SNIA®

Advancing storage & information technology

DNA Data Storage Sector Zero

Version 1.0

ABSTRACT: This specification defines the recommended method and embodiment for storing basic vendor and CODEC information (sector zero contents) within a DNA data storage archive for the purpose of enabling an archive reader to then consume archive metadata (sector one) and data contents.

This document has been released and approved by SNIA. SNIA believes that the ideas, methodologies, and technologies described in this document accurately represent SNIA goals and are appropriate for widespread distribution. Suggestions for revisions should be directed to https://www.snia.org/feedback/.

SNIA Standard

November 11, 2023

USAGE

Storage Networking Industry Association (SNIA) hereby grants permission for individuals to use this document for personal use only, and for corporations and other business entities to use this document for internal use only (including internal copying, distribution, and display) provided that:

1. Any text, diagram, chart, table or definition reproduced shall be reproduced in its entirety with no alteration, and,

2. Any document, printed or electronic, in which material from this document (or any portion hereof) is reproduced shall acknowledge SNIA copyright on that material, and shall credit SNIA for granting permission for its reuse.

Other than as explicitly provided above, you may not make any commercial use of this document or any portion thereof, or distribute this document to third parties. All rights not explicitly granted are expressly reserved to SNIA.

Permission to use this document for purposes other than those enumerated above may be requested by e-mailing tcmd@snia.org. Please include the identity of the requesting individual and/or company and a brief description of the purpose, nature, and scope of the requested use.

All code fragments, scripts, data tables, and sample code in this SNIA document are made available under the following license:

DISCLAIMER

The information contained in this publication is subject to change without notice. SNIA makes no warranty of any kind with regard to this specification, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. SNIA shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this specification.

1 Overview

Deoxyribonucleic acid (DNA), when used as a data storage media, lacks key properties found in traditional data storage media (e.g., tape, SSD, HDD). DNA data storage is commonly based on short strings of DNA called oligonucleotides (oligos) that are mixed together without a specific physical ordering scheme, whereas traditional data storage media types are based on magnetic media (e.g., hard disk drives and tape) or silicon microchips (e.g., NAND). DNA storage media lacks a dedicated and integrated controller serving as a natural point of interconnect and interface between the consuming system and the media. Further, DNA storage media lacks an organizational means by which proximity of one media subcomponent can be understood in relation to another media subcomponent.

The goal of this specification is to provide a standard means of embedding two vital pieces of information into the archive to enable an archive reader to understand:

- Who (vendor) wrote the archive
- How (CODEC) to read the metadata pertaining to the archive (sector one)

2 Sector Zero vs Sector One

To provide clarity and context, sector zero is intended to provide readers with the ability to access sector one. Sector zero indicates who wrote the archive, and with which CODEC, and sector one provides details about the logical structure of the archive and the CODEC used to read the archive contents.

The idea behind this division comes from the need to represent some starter data without the need of a CODEC. Sector zero shall fit into a single oligonucleotide and sector one shall span multiple oligonucleotides as it naturally contains much richer data and will need a CODEC to assemble the oligos correctly and decode them into readable data. This specification focuses only on sector zero.

3 Assumptions

The intent of sector zero is to supply the archive reader with the minimal amount of information that will help them read sector one and the rest of the archive, specifically sector zero contains information pertaining to who wrote the archive and what CODEC should be used to read the contents of sector one, which is the metadata needed to read the archive contents themselves. The following are the assumptions an archive reader and writer should make in following this process.

- It is assumed that each oligonucleotide contains 150 bases.

- It is assumed that sector zero payload is contained fully within a single oligonucleotide.

- It is assumed that the sector zero oligonucleotide will be comprised only of bases found in natural DNA, that is adenine (A), cytosine (C), guanine (G), and thymine

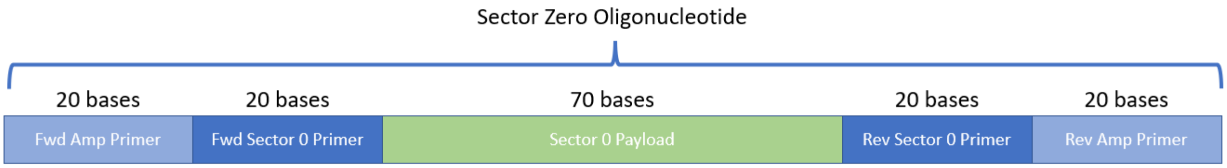
(T).

- It is assumed that the vendor/CODEC lookup table will be documented and accessible as part of the spec.

- It is assumed that the amplifying primers of sector zero are universally unique, specified by the alliance, and should not be used within other parts of the archive.

- It is assumed that, of the 150 bases:

  - 20 bases are used for each of the forward and reverse amplifying primer.
  - 20 bases are used for each of the forward and reverse sector zero primer.
  - 70 bases are used for sector zero contents.



It is further assumed that consumption of sector zero may not be mandatory for cases where the vendor and CODEC are known a priori to consuming the archive. However, for cases where the vendor and CODEC are not known a priori, sector zero serves as a reliable mechanism for identifying these key pieces of metadata. It is recommended that sector zero be included within every archive, regardless of whether or not the information contained therein is conveyed to the archive reader in another manner.

4 Sector Zero Schema

The 70 bases contained within the payload section of the sector zero oligonucleotide shall have the following schema:

- The first 35 bases, represented as a string of single-letter characters (e.g., ACGT) shall serve as a key to identify the vendor from the authoritative table described below

- The second 35 bases, represented as a string of single-letter characters shall serve as a key to identify the CODEC from the authoritative table described below.

If the archive writer has opted to not register their vendor information to receive a vendor ID, the archive writer should use the following value in the first 35 bases:

- `ACGACACAGTGATCATGCAGTCTCTATAGAGATCT`

If the archive writer has opted to not register their CODEC information to receive a CODEC ID, the archive writer should use the following value in the last 35 bases:

- ATAGTCTCTGATCACTCACGTATGTGCGTGAGCTA

5 Reading Sector Zero

Generally sector one should only need to be read in cases where the external representation of sector one (refer to the sector one specification) is not available. In many cases, archive owners will understand a priori the contents, CODEC used, and other metadata, as this information is likely stored in an external archival management system.

In cases where sector zero needs to be read from DNA, use the following process:

1. Amplify sector zero payload using sector zero primer sequence listed below.
2. Refer to the authoritative table of vendors and CODECs using the data found within the sector zero payload to identify the vendor and sector one CODEC (see section 1.6)

Note that the diagrams above indicate the presence of a forward and reverse amplifying primer. These should only be used in cases where you need to amplify not only the sector zero contents, but also the primers that encapsulate those contents. The amplifying primer for sector zero may not be unique to sector zero.

For sector zero, the following primers should be used. These primers have been selected to mitigate common issues related to GC content, homopolymers, self-dimers, and homo-dimers.

- Forward primer: GCCTCGGTACACGGTATGAG
- Reverse primer: ATGCTCCAGTTCGGTCAGTG

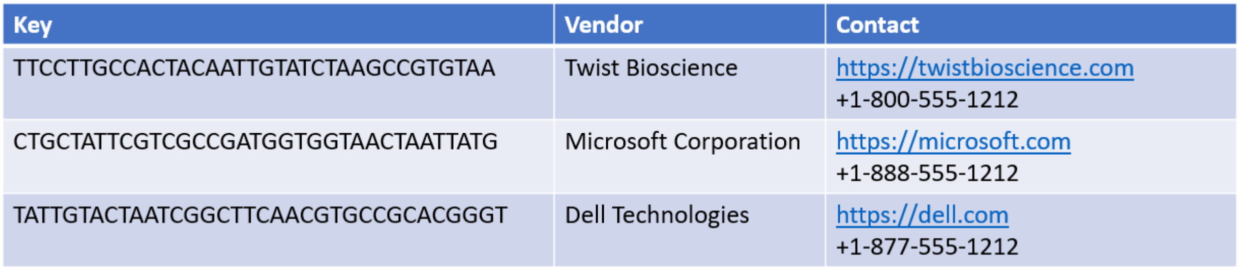
6 Authoritative Tables of Vendors and CODECs

The governing body shall maintain two authoritative tables, one of vendors and one for CODECs, to enable archive readers to ascertain who wrote an archive and which CODEC was used when writing the logical structure and metadata (sector one).

The vendor table shall provide a lookup wherein the first 35 bases from the sector zero payload serve as a key. The associated value will provide metadata about the vendor responsible for writing the archive.

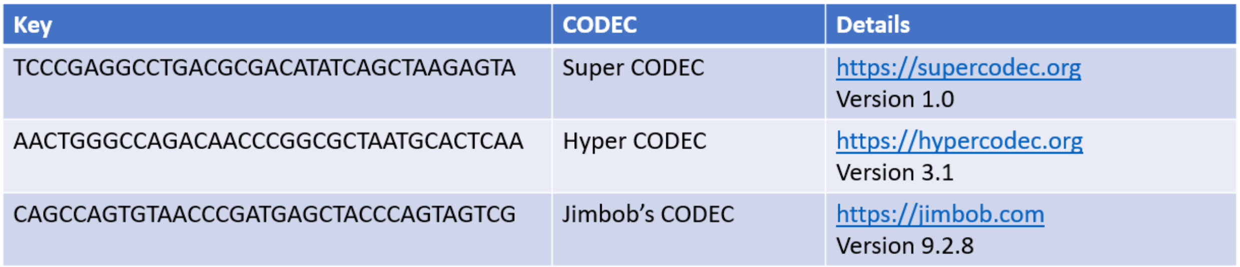
The list of properties found in these tables shall be defined separate from this specification, and documentation related to accessing the service is forthcoming. It is assumed that these tables will be made available in several formats, including document, print, online, and accessible via API. Note that the values in the table are not actual values, but examples.

The values used for keys in these tables will be managed by the governing body, be 35 bases in length, and will be generated ensuring a minimum edit distance of at least 10 between any two keys.



Key	Vendor	Contact
TTCCTTGCCACTACAATTGTATCTAAGCCGTGTAA	Twist Bioscience	https://twistbioscience.com +1-800-555-1212
CTGCTATTCGTCGCCGATGGTGGTAACTAATTATG	Microsoft Corporation	https://microsoft.com +1-888-555-1212
TATTGTACTAATCGGCTTCAACGTGCCGCACGGGT	Dell Technologies	https://dell.com +1-877-555-1212

The CODEC table shall provide a lookup wherein the second 35 bases from the sector zero payload serve as a key. The associated value will provide details about the CODEC used to write sector one into the archive. Note that the values in the table are not actual values, but examples.



Key	CODEC	Details
TCCCGAGGCCTGACGCGACATATCAGCTAAGAGTA	Super CODEC	https://supercodec.org Version 1.0
AACTGGGCCAGACAACCCGGCGCTAATGCACTCAA	Hyper CODEC	https://hypercodec.org Version 3.1
CAGCCAGTGTAACCCGATGAGCTACCCAGTAGTCG	Jimbob's CODEC	https://jimbob.com Version 9.2.8

It is important to note that vendors are encouraged, but not required, to supply CODEC details to the governing body. If a vendor chooses to not supply this information, the governing body suggests that they maintain and make readily available CODEC information for their archive readers. Further, the governing body suggests clients working with such archive writers be fully aware that should this information not be published, it could render an entire archive completely unreadable if the CODEC for sector one is unable to be discerned.

For cases where the authoritative service is not used, is unavailable, or, you are working with a local copy of the vendor IDs and CODEC IDs, the following process should be followed:

- If the total payload retrieved from the archive is not equal to 70 bases, use the left-most 35 bases as the vendor ID and the right-most 35 bases as the CODEC ID. Insertions and deletions could happen anywhere on the payload, and the assumption is that, in many cases, there will be a fairly even distribution of these across the payload.

    - If the total payload is less than 70 bases, the tail end of the vendor ID will overlap with the start of the CODEC ID
    - If the total payload is greater than 70 bases, the middle bases will be ignored

- Perform a direct comparison of your input value (vendor ID or CODEC ID) with each of the respective values in the table. If a direct match is found, use it.

- For cases where there is no direct match:

    - Calculate edit distance between input value and each value in the table.
    - Sort the edit distance in ascending order (e.g. 0, 1, 2, … n)
    - Use the value with the lowest edit distance.

A sample project highlighting how to identify the closest string using the Wagner-Fischer algorithm (Levenshtein edit distance) can be found here.

7 Summary

The embodiment defined in this specification enables archive writers to store metadata within the archive to service a future archive reader, specifically providing them with a means of identifying 1) who wrote the archive, and 2) which CODEC was used when writing sector one (logical metadata). This is referred to as sector zero.

Sector zero serves as a starting point for archive readers that don't have knowledge of who wrote the archive or how sector one was encoded. Sector zero is designed to fit into a single oligonucleotide, providing a 35-character string comprised of natural DNA bases (e.g. ACGT) for the vendor and similar string for the CODEC. With these two 35-character string values, the reader can consult the authoritative tables supplied by the governing body to identify the vendor and the CODEC.

Having vendor information enables readers to contact the entity that wrote the archive. Having CODEC information enables readers to begin reading sector one, which is encoded and spans multiple oligonucleotides, to understand the logical structure of the archive and how to consume its contents.