3 Greedy contig assembly (Projekt 3)

3.1 Implementation

3.1.1 Draft implementation in R

What I did

- Read the project description and the explanation of contig assembly in chapter 3 of the lecture
- Look up the data structures in R (Biostrings) for hashtables and DNA strings
- Draft implementation of the contig assembly algorithm from the lecture in R using PDict and DNAStringSet from Biostrings
- Test the implementation using "simple_extension_test.fna", "repeat_with_missing_test.fna", "contigSnip.fna" and "errorfreeReads.fna"
- Extend the implementation to generate several contigs, spanning the complete genome
- Merge the generated contigs (manually) and look up the result in NCBI Blast (⇒ "Heliobacter acinonychis str. Sheeba plasmid pHac1")
- Try the R implementation on "largeErrorfreeReads.fna" ⇒ fail, would take about a month to assemble all the contigs

3.1.2 Implementation in C/C++

- Store reads in a 64bit unsigned integer, this way you only need 2bit per nucleotide
- Store hash table in chained lists using pointers with a hashkey size of 24bit
- Implement a "DNAstring" data structure (chained lists of 64bit unsigned integers using pointers)
- Implement the DNA specific functions using bitwise operations (bit shifts, masking, ...) and pointer arithmetic (reverse, ...)
- Test the implementation → much faster: 12min in R vs. 0.1sec in C for reading, assembling and merging "errorfreeReads.fna", about 45min for reading and assembling "largeErrorfreeReads.fna"

3.2 Testing

3.2.1 errorfreeReads.fna

If contig-merging is enabled my program generates 4 contigs which span the complete genome; one of these contigs has a length of 3606bp, which means that there are only 11bp of the full genome of "Heliobacter acinonychis str. Sheeba plasmid pHac1" missing. But as you can see on Figure 1 on page 2 this missing part is covered by another contig.

```
>gi|109715474|emb|AM260523.1| Helicobacter acinonychis str. Sheeba plasmid pHac1 complete genome, strain Sheeba
ATTTACATTAGCAAAAAAGGGTGTTTTTTCAAGACACAAAAAAAGGGCAGGATAACACTCAAAAAGAAAAAAATGT
GCACTTTTTAAGTGGGGGTGTCTTACGCAGACCACTAGATAAGGGGGTTACAGCGTTCCTTAAAAAAAACTTAAAA
ATGAGACCTTTTTTTTGGCTCTTTTGGTAAAAAACTTGGATATAATGGGGTTGTTATGATTGATTGAGTTAAGGAG
ATAAGATGTTAAAAAAATGTAGAAATGGGGGCTAATTTTTATAAAAAGCTTGGCAAATTAGAAAAACAATTAGCTAA
ATATCAAAGTAAAGTTTTAGAAATTAAAATACAAATAAAAGAGATTAAAAAGCAATATTCTCAAGCTAAGAAAGAT
GAGAATTTAGCTAGTAAATATATTCCTAATGATGAGTTAAAAAAAGATTTATTAGATATTGATAATCCTAATACGA
CTGAAAACTTTAAGCCTAAAAATGCTAGTGAAATTTGGCAATTCGCTCTTAATTCTTAGGGGATTGAATGCTAGAG
ATTGAGTTAAAAAAGAAATTCACTAAAGATTTAAAAAAGTATATTTTAAATCAAAAAATTGAGTTAGAAGTTTTTG
ACTTAGTGGTTGAAAATTTAAGAAATCAAATTCCATTGGATAAAAGATTTAAAGACCACGCTTTAAGTGGAACATA
CAAAGGCTGTAGAGAGTGCCACATTAAACCTGATGTTTTGCTTGTGTATAGAGTGAAAGATAATGTTTTGACCTTA
GTCAGGCTTGGCAGTCATAGCGAGTTGTTTTAGACAAATTTCTTTCTAAAATCGCCTAGAATGAATTTTCAATCCT
AATCCAACCTAATCATCGTTTTTTAAGTTAAGCGATGAAATAGGGCTATTTTTAGCTCATTATTGGGGTATTTTGT
TTTTAGTTGAAAGTTAGCTTTCAAAATTCAAAACAACTTCATTTCCCTTTTAGTCGCTAACCATTTAGCCAATCTA
ACTAGTTTAGCATCTAAAAGCGCATATGACTTCGGCTTCCAATCCAACCCTACTAAAACCGCCTAGCGAGCGTTAG
CGAGCAAATAAGCGGTTTTAGACCGATTGTTTGCTGACAAGCAAACACCAACAAGCGAAGCGTTAGCGAGCATAGA
CAAAAGCGCATCGCAGTTTGAAAGCGTAGGCGTTAGCCGAAGCTGTTTTGCGTTAGCAAATCAAACAAGATAGCGC
AAACCTGGCGTTAGACTAAAAAACCCCTAAAAACTAAAACTCCAAAATTATGTAGCGCGTCATGCGCGTTGTTTTT
TTACTTTTTTAAATAGCATGCTGTTTTTTACATGTTTTACTCGCATGCGCGCGCGTGAGGGATTGGGGGTTGGAAG
AGCCTAAATAACGAAGCTGTATGGTTTCTCATTTTTGGGTGAAAATGAATAAAGGGGAACTTCTTGCCAACGATAA
GGGAACTTCTTGCCAACGATAAGGGAACTTCTTGCCAACGATAAGGGAACTTCTTGCCAACGATAAGGGAACTTAA
AAAGTTAATAGTTACCATATTAGCAATTTTTTAGCTATAATAACAGCCTAAGGCGGTATAACATGGTTGCTAATCC
TAGTGTTAACAAATTTGGAGCAATTAGCTTTAAAAGCTAGTGGGTTGGGAGTTTGTAGCGGGTAGCACTCCGTTAG
GAGGCACACCATGAAAGCATTTTTAATAGTAGTGATTTTAGTGGTAATCTTAACACAGCCACTTATATTAAAAAAA
ACCTTAGCGTTTTAATAATCCTTATAAGTCCGGCAAGACTTTTTAAGGGTTTCACTCCTATTTATTATATCGTCTT
TTGAAAAATAAGCATTAAAAGGCGGTTAAATGCCCATGAATACGAATTTTGAACAGCTTATAAAACAAGAATTGGA
ATTACGAAAATTATTAGAAGAATTAGACACACTCCCACAAACCCCACAAATTAAACTACAAAAACAAAAAATACAA
ACTTACATAGACAAGATAACTCCAAGTATTTTGAGCGGTTTTAATCAAAAATTCAAAGAAATTACAAAAAATTCAC
CAAATGAGTTTGAAAAAGAAACCCAAGAACCCACACCACTCAAAGAATCGCAAACCACGATCACACCATGCAAAGA
TTTAGTGGTTACAACCCCTAAAGATAAAACCTATATCACCTACCACAACAACGCTAATAAGGTTAATCTAGGGAAA
TTGAGCGAAAGGGAAGCCAATCTTTTATTCGCTATTTTTCAAAGGCTTAAAGATCAAGGGAATACCCTTATTCGTT
TTGAACCGCAAGATTTAAAGCGCATGCTAAACATAGATATTTCTAACGAGCGCTTATCAGAAGTTGTTATTAAGCT
ATGGGATAGCATTAAAACCGCTGATTTTTGGAAAATTAGCGAAACAGAAACTTCAATCATTCAAGAAAATTACATG
CTTTTTAGTCGGTGTAAAATTGAATTGAATAAACCTAGTAAAGATTTAAAGTATTTAGAAATCCAACTCAATGATA
GCTATCAATACTTGCTCAACAATCTAGGAATGGGTCAATACACTTCTTTCAATCTCTTAGAATTTCAAAGAGTGAG
GGGTAAATACGCTAAAACGCTCTATCGCTTGCTCAAGCAATACAAAAGCACAGGGATTTTAAGCGTGGAATGGTCT
CAATTCAGGGAGCTTTTAGACATTCCAAAAGACTATGAAATGCGAAACATCGATCAAAAAGTCTTAACTCCAGCCA
TTAAAGAACTCCATAAAATCTACCCTTTTGAACACTTGAGCTACAAGAAAGAACGCCGAAGCCATGACAAGCGCAA
GGTAACCCACATTGATTTTTACTTTGAACAATTGCCTAAAGGCGGAACCAAGAAACAAAAACAAGCGGACAAGCAA
CGCACTCAAAGAGATATTAAGCTCATAGCGTGGGACATTAAAAGACAAGGCATTATCAAACGCTCTAAAGAAACCC
TAGGAGTTAGGGAAATGGATTTGAAAAGTTTGATAGGCTCTCTTTTTGAAAACCCTAATGGGACTATTTTAAAAAT
TGAAAACATCGCTAAAGAAAAAAACCAATTCTTCATGCATGTTTCTTACCCTAACAGAAAAAGTGAACCGCAAAAA
TTGGCTGTATCTGACAAACGATACGCTTTAGAATTTTTGTATGTCAGTGGAGGATACACCTTCAAAAAAGACAATT
TGTTACAAGAAACCTTTACGCATAGCATCCACCCTATCACTAACGAACCCATCAAAGAGTTTGCAAAATACATAGG
CAAAACGATCAACATCACTAACCACAATGTAGATCAATGCCCTGATGGGATATCAAACTGCTTAAAAATCACTAAC
ATTGTCAAATTGAATGACAATCAAATCCAAATCTCTCTCCAAGATGTGGATAAACCTGACAAACTCCTAAAACCTT
TCATCGCTAAAGATGAAAAACATTTGAAGAATTGGTTTAAGAAATATTACCGCTGAAAATGTTTGTGAAAAACGCT
TGAGCGATCAATCTAATGATCAAAACTAAACTACGAACAACCTATTTGAAAGTTTAGGAAAAAAATTCTTCACATC
TTAGACTGCTAACTAAAAAACTTCAATTTTTTCTTTCTAACTAACTTCTGGCAAAAAACACGATCCAAAAAAGGGG
CAAATTTCACACC
```

Contig 1 (length 71bp) Contig 2 (length 3606bp) Contig 3 (length 37bp) Contig 4 (length 32bp)

green: overlay between Contig 1 and Contig 2 *orange: overlay between Contig 2 and Contig 4*

Figure 1: Map showing which parts of the complete genome are covered by the generated contigs

Actually these contigs should have been merged together, but due to a bug in the implementation of the contig-merging-algorithm (which is not yet fixed until now) there are still several contigs. This (and the fact that the merging-algorithm is only an optional task) is the cause why the contig-merging is disabled by default; if you want to enable it you have to use the MERGE-symbol when compiling. [1]

percent of data	# contigs (merge disabled)	longest contig (merge disabled)	# contigs (merge enabled)	longest contig (merge enabled)
100%	7	2291bp	4	3606bp
40%	60	857bp	49	858bp
20%	201	270bp	170	423bp
10%	456	125bp	422	194bp
5%	691	67bp	647	86bp

3.2.2 largeErrorfreeReads.fna

percent of data	# contigs	longest contig	runtime
100%	82272	3057bp	48m16.062s

enabling the merge-algorithm will fail on this huge dataset (I guess because of the bug mentioned earlier)...

I looked up the longest contig (3057bp) on NCBI Blast and so I assume that the reads of "largeErrorfreeReads.fna" belong to the genome of "Helicobacter acinonychis str. Sheeba" (AM260522.1), which has 1.5Mbp.

[1] g++ ... -DMERGE