ACOMA

Adaptive Codec for Organic Molecular Archives

Latchesar Ionkov

Bradley Settlemyer

Dominic Manno

NNSA
National Nuclear Security Administration

Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

LA-UR-20-21586

Goals

- Provide high bit density
- Adapt to oligos of different lengths
- Adapt to different errors
  - Error types
  - Error rates for different types
  - Spatial distribution of errors
  - Missing whole oligos

DNA Codec Challenges

- Short sequences (at the moment)
- Need for oligo identification
- “Structural” oligo restrictions
  - Homopolymer length
  - CG Content
- Sources of errors
  - Synthesis
  - Amplification
  - Sequencing
- Types of errors
  - Substitutions
  - Deletions
  - Insertions

Bit Packing

- How to encode as many bits as possible per nucleotide (nt)?
- Deal with “structural” errors only
- Theoretical maximum: 2 b/nt
- Existing methods:
  - Goldman et al.: 1.58 b/nt
  - Grass et al.: 1.78 b/nt
  - Erlich et al.: 1.98 b/nt

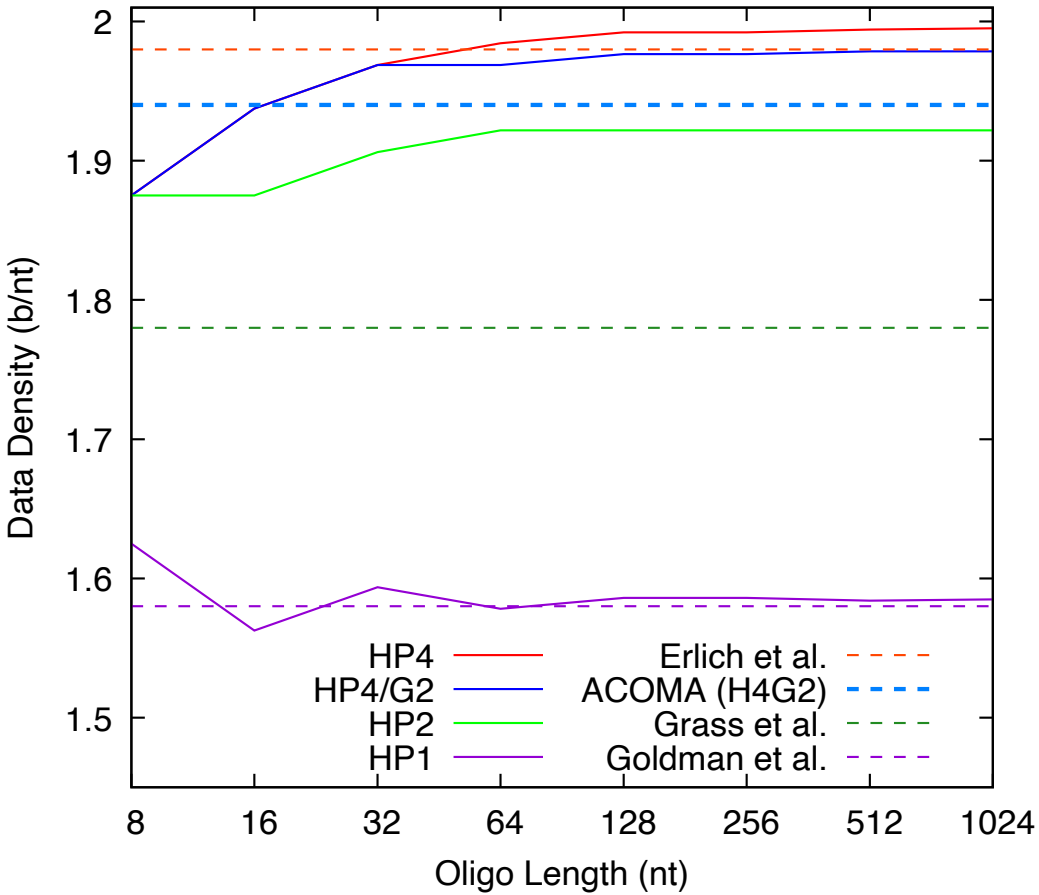
Bit Packing

Oligo counting

- Order oligos
- Reject the ones that fail criteria

~~*AAAAA*~~	~~*X*~~
AAAAT	*0*
AAAAC	*1*
…	
AAGGC	*61*
~~*AAGGG*~~	~~*XX*~~
ATAAA	*62*
…	
GGCGG	*980*
~~*GGGAA*~~	~~*XXX*~~
…	
~~*GGGGG*~~	~~*XXX*~~

Total: 981 values (1.8 b/nt)



ACOMA Codec

- **Level 0**: bit packing
  - How to convert between bits and nts
- **Level 1**: single oligo layout
  - Oligo identification (+ metadata error correction)
  - Data
- **Level 2**: error correction with multiple oligos
  - Error correction for data

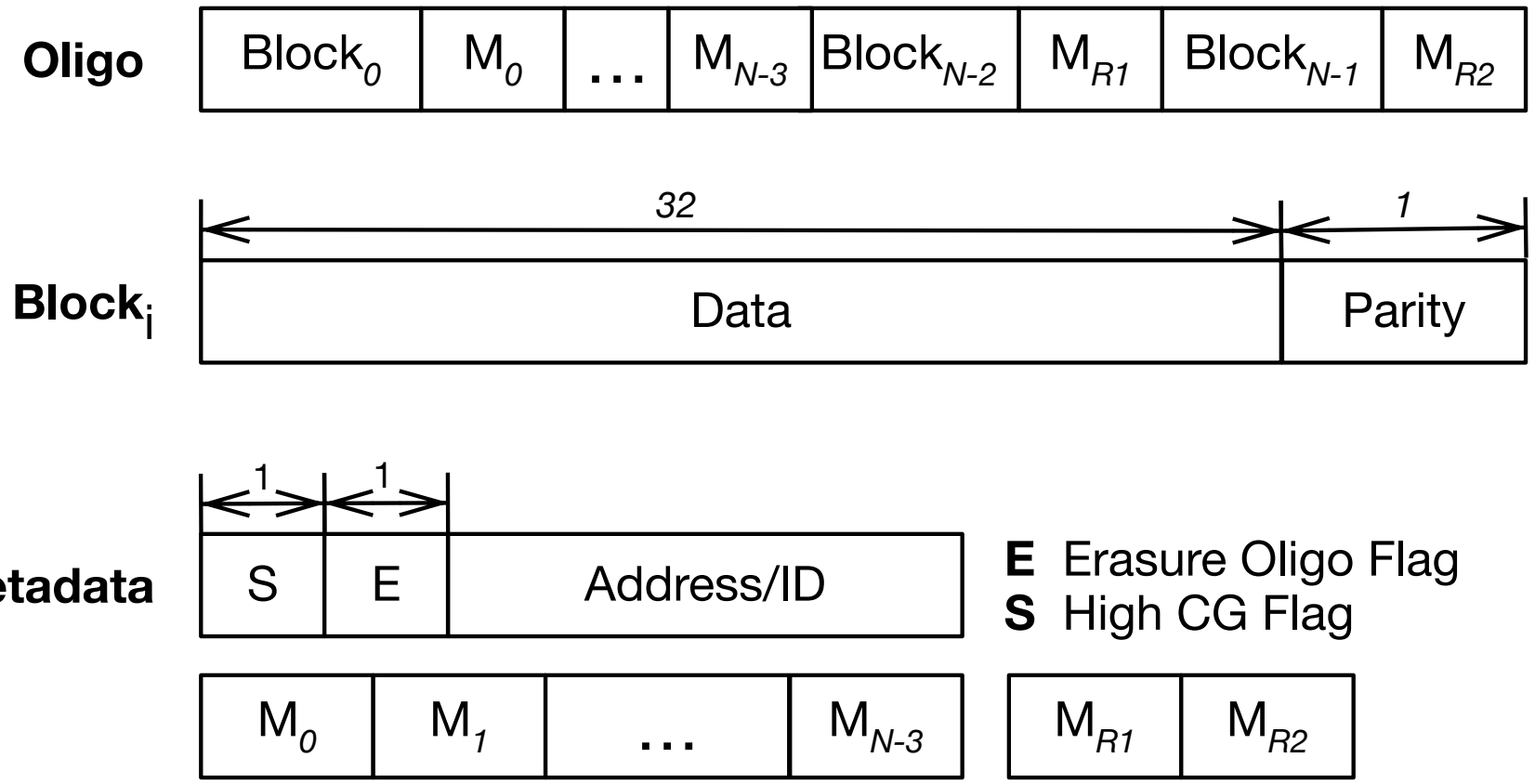
ACOMA Bit Packing (Level 0)

- Same as oligo counting
- Homopolymer restrictions (H4G2):
  - Maximum 4 for A, T, and C
  - Maximum 2 for G
- High CG content: handled at a higher codec level
- It takes exponentially longer time to encode/decode
- Limit up to 17 nts (32 bits data + 1 parity bit)
  - Minimum 10,315,700,031 values (> 2^{33})
  - Data density: 33/17 = **1.94 b/nt**
- Oligo construction:
  - Allows adding up to 17 nts fragments at the end
  - Preserves the homopolymer length restrictions

ACOMA Oligo Layout (Level 1)

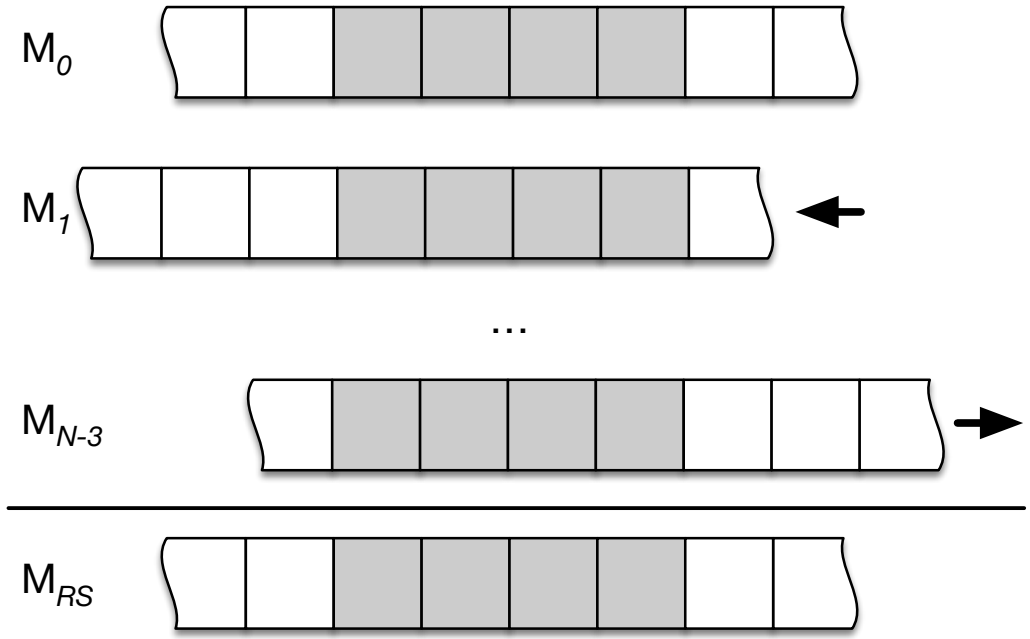
- Data
  - ***N*** 33 bit blocks (32 bit + parity)
- Metadata
  - High CG flag (1 bit)
  - Erasure Oligo flag for Level 2 (1 bit)
  - Address/ID
- Distributed metadata
  - Split into ***N-1*** blocks and place between the data blocks
  - Additional block(s) with Reed-Solomon erasure at the end
  - Used to discover deletions/insertions and align data blocks

ACOMA Oligo Layout (Level 1)



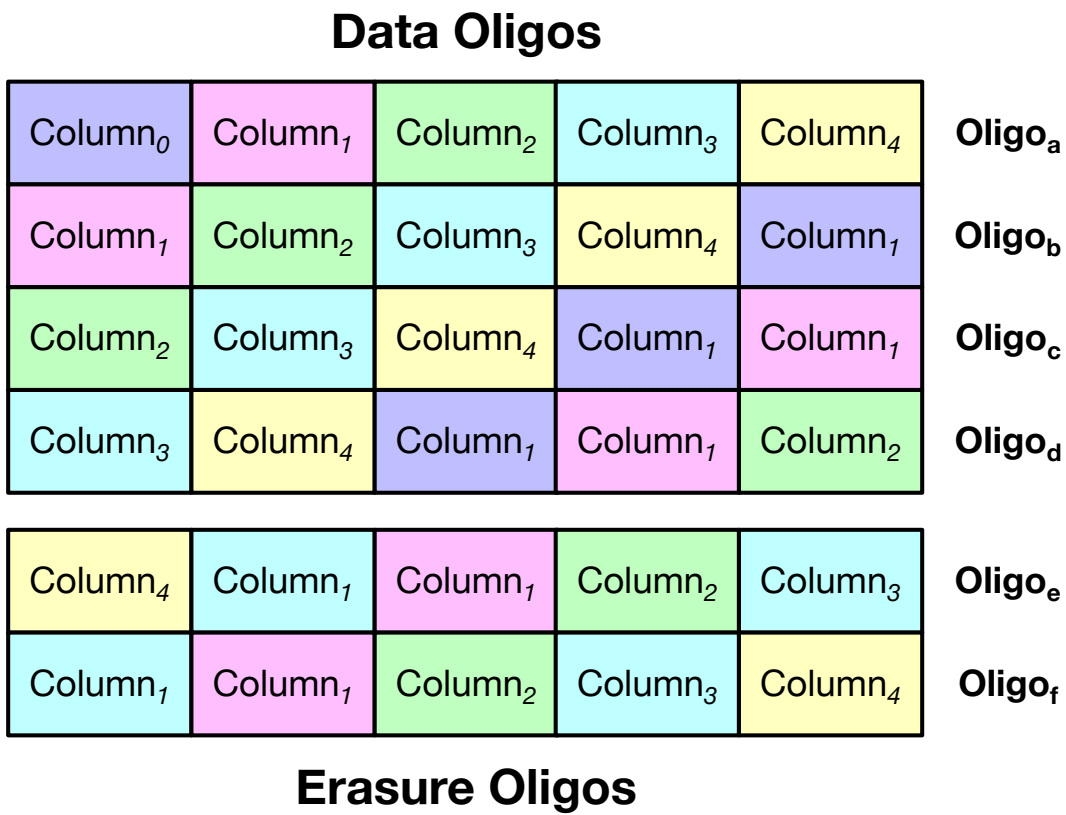
Block Alignment

- Work in progress
- Try to detect insertions and deletions by shifting the metadata blocks
- Shift by 1 left/right and try to match the erasure code
- Needs more analysis on how well it can align blocks *and* correct the metadata



ACOMA Erasure Coding (Level 2)

- 2D erasure encoding
- Group *M* oligos together
- Encode erasure codes for data blocks from different oligos
- Columns consist of blocks from different positions (to correct for spatial bias of errors)
- Reed-Solomon erasure coding
- Parity bit in data blocks used to detect errors



ACOMA Examples

ACOMA Parameters for 107 nt oligos

- 5 data blocks, 17 nts each = 85 nts: **1.94 b/nt**
- 3*4 nts metadata blocks + 2*5 nts RS blocks = 22 nts
- Metadata
  - 4 nts -> 186 values
  - 186^3 = 6,434,856 values
  - 2 bits for High-CG and Erasure-Oligo flags
  - 1608714 addresses/IDs left (30.7 MB)
- Level 1: **1.50 b/nt**
- Erasure Coding
  - 2 (1) Erasure Oligos for every 4 data oligos
  - 4*5*32 bits = 640 data bits per group for 606 (505) nts
- Level 2: **0.996 (1.196) b/nt**

ACOMA Parameters for 110 nt oligos

- 5 data blocks, 17 nts each = 85 nts: **1.94 b/nt**
- 3*5 nts metadata blocks + 2*5 nts RS block = 25 nts
- Metadata
  - 5 nts -> 733 values
  - 733^3 = 393,832,837 values
  - 2 bits for High-CG and Erasure-Oligo flags
  - 98,458,209 addresses/IDs left (1.8 GB)
- Level 1: **1.45 b/nt**
- Erasure Coding
  - 2 (1) Erasure Oligos for every 4 data oligos
  - 4*5*32 bits = 640 data bits per group for 630 (525) nts
  - Level 2: **0.969 (1.163) b/nt**

ACOMA Parameters for 204 nt oligos

- 10 data blocks, 17 nts each = 170 nts: **1.94 b/nt**
- 8*3 nts metadata blocks + 2*5 nts RS blocks = 34 nts
- Metadata
  - 3 nts -> 47 values
  - 47^8 = 23,811,286,661,761 values
  - 2 bits for High-CG and Erasure-Oligo flags
  - 5,952,821,665,440 addresses/IDs left (216 TB)
- Level 1: **1.57 b/nt**
- Erasure Coding
  - 2 (1) Erasure Oligos for every 4 data oligos
  - 4*10*32 bits = 1280 data bits per group for 1182 (985) nts
  - Level 2: **1.045 (1.254) b/nt**

Conclusions

- Adaptive data packing
  - Can use any oligo acceptance criteria, as long is it has reasonable locality
  - Demonstrated good data density with homopolymer restrictions
- Works with different oligo lengths
- Detects deletions and insertions and can align data and metadata blocks
- 2D erasure coding that corrects for spatial error bias and allows multiple erasure oligos

Extra

Implementation: Fast Bit Packing/Unpacking

- Oligo counting is not fast (and exponentially slower)
  - It takes 85 µs on average to encode 13 bit value
  - It takes 2 minutes on average to encode 33 bit value

- Solution: lookup tables
  - For each 4 nts prefix (256 values) generate a table that has the encoding for the first 20 bits of the value (1,048,576 32-bit values, 4 MB)
  - Total space used 1 GB
  - Algorithm:
    1. Pick table based on the previous 4 nts
    2. Lookup in the table the starting oligo using the first 20 bits
    3. Count for the last 13 bits
  - It takes about 1 hour to generate tables, save to disk

Rosetta Stone

- Very long-term storage needs a way to bootstrap
- Oligo(s) that define the ACOMA parameters:
  - Number data blocks per oligo
  - Length of data block (nt)
  - Length of metadata blocks (nt)
  - Number data oligos per erasure group
  - Number of erasure oligos per erasure group
  - **Oligo admission/rejection criteria for bit packing!**
    - Function that receives an oligo and returns true/false
    - Wasm or Knuth's (M)MIX?
  - Encoded with known ACOMA parameters
  - Would be great if it all fits into one oligo

Notes

- Extra data in data block (can we use it for something?)
  - Actual data density is 1.96 b/nt, not 1.94 b/nt ($\log_2(10315700031) = 33.26$)
  - Depending on the 4 nts before the data block, there are some extra values that are lost per data block
  - There is at least an extra bit available for each 4 data blocks
  - Although the minimum number of values in 17 nts is 10,315,700,031 the average number is 13,388,713,708. Can we opportunistically use the extra values, if available?
- Minimum metadata block length
  - 3 nts probably the shortest reasonable length
  - For longer oligos (> 200 nt), instead of making the metadata blocks too small, it may make more sense to add more RS blocks
  - Example: For the 197 nt oligo example, if we have 7 metadata blocks and 2 RS blocks, we'll be able to address 4.6 TB of data.

Implementation: High CG bit

- Problem
  - H4G2 encoding is asymmetric, we can't just use the complementary if high CG

- Solution
  - Encode twice and use the one that has low CG
  - Is it possible that both are high CG?