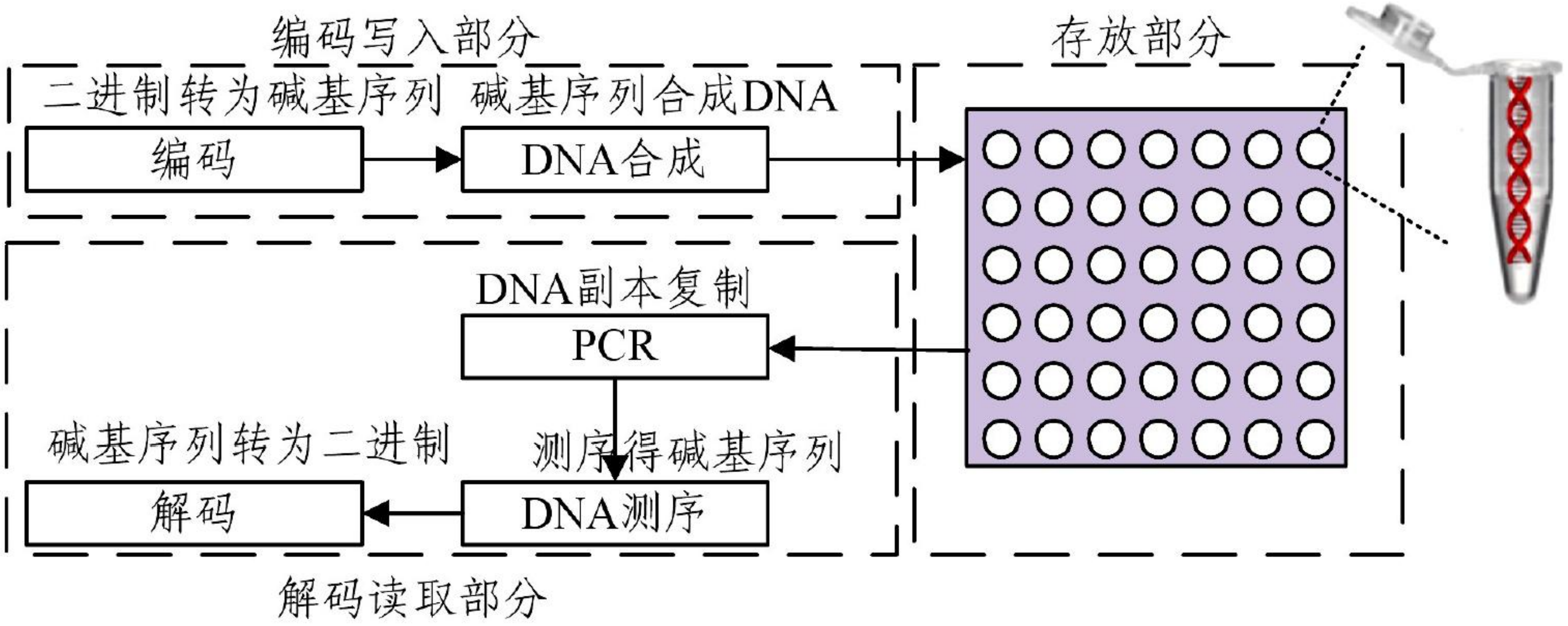
DNA存储解码课程设计

DNA存储系统概述



DNA 存储流程图

DNA存储系统概述

- 相比于传统存储系统的优势：

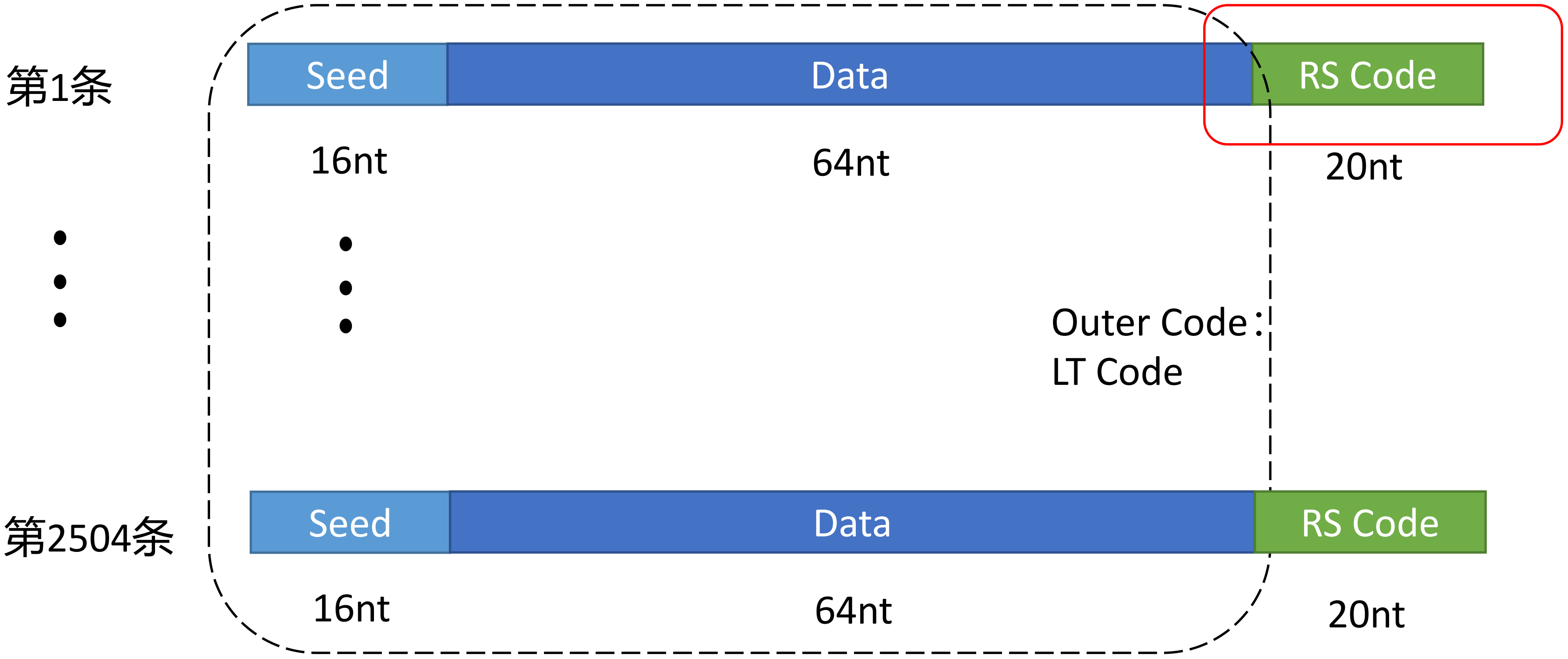
  1）存储密度大、2）能耗低、3）存储周期长、等等

- 未来工程应用需要解决的问题—DNA测序的碱基错误：

1. DNA测序存在碱基测序错误（10%-30%乃至更高的错误）
2. 碱基测序错误包括：
   A. 替换错误 （DNA测序序列中一个或者多个碱基被替换为其它碱基）
   B. 插入错误 （DNA测序序列插入一个或者多个不存在的碱基）
   C. 删除错误 （DNA测序序列中一个或者多个碱基被删除）

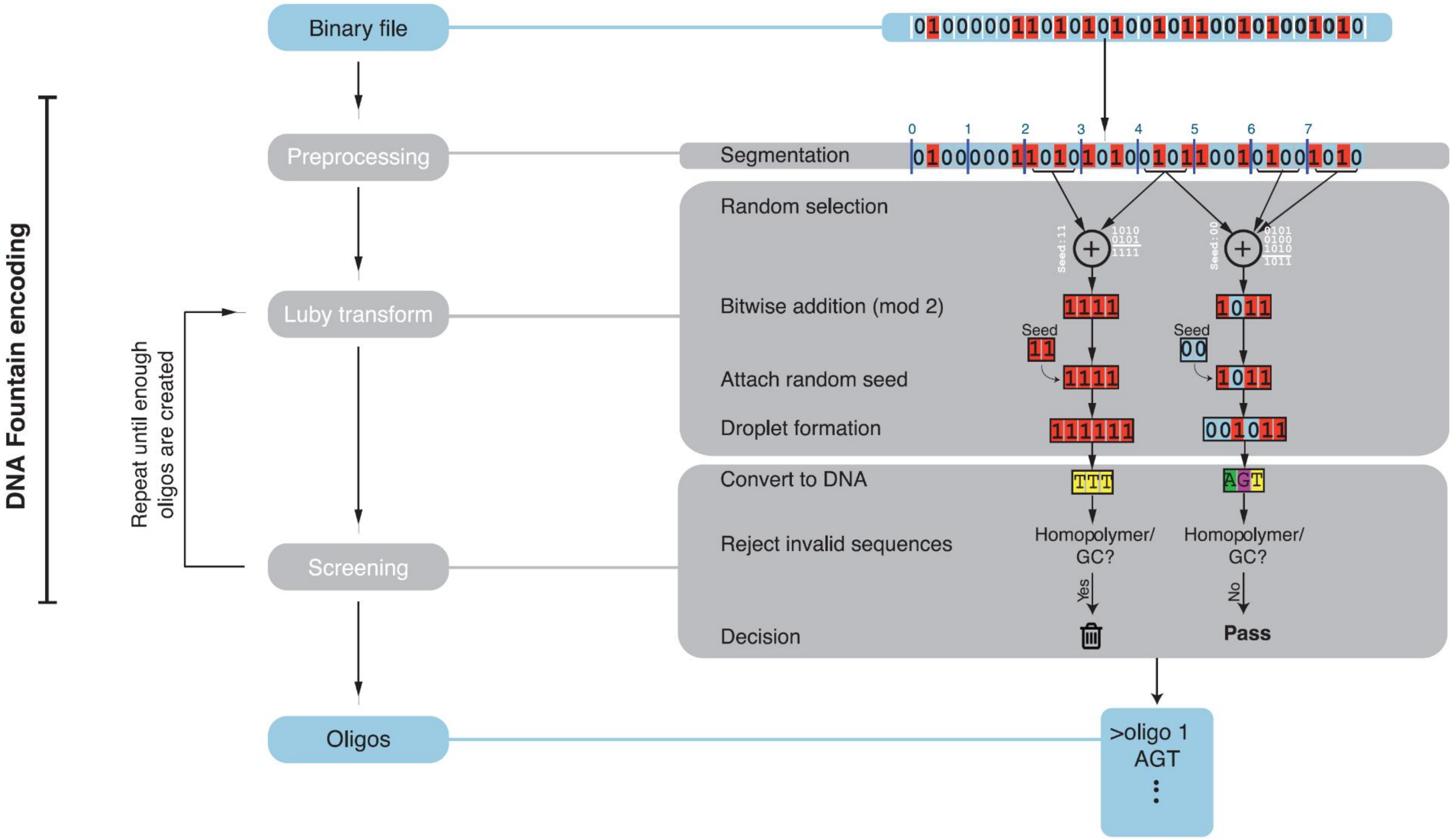
编码方案

- **神奈川冲浪编码方案：** 2504条序列长度100碱基，地址16碱基，数据64碱基，Reed-Solomon编码20碱基，文件切割成1494条编码片段。该编码特点是只需要2025条序列就可解码，每条序列可以检测出20个替换错误，纠正10个替换错误。



Inner Code：
Reed-Solomon

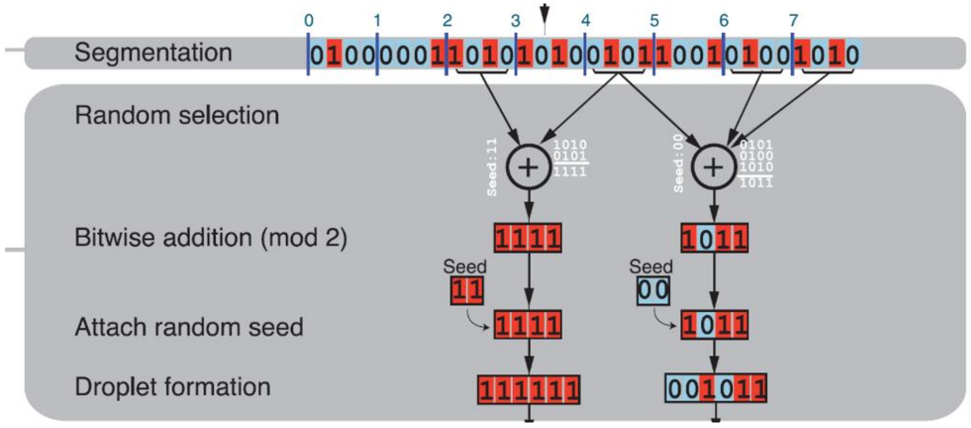
编码方案



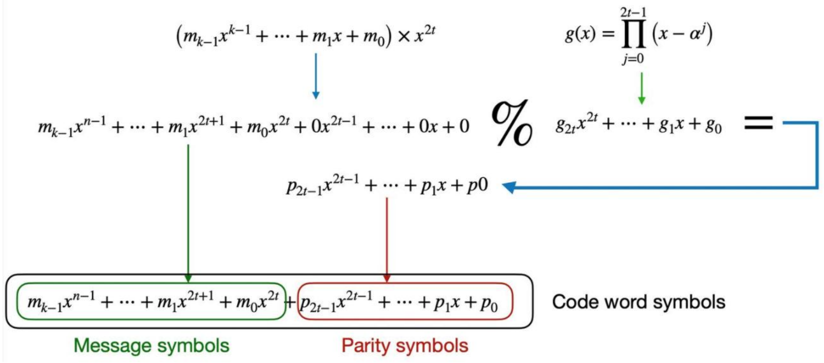
- Erlich, Yaniv, Zielinski, et al. DNA Fountain enables a robust and efficient storage architecture.[J]. Science, 2017.
- Erlich Y , Zielinski D . Capacity-approaching DNA storage. 2016.

编码方案

- Outer Code: LT Code
- Inner Code: Reed Solomon Code



LT Code


$(m_{k-1}x^{k-1}+\cdots+m_1x+m_0)\times x^{2t}$
$g(x)=\prod_{j=0}^{2t-1}(x-\alpha^j)$
$m_{k-1}x^{n-1}+\cdots+m_1x^{2t+1}+m_0x^{2t}+0x^{2t-1}+\cdots+0x+0$
%
$g_{2t}x^{2t}+\cdots+g_1x+g_0$
=
$p_{2t-1}x^{2t-1}+\cdots+p_1x+p0$
$m_{k-1}x^{n-1}+\cdots+m_1x^{2t+1}+m_0x^{2t}$
+
$p_{2t-1}x^{2t-1}+\cdots+p_1x+p_0$
Code word symbols
Message symbols
Parity symbols

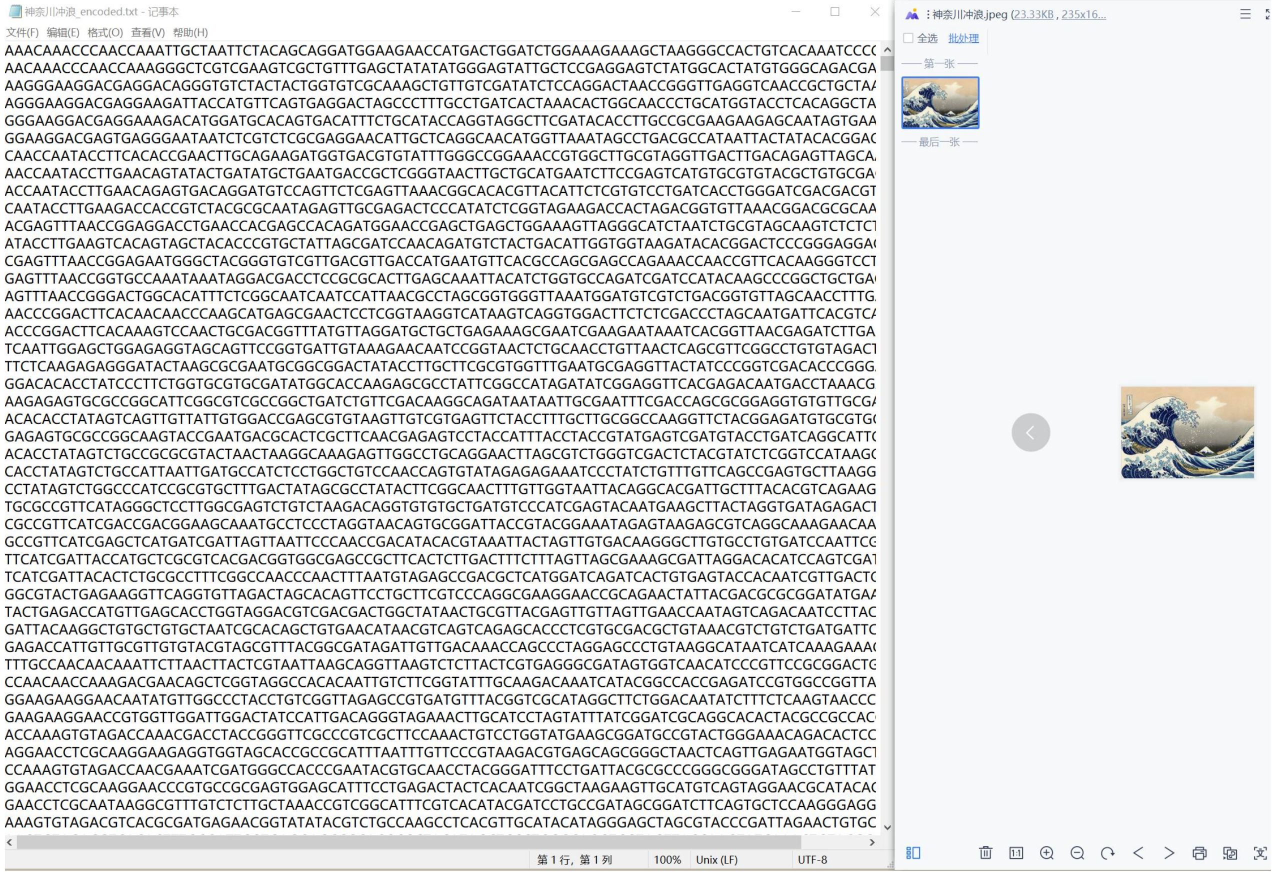
Reed-Solomon Code

- M. Luby, “LT-codes,” in Proc. 43rd Annu. IEEE Symp. Foundations of Computer Science (FOCS), Vancouver, BC, Canada, Nov. 2002, pp. 271–280.

- Mceliece R J . The theory of information and coding:a mathematical framework for communication[M]. Addison-Wesley Pub. Co. Advanced Book Program, 1977.

编码结果

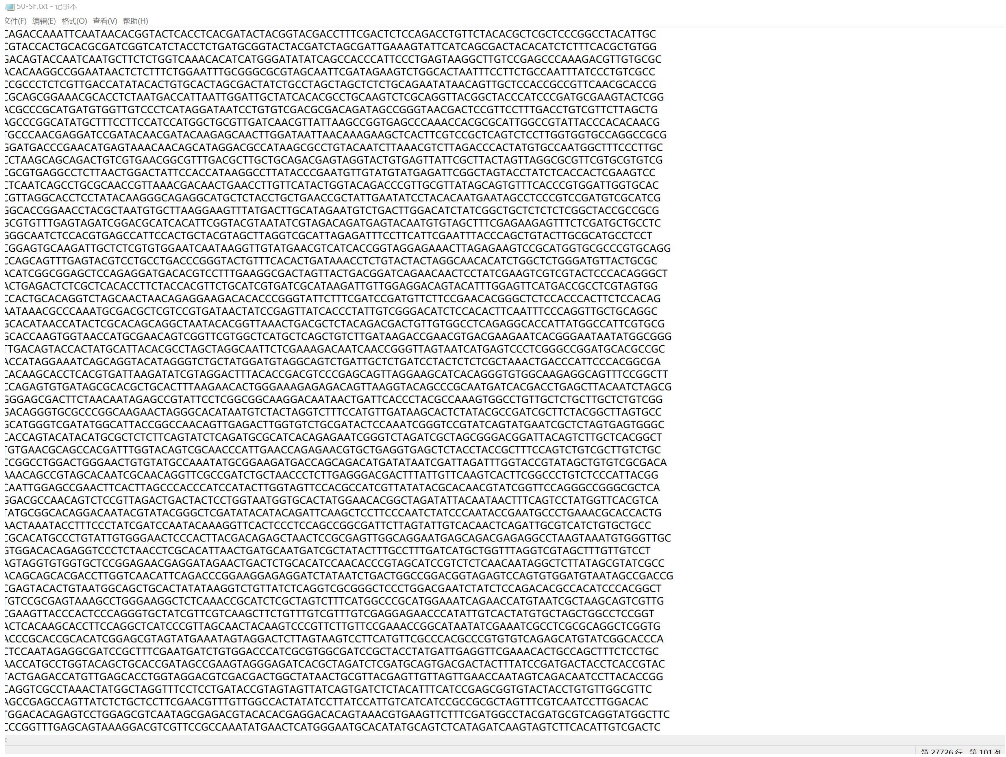
- DNA编码结果：


神奈川冲浪_encoded.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
第 1 行，第 1 列
100%
Unix (LF)
UTF-8
⁝神奈川冲浪.jpeg (23.33KB , 235x16...
□ 全选 批处理
第一张
最后一张

```
AAACAAACCCAACCAAATTGCTAATTCTACAGCAGGATGGAAGAACCATGACTGGATCTGGAAAGAAAGCTAAGGGCCACTGTCACAAATCCC
AACAAACCCAACCAAAGGGCTCGTCGAAGTCGCTGTTTGAGCTATATATGGGAGTATTGCTCCGAGGAGTCTATGGCACTATGTGGGCAGACGA
AAGGGAAGGACGAGGACAGGGTGTCTACTACTGGTGTCGCAAAGCTGTTGTCGATATCTCCAGGACTAACCGGGTTGAGGTCAACCGCTGCTA
AGGGAAGGACGAGGAAGATTACCATGTTCAGTGAGGACTAGCCCTTTGCCTGATCACTAAACACTGGCAACCCTGCATGGTACCTCACAGGCTA
GGGAAGGACGAGGAAAGACATGGATGCACAGTGACATTTCTGCATACCAGGTAGGCTTCGATACACCTTGCCGCGAAGAAGAGCAATAGTGAA
GGAAGGACGAGTGAGGGAATAATCTCGTCTCGCGAGGAACATTGCTCAGGCAACATGGTTAAATAGCCTGACGCCATAATTACTATACACGGA
CAACCAATACCTTCACACCGAACTTGCAGAAGATGGTGACGTGTATTTGGGCCGGAAACCGTGGCTTGCGTAGGTTGACTTGACAGAGTTAGCA
AACCAATACCTTGAACAGTATACTGATATGCTGAATGACCGCTCGGGTAACTTGCTGCATGAATCTTCCGAGTCATGTGCGTGTACGCTGTGCGA
ACCAATACCTTGAACAGAGTGACAGGATGTCCAGTTCTCGAGTTAAACGGCACACGTTACATTCTCGTGTCCTGATCACCTGGGATCGACGACGT
CAATACCTTGAAGACCACCGTCTACGCGCAATAGAGTTGCGAGACTCCCATATCTCGGTAGAAGACCACTAGACGGTGTTAAACGGACGCGCAA
ACGAGTTTAACCGGAGGACCTGAACCACGAGCCACAGATGGAACCGAGCTGAGCTGGAAAGTTAGGGCATCTAATCTGCGTAGCAAGTCTCTC
ATACCTTGAAGTCACAGTAGCTACACCCGTGCTATTAGCGATCCAACAGATGTCTACTGACATTGGTGGTAAGATACACGGACTCCCGGGAGGA
CGAGTTTAACCGGAGAATGGGCTACGGGTGTCGTTGACGTTGACCATGAATGTTCACGCCAGCGAGCCAGAAACCAACCGTTCACAAGGGTCC
GAGTTTAACCGGTGCCAAATAAATAGGACGACCTCCGCGCACTTGAGCAAATTACATCTGGTGCCAGATCGATCCATACAAGCCCGGCTGCTGA
AGTTTAACCGGGACTGGCACATTTCTCGGCAATCAATCCATTAACGCCTAGCGGTGGGTTAAATGGATGTCGTCTGACGGTGTTAGCAACCTTTG
AACCCGGACTTCACAACAACCCAAGCATGAGCGAACTCCTCGGTAAGGTCATAAGTCAGGTGGACTTCTCTCGACCCTAGCAATGATTCACGTC
ACCCGGACTTCACAAAGTCCAACTGCGACGGTTTATGTTAGGATGCTGCTGAGAAAGCGAATCGAAGAATAAATCACGGTTAACGAGATCTTGA
TCAATTGGAGCTGGAGAGGTAGCAGTTCCGGTGATTGTAAAGAACAATCCGGTAACTCTGCAACCTGTTAACTCAGCGTTCGGCCTGTGTAGAC
TTCTCAAGAGAGGGATACTAAGCGCGAATGCGGCGGACTATACCTTGCTTCGCGTGGTTTGAATGCGAGGTTACTATCCCGGTCGACACCCGGG
GGACACACCTATCCCTTCTGGTGCGTGCGATATGGCACCAAGAGCGCCTATTCGGCCATAGATATCGGAGGTTCACGAGACAATGACCTAAACG
AAGAGAGTGCGCCGGCATTCGGCGTCGCCGGCTGATCTGTTCGACAAGGCAGATAATAATTGCGAATTTCGACCAGCGCGGAGGTGTGTTGCG
ACACACCTATAGTCAGTTGTTATTGTGGACCGAGCGTGTAAGTTGTCGTGAGTTCTACCTTTGCTTGCGGCCAAGGTTCTACGGAGATGTGCGTG
GAGAGTGCGCCGGCAAGTACCGAATGACGCACTCGCTTCAACGAGAGTCCTACCATTTACCTACCGTATGAGTCGATGTACCTGATCAGGCATT
ACACCTATAGTCTGCCGCGCGTACTAACTAAGGCAAAGAGTTGGCCTGCAGGAACTTAGCGTCTGGGTCGACTCTACGTATCTCGGTCCATAAG
CACCTATAGTCTGCCATTAATTGATGCCATCTCCTGGCTGTCCAACCAGTGTATAGAGAGAAATCCCTATCTGTTTGTTCAGCCGAGTGCTTAAGG
CCTATAGTCTGGCCCATCCGCGTGCTTTGACTATAGCGCCTATACTTCGGCAACTTTGTTGGTAATTACAGGCACGATTGCTTTACACGTCAGAAG
TGCGCCGTTCATAGGGCTCCTTGGCGAGTCTGTCTAAGACAGGTGTGTGCTGATGTCCCATCGAGTACAATGAAGCTTACTAGGTGATAGAGAC
CGCCGTTCATCGACCGACGGAAGCAAATGCCTCCCTAGGTAACAGTGCGGATTACCGTACGGAAATAGAGTAAGAGCGTCAGGCAAAGAACAA
GCCGTTCATCGAGCTCATGATCGATTAGTTAATTCCCAACCGACATACACGTAAATTACTAGTTGTGACAAGGGCTTGTGCCTGTGATCCAATTC
TTCATCGATTACCATGCTCGCGTCACGACGGTGGCGAGCCGCTTCACTCTTGACTTTCTTTAGTTAGCGAAAGCGATTAGGACACATCCAGTCGA
TCATCGATTACACTCTGCGCCTTTCGGCCAACCCAACTTTAATGTAGAGCCGACGCTCATGGATCAGATCACTGTGAGTACCACAATCGTTGACT
GGCGTACTGAGAAGGTTCAGGTGTTAGACTAGCACAGTTCCTGCTTCGTCCCAGGCGAAGGAACCGCAGAACTATTACGACGCGCGGATATGA
TACTGAGACCATGTTGAGCACCTGGTAGGACGTCGACGACTGGCTATAACTGCGTTACGAGTTGTTAGTTGAACCAATAGTCAGACAATCCTTAC
GATTACAAGGCTGTGCTGTGCTAATCGCACAGCTGTGAACATAACGTCAGTCAGAGCACCCTCGTGCGACGCTGTAAACGTCTGTCTGATGATTC
GAGACCATTGTTGCGTTGTGTACGTAGCGTTTACGGCGATAGATTGTTGACAAACCAGCCCTAGGAGCCCTGTAAGGCATAATCATCAAAGAAA
TTTGCCAACAACAAATTCTTAACTTACTCGTAATTAAGCAGGTTAAGTCTCTTACTCGTGAGGGCGATAGTGGTCAACATCCCGTTCCGCGGACT
CCAACAACCAAAGACGAACAGCTCGGTAGGCCACACAATTGTCTTCGGTATTTGCAAGACAAATCATACGGCCACCGAGATCCGTGGCCGGTTA
GGAAGAAGGAACAATATGTTGGCCCTACCTGTCGGTTAGAGCCGTGATGTTTACGGTCGCATAGGCTTCTGGACAATATCTTTCTCAAGTAACCC
GAAGAAGGAACCGTGGTTGGATTGGACTATCCATTGACAGGGTAGAAACTTGCATCCTAGTATTTATCGGATCGCAGGCACACTACGCCGCCAC
ACCAAAGTGTAGACCAAACGACCTACCGGGTTCGCCCGTCGCTTCCAAACTGTCCTGGTATGAAGCGGATGCCGTACTGGGAAACAGACACTCC
AGGAACCTCGCAAGGAAGAGGTGGTAGCACCGCCGCATTTAATTTGTTCCCGTAAGACGTGAGCAGCGGGCTAACTCAGTTGAGAATGGTAGC
CCAAAGTGTAGACCAACGAAATCGATGGGCCACCCGAATACGTGCAACCTACGGGATTTCCTGATTACGCGCCCGGGCGGGATAGCCTGTTTAT
GGAACCTCGCAAGGAACCCGTGCCGCGAGTGGAGCATTTCCTGAGACTACTCACAATCGGCTAAGAAGTTGCATGTCAGTAGGAACGCATACA
GAACCTCGCAATAAGGCGTTTGTCTCTTGCTAAACCGTCGGCATTTCGTCACATACGATCCTGCCGATAGCGGATCTTCAGTGCTCCAAGGGAGG
AAAGTGTAGACGTCACGCGATGAGAACGGTATATACGTCTGCCAAGCCTCACGTTGCATACATAGGGAGCTAGCGTACCCGATTAGAACTGTGC
```

DNA测序

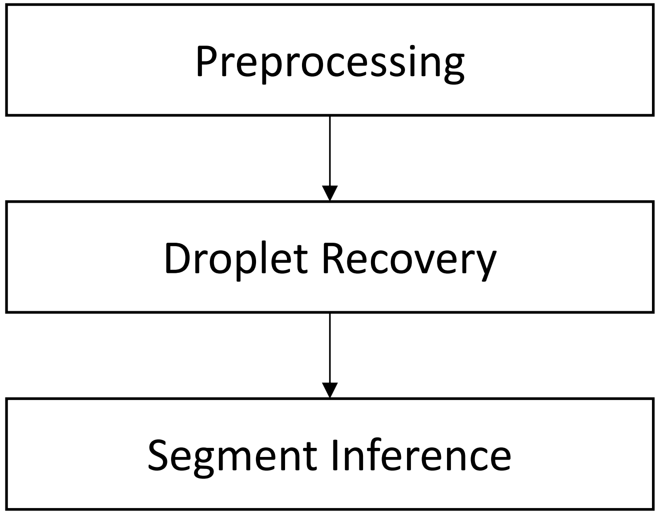
- 对DNA存储系统中储存的DNA进行测序，我们得到了文件“50-SF”文件，它包括27726个测序序列，这些序列包含了神奈川冲浪全部的信息。同时，测序结果存在插入，删除，替换错误。



```
CAGACCAAATTCAATAACACGGTACTCACCTCACGATACTACGGTACGACCTTTCGACTCTCCAGACCTGTTCTACACGCTCGCTCCCGGCCTACATTGC
CGTACCACTGCACGCGATCGGTCATCTACCTCTGATGCGGTACTACGATCTAGCGATTGAAAGTATTCATCAGCGACTACACATCTCTTTCACGCTGTGG
GACAGTACCAATCAATGCTTCTCTGGTCAAACACATCATGGGATATATCAGCCACCCATTCCCTGAGTAAGGCTTGTCCGAGCCCAAAGACGTTGTGCGC
ACACAAGGCCGGAATAACTCTCTTTCTGGAATTTGCGGGCGCGTAGCAATTCGATAGAAGTCTGGCACTAATTTCCTTCTGCCAATTTATCCCTGTCGCC
CCGCCCTCTCGTTGACCATATACACTGTGCACTAGCGACTATCTGCCTAGCTAGCTCTCTGCAGAATATAACAGTTGCTCCACCGCCGTTCAACGCACCG
CGCAGCGGAAACGCACCTCTAATGACCATTAATTGGATTGCTATCACACGCCTGCAAGTCTCGCAGGTTACGGCTACCCATCCCGATGCGAAGTACTCGG
ACGCCCGCATGATGTGGTTGTCCCTCATAGGATAATCCTGTGTCGACGCGACAGATAGCCGGGTAACGACTCCGTTCCTTTGACCTGTCGTTCTTAGCTG
AGCCCGGCATATGCTTTCCTTCCATCCATGGCTGCGTTGATCAACGTTATTAAGCCGGTGAGCCCAAACCACGCGCATTGGCCGTATTACCCACACAACG
TGCCCAACGAGGATCCGATACAACGATACAAGAGCAACTTGGATAATTAACAAAGAAGCTCACTTCGTCCGCTCAGTCTCCTTGGTGGTGCCAGGCCGCG
GGATGACCCGAACATGAGTAAACAACAGCATAGGACGCCATAAGCGCCTGTACAATCTTAAACGTCTTAGACCCACTATGTGCCAATGGCTTTCCCTTGC
CCTAAGCAGCAGACTGTCGTGAACGGCGTTTGACGCTTGCTGCAGACGAGTAGGTACTGTGAGTTATTCGCTTACTAGTTAGGCGCGTTCGTGCGTGTCG
CGCGTGAGGCCTCTTAACTGGACTATTCCACCATAAGGCCTTATACCCGAATGTTGTATGTATGAGATTCGGCTAGTACCTATCTCACCACTCGAAGTCC
CTCAATCAGCCTGCGCAACCGTTAAACGACAACTGAACCTTGTTCATACTGGTACAGACCCGTTGCGTTATAGCAGTGTTTCACCCGTGGATTGGTGCAC
CGTTAGGCACCTCCTATACAAGGGCAGAGGCATGCTCTACCTGCTGAACCGCTATTGAATATCCTACACAATGAATAGCCTCCCGTCCGATGTCGCATCG
GGCACCGGAACCTACGCTAATGTGCTTAAGGAAGTTTATGACTTGCATAGAATGTCTGACTTGGACATCTATCGGCTGCTCTCTCTCGGCTACCGCCGCG
GCGTGTTTGAGTAGATCGGACGCATCACATTCGGTACGTAATATCGTAGACAGATGAGTACAATGTGTAGCTTTCGAGAAGAGTTTCTCGATGCTGCCTC
GGGCAATCTCCACGTGAGCCATTCCACTGCTACGTAGCTTAGGTCGCATTAGAGATTTCCTTCATTCGAATTTACCCAGCTGTACTTGCGCATGCCTCCT
CGGAGTGCAAGATTGCTCTCGTGTGGAATCAATAAGGTTGTATGAACGTCATCACCGGTAGGAGAAACTTAGAGAAGTCCGCATGGTGCGCCCGTGCAGG
CCAGCAGTTTGAGTACGTCCTGCCTGACCCGGGTACTGTTTCACACTGATAAACCTCTGTACTACTAGGCAACACATCTGGCTCTGGGATGTTACTGCGC
ACATCGGCGGAGCTCCAGAGGATGACACGTCCTTTGAAGGCGACTAGTTACTGACGGATCAGAACAACTCCTATCGAAGTCGTCGTACTCCCACAGGGCT
ACTGAGACTCTCGCTCACACCTTCTACCACGTTCTGCATCGTGATCGCATAAGATTGTTGGAGGACAGTACATTTGGAGTTCATGACCGCCTCGTAGTGG
CCACTGCACAGGTCTAGCAACTAACAGAGGAAGACACACCCGGGTATTCTTTCGATCCGATGTTCTTCCGAACACGGGCTCTCCACCCACTTCTCCACAG
AATAAACGCCCAAATGCGACGCTCGTCCGTGATAACTATCCGAGTTATCACCCTATTGTCGGGACATCTCCACACTTCAATTTCCCAGGTTGCTGCAGGC
GCACATAACCATACTCGCACAGCAGGCTAATACACGGTTAAACTGACGCTCTACAGACGACTGTTGTGGCCTCAGAGGCACCATTATGGCCATTCGTGCG
GCACCAAGTGGTAACCATGCGAACAGTCGGTTCGTGGCTCATGCTCAGCTGTCTTGATAAGACCGAACGTGACGAAGAATCACGGGAATAATATGGCGGG
TTGACAGTACCACTATGCATTACACGCCTAGCTAGGCAATTCTCGAAAGACAATCAACCGGGTTAGTAATCATGAGTCCCTCGGGCCGGATGCACGCCGC
ACCATAGGAAATCAGCAGGTACATAGGGTCTGCTATGGATGTAGGCAGTCTGATTGCTCTGATCCTACTCTCTCGCTAAACTGACCCATTCCCACGGCGA
CACAAGCACCTCACGTGATTAAGATATCGTAGGACTTTACACCGACGTCCCGAGCAGTTAGGAAGCATCACAGGGTGTGGCAAGAGGCAGTTTCCGGCTT
CCAGAGTGTGATAGCGCACGCTGCACTTTAAGAACACTGGGAAAGAGAGACAGTTAAGGTACAGCCCGCAATGATCACGACCTGAGCTTACAATCTAGCG
GGGAGCGACTTCTAACAATAGAGCCGTATTCCTCGGCGGCAAGGACAATAACTGATTCACCCTACGCCAAAGTGGCCTGTTGCTCTGCTTGCTCTGTCGG
GACAGGGTGCGCCCGGCAAGAACTAGGGCACATAATGTCTACTAGGTCTTTCCATGTTGATAAGCACTCTATACGCCGATCGCTTCTACGGCTTAGTGCC
GCATGGGTCGATATGGCATTACCGGCCAACAGTTGAGACTTGGTGTCTGCGATACTCCAAATCGGGTCCGTATCAGTATGAATCGCTCTAGTGAGTGGGC
CACCAGTACATACATGCGCTCTCTTCAGTATCTCAGATGCGCATCACAGAGAATCGGGTCTAGATCGCTAGCGGGACGGATTACAGTCTTGCTCACGGCT
TGTGAACGCAGCCACGATTTGGTACAGTCGCAACCCATTGAACCAGAGAACGTGCTGAGGTGAGCTCTACCTACCGCTTTCCAGTCTGTCGCTTGTCTGC
CCGGCCTGGACTGGGAACTGTGTATGCCAAATATGCGGAAGATGACCAGCAGACATGATATAATCGATTAGATTTGGTACCGTATAGCTGTGTCGCGACA
AAACAGCCGTAGCACAATCGCAACAGGTTCGCCGATCTGCTAACCCTCTTGAGGGACGACTTTATTGTTCAAGTCACTTCGGCCCTGTCTCCCATTACGG
CAATTGGAGCCGAACTTCACTTAGCCCACCCATCCATACTTGGTAGTTCCACGCCATCGTTATATACGCACAACGTATCGGTTCCAGGGCCGGGCGCTCA
GGACGCCAACAGTCTCCGTTAGACTGACTACTCCTGGTAATGGTGCACTATGGAACACGGCTAGATATTACAATAACTTTCAGTCCTATGGTTCACGTCA
TATGCGGCACAGGACAATACGTATACGGGCTCGATATACATACAGATTCAAGCTCCTTCCCAATCTATCCCAATACCGAATGCCCTGAAACGCACCACTG
AACTAAATACCTTTCCCTATCGATCCAATACAAAGGTTCACTCCCTCCAGCCGGCGATTCTTAGTATTGTCACAACTCAGATTGCGTCATCTGTGCTGCC
CGCACATGCCCTGTATTGTGGGAACTCCCACTTACGACAGAGCTAACTCCGCGAGTTGGCAGGAATGAGCAGACGAGAGGCCTAAGTAAATGTGGGTTGC
GTGGACACAGAGGTCCCTCTAACCTCGCACATTAACTGATGCAATGATCGCTATACTTTGCCTTTGATCATGCTGGTTTAGGTCGTAGCTTTGTTGTCCT
AGTAGGTGTGGTGCTCCGGAGAACGAGGATAGAACTGACTCTGCACATCCAACACCCGTAGCATCCGTCTCTCAACAATAGGCTCTTATAGCGTATCGCC
ACAGCAGCACGACCTTGGTCAACATTCAGACCCGGAAGGAGAGGATCTATAATCTGACTGGCCGGACGGTAGAGTCCAGTGTGGATGTAATAGCCGACCG
CGAGTACACTGTAATGGCAGCTGCACTATATAAGGTCTGTTATCTCAGGTCGCGGGCTCCCTGGACGAATCTATCTCCAGACACGCCACATCCCACGGCT
TGTCCGCGAGTAAAGCCTGGGAAGGCTCTCAAACCGCATCTCGCTAGTCTTTCATGGCCCGCATGGAAATCAGAACCATGTAATCGCTAAGCAGTCGTTG
CGAAGTTACCCACTCCCAGGGTGCTATCGTTCGTCAAGCTTCTGTTTGTCGTTTGTCGAGGAGAACCCATATTGTCACTATGTGCTAGCTGGCCTCCGGT
ACTCACAAGCACCTTCCAGGCTCATCCCGTTAGCAACTACAAGTCCCGTTCTTGTTCCGAAACCGGCATAATATCGAAATCGCCTCGCGCAGGCTCGGTG
ACCCGCACCGCACATCGGAGCGTAGTATGAAATAGTAGGACTCTTAGTAAGTCCTTCATGTTCGCCCACGCCCGTGTGTCAGAGCATGTATCGGCACCCA
CTCCAATAGAGGCGATCCGCTTTCGAATGATCTGTGGACCCATCGCGTGGCGATCCGCTACCTATGATTGAGGTTCGAAACACTGCCAGCTTTCTCCTGC
AACCATGCCTGGTACAGCTGCACCGATAGCCGAAGTAGGGAGATCACGCTAGATCTCGATGCAGTGACGACTACTTTATCCGATGACTACCTCACCGTAC
TACTGAGACCATGTTGAGCACCTGGTAGGACGTCGACGACTGGCTATAACTGCGTTACGAGTTGTTAGTTGAACCAATAGTCAGACAATCCTTACACCGG
CAGGTCGCCTAAACTATGGCTAGGTTTCCTCCTGATACCGTAGTAGTTATCAGTGATCTCTACATTTCATCCGAGCGGTGTACTACCTGTGTTGGCGTTC
AGCCGAGCCAGTTATCTCTGCTCCTTCGAACGTTTGTTGGCCACTATATCCTTATCCATTGTCATCATCCGCCGCGCTAGTTTCGTCAATCCTTGGACAC
TGGACACAGAGTCCTGGAGCGTCAATAGCGAGACGTACACACGAGGACACAGTAAACGTGAAGTTCTTTCGATGGCCTACGATGCGTCAGGTATGGCTTC
CCCGGTTTGAGCAGTAAAGGACGTCGTTCCGCCAAATATGAACTCATGGGAATGCACATATGCAGTCTCATAGATCAAGTAGTCTTCACATTGTCGACTC
```

解码方案

- Divided into 3 steps: Preprocessing, droplet recovery, and segment inference



Preprocessing

Droplet Recovery

Segment Inference

Preprocessing

- Stitch the paired-end reads using **PEAR**
- Retain only sequences whose length is 60/100nt
- Collapse identical sequence and store the collapsed sequence and number of occurrences in the data
- Sort the sequences based their abundance

[1]Zhang, Jiajie, Kobert, et al. **PEAR**: a fast and accurate Illumina Paired-End reAd mergeR.[J]. Bioinformatics, 2014.

Droplet Recovery

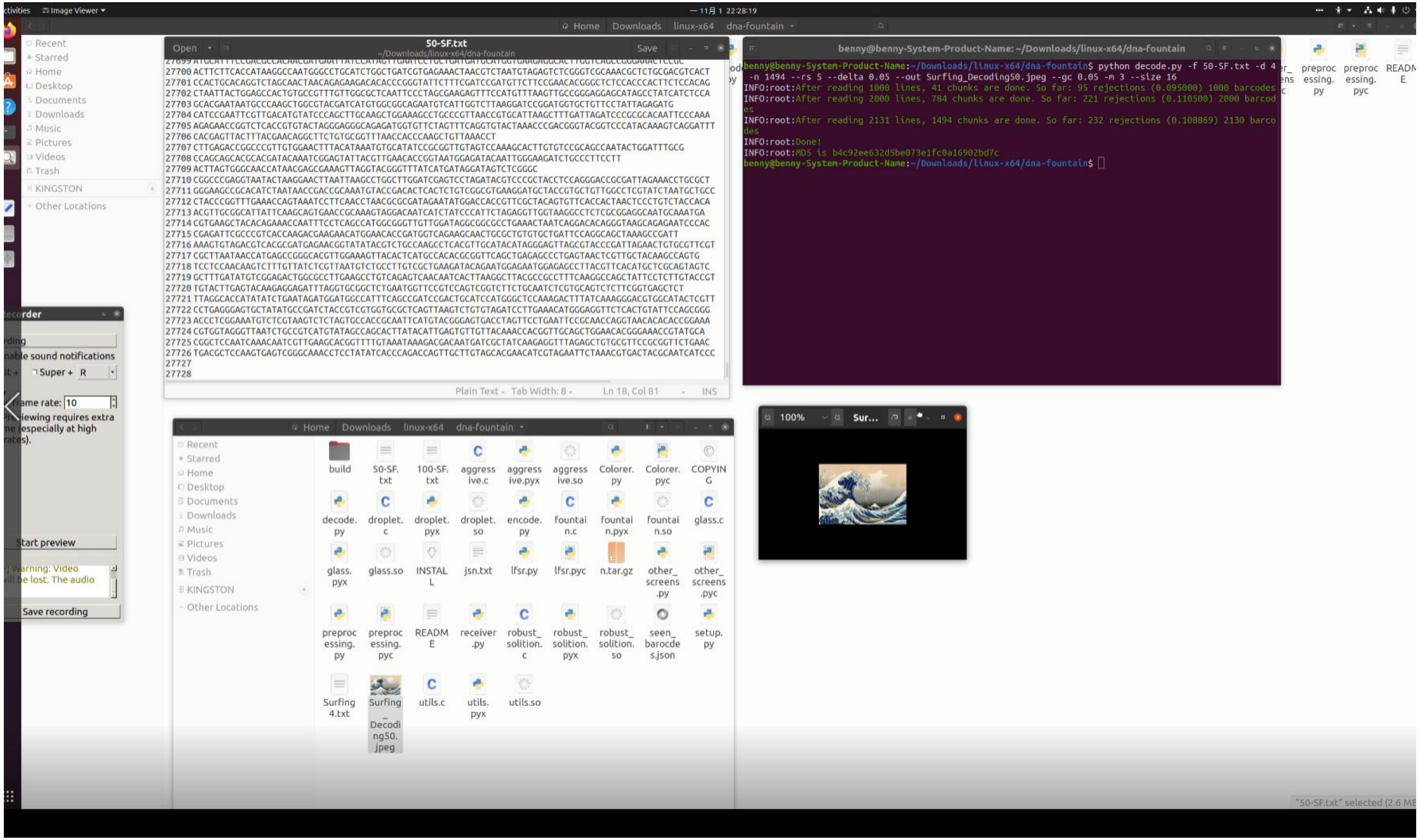
- Translating {A,C,G,T} to {0,1,2,3}
- Extract Seed, data payload, and the RS code from the sequence
- Exclude the sequence with error, which is founded by RS code
- Attempt to correct the substitution error with RS code

By adding t check symbols to the data, a Reed-Solomon code can detect any combination of up to and including t erroneous symbols, or correct up to and including $\lfloor t/2 \rfloor$ symbols - Wikipedia

Segment Inference

- Generate a list of segment identifiers
- Run a message passing algorithm, which works as follows:
  - If the droplet contains inferred segments, the algorithm will XOR these segments from the droplet and remove them from the identity list of droplet
  - If the droplet has only one segment left in the list, the algorithm will set the segment to the droplet’s data payload
  - Recursively propagate the new inferred segment to all previous droplets until no more updates are made
  - If the file is not recovered, the decoder will move to the next sequence in the file and execute the droplet inference and segment inference

解码结果



```
ACTTCTTCACCATAAGGCCAATGGGCCTGCATCTGGCTGATCGTGAGAAACTAACGTCTAATGTAGAGTCTCGGGTCGCAAACGCTCTGCGACGTCACT
CCACTGCACAGGTCTAGCAACTAACAGAGAAGACACACCCGGGTATTCTTTCGATCCGATGTTCTTCCGAACACGGGCTCTCCACCCACTTCTCCACAG
CTAATTACTGGAGCCACTGTGCCGTTTGTTGGCGCTCAATTCCCTAGCGAAGAGTTTCCATGTTTAAGTTGCCGGGAGGAGCATAGCCTATCATCTCCA
GCACGAATAATGCCCAAGCTGGCGTACGATCATGTGGCGGCAGAATGTCATTGGTCTTAAGGATCCGGATGGTGCTGTTCCTATTAGAGATG
CATCCGAATTCGTTGACATGTATCCCAGCTTGCAAGCTGGAAAGCCTGCCCGTTAACCGTGCATTAAGCTTTGATTAGATCCCGCGCACAATTCCCAAA
AGAGAACCGGTCTCACCGTGTACTAGGGAGGGCAGAGATGGTGTTCTAGTTTCAGGTGTACTAAACCCGACGGGTACGGTCCCATACAAAGTCAGGATTT
CACGAGTTACTTTACGAACAGGCTTCTGTGCGGTTTAACCACCCAAGCTGTTAAACCT
CTTGAGACCGGCCCGTTGTGGAACTTTACATAAATGTGCATATCCGCGGTTGTAGTCCAAAGCACTTGTGTCCGCAGCCAATACTGGATTTGCG
CCAGCAGCACGCACGATACAAATCGGAGTATTACGTTGAACACCGGTAATGGAGATACAATTGGGAAGATCTGCCCTTCCTT
ACTTAGTGGGCAACCATAACGAGCGAAAGTTAGGTACGGGTTTATCATGATAGGATAGTCTCGGGC
CGGCCCGAGGTAATACTAAGGAACTTAATTAAGCCTGGCTTGGATCGAGTCCTAGATACGTCCCGCTACCTCCAGGGACCGCGATTAGAAACCTGCGCT
GGGAAGCCGCACATCTAATAACCGACCGCAAATGTACCGACACTCACTCTGTCGGCGTGAAGGATGCTACCGTGCTGTTGGCCTCGTATCTAATGCTGCC
CTACCCGGTTTGAAACCAGTAAATCCTTCAACCTAACGCGCGATAGAATATGGACCACCGTTCGCTACAGTGTTCACCACTAACTCCCTGTCTACCACA
ACGTTGCGGCATTATTCAAGCAGTGAACCGCAAAGTAGGACAATCATCTATCCCATTCTAGAGGTTGGTAAGGCCTCTCGCGGAGGCAATGCAAATGA
CGTGAAGCTACACAGAAACCAATTTCCTCAGCCATGGCGGGTTGTTGGATAGGCGGCGCCTGAAACTAATCAGGACACAGGGTAAGCAGAGAATCCCAC
CGAGATTCGCCCGTCACCAAGACGAAGAACATGGAACACCGATGGTCAGAAGCAACTGCGCTGTGTGCTGATTCCAGGCAGCTAAAGCCGATT
AAAGTGTAGACGTCACGCGATGAGAACGGTATATACGTCTGCCAAGCCTCACGTTGCATACATAGGGAGTTAGCGTACCCGATTAGAACTGTGCGTTCGT
CGCTTAATAACCATGAGCCGGGCACGTTGGAAAGTTACACTCATGCCACACGCGGTTCAGCTGAGAGCCCTGAGTAACTCGTTGCTACAAGCCAGTG
TCCTCCAACAAGTCTTTGTTATCTCGTTAATGTCTGCCTTGTCGCTGAAGATACAGAATGGAGAATGGAGAGCCTTACGTTCACATGCTCGCAGTAGTC
GCTTTGATATGTCGGAGACTGGCGCCTTGAAGCCTGTCAGAGTCAACAATCACTTAAGGCTTACGCCGCCTTTCAAGGCCAGCTATTCCTCTTGTACCGT
TGTACTTGAGTACAAGAGGAGATTTAGGTGCGGCTCTGAATGGTTCCGTCCAGTCGGTCTTCTGCAATCTCGTGCAGTCTCTTCGGTGAGCTCT
TTAGGCACCATATATCTGAATAGATGGATGGCCATTTCAGCCGATCCGACTGCATCCATGGGCTCCAAAGACTTTATCAAAGGGACGTGGCATACTCGTT
CCTGAGGGAGTGCTATATGCCGATCTACCGTCGTGGTGCGCTCAGTTAAGTCTGTGTAGATCCTTGAAACATGGGAGGTTCTCACTGTATTCCAGCGGG
ACCCTCGGAAATGTCTCGTAAGTCTCTAGTGCCACCGCAATTCATGTACGGGAGTGACCTAGTTCCTGAATTCCGCAACCAGGTAACACACACCGGAAA
CGTGGTAGGGTTAATCTGCCGTCATGTATAGCCAGCACTTATACATTGAGTGTTGTTACAAACCACGGTTGCAGCTGGAACACGGGAAACCGTATGCA
CGGCTCCAATCAAACAATCGTTGAAGCACGGTTTTGTAAATAAAGACGACAATGATCGCTATCAAGAGGTTTAGAGCTGTGCGTTCCGCGGTTCTGAAC
TGACGCTCCAAGTGAGTCGGGCAAACCTCCTATATCACCCAGACCAGTTGCTTGTAGCACGAACATCGTAGAATTCTAAACGTGACTACGCAATCATCCC
```

benny@benny-System-Product-Name: ~/Downloads/linux-x64/dna-fountain

```
benny@benny-System-Product-Name:~/Downloads/linux-x64/dna-fountain$ python decode.py -f 50-SF.txt -d 4
 -n 1494 --rs 5 --delta 0.05 --out Surfing_Decoding50.jpeg --gc 0.05 -m 3 --size 16
INFO:root:After reading 1000 lines, 41 chunks are done. So far: 95 rejections (0.095000) 1000 barcodes
INFO:root:After reading 2000 lines, 784 chunks are done. So far: 221 rejections (0.110500) 2000 barcod
es
INFO:root:After reading 2131 lines, 1494 chunks are done. So far: 232 rejections (0.108869) 2130 barco
des
INFO:root:Done!
INFO:root:MD5 is b4c92ee632d5be073e1fc0a16902bd7c
benny@benny-System-Product-Name:~/Downloads/linux-x64/dna-fountain$ 
```

课程设计

- 设计解码程序对DNA编码测序文件“50-SF”进行解码
- 成功恢复出神奈川冲浪图片