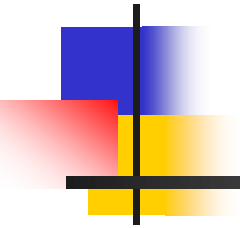
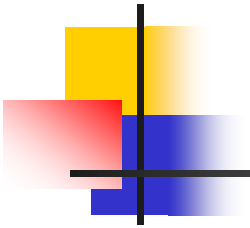
Pattern Search in Noncoding RNA



Chen-Hanson Ting

SVFIG

July 27, 2019



Summary

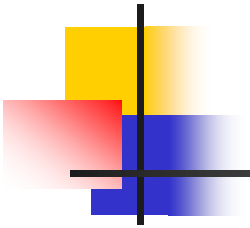
- Failure in 2018
- New work in 2019
- Bioinformatics, DNA / RNA
- Pearls
- Necklaces
- Examples

Failure in 2018

- **I hoped to find microRNAs in the DNA genomes of simple cells like bacteria and nematode.**
- **Repeated patterns of 20bp found in DNA genomes did not match know microRNAs.**

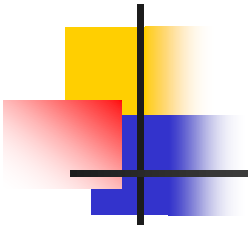
New Works in 2019

- **Long-noncoding RNA (lncRNA) are better places to look for 20 bp microRNA-like patterns.**
- **lncRNAs are expressed genes which may control cell functions.**



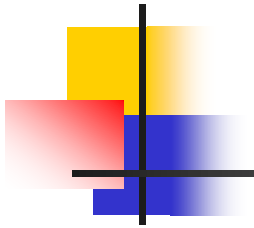
Bioinformatics, DNA

- **Genome: DNA sequences contained in chromosomes.**
- **Coding DNA, <2% of genome expressed as mRNA to produce proteins.**
- **Non-Coding DNA, >98% of genome, or junk DNA.**



Bioinformatics, RNA

- mRNAs, messenger RNA to produce proteins
- tRNA, transfer RNA
- rRNA, ribosomal RNA
- lncRNA, long non-coding RNA, 200 bp or more
- miRNA, microRNA 18-22 bp
- siRNA, snoRNA, piRNA, srRNA



Bioinformatics

- Information is transcribed from DNA to RNA.
- mRNA produce proteins.
- tRNA and rRNA assist protein production.
- miRNA control mRNA expression
- lncRNA, ???

Cell Computer

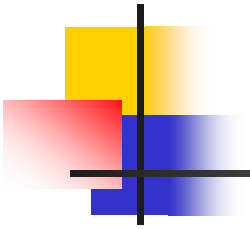
- A cell computer uses miRNAs as instructions.
- lncRNAs contain lists of miRNA.
- Some miRNAs transcribe mRNAs to produce proteins.
- Some miRNAs transcribe lncRNAs to perform complicated functions.

Cell Computer

- **To prove the existence of cell computers, I have to demonstrate that lncRNAs constain lists of miRNAs.**
- **Known miRNAs are not enough to prove the above hypothesis.**

Cell Computer

- **The collections of lncRNA are complete enough to prove my hypothesis.**
- **Exhaustive search of 20 bp patterns in all lncRNAs yields a collection of pearls.**
- **lncRNAs contain lists of pearls, as necklaces.**



What is information?

- Repeated patterns
- 3 bp code for amino acids
- Consecutive amino acid code for proteins
- 20 bp repeated code as pearls

Long Non-Coding RNA

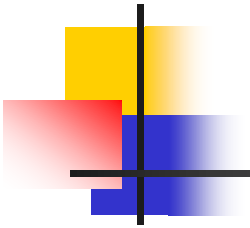
- More than 80% of expressed RNA are long non-coding lncRNAs.
- Most abundant in testis and neural tissues.
- 80% of lncRNA are tissue specific.
- 270,044 RNA transcripts in human.

Ensembl

- European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.
- Release 97, July 2019
- GRCh38_ncrna.fa, 77,596KB
- GRCh38_cdna.fa, 361,405KB

Human Genome

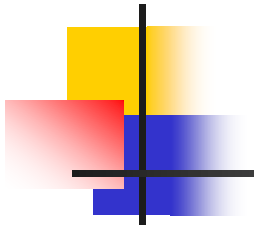
- 3,088,286,481 bp
- Coding DNA <2%
- 203,903 transcripts
- 20,376 genes
- 57,624 non-coding ncRNA
- 189,154 cDNA



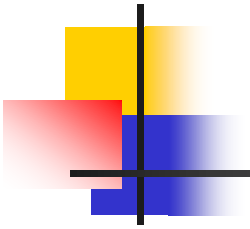
Nematode

- Genome, 99,147KB
- 7 chromosomes
- Non-coding RNA, 1,870KB
  - 3154 entries

Exhaustive Search of Pearls



- **Identify all pearls, unique and repeated 20-base patterns in genomes.**
- **Identify all necklaces, which are clusters of adjacent pearls.**
- **Pearls are related to microRNAs.**
- **Necklaces are related to noncoding RNAs**

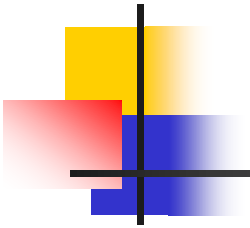


Advanced Forth Search

- **Break lncRNA data file into 4096 threads, each associated with an unique 6-base pattern.**
- **Search repeated 20-base patterns in each thread and search time is greatly reduced.**

Advanced Forth Search

- **Big lncRNA files are searched in 4M bp chunks.**
- **Pearls in each chunk are identified.**
- **All unique pearls are combined, and each assigned an ID.**



Pearls

- **Huge numbers of repeated patterns in consecutive locations, caused by duplicated genes. These patterns must be deleted.**
- **20 base patterns outside of genes are pearls.**

Pearls and Necklaces

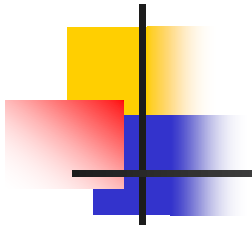
- **Pearls addresses are identified in lncRNA data file.**
- **lncRNA annotations are inserted back into the pearl address list.**
- **Necklaces can be easily identified in the pearl address list.**

Nematode

- **lncRNA data file is only 1.8 MB, with 3155 lncRNAs.**
- **Search time is 10 minutes.**
- **2705 pearls.**
- **107 miRNA matches.**

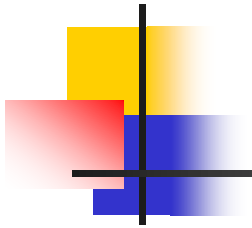
Homo sapiens

- GRCh38_ncrna.fa data file is 77,596 KB, with 57,624 lncRNAs.
- Search time is 2 hours.
- 33,550 pearls.
- 107 miRNA matches.

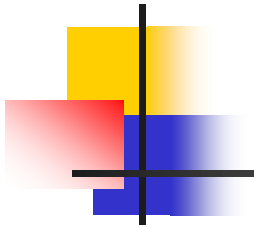


Pearls and Necklaces

- **In my cell computer model,**
  - **Pearls and microRNAs are instructions.**
  - **Protein-coding genes are primitive instructions which produces messengerRNAs.**
  - **Necklaces are high level instructions with lists of pearls.**



Questions?



Thank You!