Pseudo DNA Sequence Generation of Non-coding Distributions using Stream Cipher Mechanism

Jeffrey Zheng
School of Software, Yunnan University
August 4, 2014
Content

- Frontier of Non-Coding DNAs/RNAs
- General Comparison Model for Pseudo DNAs & Real DNAs
- Sample Cases
- Conclusion
Frontier of Non-Coding DNAs/RNAs

Ratios on Non-Coding DNAs
Tools for Analysis
Current Situation
Assumption & Question
Typical Ratios of Non-Coding DNAs/RNAs

- 3% U. Gibba
- 90% Takifugu
- 30% Arabidopsis
- 98% Human

ENCODE: over 80% of DNA in the human genome "serves some purpose, biochemically speaking". However, this conclusion is strongly criticized ...
Tools to Analyze Non-Coding DNAs/RNAs

- Frequency Distribution
- GC densities
- Repeat sub-sequences
- ...
- Machine Learning
- Bayesian Inference and Induction
- Neural Network
- Hidden Markov Model
- ...

...
A case of Non-Coding DNA: Hairpin

A DNA Sequence

A hairpin

Analysis Results in various conditions

Refined Distributions on different parameters
Current Situation

- Total DNA varies widely between organisms
- Ratios of coding DNAs and Non-coding DNAs in genomes are different significantly
- 98% human genomes are Non-coding DNAs
- Non-coding RNAs/DNAs may be drivers of complexity, they are a larger heterogeneous group

Due to various criteria, no a general classification can be used to sub-classify this group
Assumption & Question

Assumption:

A general classification of Non–Coding DNA interactions could be relevant to higher levels of pair structures between a distance on a DNA sequence.

Both 0–1 outputs & DNA segments are random sequences

Question:

Can interaction models of Stream Cipher mechanism simulate a general classification for Non–Coding DNAs?
General Comparison Model for Pseudo DNAs & DNAs

Variant Logic
DNAs & Pseudo DNAs
General Model
Main Procedure
**Variant Logic**

- An unified 0–1 logic framework base on input/output and logic functions using four Meta symbols: \{\bot, +, -, \top\}
  - 0–0 : \bot, 0–1 : +,
  - 1–0 : -, 1–1 : \top.

- Multiple Maps of Variant Phase Spaces can be visualized
Results of automated chain-termination DNA sequencing.
A Comparison Model to simulate Non–Coding DNAs in Visual Maps

- Two input sources:
  - Pseudo DNAs – Artificial Sequences using Stream Cipher on Interactions – HC256
  - Real DNAs – Human DNAs

- Variant Construction to measure & quantity input sequences on 4 meta bases {ACGT}

- Using Visual Maps to identify higher levels of global symmetries between A&T and C&G maps for both artificial & real DNAs
General Comparison Model

HC256 → 0-1 Sequences + Interaction Models → Pseudo DNA Sequences → DNA Sequences

Human ... Virus

Sample Cases on Pseudo DNA:

\[ Y = 100111001011 \]
\[ \text{mode} = 1 \]
\[ X_{r=1} = \text{TGACCTGATA} \]
\[ X_{r=2} = \text{TAACTTAGC} \]
\[ X_{r=3} = \text{CAATTCGAC} \]

Artificial DNAs vs. Real DNAs in Visual Maps

Y = 100111001011
mode = 1
\[ X_{r=1} = \text{TACGTC} \]
\[ X_{r=2} = \text{TATTCA} \]
\[ X_{r=3} = \text{CAAGAC} \]

Probability Statistics on 4 Meta symbols

Different Maps
Main Procedure

Input: Pseudo DNA/Real DNA Vector

$X^t$: GGTACTTGCAT...

Projected as Four 0-1 vectors

$M_G$: 11000001000 ...

$M_A$: 00010000010 ...

$M_T$: 00100110001 ...

$M_C$: 00001000100 ...

Calculated as four Probability Vectors

$\{p_Y\}_{d=1}^{m_c}$

Determine four pairs of map position

$\{(x_Y^t, y_Y^t)\}_{VED}$

Collected all DNA Vectors

$\forall t, X^t \in D^{N_t}$

Four Maps constructed

$\{Map_Y\}_{VED}$
Sample Cases

2700 DNA Sequences
Human DNAs vs. HC256 Pseudo DNAs
Sets of Maps
Non–Coding DNA Sequence Information

- Two Sets of $T=2700$ sequences
  - Non–Coding DNAs for Human Genomes
    - SRR027956.xxxxxxx, $N=500$bp

- For a sample point, a sequence could be

```plaintext
>SRR027962.18095784
TAATTCTTGAGTTCATGTCCCGCATCCAGGGCACACTTGTCAGAAGGGGTGGGTTCCCAAGACCTTAT
GCAGCTCTGCCTCTGTGGCTTTTCAGTGTAACAGTCACCATGGCTGCTGTCTTGGAATAGCTAGTTGAT
GCCTGTGGTATTTCTAGGCTCAGGATGAAAGCTTCCCCGCTGGCTCTAACATTCAGGGATCTTGACGTG
GCGGCCCCATTCCCCACAGCTCCTGTAGGTAAGTCACCCAGTGGGACTCTGTGTGGAGGCTCAATC
CCATATTTCTGTTGGCACTGCCCCATGTGAACCTTTGGATTTCTTTCTGATTCACTGGTCTTGGAAAGGTG
GTGTTTCAGGAAATTTATCCATTTTCTCTAGGTGGTTATGGCAGCACAAGATATTCTGAGGATCT
TTTTTTGTGTCAGTGGATCTCCTTGCAATGTCTCAATTTTGTAAATTGGTGCTTATTGGGAATCTTCTT
TTTTCTGTATAATCTAACTAGCA
```
Human DNAs vs. Pseudo DNAs

Human DNA:

Pseudo DNA: HC256

Two groups of ten 2D maps in the range of $n=3\sim20$, $k=7$, $N\equiv200\sim600$, $T=2700$; (a1-a5) Map$_A$ for the file Right; (b1-b5) Map$_A$ for the file hc256 mode = 1, $r = 1$. 
Pseudo DNAs on various conditions

Figure 5. Four groups of sixteen 2D maps in the range of n = 15, k = {2,3,4,7}, N = 500, T = 2700; (a) group (a1 - a4) four Map_A maps; (b) group (b1-b4) four Map_T maps; (c) (c1 - c4) four Map_C maps; (d) (d1 - d4) four Map_C maps for the file right.
Pseudo DNA sequences on different parameters

Figure 6. Four groups of sixteen 2D maps in the range of $n = 12$, $k = \{2,3,4,7\}$, $N \approx 500$, $T = 2700$ for the file $hc256$, $r = 1$, mode = 1: (a) group (a1 - a4) four $\text{Map}_A$ maps; (b) group (b1-b4) four $\text{Map}_T$ maps; (c) (c1 - c4) four $\text{Map}_C$ maps; (d) (d1 - d4) four $\text{Map}_C$ maps.
Two Groups of Human DNAs

Figure 7. Two groups of eight 2D maps in the range of n = 15, k = 7, N ≈ 200~600, T = 2700; (a) group (a1 - a4) four MapV maps for the file left; (b) group (b1-b4) four MapV maps for the file right.
Pseudo DNAs under Various Interactions

Figure 8. Three groups of twelve 2D maps in the range of $n=12$, $k=7$, $N=500$, $T=2700$ for the file $hc256$, $r\in\{1,2,3\}$, $mode=1$; (a) group (a1 - a4) four Map$_A$ maps $r=1$; (b) group (b1-b4) four Map$_T$ maps $r=2$; (c) group (c1 - c4) four Map$_C$ maps $r=3$.
Human DNAs vs. Pseudo DNAs

Figure 9. Three groups of twelve maps in the ranges: $N=500$, $T=2700$, $k=7$; (a) Real DNA Data; (a1-4) DNA sequences from the file right; (b-c) Simulation Data; (b1-4) Binary Sequences from the file $hc256$, $r=1$; (c1-4) Binary Sequences from the file $hc256$, $r=3$. 
Conclusion

- Using Variant Logic, Four DNA Meta States correspond Four Variant Meta States
- Pseudo DNAs can be generated under Various conditions to form Visual Maps
- Both Real & Artificial DNAs have stronger similarity
- Visual Maps may provide a General Classification for Genomic analysis on DNA Interactions
- Further Explorations are required…
References


Thanks