A theoretical model for template-free synthesis of long DNA sequence

S. M. Minhaz Ud-Dean

Abstract This theoretical scheme is intended to formulate a potential method for high fidelity synthesis of Nucleic Acid molecules towards a few thousand bases using an enzyme system. Terminal Deoxyribonucleotidyl Transferase, which adds a nucleotide to the 3’OH end of a Nucleic Acid molecule, may be used in combination with a controlled method for nucleotide addition and degradation, to synthesize a predefined Nucleic Acid sequence. A pH control system is suggested to regulate the sequential activity switching of different enzymes in the synthetic scheme. Current practice of synthetic biology is cumbersome, expensive and often error prone owing to the dependence on the ligation of short oligonucleotides to fabricate functional genetic parts. The projected scheme is likely to render synthetic genomics appreciably convenient and economic by providing longer DNA molecules to start with.

Keywords Long DNA synthesis · Enzyme system modelling · Synthetic gene · pH regulated synthesis · Terminal deoxyribonucleotidyl transferase · Genetic parts fabrication

Abbreviations
3’AcDN 3’ Acetyl deoxyribo nucleoside
3’AcDNA 3’ Acetyl deoxyribo nucleic acid
3’AcNDP 3’ Acetyl deoxyribo nucleoside di phosphate
3’AcNMP 3’ Acetyl deoxyribo nucleoside mono phosphate
3’AcNTP 3’ Acetyl deoxyribo nucleoside tri phosphate
Ac Acetyl group
AcOH Acetic acid (in equilibrium with acetate)
AE Acetylesterase, deacetylase (EC: 3.1.1.6)
AP Acid phosphatase (EC: 3.1.3.2)
DNA Deoxyribo nucleic acid
dNTP Deoxyribo nucleoside tri phosphate
ds Double strand
ID Inorganic diphosphatase (EC: 3.6.1.1)
Kb Kilobase
Mb Megabase
NA Nucleic acid
NTP Ribo nucleoside tri phosphate
P Phosphate group
PCR Polymerase chain reaction
RNA Ribo nucleic acid
ss Single strand
TdT Terminal deoxynucleotidyl transferase, DNA nucleotidyl exotransferase (EC: 2.7.7.31)

Introduction

To a great extent the progress in biotechnology and basic biomedical research has been centralized on the major advances in DNA synthesis and sequencing. Elucidation of the genetic code (Khorana 1968), production of synthetic gene (Agarwal et al. 1974), widespread use of PCR (Kleppe et al. 1971; Saiki et al. 1988), sequencing of the human genome (Venter et al. 2001; Lander et al. 2001), and the synthesis of whole genome of a microorganism (Gibson et al. 2008) are powerful examples. These
applications along with many others have pivotally depended on the ability to synthesize short oligonucleotides, used as primers, typically ssDNA 10–80 base in length (Caruthers 1985).

The widespread application of synthetic biology is essentially limited by the complexity and cost of assembling short oligonucleotides into longer functional DNA (Endy 2005). In addition to whole genomes (Hutchison et al. 1999; Smith et al. 2003; Gibson et al. 2008), construction of entire biochemical pathways (Martin et al. 2003; Mehl et al. 2003; Kodumal et al. 2004) and genetic circuitry (Elowitz and Leibler 2000; Sprinzak and Elowitz 2005) are illustrative of synthetic biology requiring synthesis of far more than a single gene. Therefore, the rapid availability of predefined DNA, more than 1 Mb in length at a cost per base comparable to or less than oligonucleotides is greatly appealing. Despite important progress in this direction through building large numbers of genes by harnessing the massively parallel form of oligonucleotide synthesis to produce microarrays (Richmond et al. 2004; Tian et al. 2004), still the procedures depend on the chemical synthesis of short oligos. Thus the availability of longer DNA molecules is in principle restricted by the inherent limit of the chemical processes for oligonucleotide synthesis.

The error rate of synthesizing oligonucleotides is also of immense importance. At an error rate of 1 in 600 bp, it is required to sequence 10 clones to get a DNA construct of 1 Kb and 100 clones for 2 Kb (Baedeker and Schulz 1999; Withers-Martinez et al. 1999; Hoover and Lubkowski 2002; Chalmers and Curnow 2001). Thus large target fabrication is extremely difficult and error prone at this rate (Cello et al. 2004). The error was reduced to 1 in 1,400 bp by Tian et al. (2004) using microchip based multiplex synthesis. Protein mediated error correction has been utilized to reduce the rate further to 1 in 10,000 bp (Carr et al. 2004).

If the starting material is precise DNA molecules 1–10 Kb in length rather than ~50 bases, the speed and over all efficiency of synthetic genomics would tremendously increase because synthesis of longer constructs would be possible with fewer clones and sequencing. An extensive application of synthetic genes would be possible with a more convenient and reliable process for their fabrication. Thus an investigation for an efficient synthetic scheme is worth pursuing.

An attempt to delineate the necessary properties of a successful long DNA synthesis process leads to some salient features. Firstly, the error rate should be less than $10^{-4}$ in order to allow convenient synthesis of DNA constructs around $10^5$ bp in length. Second, the maximum synthesizable length of DNA should be greater than $10^5$ as current protocols allow transformation with DNA constructs of up to $3 \times 10^5 - 20 \times 10^5$ bp in length (Glick and Pasternak 2003). These features indicate that preferably an enzymatic system might be successful. Moreover, enzymatic systems would in principle be more specific and efficient than chemical processes (Sitnitsky 2006). However, the enzymes those can synthesize DNA without templates are not controlled. For example TdT will add a homopolymeric tract of nucleotides when provided with a primer and a single type of NTP (Bollum 1978; Ratliff 1981). When provided with a mixture of nucleotides and a primer, TdT act as a random-sequence generator (Bollum 1978; Ratliff 1981). The physiological role of TdT conform to this random sequence generation as it provides additional variations in hematopoetic cells through acting as a somatic mutator, diversifying the amino acid sequence in the variable region of immunoglobulin molecules (Ratliff 1981).

TdT has already been utilized in genomics. It has been used for the production of synthetic homo and heteropolymers (Bollum 1974), homopolymeric tailing of linear duplex DNA (Deng and Wu 1983; Eschenfeldt et al. 1987), oligodeoxyribonucleotide and DNA labeling (Deng and Wu 1983; Tu and Cohen 1980; Vincent et al. 1982; Kumar et al. 1988; Igloi and Schiefermayr 1993), rapid amplification of cDNA ends (Frohman et al. 1988) and in situ localization of apoptosis (Gorczyca et al. 1993). Scheele and Fukuoka (1992, 1997) used TdT to add homopolymeric oligo dC tract to 3′ end of ss linear DNA in order to facilitate synthesis of ds linear DNA using oligo dG primer. However, no attempt to use TdT to synthesize a defined DNA sequence has been reported. Perhaps the random sequence generation by TdT in an uncontrolled system has prevented this.

Here a theoretical model for predefined long DNA synthesis based on TdT is proposed. The scheme depends on the addition of 3′AcdNTP to the 3′-OH end of a DNA. The Ac group is suggested to prevent polymer formation and thus prevent the chain extension to a single base in each cycle (Fig. 1). After extension the pH of the system is decreased to activate AE and AP while TdT would be deactivated. AP would hydrolyze excess 3′AcdNTP while AE would remove the Ac groups. Then dialysis would render the small molecules out of the system. Dialysis would also elevate the pH to activate TdT and deactivate AE and AP. Thus a new cycle of extension can be initiated by addition of 3′AcdNTP. Theoretical analysis based on available data suggests the scheme to be highly promising in synthesizing long DNA molecules.

**Methods**

The scheme

Synthesis of a predefined DNA sequence requires the addition of a single known nucleotide at each step to a
given nucleic acid polymer. Thus the problem is essentially reduced to formulating a system such that:

\[
\text{Enzyme Systems DNA}_n \rightarrow \text{DNA}_{n+1}
\]

is allowed, while:

\[
\text{Enzyme Systems DNA}_n \rightarrow \text{DNA} > n+1
\]

is forbidden. Also the allowed reaction should be virtually complete such that no unextended DNA can go through the second round. Searching for enzymes able to extend DNA molecules in the BRENDA enzyme database (Schomburg et al. 2002, 2004; Barthelmes et al. 2007) TdT (EC: 2.7.7.31) was selected to be a potential enzyme because it can add nucleotides to the 3'-OH group of DNA several Kb long and can add modified bases.

In order to prevent formation of a homopolymeric tract the 3'-OH group of the incoming nucleotide should be blocked. Methyl and acetyl groups are extensively used as OH protecting groups. Due to the promising convenience of deprotection by acetyleraserase (EC: 3.1.1.6); acetyl group is expected to be a better candidate. Thus, instead of dNTP the DNA primer is to be elongated by addition of a 3'AcNTP catalyzed by TdT. There is lack of \(K_m\) and \(K_{cat}\) data of TdT with 3'AcdNTP. Nevertheless, the accommodation of modified bases by TdT indicates a high possibility of addition of a single 3'AcdNTP to 3'-OH end of a DNA.

Now the completion of the reaction remains a problem. TdT has \(K_m\) value of 3 \(\times\) 10^{-4} mM for oligonucleotide primers at pH 8.2 in the presence of Mn^{2+} (Coleman 1977) and turnover number 0.833 for ATP (Bollum 1974). Thus TdT should be present in a high molar ratio with respect to primers in order to ensure the completeness of the reaction. Further, the polymerization reaction would be energetically favorable as pyrophosphate may be quickly removed by hydrolysis.

A new cycle of base addition must be preceded by removal of excess 3'AcdNTP and deprotection of the 3'OH group of the extended DNA. Acid phosphatase (EC: 3.1.3.2) is a suitable candidate for hydrolysis of excess 3'AcdNTP while acetyleraserase is preferred for deprotection.

\[
\text{DNA}_n + 3'\text{AcdNTP} \xrightarrow{TdT} 3'\text{AcDNA}_{n+1} + \text{PP}
\]

\[
\text{PP} \xrightarrow{3'}2\text{P}
\]

\[
3'\text{AcdNTP} \xrightarrow{\text{AP}} 3'\text{AcdNDP} \xrightarrow{\text{AP}} 2\text{P} + 3'\text{AcdNMP}
\]

\[
3'\text{AcDNA}_{n+1} \xrightarrow{\text{AE}} \text{AcOH} + \text{DNA}_{n+1}
\]

However the process raises problem regarding precise temporal order. As depicted in Fig. 2, if the protective Ac group is removed before complete hydrolysis of 3'AcdNTP, additional bases may add to the extended DNA. Moreover AE deacetylates wide range substrates. Thus 3'AcdNTP may become deacetylated before incorporation into DNA. This event would also fail the objective. Thus TdT must be inactive when AE is active and vice versa. Bovine TdT has a pH optimum at 7.5 and has limited activity below pH 6.9 (Coleman 1977) however it is stable at pH down to 4.5. On the other hand Aspergillus niger AE has a pH optimum of 5.5 with limited activity over pH 6 (Kormelink et al. 1993). A. niger AE is however stable at pH 8. Therefore, the elongation step should be catalyzed by TdT at pH 7.5 while the deprotection step should take place at pH 5.5.

Penicillium chrysogenum acid phosphatase has pH optimum at pH 5.5 (Haas et al. 1991). Therefore it is expected to be an efficient scavenger of excess 3'AcdNTP after elongation. E. coli inorganic diphosphatase (EC: 3.6.1.1) has pH optimum at pH 7.5 (Vainonen et al. 2005). Thus it can be utilized to hydrolyze pyrophosphate formed during elongation by TdT.
The rapid change is pH demanded by such a pH controlled system would be very difficult to achieve in a traditional reactor system. A nanoreactor would provide distinctive advantage of rapid condition alteration. Thus, the enzyme system in a nanoreactor coupled with pH regulation is a promising one for template free long DNA synthesis. 

Thermodynamics of the scheme

The free energy of the intended reactions would depend on the concentration of the reactants and products. The choice of reactant concentration is determined by the intended concentration of the product. DNA concentration of $10^{-12}$ M is enough for most molecular biology protocols (Sambrook and Russel 2001). In order to synthesize sufficient amount of DNA for molecular biology protocols, here we arbitrarily set the intended product concentration to be $10^{-9}$ M. For the sake of the completeness of the reaction the initial concentrations in Table 1 are proposed.

In the reaction:

$$DNA_n + 3'AcNTP \xrightarrow{TdT, pH_{7.5}} 3'AcDNA_{n+1} + PP$$

$$3'AcNTP \xrightarrow{PP, pH_{5.5}} P + 3'AcNTP$$

$$2P + 3'AcNMP$$

$$3'AcDNA_{n+1} \xrightarrow{AE, pH_{5.5}} AcOH + DNA_{n+1}$$

A phosphodiester bond is synthesized while an ester bond between $\alpha$ and $\beta$ phosphate of the $3'AcNTP$ is hydrolyzed. The $\Delta G^{\circ}$ of synthesizing a phosphodiester bond

### Table 1: Proposed initial concentration of reactants in the synthetic scheme

| Species          | Initial concentration (M) |
|------------------|---------------------------|
| Primer (DNA$_n$) | $10^{-9}$                 |
| 3'AcNTP          | $10^{-8}$                 |
| TdT              | $10^{-8}$                 |
| PP               | $<10^{-12}$               |
| P                | $<10^{-12}$               |
| 3'AcDNA$_{n+1}$  | $<10^{-15}$               |
| DNA$_{n+1}$      | $<10^{-15}$               |
| AcOH             | $10^{-6}$                 |
| H$_2$O           | 55.5                      |
| 3'AcNMP          | $10^{-8}$                 |
| TdT              | $10^{-8}$                 |
bond is +22.2 KJ mol⁻¹ (Dickson et al. 2000) while $\Delta G^{\text{eq}}$ of hydrolyzing bond between $\alpha$ and $\beta$ phosphate of nucleotides is $-32.2$ KJ mol⁻¹ (Voet and Voet 2004). Thus the $\Delta G^{\text{eq}}$ of the above reaction would be very close to $-10$ KJ mol⁻¹. For the concentrations of reactants in table (1) the value of $\Delta G$ is $-69.3$ KJ mol⁻¹.

In the next reaction:

$$\text{ID, pH}7.5 \xrightarrow{\text{PP}} 2\text{P}$$

An ester bond between two phosphate groups of a pyrophosphate is hydrolyzed. $\Delta G^{\text{eq}}$ of this reaction is $-33.5$ KJ mol⁻¹ (Voet and Voet 2004). For the concentrations of reactants in table (1) the value of $\Delta G$ is $-115.1$ KJ mol⁻¹.

The following reaction:

$$3'\text{AcdNTP} \xrightarrow{\text{AP, pH}5.5} \text{P} + 3'\text{AcdNDP} \xrightarrow{\text{AP, pH}5.5} 2\text{P} + 3'\text{AcdNMP}$$

Two high energy phosphate bonds are hydrolyzed. The $\Delta G^{\text{eq}}$ of this reaction is $-65.7$ KJ mol⁻¹ (Voet and Voet 2004). For the concentrations of reactants in table (1) the value of $\Delta G$ is $-181.4$ KJ mol⁻¹.

The following reaction:

$$3'\text{AcDNA}_{n+1} \xrightarrow{\text{AE, pH}5.5} \text{AcOH} + \text{DNA}_{n+1}$$

breaks an ester bond and a H–OH bond while creates a CH$_3$C(O)–OH and a RO–H bond. The bond dissociation energies of ester, H–OH, CH$_3$C(O)–OH and RO–H bonds are 433.1, 497.4, 456.2 and 425.1 KJ mol⁻¹ respectively (Luo 2007). Thus $\Delta H^\circ$ of this reaction would be close to $+49.2$ KJ mol⁻¹. Both the products of the reaction are soluble in water, thus the entropy of the reaction would be positive in aqueous media. Therefore, $\Delta G^\circ$ of this reaction would be less than $\Delta H^\circ$. For the concentrations of reactants in table (1) the value of $\Delta G$ should be less than $-44.2$ KJ mol⁻¹. Moreover, the activation energy of hydrolysis of acetate ester in water is around $+40$ KJ mol⁻¹ at temperature 35–37°C (Aksnes and Libanu 1991). Therefore, the reactants would be kinetically stable in absence of an active catalyst.

### Results

The study on enzyme properties and possible combinations suggests that an enzymatic synthesis of predefined long DNA molecule is theoretically plausible. Based on the available data the process depicted in Fig. 3 is expected to be successful. AcOH would be added to decrease the pH to 5.5 while dialysis in a buffer of pH 7.5 would serve multiple functions, including removal of AcOH, phosphate groups, nucleosides and other small molecules in addition to restoring the pH to 7.5.

Incubation time T1 would depend on the relative concentration of TdT, primer, 3’AcdNTP and ID. The total time required for completing a step of the synthetic scheme would be determined according to the following equation:

$$t_{\text{step}} = \frac{0.6a^2}{D} + \frac{1}{K_{\text{cat}E_0}} \left( K_M \ln \frac{|S_1|}{|S_2|} + (|S_1| - |S_2|) \right)$$

where, $a$ is radius of the reactor, $K_{\text{cat}}$ is turnover number of the enzyme, $D$ is diffusion coefficient of H⁺ ions, $|S_1|$ is initial substrate concentration, $|S_2|$ is intended final substrate concentration, $|E_0|$ is initial concentration of enzyme, $K_M$ is Michaelis constant of the enzyme for the substrate.

In order to complete pH switching within less than 0.1 s the radius of the reactor should be less than $3.94 \times 10^{-3}$ cm with a volume less than $1.92 \times 10^{-7}$ ml or 0.19 nl. As an enzyme catalyzed reaction is designed to take place in nanoliter volume, the reactor is named nanobioreactor. Use of nanobioreactors would greatly reduce the time of pH switching and thus lead to faster synthesis.

Incubation time T2 would depend on the size of the reactor, $K_{\text{cat}AP}$ and $K_{\text{cat}AE}$. The smaller the reactor the less time would be necessary to activate AE and AP through shift in pH. Moreover, a slow pH shift may lead to a transitory period with both TdT, ID and AP, AE enzyme pairs active; thus increase error rate. Therefore a rapid pH change essential. Higher enzyme activity would decrease the time required for complete hydrolysis of excess bases and deprotection of 3’-OH groups. Incubation time T3

---

**Fig. 3** Proposed protocol for DNA synthesis utilizing the scheme

- Primer + TdT + AE + AP + ID
- Add 3’AcdNTP
- Inactivate
- Dialysis
- Incubate T1 (pH=7.5)
- Add AcOH
- Incubate T2 (pH=5.5)
- N cycles
- Activate
- Dialysis
- Incubate T3
- Long DNA molecule

© Springer
would depend mostly on the time required to restore the pH back to 7.5. Thus the size of the reactor would again be an important factor.

Number of possible cycles N would depend on the error rate, which would in turn depend on the completeness of reaction at each step. As reaction of each step in the cycle is thermodynamically favorable, the completeness is in principle warranted, provided with sufficient time. Since the $K_M$ of TdT for oligonucleotide primer is in the range of $10^{-4}$, at equivalent molar ratio with primer TdT is expected to make less than 1 in 10,000 errors. The error rate may however increase if incubation time $T_1$ is longer due to slow pH shift. Thus faster pH shift would maintain high fidelity.

**Discussion**

The scheme proposed here is a potentially efficient enzymatic method for *de novo* DNA synthesis. From available data on the thermodynamics of the reactions and properties of the involved enzymes, the scheme is theoretically feasible. However, experimental demonstration would be crucial for its practical application.

A major difficulty may be the low turnover number of TdT with 3’AcdNTP as substrate. From the reported results of incorporation of modified bases by TdT (Deng and Wu 1983; Tu and Cohen 1980; Vincent et al. 1982; Kumar et al. 1988; Igloi and Schiefermayr 1993), it is inferred that TdT would accommodate an acetyl group. Recombinant enzymes are likely to be available to circumvent the problem in case native TdT does not incorporate 3’AcdNTP. Recombinant enzymes may also lead to a TdT with higher affinity to primers further limiting the error rate.

Nanobioreactors with capabilities of rapid change in pH would offer great advantage not only for this scheme but also for synthetic methods requiring sequential temporal activity switching of different sets of enzymes. Enzymatic correction (Carr et al. 2004) may be used on the longer DNA sequences. This combination may limit the error further down to 1 in 150,000. The availability of longer DNA for gene assembly would significantly increase the power and range of synthetic genomics. Not only small microorganism genomes, large genomes of higher organisms may also become amenable to synthesis with longer starting material.

**Acknowledgement**  This study was carried out with support from Ashtabhuj Systems.

**Conflict of interest**  The author declares that he has no conflict of interest.

**Open Access**  This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.
