Error-free synthetic DNA by molecular dictation

Authors:

Philipp Knyphausen\textsuperscript{1,2}, Laurens Lindenburg\textsuperscript{1,3} & Florian Hollfelder\textsuperscript{*1}

Affiliation:

\textsuperscript{1}Department of Biochemistry, University of Cambridge, 80 Tennis Court Rd, Cambridge CB2 1GA, UK.

Current addresses:

\textsuperscript{2}CRISPR and Genome Engineering, Biologics Research, Bayer AG, Cologne, Germany.

\textsuperscript{3}Genmab BV, Uppsalalaan 15, 3584 CT Utrecht, Netherlands

*Correspondence: fh111@cam.ac.uk (F. Hollfelder)

Orcids:

PK: 0000-0002-2544-7920

LL: 0000-0001-6775-8041

FH: 0000-0002-1367-6312

Lab website: https://www.bioc.cam.ac.uk/hollfelder

Twitter: @hollfelderlab

Keywords:

DNA synthesis, error correction, synthetic biology
Abstract

Synthetic DNA is the linchpin of the rapidly accelerating biotechnological era and is perhaps the most promising candidate for long-term digital data storage. Despite huge advances, manufacturing error-free DNA at low cost and high throughput remains challenging. Borrowing from well-established sequencing-by-synthesis technologies, we describe a new solution for DNA error-correction.

Synthetic DNA for a biotechnological future

There is growing demand for synthetic DNA in a number of technological areas, such as synthetic biology, nanotechnology and digital data storage. However, while a steep decline in DNA sequencing costs, mainly enabled by reversible terminator chemistry-based DNA sequencing-by-synthesis (SBS) [1], facilitated a plethora of biological breakthroughs in the past decade, DNA synthesis has not yet undergone such rapid progress. The reason for this ‘read-write gap’ boils down to the problem of noise or inaccuracy, which can be dealt with relatively easily in DNA sequencing – e.g. through over-sampling – but is problematic for most applications where synthetic DNA is needed: just one mutation can have grave functional consequences [2]. Thus, errors in synthetic DNA have costly consequences in the physical world, for example, by prolonging screening campaigns to identify a desired phenotype.

De novo DNA synthesis: Old chemistry pushed to its limits

With current methods, ‘long’ DNA molecules (>300 bp) cannot be de novo synthesized directly but have to be assembled from multiple shorter oligonucleotides. Phosphoramidite chemistry, a ~40 years-old and still absolutely dominant DNA synthesis
method, has a relatively high error rate (about $10^{-2}$ to $10^{-3}$ per cycle), limiting synthesis lengths and/or requiring sophisticated purification procedures [3].

Post-synthesis oligonucleotide purification is only partially effective against substitutions, short deletions and unwanted base modifications, which can result from the rather harsh chemistry. Even more challenging is purification of material from micro-array-based, high-throughput systems, where oligonucleotides on the order of 10-30k are synthesized on a single chip and subsequently assembled into hundreds to thousands of longer (>500 bp) gene products. Fragment assembly has been performed in a (sub-) pooled fashion [4] or by construct-specific compartmentalization [5]. Unfortunately, all these assembly strategies introduce new sources of errors.

**Enzymatic DNA synthesis to the rescue?**

These length limitations and the environmental impact of phosphoramidite synthesis have spurred renewed interest in enzymatic methods for *de novo* DNA synthesis, employing the activity of template-independent DNA polymerases such as terminal nucleotidyltransferase, TdT. However, despite a surge of new companies and capital investment, no published study or commercial provider has demonstrated an enzymatic DNA synthesis process that can rival phosphoramidite-based synthesis [6]. Thus, as long as enzymatic DNA synthesis methods are still immature, errors will remain a significant challenge for the foreseeable future, especially for the most ambitious synthesis projects: the assembly of entire genomes [7].

**The state-of-the art: How to correct errors in synthetic DNA**

Current correction methods can be categorized into either enzymatic removal of error-containing sequences or so-called ‘dial-out’ techniques. Enzymatic error correction relies on the activity of DNA-binding or -cleaving enzymes that recognize mis-
matched regions of double-stranded DNA (dsDNA). Assemblies of multiple oligos are heated and allowed to re-anneal, thereby producing a mixture of perfectly matched dsDNA and dsDNA containing mismatches or bulges at erroneous sites (as most errors occur randomly, two matching errors finding each other is unlikely). Treatment of this mixture with mismatch-detecting enzymes (such as MutS) allows for the removal of the undesired error-containing species. However, even in a low-complexity test set-up, only about 60% of the resulting sequences obtained are perfect [8].

‘Dial-out’ describes the coupling of microarray-based oligo synthesis and gene assembly with next-generation-sequencing (NGS) of the assembly output. By including molecular barcodes for each assembly, perfect sequences can be selectively amplified with corresponding barcode-specific primers [9]. Alternatively, laser-based picking of correct clones from an NGS microchip has been demonstrated [10]. Dial-out methods are ultimately limited by the read length and accuracy of the sequencing method being used (originally ~500 bp, when used with short-read SBS). With improving accuracy of long-read sequencing methods (e.g. nanopore or SMRT sequencing), dial-out techniques may become more viable in the future.

A new solution for error correction in synthetic DNA: Dictate the sequence

A curious omission in the repertoire of error correction methods is that the combination of nucleotide chemistry and template-dependent DNA polymerase fidelity that enabled the success of SBS-based approaches has not been leveraged for error correction of synthetic DNA. We term synergistic application of these well-established strands of research ‘CRESS’ (controlled reverse-strand synthesis). CRESS relies on the controlled, template-dependent, enzymatic synthesis of a strand complementary to an initial, chemically or enzymatically synthesized strand, followed by subsequent enrichment of error-free DNA. Depending on the nucleotides being used, this ap-
approach can address all types of errors (deletions, insertions and substitutions), is fully compatible with microarray-based synthesis and may be used for error correction of oligonucleotides or fragment assemblies.

In CRESS, enzymatic reverse-strand synthesis is carried out in cycles on a solid support. In each cycle, nucleotides are provided according to the next expected template base (Fig. 1a). If a strand presents an incorrect template base, the synthesis of its reverse strand will ultimately fail. Thus, erroneous DNA will remain partially single-stranded, which serves to discriminate it from error-free DNA in subsequent enrichment steps.

**CRESS can correct all types of errors**

The kind of error that can be addressed with CRESS depends on the type of nucleotide being used (Table 1). With natural dNTPs, only substitutions and insertions can be addressed and reverse-strand synthesis is carried out in a semi-controlled manner. This means synthesis will not immediately terminate once a substitution or an insertion is encountered. Instead, the synthesis of these erroneous strands will ‘dephase’ because the sequence of the template will no longer be synchronized with the nucleotides provided. In case of deletion errors, dephasing would not occur since strands with such errors would just pause until the next cycle. Another drawback of natural dNTPs is poor compatibility with array-based or otherwise parallelized synthesis workflows, which require control over nucleotide incorporation at each spot. Nevertheless, natural dNTPs are attractive in terms of incorporation efficiency and cost.

Another option is the use of reversible terminator deoxynucleotides (rt-dNTPs), which are the basis of many sequencing-by-synthesis approaches (SBS, e.g. on the Illumi-
na platforms). The incorporation of rt-dNTPs happens one-by-one since each nucleotide carries a blocking group that prevents further extension by the polymerase. Blocking groups are labile under certain conditions and can be removed in a controlled manner (e.g. by applying light irradiation or reducing agents). Despite offering spatial and temporal control over the synthesis process, rt-dNTPs, like dNTPs, do not enable selective removal of deletion-containing templates during reverse-strand synthesis.

In order to address all types of sequence errors, we propose a ‘decoy’ strategy (Fig. 1b). In this approach, reverse-strand synthesis is carried out with rt-dNTPs in a controlled manner as discussed above, allowing extension of correctly templated strands. In addition, ‘decoy’ dideoxynucleotides (ddNTPs) of the three other bases are directly included in the synthesis mixture or provided separately in a subsequent synthesis cycle. For any strand presenting a wrong template base, the reverse-strand will thus irreversibly terminate. With the decoy strategy, even the most challenging deletions – those occurring within homopolymer stretches – are caught eventually due to an incorrect number of identical bases in a row.

**CRESS for ‘trits’-based DNA data storage**

CRESS-based error correction with natural dNTPs and decoy ddNTPs may be immediately useful for a recently described DNA storage method, where information is encoded as the three possible transitions between homopolymer stretches i.e. ‘trits’ [11]. In this method, DNA synthesis is carried out with two competing enzymes, TdT and apyrase, with the activity of the latter eventually stopping any further extension due to substrate depletion. For error correction of these synthesis products, erroneous strands could be removed by providing decoy ddNTPs and matching natural dNTPs in an alternating fashion (Fig. 1c). CRESS would thus improve the accuracy
of the DNA-encoded digital data storage system, while requiring only relatively simple set of cheap chemicals (dNTPs and ddNTPs).

**Selective enrichment or removal to retrieve the correct sequence of interest**

After enzymatic reverse-strand synthesis has been carried out, erroneous strands are selectively depleted or, conversely, error-free strands are selectively enriched. Nucleases with specificity for ssDNA are an obvious choice for selective degradation of erroneous strands. In turn, dsDNA-specific endonucleases or a polymerase chain reaction are suitable for selective elution or amplification of error-free DNA, respectively.

**What are the limitations and implications of CRESS?**

An important consideration for CRESS are inefficiencies in rt-dNTP chemistry (e.g. incomplete terminator removal), which also contribute to the ~250 bp read length limitation of SBS technologies [12]. For rt-dNTPs-based CRESS, such factors may result in incomplete reverse-strand synthesis of error-free templates, thereby decreasing overall post-error correction yield. Depending on the sensitivity of the enrichment or depletion steps, longer maximal synthetic lengths may be achieved with CRESS than with SBS. While each sequence must essentially be synthesized twice in CRESS, it is not clear whether this will reduce commercial viability. Finally, an attractive option could be the application of CRESS to non-natural xenobiotic (not necessarily phosphodiester-linked) nucleic acid backbones [13], for which synthesis may be highly efficient. Such a synthetic process could, for example, circumvent the challenge of improving the already highly optimized phosphoramidite chemistry even further.

**Conclusion**
CRESS has the potential of becoming a versatile approach for error correction of synthetic DNA. It relies on proven chemistries and the accompanying optimized, high fidelity, template-dependent DNA polymerases from SBS approaches, minimizing obstacles to its implementation. Moreover, CRESS can be adapted to different synthesis platforms (chemical or enzymatic, with or without massively parallelized oligonucleotide synthesis) and expected error profiles. The cost of highly accurate de novo DNA synthesis will have to fall considerably to make ambitious goals such as DNA-based data storage and whole-genome synthesis to become an everyday reality. Efficient error correction methods will remain highly relevant for the foreseeable future and CRESS may play an important role in this area.

Acknowledgements
We acknowledge support by the EU (PK: FP7 IAPP HOTDROPS, 324439; LL: H2020 Marie-Curie Fellowship, 659029; FH: H2020 ERC Advanced Investigator, 695669).

References


**Table 1.** Error types addressed with different nucleotide mixtures.

<table>
<thead>
<tr>
<th>\ Nucleotide Error type</th>
<th>Single-bp level synthesis</th>
<th>Trits-based storage**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dNTPs rt-dNTPs Natural dNTPs w/ ddNTP decoy rt-dNTPs w/ ddNTP decoy</td>
<td>Natural dNTPs w/ ddNTP decoy</td>
</tr>
<tr>
<td>Substitutions</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓</td>
</tr>
<tr>
<td>Insertions</td>
<td>✓ ✓ ✓ ✓</td>
<td>not relevant</td>
</tr>
<tr>
<td>Deletions</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓</td>
</tr>
<tr>
<td>Deletions in homopolymer stretches</td>
<td>✓</td>
<td>not relevant</td>
</tr>
</tbody>
</table>

*Deletion errors are caught at the end of homopolymer stretches.

**Trits-based storage system as described by Lee et al. (2019), see main text.*
Figure 1. (A) Schematic overview of controlled reverse-strand synthesis for the enrichment of error-free DNA. From left to right: DNA from different sources can be used as starting material and is immobilized on a solid phase if not already the case. A primer is annealed and extended at the 3'-end using a high-fidelity DNA polymerase and by providing the expected nucleotide at each cycle. The nucleotides being used can be (combinations of) reversible terminators, natural dNTPs and/or irreversible terminators, depending on the expected types of errors (see main text). The controlled reverse-strand synthesis leads to lagging or premature termination of erroneous templates and leaves them vulnerable to single-strand specific enzymatic deple-
tion. (B) Detailed scheme of the CRESS approach: Primers are hybridized to immobilized de novo synthesized ssDNA and extended through cycles of deprotection and nucleotide addition. As an example, 3’ immobilization is shown, as is the case after phosphoramidite-based synthesis, but 5’ immobilization is also possible. *Left-hand panel:* thus far, all depicted strands have successfully been extended enzymatically (as no error in the de novo synthesized strand was encountered by this stage) and are reversibly terminated. *Middle panel:* in preparation for the next CRESS cycle, the newly incorporated nucleotides are deprotected. *Right-hand panel:* in the next CRESS cycle, a mixture of the correct, template-complementary, reversibly terminated nucleotide (rt-dCTP) is provided together with a mixture of three decoy, irreversibly terminated nucleotides (ddATP, ddGTP and ddTTP). The correct strand is enzymatically extended by the reversibly terminated nucleotide and therefore available for extension in the next cycle while the erroneous strands are irreversibly terminated by the decoy ddNTPs. (C) CRESS workflow adapted to synthesis products of trits-based DNA data storage system described by Lee et al. (2019). *Left-hand panel:* Hybridized primers were extended through an initiating cycle with dATP. *Middle panel:* After extension with decoy dideoxynucleotides (ddATP, ddCTP, ddTTP), erroneous strands presenting wrong template bases are irreversibly terminated while the correct strand remains unchanged. *Right panel:* Addition of dGTP led to the extension of the correct template strand until the next transition.