Minimum Information About a DNA Construct (MIADNA)

Jonathan Calles
Stanford Bioengineering, jecalles@stanford.edu

Marc Salit
JIMB, msalit@stanford.edu

Anybody Else?
Affiliation TBD, email@TBD.xxx

July 16, 2019

Abstract

Here, we propose a minimum information standard to describe DNA. Our goal is to facilitate the distributed production and use of DNA by standardizing the description of the following: the design of DNA, the processes used to make DNA, the measurement of DNA products, and the downstream use of the final product.

Contents

1 Background

Previously, the tasks of designing, building, testing, and using DNA constructs were performed by the same individual or group. This obviated the need for a consistent and standard method for communicating information between individuals performing each task. As the scope and scale of genetic engineering expands, however, parties are becoming specialized at individual tasks along the DNA construction pipeline. This division of labor enables more rapid development of increasingly complicated genetic designs but requires communicating complicated and often highly technical information between parties in a rapid, reliable, and consistent manner. Here, we introduce the Minimum Information About a DNA Construct (MIADNA) standard, which is intended to standardize the communication of any information relevant to the design, construction, validation, and use of DNA constructs.

2 Stakeholders

A biotechnology sector supporting fully disaggregated production of DNA separates the processes of design, construction, verification, and use such that no one party would necessarily perform all these processes individually. We therefore define four roles that parties involved in the DNA production pathway may take on.

A **DNA designer** is an individual or group who designs DNA sequences encoding a set of biological functions specified by a downstream user.

A **DNA constructor** is an individual or group who makes a DNA construct from a design. This party may perform DNA synthesis, library construction, gene assembly, or transformation. This party may also prepare the product for storage and transport, or perform any other process involved with physically realizing the DNA object.

A **DNA verifier** is an individual or group who measures the final product made by DNA constructors. This party validates the identity and purity of the final product and reports their results to relevant parties.

A **DNA user** is an individual or group who receives constructed and validated DNA for downstream use. This party may or may not be involved with the design, construction, or validation of the DNA.

3 Scope of MIADNA Standard

MIADNA aims to standardize the design of DNA sequences. A MIADNA compliant design should completely specify a DNA sequence from design, through construction and validation, to use. MIADNA describes three DNA objects: a **requested sequence** produced by DNA designers, a **designed sequence** to be made by DNA constructors, and a **synthesized sequence** that was actually made by the DNA construction process. MIADNA also describes three processes: the **design rules** describing constraints imposed by DNA constructors on input sequences, the **making** of the physical DNA,

and the **characterization** of the final product. The **Requested Sequence** object is the annotated nucleotide sequence of a DNA construct(s) to be synthesized, produced by a DNA designer. The Requested Sequence also includes designer imposed constraints on the final product, specifying how tolerant each genetic component is to change.

The **Design Rules** specify the design constraints imposed by DNA constructors on DNA sequences to be made. These constraints are specific to the processes used to make the DNA.

The **Designed Sequence** object describes the nucleotide sequence that a DNA constructor will attempt to make. The Designed Sequence results from modifying the Requested Sequence to satisfy both user tolerances and the constraints specified in the Design Rules.

Making describes the construction processes used by DNA constructors to physically realize the Designed Sequence. Making also logs the individual steps of the construction process for future reference.

Characterization describes the methods used to analyze the resulting Synthesized Sequence, as well as the results of these analyses.

The **Synthesized Sequence** object describes the nucleotide sequence made by DNA constructors, as well as the format of the final product (e.g., lyophilized DNA, E.coli stab, etc.) and instructions for handling the product (i.e., MSDS, storage conditions, etc.) .

4 Requested Sequence: What you wish to make

The Requested Sequence object describes the nucleotide sequence of the DNA construct to be synthesized, as well as metadata pertaining to the identity and biological function of the individual genetic components comprising the construct. These metadata include gene and part annotations, user-specified design tolerances, and results from IP and biosafety/biosecurity screens. The Requested Sequence is produced by a DNA designer for production by a DNA constructor.

4.1 Sequence

The "Sequence" field describes the nucleotide sequence that the DNA designer wishes to be made. This sequence should be described using a format that is both machine and human readable, either at the level of individual nucleotides (e.g., FASTA format) or abstracted to the level of biological parts (e.g., SBOL language).

4.2 Annotation

The "Annotation" field describes the Requested Sequence at the level of parts and components. This field includes names and biological functions assigned to component of the Requested Sequence, as well as metadata documenting the design tolerances of each component. This data should be stored in an accepted standard format such as GenBank files when possible. Designers should use standard ontologies when appropriate; see the Open Biological and Biomedical Ontology (OBO) Foundry for a comprehensive list of appropriate ontologies (http://www.obofoundry.org/).

A DNA designer may wish to flag segments of the Requested Sequence whose function are particularly intolerant to changes to the specified nucleotide sequence (e.g., restriction sites, catalytic sites of enzymes, etc.). Conversely, some segments of the Requested Sequence may have less restrictive sequence constraints, only requiring for example a particular length or GC content. Some segments of the Requested Sequence may be even be interchangeable with functionally similar or equivalent parts (e.g., promoters with similar expression profiles, restriction enzyme sites that are compatible with the same assembly method). MIADNA annotation metadata should support user specified design tolerances for the nucleotide sequence of the construct, aiding downstream constructors in optimizing a Requested Sequence for their particular synthesis and assembly protocols.

5 Design Rules: Process transforming requested sequences to what can be made

Due to the constraints of DNA construction processes, not every nucleotide sequence is makeable as designed. DNA designs may need to be altered to physically realize the final product. The Design Rules object documents the synthesis and assembly methods used to realize the final product, and the tolerances and limitations of these processes as implemented by the DNA constructors. The Design Rules inform how a Requested Sequence is converted into a makeable Designed Sequence.

5.1 Design Tolerances

The "Design Tolerances" field documents constraints specific to the processes used by DNA constructors that limit what nucleotide sequences can be made. This field includes parameters such as GC limits, mono- and polynucleotide repeat limits, synthesis fragment length limits, and forbidden sequences (e.g., incompatible with restriction enzymes used in assembly process).

5.2 Screening

Ab initio gene synthesis has the potential to facilitate sourcing dangerous or illicit biological materials [] [Nouri and Chyba, Nat. Biotech. 2009]. As a result, it has become common practice for DNA synthesis groups to screen incoming orders against externally curated databases of controlled sequences (see: International Gene Synthesis Consortium and their Harmonized Screening Protocol).

More practically, DNA constructors may wish to screen incoming Requested Sequences in order to inform the optimal synthesis and assembly of their backlog. For example, a DNA design may include some commonly used components that a DNA constructor might have in inventory (e.g., the green fluorescent protein GFP, or a standard promoter), as well as other components that might need to be purpose-built. Additionally, for some sets of

Requested Sequences varying only at a particular locus of interest, efficient library preparations may be more appropriate than *de novo* synthesis of individual gene variants.

The "Screening" field describes the results of any pre-construction bioinformatic screenings performed by DNA constructors, as well as metadata pertaining to the screening methods used.

6 Designed Sequence: What will be made

The Designed Sequence object describes the nucleotide sequence that DNA constructors aim to make, as well as metadata pertaining to the construction of the physical object. The Designed Sequence results from applying all changes specified by the Design Rules to the Requested Sequence in order to satisfy design tolerances set by both objects.

6.1 Design Log

6.2 Changes From Requested Sequence

If a Requested Sequence object complies with the design tolerances imposed by the Design Rules object, then the Requested Sequence can be converted to a Designed Sequence object without any changes. However, some Requested Sequences may not be realizable as designed and must be altered within the Specified Tolerances specified by the DNA designer under the "Annotation" field. The "Changes from Requested Sequence" field documents what alterations are applied to the Requested Sequence in order to satisfy the tolerances set by the Design Rules. These may include artifacts of the construction processes (e.g., assembly scars), changes to nucleotide sequences of individual genetic components (e.g., synonymous recoding in protein-coding sequences), substitution of equivalent or comparable genetic components (e.g., changing designed restriction sites to comply with assembly process), or other alterations. These alterations should be reported in a

machine and human readable format such as an annotated variant call list using the VCF format (https://github.com/samtools/hts-specs).

7 Making

The making object describes the processes used by the DNA constructor to make the Synthesized Sequence. It also logs the production of the Synthesized Sequence.

7.1 Synthesis Process Description

Some Designed Sequences may specify genetic components that for one reason or another must be chemically synthesized *ab initio*. The "Synthesis Process Description" field documents the processes used to chemically synthesize these segments and metadata related to these processes, including expected yield, purity, error-rate, etc.

7.2 Assembly Process Description

Unless the Designed Sequence is synthesized de novo in its entirety, the constitutive components of the DNA construct must be assembled together. The "Assembly Process Description" field documents the assembly processes used to assemble the final product, as well as metadata related to these processes.

7.3 Production Log

The "Production Log" field documents the step-by-step construction of the DNA. This field may include any automatically generated, process specific event logs, as well as the results of any intermediate quality checks and timestamps for individual steps. The Production Log should contain enough information for a competent biotechnician to debug any errors that may arise

during DNA construction or afterwards during downstream use of the final product. A DNA constructor who is wary of revealing IP relating to specific processes used during DNA construction may wish to leave this field blank, or only leave it filled during internal use.

8 Characterization: How the physical product was made

The Characterization object documents the measurement methods used by the DNA verifier to observe the sequence, yield, and purity of the Synthesized Sequence, as well as metadata pertaining to these methods.

8.1 Sequence Measurement

The "Sequence Measurement" field describes any measurements performed on the genetic material produced by the DNA construction process, as well as metadata pertaining to those processes.

8.2 Verified Sequence

The "Verified Sequence" subfield documents the sequencing process used to verify the final product. It should also store metadata pertaining to the sequence verification protocol, including expected accuracy and any limitations or biases of the protocol or of downstream data processing as performed by the DNA verifier.

8.3 Clonality

The clonality of a set of biological object refers to the degree to which those objects are genetically identical within the set. For example, a bacterial cul-

ture is said to be "monoclonal" if all individuals in the culture arose from the asexual replication of a single individual and are thus genetically identical, save for the few mutations arising during said replication events. The "Clonality" subfield documents the expected clonality of the Designed Sequence given the synthesis and assembly methods used. Clonality also documents the methods used by the DNA verifier to assess the clonality of the final product, as well as metadata related to those methods.

8.4 Yield Measurement

The "Yield Measurement" field documents the amount of physical DNA produced (i.e., by mass and by moles) as well as the process used to measure the amount of DNA and metadata pertaining to that process.

8.5 Purity Measurement

The "Purity Measurement" field documents the chemical purity of the final product, i.e., what fraction of the final product is composed of Designed Sequence as opposed to other nucleotide sequences resulting from errors during the synthesis and assembly processes. This field also documents the method by which chemical purity is measured and metadata pertaining to that method.

8.6 Contaminants

The "Contaminants" field documents the identity and amount of material in the final product that does not correspond to intentionally synthesized and assembled DNA (contaminating materials). These contaminating materials may include the following: nucleic acids, proteins, or other biological molecules; organic or inorganic small molecules; bacteria or other single-cellular organisms; bacterial or fungal spores; or other materials unintentionally introduced during the synthesis and assembly processes. Contaminating materials do not include genetic materials produced intentionally by DNA

constructors whose nucleotide sequence varies from the Designed Sequence due to errors during the synthesis or assembly processes. The Contaminants field also documents the processes used to identify contaminating materials and their relative amounts, as well as metadata pertaining to these processes.

8.7 Other Measurements

The "Other Measurements" field documents any additional processes used to characterize the final product, as well as metadata pertaining to these processes.

9 Synthesized Sequence: What was made

The "Synthesized Sequence" object describes the physical product made by the DNA constructor for the DNA user, including the measured nucleotide sequence and clonality of the final product, the format of the final product (e.g., lyophilized, bacterial stab, etc.), and recommended storage and handling conditions.

9.1 Measured Sequence

The "Measured Sequence" field describes the results of the Sequence Measurement performed by the DNA validator. This includes the observed nucleotide sequence and clonality of the final product.

9.2 Molecules

The "Molecules" field describes the inorganic, organic, and biological molecules intentionally included in the final product. This includes the identities and quantities of all relevant compounds, as well as the methods used to verify these compounds and metadata related to these methods. Molecules does

not include contaminating materials introduced unintentionally during DNA construction (see: Characterization/Contaminants).

9.3 Organism

If the final product is stored in an in vivo format (e.g., as an E. coli stab), then the "Organism" field optionally characterizes the organism used. This field should include the species and strain used to carry the final product, as well as any additional information known for that strain (e.g., growth characteristics, genotype, etc.).

9.4 User Guide

The "User Guide" field documents recommended handling instructions for DNA users. If the Synthesized Sequence is integrated into a plasmid, or is synthesized as pDNA, then the User Guide field should include the annotated sequence of the plasmid as well as metadata describing its specifications (e.g., origins of replication and termination, selective markers, multiple cloning sites, copy number, etc.). The User Guide field should also include instructions for the handling of the Synthesized Sequence in its delivered format, including safety information and protocols for downstream cloning or use of the Sequence.