Code in Long Noncoding RNA

Chen-Hanson Ting

SVFIG

September 28, 2019

Summary

- Long noncoding RNA
- Pre-processing lncRNA data
- Exhaustive search
- Pearls and Necklaces
- The Central Dogma
- A cell computer
- The General Dogma

Genomic Information

- Coding DNA or genes: segments of DNA which encode proteins.
- Coding RNA or messenger RNA: RNA which are translated into proteins.
- Junk DNA: 98.5% of human DNA which do not code proteins.

Long Non-coding RNA

- lncRNA: RNA molecules 200 bp or more, which do not encode proteins, or which do not serve known or useful functions.
- lncRNA exclude mRNA, tRNA, rRNA, and small RNA like microRNA, siRNA, snoRNA, and many others.

Long Non-coding RNA

- lncRNA had to contain useful information; otherwise they would not be transcribed.
- Cells are known to be efficient in utilizing its resources. It is hard to imaging that lncRNA are transcribed for no good reason.

Long Non-Coding RNA

- lncRNA must contain information highly condensed, and actually used by a cell.
- What kind of information do we expect in lncRNA?
- How do we search for information store in lncRNA?

lncRNA Databases

Name	Size(KB)	RNA
GRCh38_ncrna.fa	64,249	67,419
LNCipedia_5_2.fasta	196,560	102,369
NONCODEv5.fa	284,922	165,911
GRCh38_cdna.fa	361,405	139,155
lncRNA_lncbook.fa	400,768	208,848

Information in lncRNA

- Information are repeated patterns.
- For lncRNA, I arbitrarily select 20 bp patterns, like microRNA.
- Goal is to find all repeated patterns with 20 bp or more in lncRNA databases.

Information Search

- We are dealing with huge databases, up to 400 MB long.
- lncRNA databases are first formatted in records with tab-separated fields, suitable for text processing in Python 3.7.4, and in Excel 2010.

Pre-Processing of Data

- lncRNA databases in fasta format are converted to data file with records separated into three fields: name field, length field, and lncRNA data field.
- Records with duplicated lncRNA data fields are removed.

Pre-Processing of Data

- Records are sorted in ascending length.
- Shorter lncRNA are removed if they are embedded in longer lncRNA.
- Records are sorted in ascending names.

Pre-Processing of Data

- lncRNA data fields are combined into a single data file and an index file.
- The index file has three field: a pointer field pointing to the beginning of this lncRNA in the data file, a name field, and a length field.

Exhaustive Search

- All lncRNA data are stored in a flat lncRNA data file.
- ALL repeated 20 bp patterns, or ‘Repeats’, in this data file are identified.
- Consecutive Repeats are packed into ‘Pearls’.
- Clusters of Pearls are ‘Necklaces’.

Pearls File

- Pearls are saved in records with three field: pointer field pointing to the beginning of this Pearl in the data file, length field, and pearl text field.
- Pearls are sorted in ascending pointer, and records in index file are merged to identify Necklaces.

Pearls in lncRNA

- **Pearls are repeated patterns 20 bp or more. They are arbitrarily classified as:**
  - **longPearls: 200 bp or longer.**
  - **shortPearls: 50-199 bp.**
  - **microPearls: 20-49 bp.**

longPearls in lncRNA

- longPearls are repeated patterns 200 bp or more.
- They are caused most often by redundancies in lncRNA database.
- If redundancy are removed, they are likely functional lncRNA.

Necklaces

- Clusters of microPearls within each lncRNA can now be identified as Necklaces.
- Lots of Necklaces are seen in the Pearls file.
- Lots of longPearls persist in the Pearls file.

Redundancy Removal

- Redundancies in lncRNA database were removed by eliminating shorter lncRNA if they are embedded in longer lncRNA.
- Redundancies were further removed by searching shorter longPearls embedded in longer longPearls.

Redundancy Removal

Name	Size (KB)	RNA	Size (trim)	RNA (trim)	Pearls
GRCh38_ncrna	64,249	67,419	40,007	32,841	99,250
LNCipedia_5_2	196,560	102,369	127,160	83,526	351,623
NONCODEv5	284,922	165,911	191,669	116,739	610,040
GRCh38_cdna	361,405	139,155	180,974	102,927	371,351
lncbook	400,768	208,848	293,014	199,756	879,608

Pearls and Necklaces

- There are enormously large numbers of microPearls and shortPearls.
- Lots of microPearls and shortPearls form Necklaces.
- Pearls and Necklaces are code (information) in lncRNA.

Pearls and Necklaces

Pearls and Necklaces

- I expected that Pearls in various lncRNA databases would be highly correlated, but they were not.
- I think these lncRNA databases are incomplete in themselves. I wish microbiologists will do a better job in providing a better collection.

The General Dogma

Pearls and Necklaces

- If Pearls were identified with microRNA, and that Necklaces were lists of microRNA in lncRNA, a General Dogma could be proposed to explain the functioning of cell computers.

The Central Dogma

- The Central Dogma originally stated by Crick asserted that genetic information are transferred from DNA to RNA, and from RNA to proteins.
- Information in proteins were not transferred from protein to other proteins, nor back to RNA and DNA.

The Central Dogma

The most popular version of the Central Dogma is:

DNA chromosomes

v

Coding DNA produce coding RNA

v

Coding RNA produce proteins

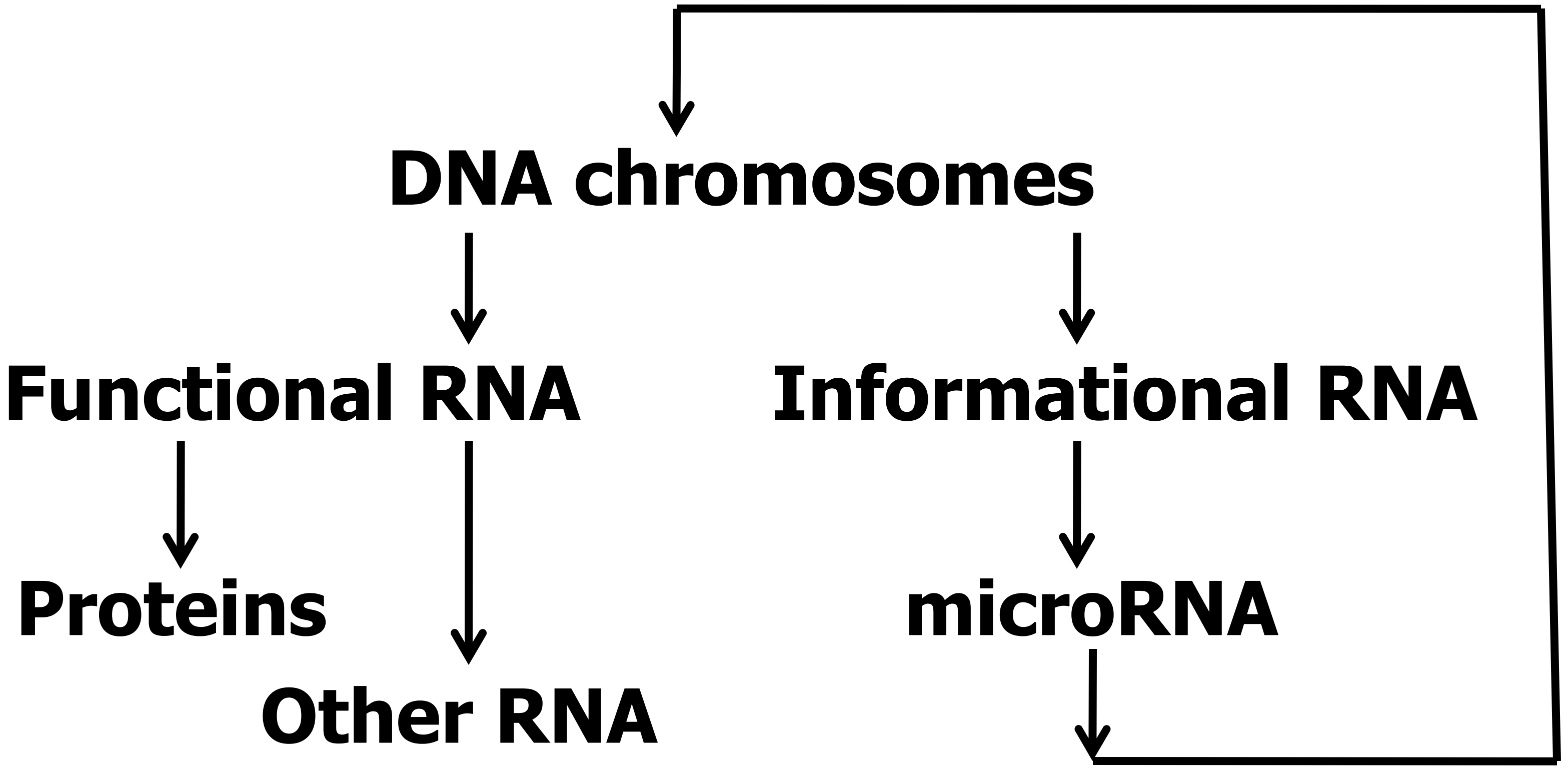
The Central Dogma

- It does not explain the huge amount of non-coding DNA.
- It does not explain the huge number of lncRNA and microRNA.
- It does not explain cell functions other than protein production.

The General Dogma

- A microRNA causes transcription of a RNA.
- A functional RNA performs its assigned function, including protein production.
- An informational RNA releases a cluster of microRNA.
- MicroRNA may be released to cause multicellular functions.

The General Dogma



DNA chromosomes

Functional RNA

Informational RNA

Proteins

Other RNA

microRNA

Cell Computer

- Each cell behaves like a computer.
- All code are stored in DNA.
- Useful code are transcribed and edited in RNA.
- Functional RNA perform specified functions.
- Informational RNA release clusters of microRNA.

Cell Computer

- A living cell is a Turing Machine using microRNA as instructions.
- A set of primitive microRNA cause transcription of functional RNA.
- Other microRNA cause transcription of informational RNA.
- Informational RNA release clusters of microRNA.

Cell Computer

- Each microRNA causes transcription of a RNA.
- Functional RNA perform specific functions, including mRNA, tRNA, rRNA, etc.
- Informational RNA produce clusters of microRNA.

Cell Computer

- Recursively transcribing functional RNA through microRNA-DNA-informational RNA-microRNA loops may account for the complexity of a living cell, and its precise control.

Forth Computer

- A cell computer is very similar to a Forth computer.
- In a Forth computer:
  - Primitive instructions perform specific functions.
  - Compound instructions contain lists of primitive instructions and other compound instructions.

Forth Computer

- **Forth, similar to LISP, has been proven that recursively processing nested lists can solve any computable problem.**
- **Necklaces as clusters of Pearls support the General Dogma as a plausible mechanism for the functioning of living multicellular organisms.**

The General Dogma

- A dogma is a belief system supported by insufficient evidences, but is open to further improvements.
- The Central Dogma guided 3 generations of microbiologists.
- The General Dogma may be useful for the next 3 generations.

Challenge to Microbiology

- Are microPearls microRNA?
- Does a microRNA cause transcription of a RNA?
- How do lncRNA release microRNA?
- Are longPearls redundant lncRNA?
- Could some longPearls be functional RNA?

Challenge to Microbiology

- The current lncRNA and microRNA databases are not complete.
- lncRNA and microRNA are very elusive. They function at very low concentrations and have spatial and temporal dependencies.
- Living cells are very delicate, and cannot withstand attacks by our very crude instrumentation.

Conclusions

- **Exhaustive pattern search is a very powerful tool to analyze large genome databases.**
- **Pearls and Necklaces are common in lncRNA databases.**
- **Pearls and Necklaces allows a plausible mechanism of The General Dogma for living cells.**

Questions?

Thank You!

Long Noncoding RNA

- **GRCh38_ncrna.fa**
  - **65,790,873 bp**
  - **67,419 lncRNA**
- European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom.

Long Noncoding RNA

- **LNCipedia_5_2.fasta**
  - **192,690,141 bp**
  - **127,802 transcripts**
  - **56,946 genes**
- Ghent University - VIB, Life Sciences Research Institute in Flanders, Belgium.

Long Noncoding RNA

- **NONCODEv5_human.fa,**
  - **278,614,288 bp**
  - **165,911 lncRNA**
- Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Long Noncoding RNA

- **GRCh38_cdna.fa**
  - **316,791,371 bp**
  - **139,155 lncRNA**
- European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom.

Long Noncoding RNA

- **lncRNA_lncbook.fa**
  - **405,815,189 bp**
  - **268,848 lncRNA**
- BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China