 to 0.7 mol fraction of dGTP. When the G+C rich template S2 is used (35% G) the maximum elongation rate occurs at *X*G = 0.300 and this molar fraction coincides with the minimum error frequency of 4.05/106 (Fig. 5). However, the optimum value *X*G = 0.300 does not correspond to the nucleotide composition of the template. Although the optimum molar fraction is more than the equimolar value, an increase to *X*G = 0.35 impedes the extension rate of the other three nucleotides. Interestingly, the values of the error frequencies of S1 and S2 at their optimum compositions are the same, but at *X*G = 0.925 the error rate of S1 (39.60/106) exceed the error rate of S2 (35.20/106).

5. The role of kinetics in product distribution

5.1. Example I: poly-T extension

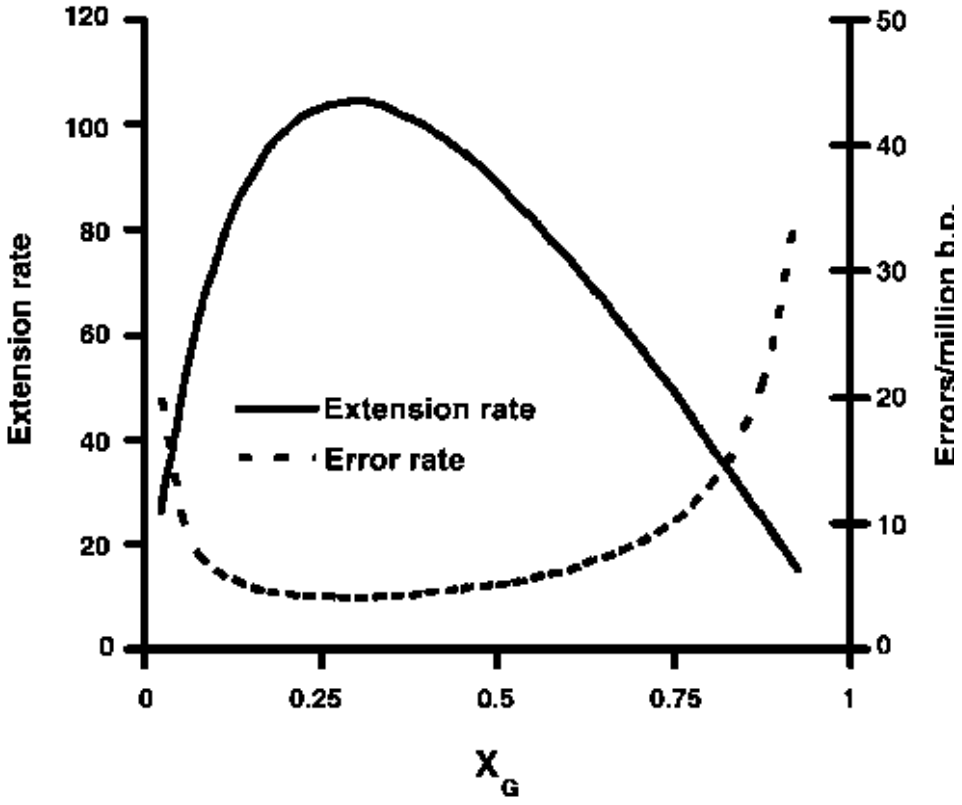

Fig. 5. Elongation rate and error frequency vs. XG for sequence S2.

Consider a template that consists only of T and a pool that only contains dATP. In this case only the correct nucleotide can attach. Let, the total elongation time be tTot, then the total number of events is N and it consists of nS insertion events and nU unsuccessful events. The unsuccessful events are a combination of F, D1 and D2. The time for a successful insertion is τS (cf. Eq. (12)). From Eq. (14), the time for an unsuccessful event is $\tau - \tau S$ and the probability of its occurrence is $1-P$S. tTot = nSτS + nU($\tau - \tau$S). Thus, nS may vary between 0 and tTot/τS and accordingly nU may vary between (tTot $-n$Sτ)/$\tau - \tau$S. After a large number of templates have been tracked, the following probability distribution is found:

$$\prod(n_S) = \frac{N!}{n_S!(N - n_S)!} P_S^{n_S}(1 - P_S)^{N-n_S}. \qquad (19)$$

In Fig. 6 the distributions are shown for two cases, (1) $\tau_S = 0.3\,\text{ms}, \tau = 1\,\text{ms}$ and $P_S = 0.33$ and (2) $\tau_S = 0.3\,\text{ms}, \tau = 0.75\,\text{ms}$ and $P_S = 0.33$. The maximum extension is 100 nucleotides, thus for $n_S > 100$, $\sum_{100}^{n_S} \prod(n_S)$ accumulates

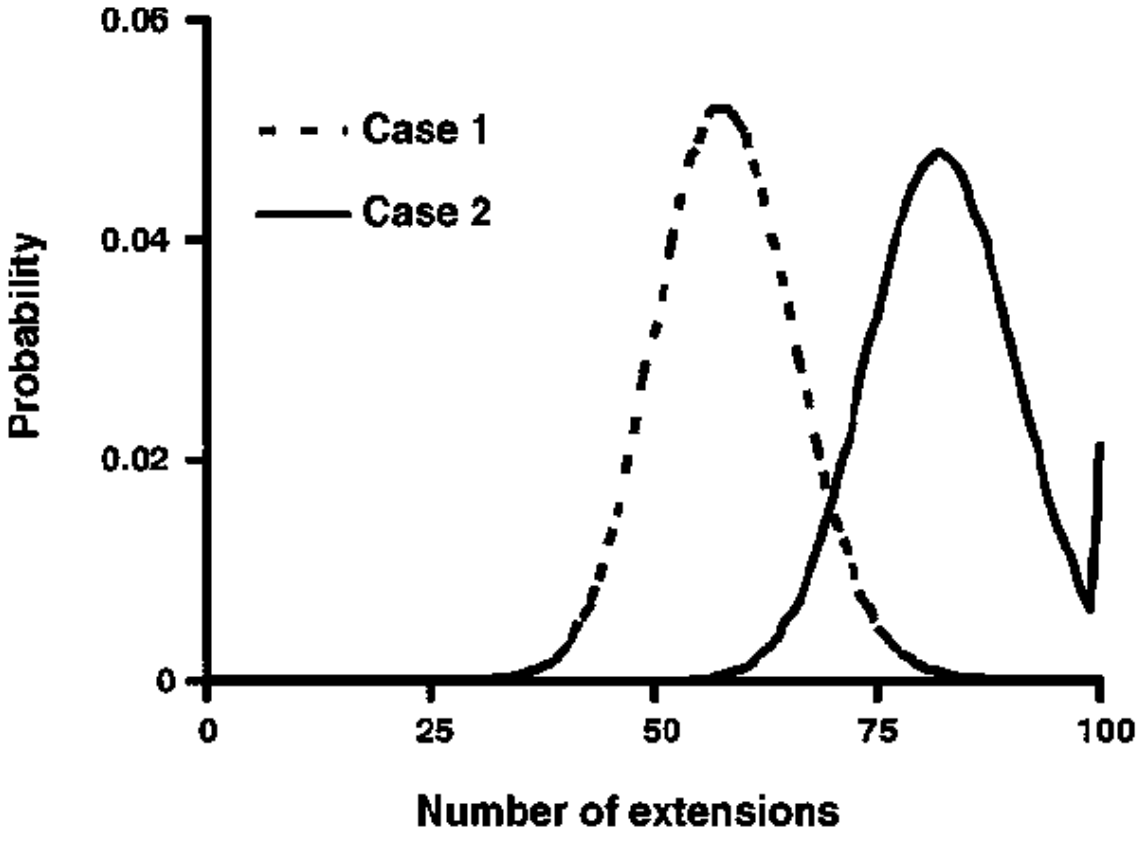

Fig. 6. Probability distribution of successful events vs. n_S.

at _(100). In case (1) the highest probability occurs at nS = 58. The 5 and 95% intervals are at nS = 46 and 71, respectively and the probability to extend fully is negligible. In case (2) the maximum lies at nS = 82 and the 5 and 95% intervals are at ns = 69 and 96. The accumulation at nS = 100 indicates that some templates have extended fully. The reduction in passage time of unproductive events from 0.7 ms in case (1) to 0.45 ms in case (2), allows more successful attempts within the time span tTot and subsequently the distribution shifts to higher number of extensions. The calculation of distribution functions becomes complicated for general templates. The total elongation time is the sum of time spent to process correct and wrong nucleotides; tTot **=** nCτ **+** $nw\tau w$. If nC is varied between tTot/τ and 0, a first distribution _1(nC) must be calculated and for each value of nC a second distribution _2(nS) exists for the successful events. The calculation of _1(nC) requires the probability to select the correct nucleotide, but that is determined by the template sequence. Strictly speaking, the [dNTP] pool changes and these changes may affect the kinetics and therefore the distribution _1. Distributions are further influenced by temperature. The correct application of the mapping procedure requires careful bookkeeping of reactant consumption, product formation and changes in temperature. However, all the analytical tools are now in place to perform such detailed studies. A simpler, but less elegant approach to find the distribution function is a Monte Carlo simulation. Starting at state A0 a randomly generated number is compared to the probability to proceed to state A2. The outcome determines if state A2 or A1 is reached. If state A2 is reached a new random number determines the choice of nucleotide type and thus the correct parameter set (i.e. all parameters that depend on k3, k4 or $k3w$, $k4w$). Comparisons of random numbers with probabilities determine the further progression of the complex. After a change of state has occurred, the elapsed time is updated by adding the passage time for that change.

Once the elapsed time tTot has been reached, the total number of extensions is recorded. Let the total number of extensions be n*, then the distribution function is updated as follows: _(n*) =_(n*) + 1 and the process starts over. Sufficient repetitions produce a fairly smooth distribution function. In Fig. 7 the normalized distribution function is shown after two million repetitions. The parameters correspond to the values in Table 1 and an equimolar dNTP pool has been used.

The Monte Carlo simulation gives nearly the same average extension rate (103.41 nts) as calculated from Eq. (16), 103.38 nts. The error rate is calculated as 4.20 errors/106, which is slightly larger than the value of 4.09 errors/106 calculated from Eq. (17). It takes many more Monte Carlo simulations to calculate error frequencies accurately as compared to extension rates, because errors are very unlikely events. Now one can apply the distribution function to the PCR process. Let the template/primer concentration at the onset of an elongation step be *C*0 (arbitrary units). The product C0_(*n*) is the fraction of *C*0 that has extended *n* base pairs, provided *n* Probability distribution of successful insertions calculated by Monte Carlo simulation method for parameters as listed in Table 1. is less than the target length *N* of the template. For *n*≥*N*, the fully extended product is*C*1(*N*) = *C*0_*n*max *n*=*N* _(*n*).Asimilar approach has been used by Whitney et al. (2004) to simulate the elongation steps of a PCR process.

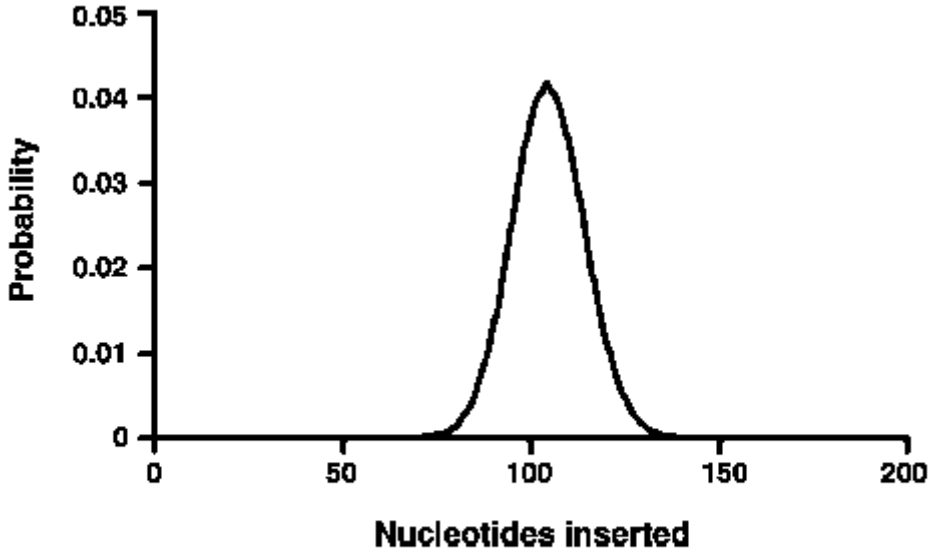

Fig. 7. Probability distribution of successful insertions calculated by Monte Carlo simulation method for parameters as listed in Table 1.

6. Relationship of model to biological observations.

Imbalances of the intracellular dNTP pool are linked with enhanced rates of biological mutagenesis, cancer, and genetic diseases. To begin with, dGTP is naturally underrepresented in eukaryotic cells and typically accounts for 5–10% of the total pool (that is, XG = 0.05–0.10) (Mathews and Ji, 1992). Biochemical and cellular experiments show that eukaryotic polymerases are slightly but significantly more mutagenic in these naturally imbalanced dNTP pools than in equimolar pools (Martomo and Mathews, 2002). This is exactly the modest outcome predicted by our model. When the pool is biased beyond the natural, it should cause substantial increases in the mutagenic rates. In fact, a wide variety of external agents do cause severe changes to cytosolic, nuclear, and mitochondrial dNTP pool compositions and increases to cellular mutation rates. Most changes that affect the nucleotide biosynthetic or salvaging pathways lead to dNTP pool imbalances such as by heat-induced cytosine deamination (Lindahl and Nyberg, 1974), by ultraviolet light and cytosine arabinoside (Das et al., 1983; Suzuki et al., 1983), by fluorouracil (Yoshioka et al., 1987), by chlorodeoxyadenosine (Wataya et al., 1988), by changes to the single-carbon transfer pathway through dietaryfolate or methotrexate treatment (James et al., 1992; Kasahara et al., 1993; Melnyk et al., 1999), by the antineoplastic tropoline (Yamato et al., 1992), by difluorodeoxycytidine (Heinemann et al., 1995), by hypoxia (Chimploy et al., 2000), and by ribonuclease reductase inhibitors such as hydroxyurea. The exceptions seem to be the nucleoside analog drugs azidothymidine and azidocytidine, which do not cause changes to the dNTP pool even though they inhibit intracellular DNA synthesis (Akerblom et al., 1982; Julias and Pathak, 1998). Finally, mitochondrial DNA deletion syndrome (MDS) is a rare disease caused by mutations to a variety of genes. Specifically, mutations in thymidine phosphorylase (Song et al., 2003) or thymidine kinase 2 (TK2) have been associated with different forms of MDS (Saada et al., 2001; Wang et al., 2003). Our results demonstrate that dNTP pool imbalances have to increase or decrease substantially before any effects will be observed on DNA polymerase elongation and error rates. The cited biological results indicate that such drastic changes to the dNTP pools do occur when cells are treated with many nucleoside analogs or when the nucleoside biosynthetic or salvaging pathways are mutated or altered in some way.

7. Conclusions

The rate expression, Eq. (16), has been derived for a multisubstrate enzyme-catalyzed DNA extension reaction. A reaction scheme, that is assumed to be generic to most polymerases,
has been used for the derivation. If we track the behavior of a single enzyme/template complex over a period of time (let us call this period a cycle), we have complete knowledge about this complex over the span of the cycle. If the 'tracking experiment' is repeated a large number of cycles, we learn more about the average behavior of a single complex. The derivation is based on the equivalence of the average behavior of a single complex over many cycles and the average behavior of many complexes over one cycle.

The practical application of Eq. (16) is straightforward. Eq. (16) may be written as follows:

$$\tau_{\text{ave}} = \frac{1}{\nu_{\text{ave}}} = \frac{1}{N} \sum_{i=\text{A,C,T,G}} \frac{N_i[x_i\tau + (1-x_i)\tau_w]}{x_i P_{\text{S}} + (1-x_i)P_{\text{S}w}}.$$

Since $P_{\text{S}} \gg P_{\text{S}w}$ and unless the extreme dNTP pool compositions $x_i \approx 0$ is used, the expression simplifies to

$$\tau_{\text{ave}} \approx \frac{\tau}{P_{\text{S}}} + \left(\sum_{i=\text{A,C,T,G}} \frac{N_i(1-x_i)}{Nx_i} \right) \frac{\tau_w}{P_{\text{S}}}.$$

The two unknown parameters $\Gamma = \tau/P_{\text{S}}$ and $\Gamma_w = \tau_w/P_{\text{S}}$

are solved from two measurements of τave at two different dNTP pool compositions (at all times dNTP's must be in excess) for a template with a known sequence. The parameters are unique to the type of polymerase, but they also depend on other factors like template structure (see Section 3) and temperature.

Temperature affects kinetic rate constants and as a first approximation we could assign Arrhenius dependence to these two parameters. If the two abovementioned measurements are repeated at a different elongation temperature, it is possible to resolve $\Gamma = \Gamma 0 e^{-E/RT}$; $\Gamma w = \Gamma w 0 e^{-Ew/RT}$.

The main results of the study are:

(1) The inverse of the reaction rate is the average time to insert a base pair and it consists of the time to insert a correct nucleotide plus delays to edit and proof-read for incorrect nucleotides and the assembly of enzyme/template complexes following polymerase detachment from the complex.

(2) The insertion times depend on the type of nucleotide to be inserted. It is further known that the composition of the template in the vicinity of the insertion site (behind and ahead of the insertion point) could play a role in the rate of insertion. In principle, the roles of sequence composition and topology could be included in the model—easily demonstrable in a phenomenological model, but more experimental data is needed for a quantitative analysis.

(3) Error frequencies can be easily calculated and once again the limitation is availability of experimental kinetic data.

(4) Insertion rates depend on temperature and the role of temperature has been demonstrated for a select set of parameters. The example has also found that minimum error frequency and optimum insertion rate coincide.

(5) Insertion rates depend on substrate composition. The effect has been demonstrated by varying the [dGTP] concentration. When [dGTP] is lowered with respect to the other [dNTP] concentrations, the average time to insert dGMP increases. The enzyme spends more time to edit and reject other nucleotides at G-sites. When the [dGTP] concentration is increased with respect to the other [dNTP] concentrations, the insertion times for the other dNMP’s increase, because the enzyme spends more time editing and rejecting dGTP from non-G insertion sites.

(6) The model lays the basis for the calculation of distribution functions. These functions are used to calculate the varying number of nucleotides that have been added to primer/template complexes. The distribution depends on the kinetic rates, dNTP pool, temperature and the total elongation time. Acknowledgements The authors (MN, HV) are grateful to financial support by the National Institutes of Health. HJV also acknowledges financial support by the Army Research Laboratory.

Appendix A. Supplementary data Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem. 2005.02.003.

References

Ahmed, A., Tollefsbol, T.O., 2003. Telomerase, telomerase inhibition and cancer. J. Anti Aging Med. 6, 315–325.

Akerblom, L., Pontis, E., Reichard, P., 1982. Effects of azidocytidine on DNA synthesis and deoxynucleotide pools of mouse fibroblast cell lines. J. Biol. Chem. 257, 6776–6782.

Briggs, GE., Haldane, J.B.S., 1925. A note on the kinetics of enzyme action. Biochem. J. 19, 338–339.

Chimploy, K., Tassotto, M.L., Mathews, C.K., 2000. Ribonucleotide reductase a possible agent in deoxyribonucleotide pool asymmetries induced by hypoxia. J. Biol. Chem. 275, 39267–39271.

Connolly, B.A., Fogg, M.J., Shuttleworth, G., Wilson, B.T., 2003. Uracil recognition by archaeal family B DNA polymerises. Biochem. Soc. Trans. 31, 699–702.

Das, S.K., Benditt, EP., Loeb, L.A., 1983. Rapid changes in deoxynucleoside triphosphate pools in mammalian cells treated with mutagens. Biochem. Biophys. Res. Commun. 114, 458–464.

Fogg, M.J., Pearl, L.H., Connolly, B.A., 2002. Structural basis for uracil recognition by archaeal family B DNA polymerises. Nat. Struct. Biol. 9, 922–927.

Greagg, M.A., Fogg, M.J., Panayotou, G., Evans, S.J., Connolly, B.A., Pearl, L.H., 1999. A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. Proc. Natl. Acad. Sci. U.S.A. 96, 9045–9050.

Grundberg-Manago, M., Oritz, PJ., Ochoa, S., 1955. Enzymatic synthesis of nucleic acid-like polynucleotides. Science 122, 907–910.

Heinemann, V., Schulz, L., Issels, RD., Plunkett, W., 1995. Gemcitabine: a modulator of intracellular nucleotide and deoxynucleotide metabolism. Semin. Oncol. 22, 11–18.

Jager, J., Pata, J.D., 1999. Getting a grip: polymerases and their substrate complexes. Curr. Opin. Struct. Biol. 9, 21–28.

James, S.J., Cross, D.R., Miller, B.J., 1992. Alterations in nucleotide pools in rats fed diets deficient in choline methionine and/or folic acid. Carcinogenesis 13, 2471–2474.

Johnson, K.A., 1993. Conformational coupling in DNA polymerase fidelity. Annu. Rev. Biochem. 62, 685–713.

Joyce, C.M., Benkovic, S.J., 2004. DNA polymerase fidelity: kinetics, structure and checkpoints. Biochem 43, 14317–14324.

Julias, JG., Pathak, V.K., 1998. Deoxyribonucleoside triphosphate pool imbalances in vivo are associated with an increased retroviral mutation rate. J. Virol. 72, 7941–7949.

Kasahara, Y., Nakai, Y., Miura, D., Kanatani, H., Yagi, K., Hirabayashi, K., Takahashi, Y., Izawa, Y., 1993. Decrease in deoxyribonucleotide triphosphate pools and induction of alkaline-labile sites in mouse bone marrow cells by multiple treatments with methotrexate. Mutat. Res. 319, 143–149.

Kiefer, J.R., Mao, C., Braman, J.C., Beese, L.S., 1998. Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. Nature 391, 304–347.

Kunkel, T.A., 2004. DNA replication fidelity. J. Biol. Chem. 279, 16895–16898.

Lehman, I.R., 2003. Discovery of DNA polymerase. J. Biol. Chem. 278, 34733–34738.

Li, Y., Waksman, G., 2001. Structural studies of the Klentaq1 DNA polymerase.

Curr. Org. Chem. 5, 871–883.

Lindahl, T., Nyberg, B., 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochem 13, 3405–3410.

Martomo, S.A., Mathews, C.K., 2002. Effects of biological DNA precursor pool asymmetry upon accuracy of DNA replication in vitro.
Mutat. Res. 499, 197–211. Mathews, C.K., Ji, J., 1992. DNA precursor asymmetries, replication fidelity, and variable genome evolution. Bioessays 14, 295– 301.
Melnyk, S., Pogribna, M., Miller, B.J., Basnakian, A.G., Pogribny, I.P., James, S.J., 1999. Uracil misincorporation, DNA strand breaks, and gene amplification are associated with tumorigenic cell transformation in folate deficient/repleted Chinese hamster ovary cells. Cancer Lett. 146, 35–44.
Michaelis, L., Menten, M.L., 1913. Die Kinetik der Invertinwirkung [Kinetics of Invertase Action]. Biochem. Z. 49, 333–369.
Miura, S., Izuta, S., 2004. DNA polymerases as targets of anticancer nucleosides. Curr. Drug Targets 5, 191–195.
Ninio, J., 1987. Alternative to the steady-state method: derivation of reaction rates from first-passage times and pathway probabilities. Proc. Natl. Acad. Sci. U.S.A. 84, 663–667.
Patel, S.S., Wong, I., Johnson, K.A., 1991. Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant. Biochem 30, 511–525.
Saada, A., Shaag, A., Mandel, H., Nevo, Y., Eriksson, S., Elpeleg, O., 2001. Mutant mitochondrial thymidine kinase in mitochondrial DNA depletion myopathy. Nat. Genet. 29, 342–344.
Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., Arnheim, N., 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science, 230.
Shuttleworth, G., Fogg, M.J., Kurpiewski, M.R., Jen-Jacobson, L., Connolly, B.A., 2004. Recognition of the pro-mutagenic base uracil by family B DNA polymerases from archaea. J. Mol. Biol. 337, 621–634.
Smith, A.J., 1980. DNA sequence analysis by primed synthesis. Methods Enzymol. 65, 560–580.
Song, S., Wheeler, L.J., Mathews, C.K., 2003. Deoxyribonucleotide pool imbalance stimulates deletions in HeLa cell mitochondrial DNA. J. Biol. Chem. 278, 43893–43896.

Suzuki, K., Miyaki, M., Ono, T., Mori, H., Moriya, H., Kato, T., 1983. UV-induced imbalance of the deoxyribonucleoside triphosphate pool in *E. coli*. Mutat. Res. 122, 293–298.

Vieille, C., Zeikus, G.J., 2001. Hyperthermophilic enzymes: sources, uses and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev. 65, 1–43.

Wang, L., Saada, A., Eriksson, S., 2003. Kinetic properties of mutant human thymidine kinase 2 suggest a mechanism for mitochondrial DNA depletion myopathy. J. Biol. Chem. 278, 6963–6968.

Wataya, Y., Hazawa, T., Watanabe, K., Hirota, Y., Yoshioka-Hiramoto, A., 1988. Detection of deoxyribonucleoside-triphosphate imbalance deathinduced DNA double strand breakage in FM3A cells by orthogonalfield- alternation gel electrophoresis (OFAGE).
Nucl. Acids Symp. Ser., 53–55.

Whitney, S.E., Alugupally, S., Nelson, R.M., Viljoen, H.J., 2004. Principles of rapid polymerase chain reactions: mathematical modeling and experimental verification.
Comp. Biol. Chem. 28, 195–209.

Yamato, M., Hirota, Y., Yoshida, S., Tanaka, S., Morita, T., Sakai, J., Hashigaki, K., Hayatsu, H., Wataya, Y., 1992. Imbalance of deoxyribonucleoside triphosphates and DNA double-strand breaks in mouse mammary tumor FM3A cells treated in vitro with an antineoplastic tropolone derivative. Jpn. J. Cancer Res. 83, 661–668.

Yoshioka, A., Tanaka, S., Hiraoka, O., Koyama, Y., Hirota, Y., Ayusawa, D., Seno, T., Garrett, C., Wataya, Y., 1987. Deoxyribonucleoside triphosphate imbalance. 5-Fluorodeoxyuridine-induced DNA double strand breaks in mouse FM3A cells and the mechanism of cell death.
J. Biol. Chem. 262, 8235–8241.