HARDWARE ACCELERATION FOR GENOMICS

Department of Electrical Engineering, Indian Institute of Technology Gandhinagar

Anuja Chaudhari
Sia Jariwala
Seemanshi Mall
Ginisha Garg

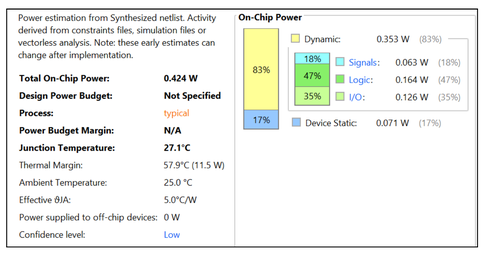
5. Results

BWT REPORTS

1. **Utilization Report**

Resource	Utilization	Available	Utilization %
LUT	1345	20800	6.47
FF	228	41600	0.55
IO	42	106	39.62

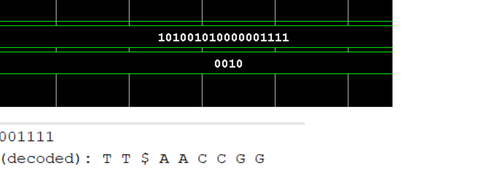
2. **Power Report**



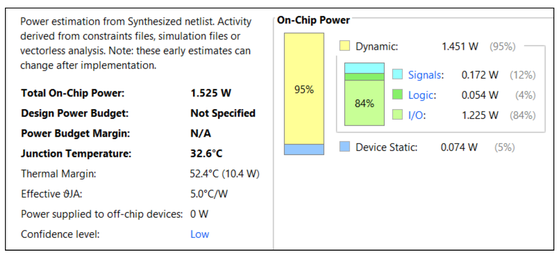
HASHING REPORTS

1. **Utilization Report**

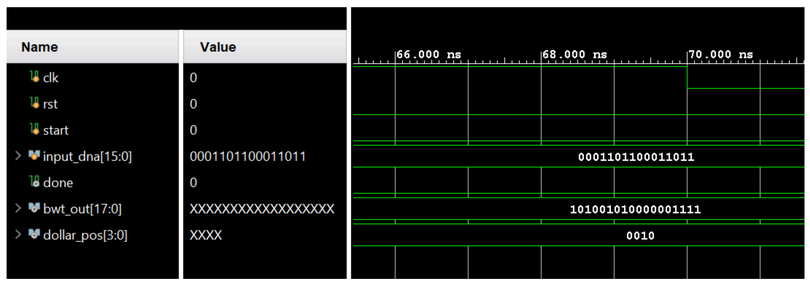
Resource	Utilization	Available	Utilization %
LUT	6	20800	0.03
FF	36	41600	0.09
IO	43	106	40.57



2. **Power Report**

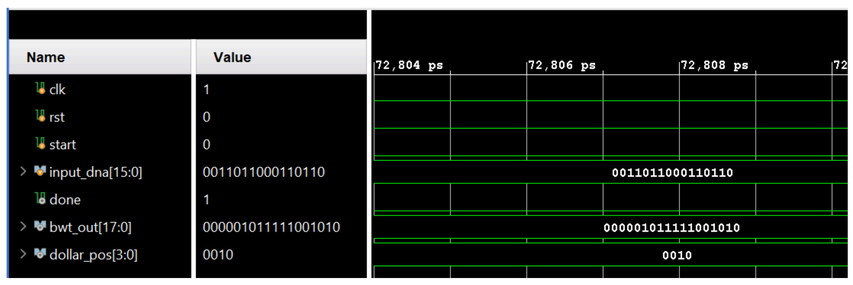


Reference Genome : ATGCATGC
Reference Genome in binary: 0001101100011011



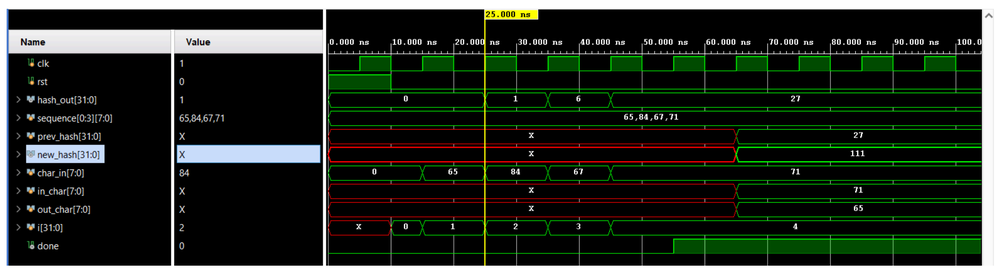

```
BWT Output (binary): 101001010000001111
BWT Output (decoded): BWT Output (decoded): T T $ A A C C G G
```

Reference Genome : ATCGATCG
Reference Genome in binary: 0011011000110110




```
BWT Output (binary): 000001011111001010
BWT Output (decoded): BWT Output (decoded): G G $ T T C C A A
```

Reference Genome: ATCGG (For Hashing)
Reference Genome in Binary: 0001101111



6. Comparison

CASE 1: REFERENCE INPUT : TGCAGTACGTAGCGATACCTAGTA
SEEDS : ["AACGTA", "TTCGGA", "CGTAGC", "TTAGCT", "GCTTAA", "AGTTAC", "CGATAG"]

Reconstruction through Backward Search

Suffix Array: [24, 23, 16, 6, 10, 20, 3, 14, 2, 17, 12, 7, 18, 13, 1, 11, 21, 4, 8, 22, 15, 5, 9, 19, 0]
BWT String: ATTTTTCGGAGACCTAAACGAGGC$
C Table: {'$': 0, 'A': 1, 'C': 8, 'G': 13, 'T': 19}
Occurrence Table:
$: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1]
A: [0 1 1 1 1 1 1 1 1 1 2 2 3 3 3 3 4 5 6 6 6 7 7 7 7]
C: [0 0 0 0 0 0 0 1 1 1 1 1 1 2 3 3 3 3 3 4 4 4 4 4 5]
G: [0 0 0 0 0 0 0 0 1 2 2 3 3 3 3 3 3 3 3 3 4 4 5 6 6]
T: [0 0 1 2 3 4 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6]
Performing seed-based search and reconstruction:
Seed: 'AACGTA'
Found at positions: []
Seed: 'TTCGGA'
Found at positions: []
Seed: 'CGTAGC'
Found at positions: [7]
Seed: 'TTAGCT'
Found at positions: []
Seed: 'GCTTAA'
Found at positions: []
Seed: 'AGTTAC'
Found at positions: []
Seed: 'CGATAG'
Found at positions: []
Reconstructed Genome :
-------CGTAGC----------

Reconstruction through Smith Waterman

Read: AACGTA → Most probable region: None
Read: TTCGGA → Most probable region: None

Read: CGTAGC
→ Most probable region (starting at 7): CGTAGC
→ Alignment:
Read : CGTAGC
Ref : CGTAGC
Read: TTAGCT → Most probable region: None
Read: GCTTAA → Most probable region: None

Read: AGTTAC
→ Most probable region (starting at 3): AGTACG
→ Alignment:
Read : AGTTAC
Ref : AG-TAC

Read: CGATAG
→ Most probable region (starting at 12): CGATAC
→ Alignment:
Read : CGATA
Ref : CGATA

Reconstructed Reference Genome: ---AGTTCGTAGCGATA-------
Reference Genome TGCAGTACGTAGCGATACCTAGTA

CASE 2: REFERENCE INPUT : ACGTACGTGCTAGCTAGCTAGCTA
SEEDS : ["ACGTAC", "TACGTG", "TGCTAG", "AGCTAG", "AGCTAG", "TAGCTA", "AGCTAG"]

Reconstruction through Backward Search

Suffix Array: [24, 23, 0, 4, 19, 15, 11, 1, 5, 21, 17, 13, 9, 20, 16, 12, 8, 2, 6, 22, 3, 18, 14, 10, 7]
BWT String: AT$TTTTAAGGGGAAATCCCGCCCG
C Table: {'$': 0, 'A': 1, 'C': 7, 'G': 13, 'T': 19}
Occurrence Table:
$: [0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
A: [0 1 1 1 1 1 1 1 2 3 3 3 3 3 4 5 6 6 6 6 6 6 6 6 6]
C: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3 3 4 5 6]
G: [0 0 0 0 0 0 0 0 0 0 0 1 2 3 4 4 4 4 4 4 4 5 5 5 6]
T: [0 0 1 1 2 3 4 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6]

Performing seed-based search and reconstruction:

Seed: 'ACGTAC'
Found at positions: [0]

Seed: 'TACGTG'
Found at positions: [3]

Seed: 'TGCTAG'
Found at positions: [7]

Seed: 'AGCTAG'
Found at positions: [15, 11]

Seed: 'TAGCTA'
Found at positions: [18, 14, 10]

Reconstructed Genome:
ACGTACGTGCTAGCTAGCTAGCTA

Reconstruction through Smith Waterman

Read: ACGTAC
→ Most probable region (starting at 0): ACGTAC
→ Alignment:
Read : ACGTAC
Ref : ACGTAC

Read: TACGTG
→ Most probable region (starting at 3): TACGTG
→ Alignment:
Read : TACGTG
Ref : TACGTG

Read: TGCTAG
→ Most probable region (starting at 7): TGCTAG
→ Alignment:
Read : TGCTAG
Ref : TGCTAG

Read: AGCTAG
→ Most probable region (starting at 11): AGCTAG
→ Alignment:
Read : AGCTAG
Ref : AGCTAG

Read: TAGCTA
→ Most probable region (starting at 10): TAGCTA
→ Alignment:
Read : TAGCTA
Ref : TAGCTA

Reconstructed Reference Genome: ACGTACGTGCTAGCTAGCTAGCTA
Reference Genome : ACGTACGTGCTAGCTAGCTAGCTA

BWT and backward search analysis table

Algorithm	Time Complexity	Space Complexity	No. of operations (Case 1)	No. of operations (Case 2)
Suffix Tree Construction	$O(n^2)$	$O(n^2)$	576	625
Suffix Array from Tree	O(nlogσ)	O(n)	25	25
BWT Construction	O(n)	O(n)	25	25
C Table	O(nlogσ)	O(σ)	130	130
OCC Table	O(n × σ)	O(n × σ)	125	125
Backward Search (per seed)	O(m)	O(1)	50	50
Reconstruction (all seeds)	O(k× m)	O(n)	12	42
Total	$O(n^2)$	$O(n^2)$	943	1023

Smith Waterman analysis table

Algorithm	Time Complexity	Space Complexity	Number of Operations (Case 1)	Number of Operations (Case 2)
BASE4_HASH	O(k)	O(1)	9	9
Rolling Hash	O(1)	O(1)	34	34
Storing Reference Genome	O(nk)	O(n)	22	22
Global Positioning	O(k*R)	O(R)	7	7
Smith Watermann for all read	$O(RL^2)$	$O(RL^2)$	258*3	258 * 7
Genome Reconstruction	O(n*b)	O(n)	81	95
Total (Approximate)	$O(RL^2)$	O(n)	927	1973

7. Conclusion

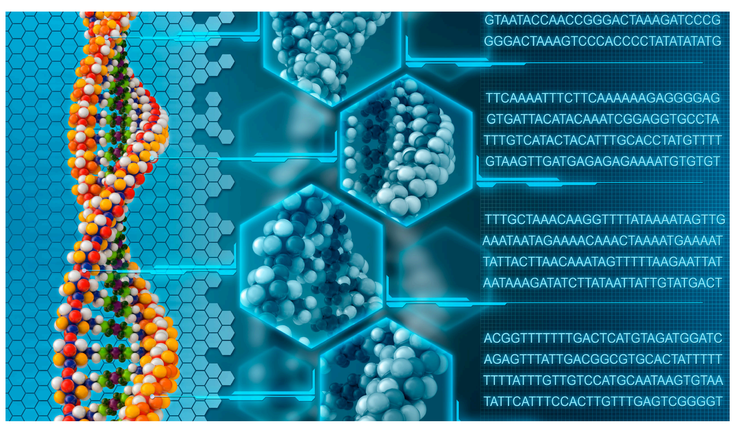
This poster demonstrates two contrasting approaches for reference genome reconstruction using short DNA seeds: the FM-index based backward search and the Smith-Waterman algorithm. Our results show that:

- Backward Search offers highly efficient, memory-light exact matching, enabling fast reconstruction of reference segments from seed matches.
- Smith-Waterman, while computationally intensive, provides greater tolerance to mismatches and gaps, reconstructing sequences even with minor errors or variations in seeds.

Together, these methods highlight the trade-offs between speed and sensitivity in genome assembly tasks. Combining both techniques or choosing based on application needs (e.g., precision vs. error tolerance) can significantly enhance bioinformatics pipelines for sequence alignment and genome reconstruction.

Future Scope:

- To further enhance performance and scalability, we plan to implement Darwin, a hardware-accelerated framework based on the seed-and-extend approach using D-SOFT for seed filtering and GACT for efficient alignment. We also aim to integrate TALCO, an accelerator optimized for long-read error correction in high-error-rate sequencing technologies using tiling. These additions will enable a high-throughput, adaptable genomics pipeline capable of supporting both short and long reads with real-time processing capabilities.



8.Acknowledgement

We wish to express our sincere gratitude for the opportunity, guidance, and support extended by Professor Joycee Mekie throughout the project course.

9.References

1. O. Mutlu, "Bioinformatics," *ETH Zurich*, 2024. [Online]. Available: https://safari.ethz.ch/projects_and_seminars/fall2024/doku.php?id=bioinformatics. [Accessed: 21-Feb-2025].
2. S. Walia, C. Ye, A. Bera, D. Lodhavia and Y. Turakhia, "TALCO: Tiling Genome Sequence Alignment Using Convergence of Traceback Pointers," 2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA), Edinburgh, United Kingdom, 2024, pp. 91-107, doi: 10.1109/HPCA57654.2024.00044. keywords: {Sequential analysis;Software algorithms;Memory management;Genomics;Throughput;System-on-chip;Bioinformatics;Genome Sequence Alignment;Hardware Accelerator;Banded Alignment;Tiling},
3. Y. Turakhia, S. D. Goenka, G. Bejerano and W. J. Dally, "Darwin-WGA: A Co-processor Provides Increased Sensitivity in Whole Genome Alignments with High Speedup," 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), Washington, DC, USA, 2019, pp. 359-372, doi: 10.1109/HPCA.2019.00050. keywords: {Genomics;Bioinformatics;Sensitivity;Software;Hardware;Acceleration;Co-processor;Comparative Genomics;Whole Genome Alignment;Gapped Filtering},