2nd International Summit on

Integrative Biology

August 4-5, 2014 Chicago, USA

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

Jeffrey Zheng
School of Software, Yunnan University

August 4, 2014

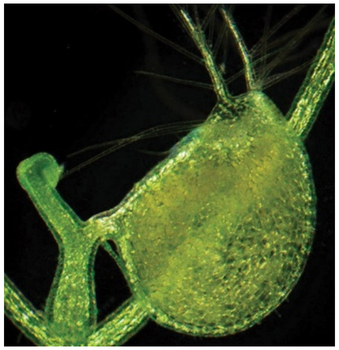
Content

- 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

30% Arabidopsis

90% Takifugu

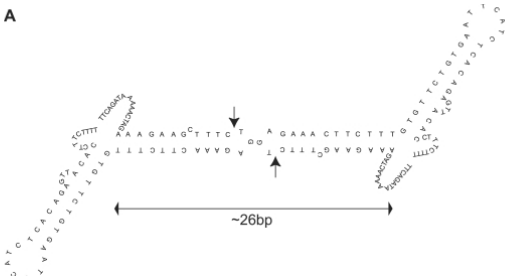
98% Human

- ENCODE: over 80% of DNA in the human genome "*serves some purpose, biochemically speaking*".
- However, this conclusion is strongly criticized ...

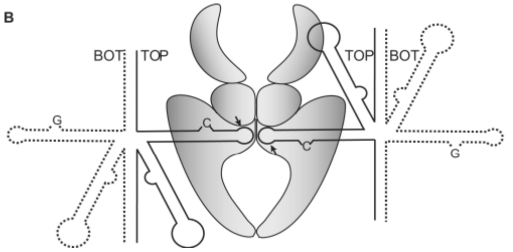
Tools to Analyze Non-Coding DNAs/RNAs

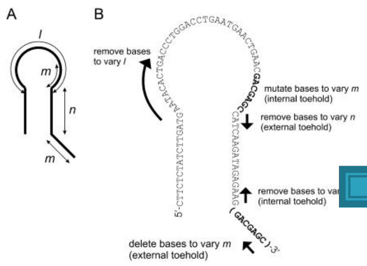
- Frequency Distribution
- GC densities
- Repeat sub-sequences
- …
- Machine Learning
- Bayesian Inference and Induction
- Neural Network
- Hidden Markov Model
- …

A case of Non-Coding DNA: Hairpin

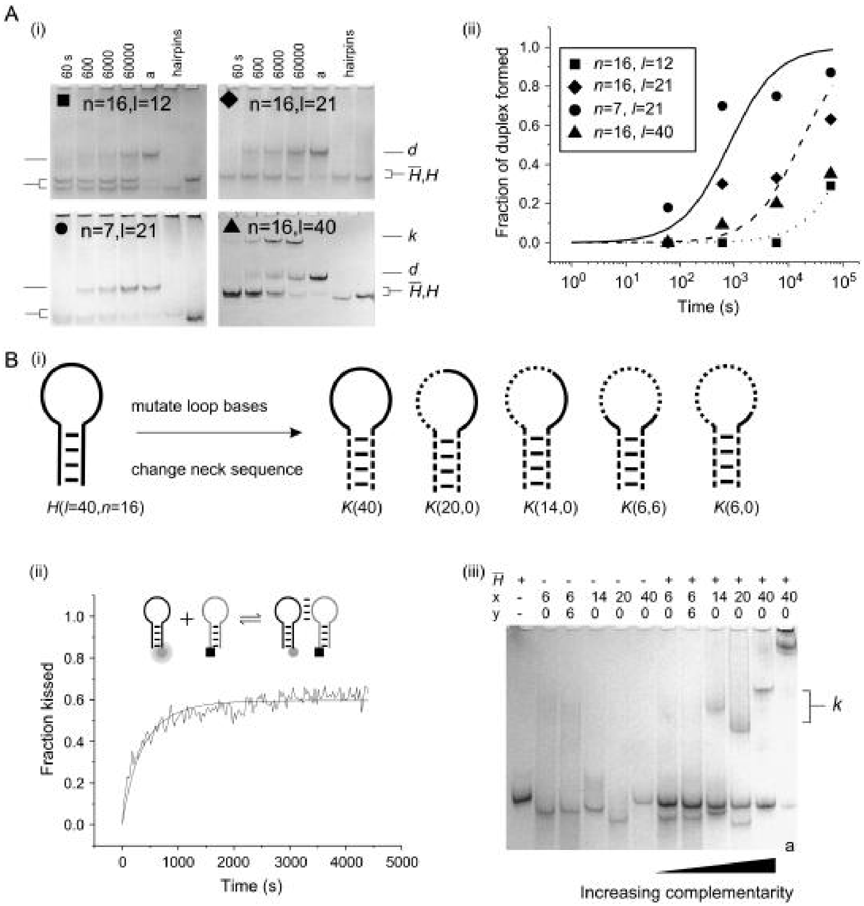


A DNA Sequence

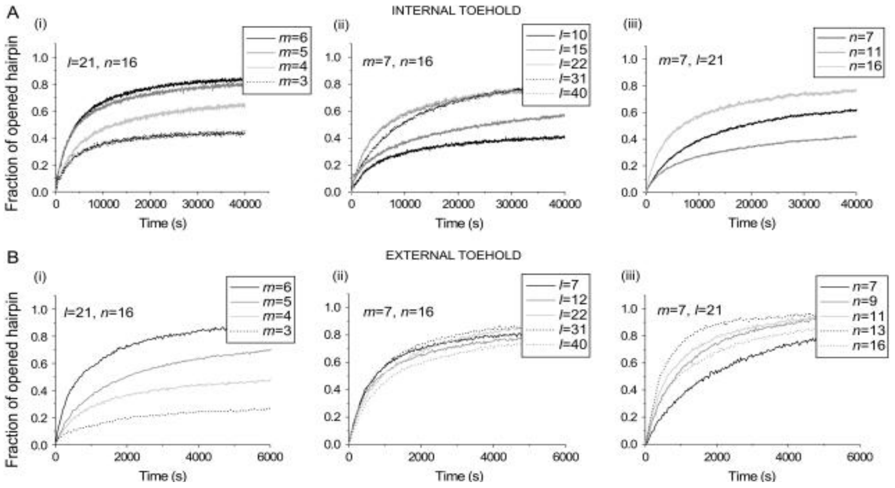



A
l
m
n
m
B
remove bases to vary l
mutate bases to vary m (internal toehold)
remove bases to vary n (external toehold)
remove bases to vary n (internal toehold)
delete bases to vary m (external toehold)

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

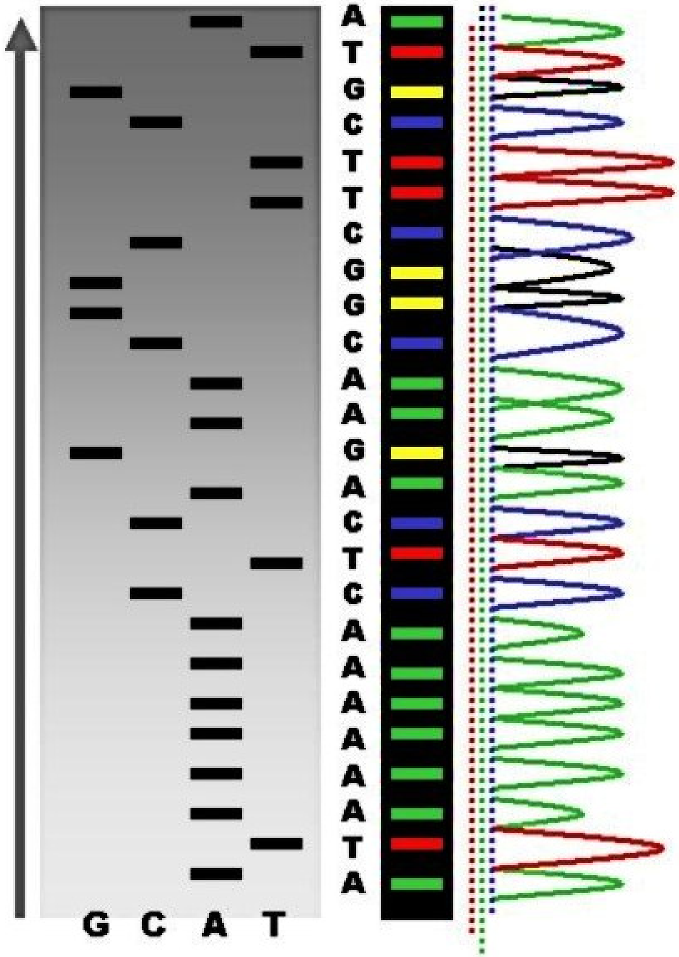
- Variant Logic
- DNAs & Pseudo DNAs
- General Model
- Main Procedure

Variant Logic

- An unified 0-1 logic framework base on input/output and logic functions using four Meta symbols: {⊥, +, −, ⊤}
  - 0-0 : ⊥ , 0-1 : + ,
  - 1-0 : − , 1-1 : ⊤ .

- Multiple Maps of Variant Phase Spaces can be visualized

Variant Logic & DNA Sequencing



Results of automated chain-termination DNA sequencing.

DNA Sequences	Variant Logic
G	0–0 : ⊥
A	0–1: +
T	1–0: –
C	1–1: ⊤

Four Meta States

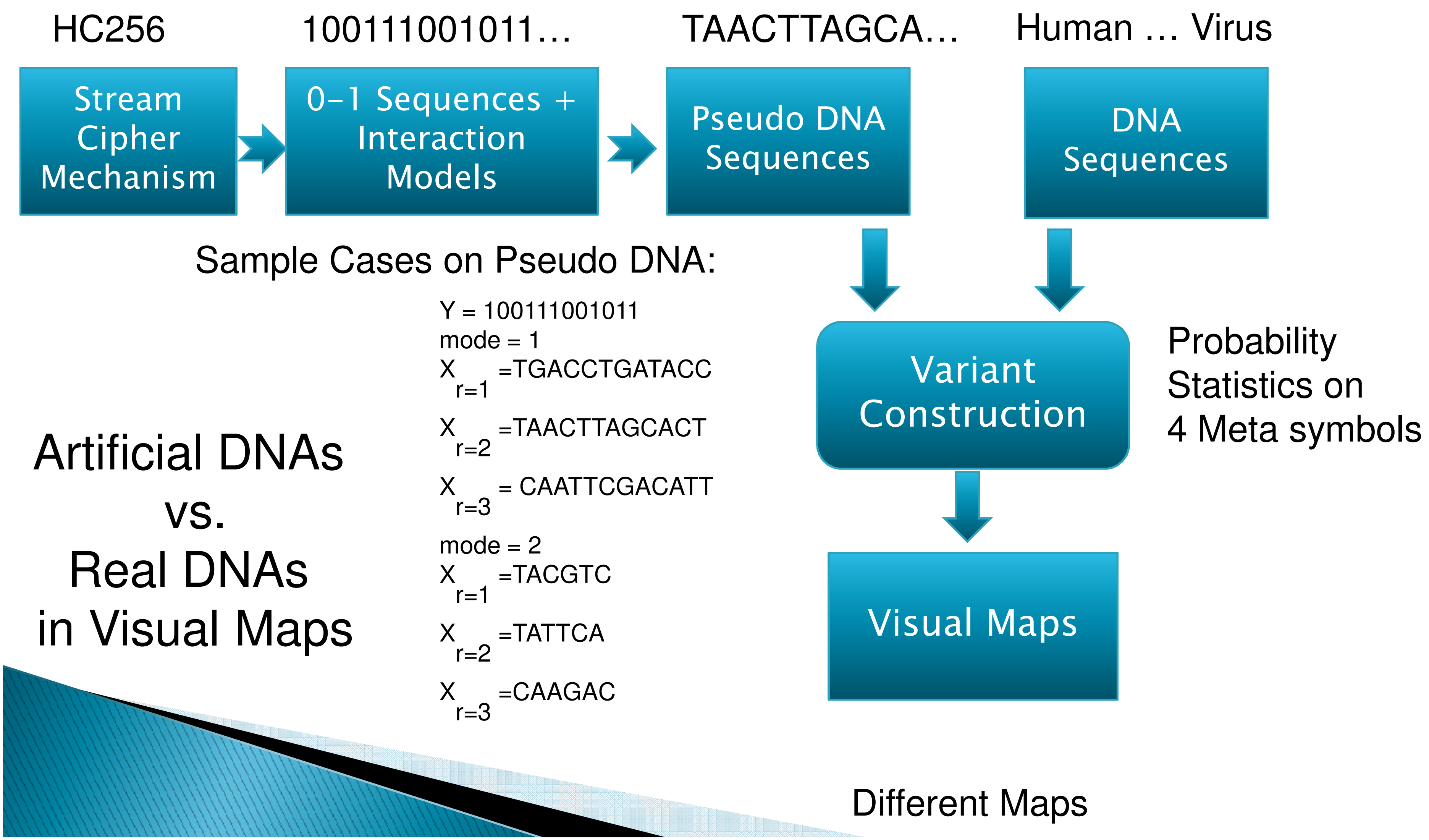
A Comparison Model to simulate Non-Coding DNAs in Visual Maps

- 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

Stream Cipher Mechanism

100111001011…

0-1 Sequences + Interaction Models

TAACTTAGCA…

Pseudo DNA Sequences

Human … Virus

DNA Sequences

Variant Construction

Probability Statistics on 4 Meta symbols

Visual Maps

Different Maps

Sample Cases on Pseudo DNA:

Y = 100111001011

mode = 1

$X_{r=1}$ =TGACCTGATACC

$X_{r=2}$ =TAACTTAGCACT

$X_{r=3}$ = CAATTCGACATT

mode = 2

$X_{r=1}$ =TACGTC

$X_{r=2}$ =TATTCA

$X_{r=3}$ =CAAGAC

Artificial DNAs
vs.
Real DNAs
in Visual Maps

Main Procedure

Input: Pseudo DNA/Real DNA Vector

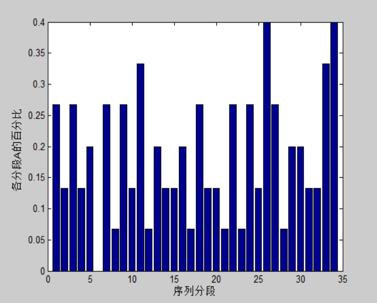
X^t: GGTACTTGCAT…

Projected as Four 0-1 vectors

M_G: 11000001000 …
M_A: 00010000010 …
M_T: 00100110001 …
M_C: 00001000100 …

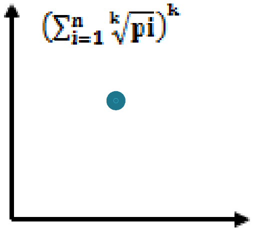
Calculated as four Probability Vectors

$$\{\rho_l^v\}_{0\le l<m_t}$$



Determine four pairs of map position

$$\{(x_v^k, y_v^k)\}_{v\in D}$$



$(\sum_{i=1}^{n} \sqrt[k]{pi})^k$

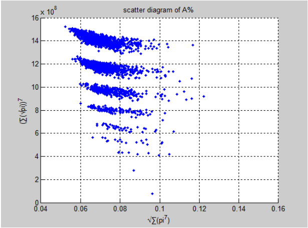
$\sqrt[k]{\sum_{i=1}^{n} pi^k}$

Collected all DNA Vectors

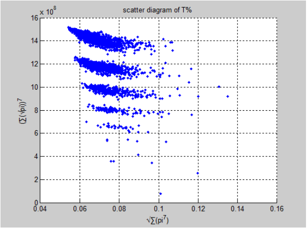
$$\forall t, X^t \in D^{N_t}$$

Four Maps constructed

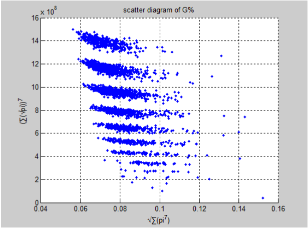
$$\{Map_v\}_{v\in D}$$



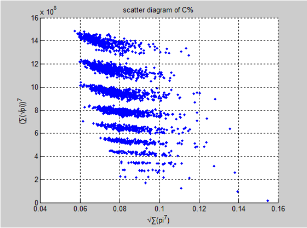
Map_A



Map_T



Map_G



Map_C

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= 500bp

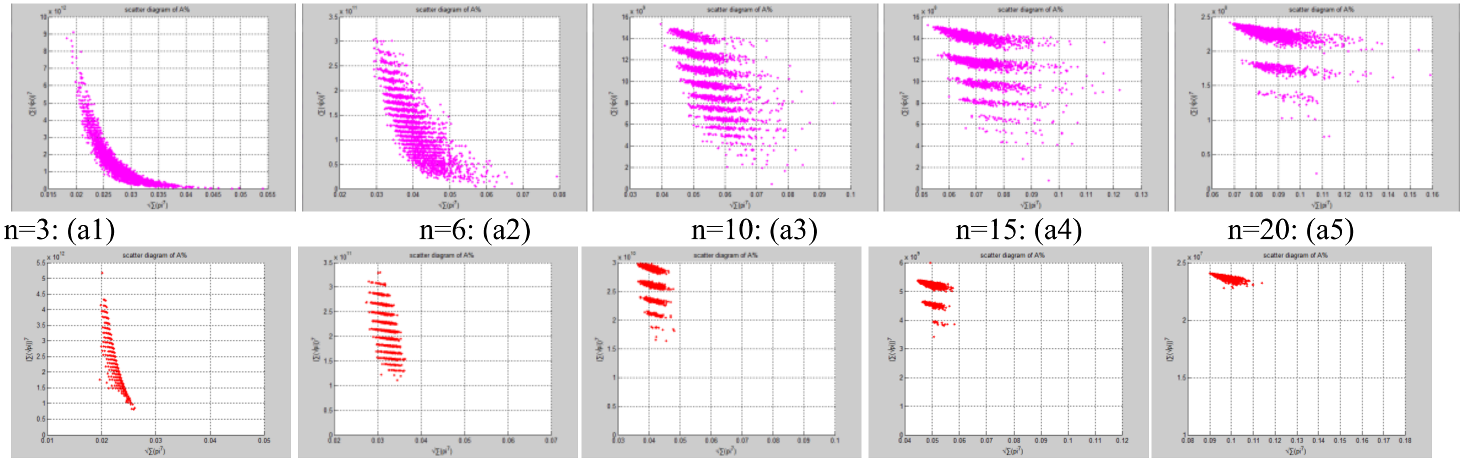
- For a sample point, a sequence could be

```
>SRR027962.18095784
TAATTCTTGAGTTCATGTCCCGCATCCAGGGCACACTTGTGCAAGGGGTGGGTTCCCAAGACCTTAT
GCAGCTCTGCCTCTGTGGCTTTGCAGTGTACAGTCACCATGGCTGCTGTCTTGGATCAGAGTTGAGT
GCCTGTGGTATTTCTAGGCTCAGGATGAAAGCTTCCCGTGGCTCTACCATTCAGGGATCTTGACGTG
GCGGCCCCATTCCCACAGCTCCTGTAGGTAGTGCCCCAGTGGGGACTCTGTGTGGAGGCTTCAATC
CCATATTTCCTGTTGGCACTGCCCTAGTGGACTTTTGATTTCTTTCTGATTCAGTCTTGGAAGGTTGT
GTGTTTCCAGGAATTTATCCATTTTCTCTAGGTTTTCTAGTTTATGCACACAAAGATATTCTGAGGATCT
TTTTTTGTGTCAGTGGTATCCTTTGCAATGTCTCATTTGTAATTTTTGATTGTGCTTATTGGAATCTTCTT
TTTTCTTGTATAATCTAACTAGCA
```

Human DNAs vs. Pseudo DNAs

Human DNA:

Pseudo DNA: HC256

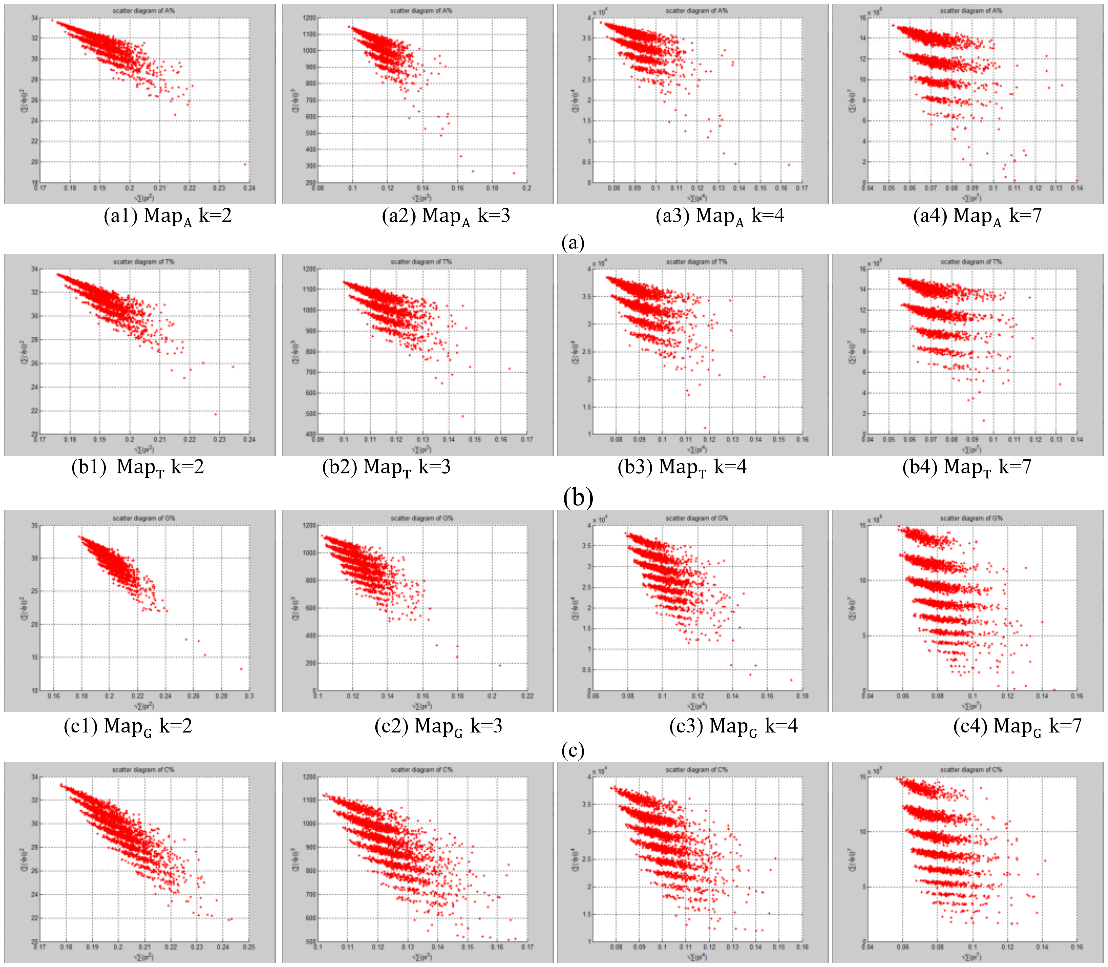


n=3: (a1) n=6: (a2) n=10: (a3) n=15: (a4) n=20: (a5)

(b1) (b2) (b3) (b4) (b5)

Two groups of ten 2D maps in the range of *$n=3\sim20$, $k=7$, $N\cong200\sim600$, $T=2700$*; (a1-a5) Map_A for the file *Right*; (b1-b5) Map_A for the file *hc256* $\mathrm{mode}=1, \mathrm{r}=1$.

Pseudo DNAs on various conditions



(a1) Map$_A$ k=2 (a2) Map$_A$ k=3 (a3) Map$_A$ k=4 (a4) Map$_A$ k=7

(a)

(b1) Map$_T$ k=2 (b2) Map$_T$ k=3 (b3) Map$_T$ k=4 (b4) Map$_T$ k=7

(b)

(c1) Map$_G$ k=2 (c2) Map$_G$ k=3 (c3) Map$_G$ k=4 (c4) Map$_G$ k=7

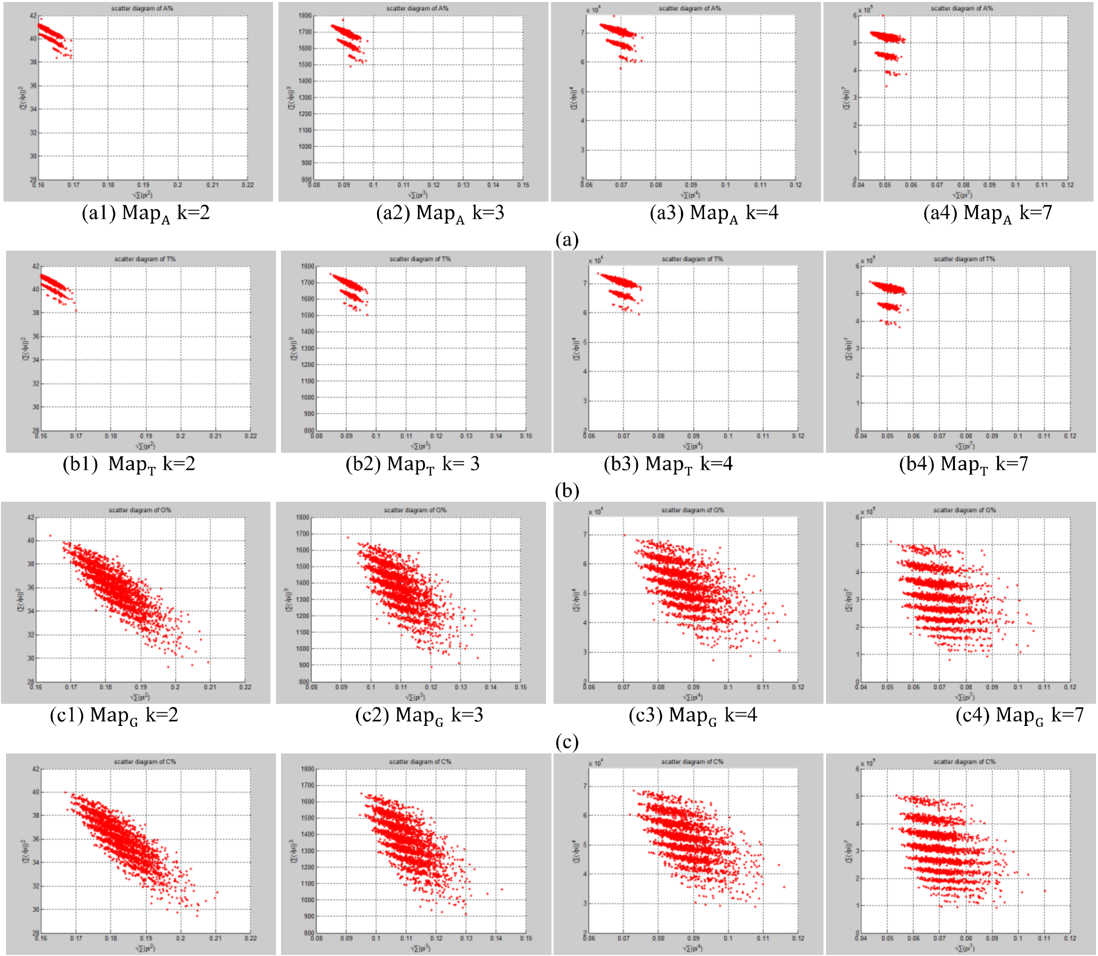
(c)

(d1) Map$_C$ k=2 (d2) Map$_C$ k=3 (d3) Map$_C$ k=4 (d4) Map$_C$ k=7

(d)

Figure 5. Four groups of sixteen 2D maps in the range of $n = 15,\ k = \{2,3,4,7\}, N \cong 500, T = 2700$; (a) group (a1 - a4) four Map$_A$ maps; (b) group (b1-b4) four Map$_T$ maps; (c) (c1 - c4) four Map$_G$ maps; (d) (d1 - d4) four Map$_C$ maps for the file *right*.

Pseudo DNA sequences on different parameters



(a1) Map_A k=2 (a2) Map_A k=3 (a3) Map_A k=4 (a4) Map_A k=7

(a)

(b1) Map_T k=2 (b2) Map_T k= 3 (b3) Map_T k=4 (b4) Map_T k=7

(b)

(c1) Map_G k=2 (c2) Map_G k=3 (c3) Map_G k=4 (c4) Map_G k=7

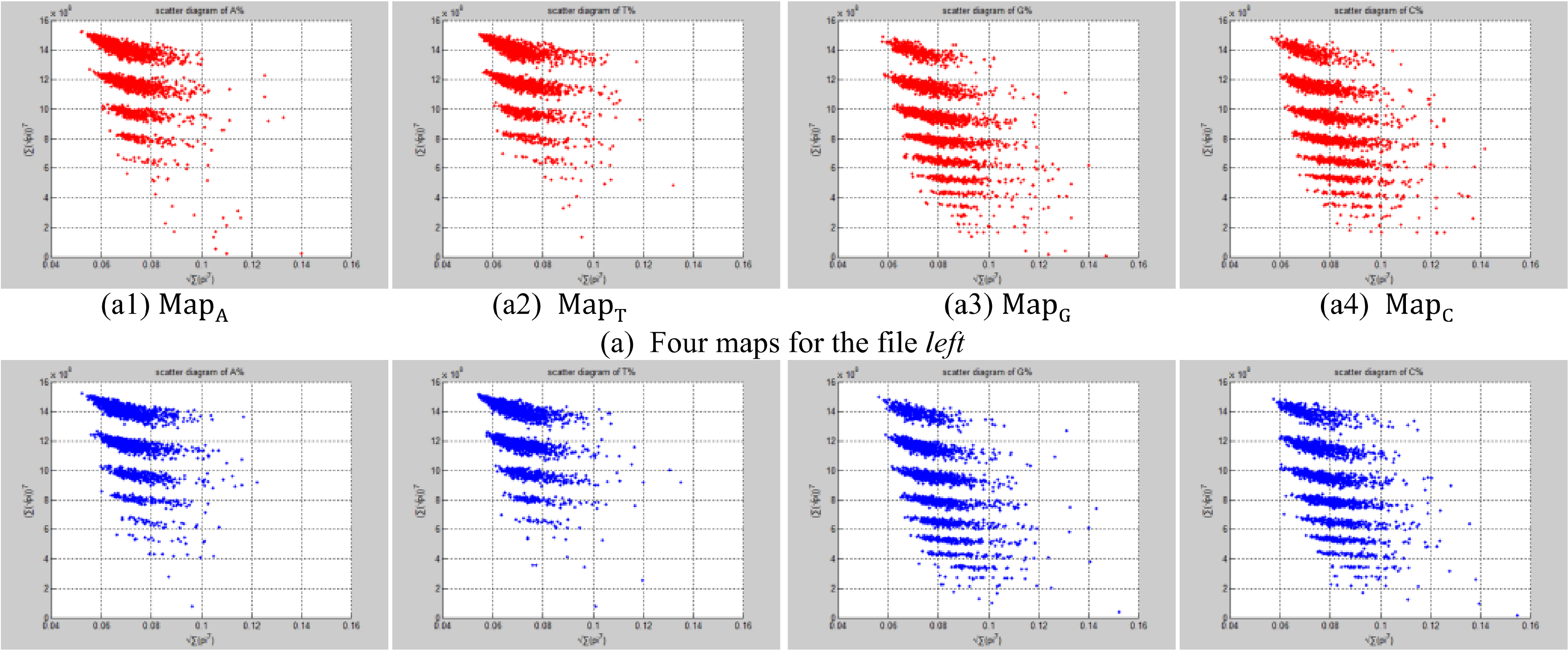
(c)

(d1) Map_C k=2 (d2) Map_C k=3 (d3) Map_C k=4 (d4) Map_C k=7

(d)

Figure 6. Four groups of sixteen 2D maps in the range of $\mathrm{n} = 12,\ \mathrm{k} = \{2,3,4,7\}, \mathrm{N} \cong 500, \mathrm{T} = 2700$ for the file *hc256*, $\mathrm{r} = 1, \mathrm{mode} = 1$; (a) group (a1 - a4) four Map_A maps; (b) group (b1-b4) four Map_T maps; (c) (c1 - c4) four Map_G maps; (d) (d1 - d4) four Map_C maps.

Two Groups of Human DNAs



(a1) Map_A (a2) Map_T (a3) Map_G (a4) Map_C

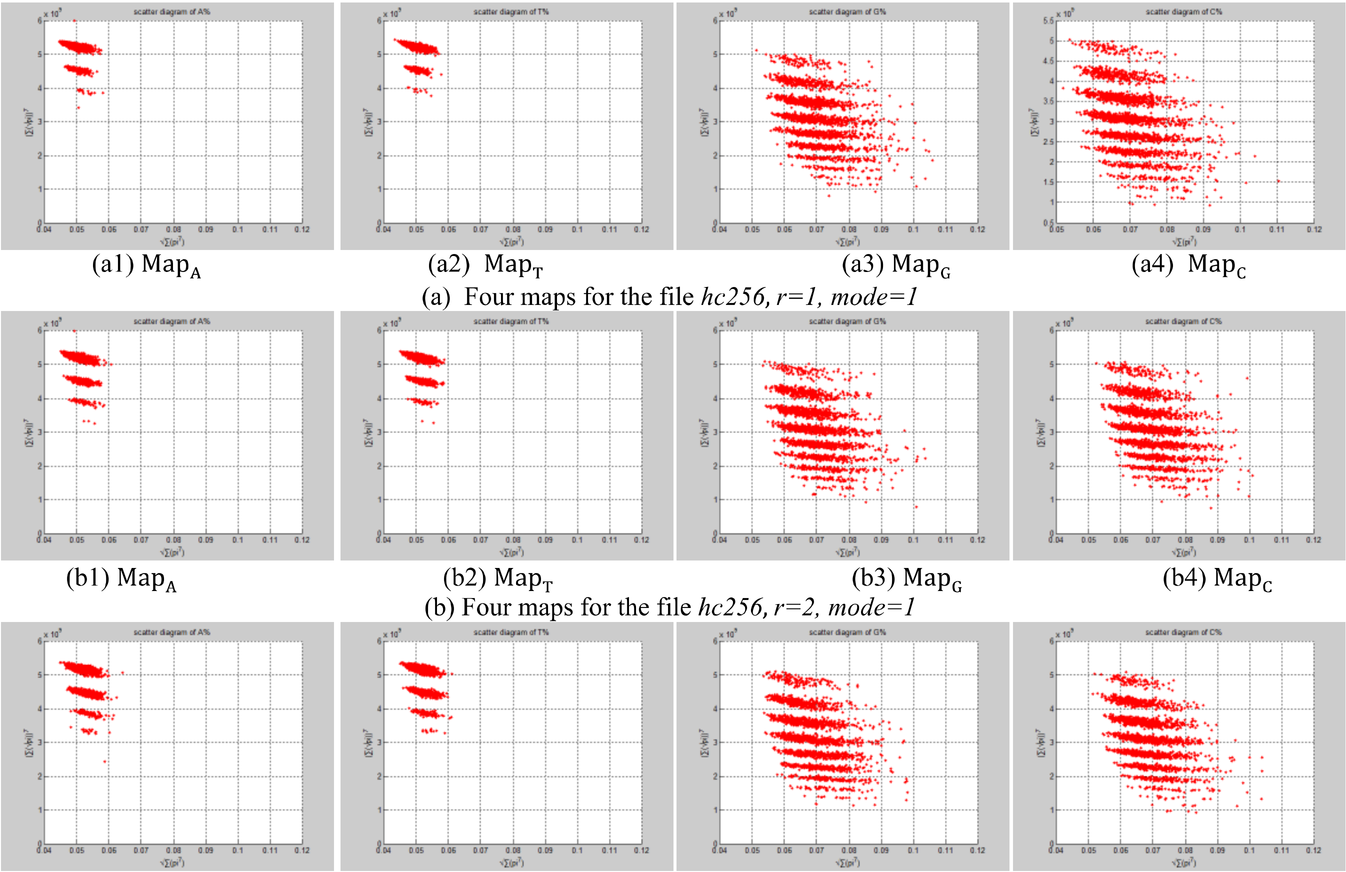
(a) Four maps for the file *left*

(b1) Map_A (b2) Map_T (b3) Map_G (b4) Map_C

(b) Four maps for the file *right*

Figure 7. Two groups of eight 2D maps in the range of $n = 15,\ k = 7, N \cong 200{\sim}600, T = 2700$; (a) group (a1 - a4) four Map_V maps for the file *left*; (b) group (b1-b4) four Map_V maps for the file *right*.

Pseudo DNAs under Various Interactions



(a1) Map_A (a2) Map_T (a3) Map_G (a4) Map_C

(a) Four maps for the file *hc256, r=1, mode=1*

(b1) Map_A (b2) Map_T (b3) Map_G (b4) Map_C

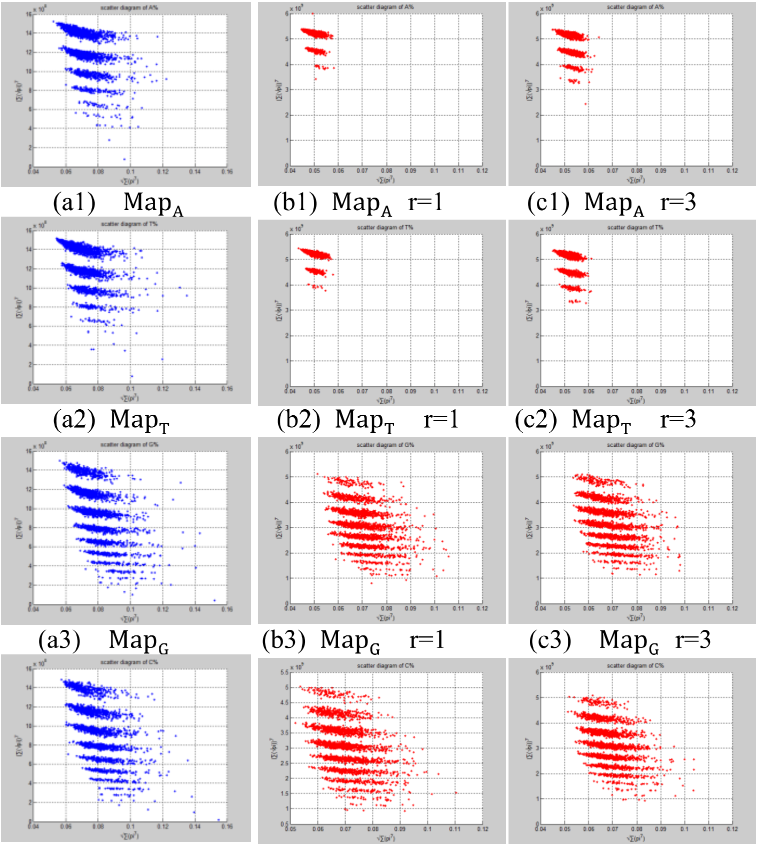
(b) Four maps for the file *hc256, r=2, mode=1*

(c1) Map_A (c2) Map_T (c3) Map_G (c4) Map_C

(c) Four maps for the file *hc256, r=3, mode=1*

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={1,2,3}, mode=1;* (a) group (a1 - a4) four Map_V maps *r=1*; (b) group (b1-b4) four Map_V maps *r=2*; (c) group (c1 - c4) four Map_V maps *r=3*

Human DNAs vs. Pseudo DNAs



(a1) Map$_A$ (b1) Map$_A$ r=1 (c1) Map$_A$ r=3

(a2) Map$_T$ (b2) Map$_T$ r=1 (c2) Map$_T$ r=3

(a3) Map$_G$ (b3) Map$_G$ r=1 (c3) Map$_G$ r=3

(a4) Map$_C$ (b4) Map$_C$ r=1 (c4) Map$_C$ r=3

(a1)-(a4) Four maps for the file *right, n=15, mode=0*

(b1)-(b4) Four maps for the file *hc256, n=12, r=1, mode=1*

(c1)-(c4) Four maps for the file *hc256, n=12, r=3, mode=1*

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

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

1. B. Banfai, H. Jia, J. Khatun et al. (2012) Long noncoding RNAs are rarely translated in two human cell lines, *Genome Research*, Cold Spring Harbor Laboratory Press, 22:1646-1657 Doi:10.1101/gr.134767.111
2. J.M. Engreitz, A. Pandya-Jones, P. McDonel et al. (2013) Large Noncoding RNAs can Localize to Regulatory DNA Targets by Exploriting the 3D Architecture of the Genome, *Proceedings of The Biology of Genomes*, Cold Spring Harbor Laboratory Press, 122
3. J. Zheng, C. Zheng and T. Kunii (2011) A Framework of Variant Logic Construction for Cellular Automata, in *Cellular Automata – Innovative Modelling for Science and Engineering*, Edited by A. Salcido, InTech Press, 325-352, 2011. http://www.intechopen.com/chapters/20706
4. J. Zheng, W. Zhang, J. Luo, W. Zhou and R. Shen (2013) Variant Map System to Simulate Complex Properties of DNA Interactions Using Binary Sequences, *Advances in Pure Mathematics*, 3 (7A) 5-24. doi: 10.4236/apm.2013.37A002
5. J. Zheng, J. Luo and W. Zhou (2014) Pseudo DNA Sequence Generation of Non-Coding Distributions Using Variant Maps on Cellular Automata," *Applied Mathematics*, 5(1) 153-174. doi: 10.4236/am.2014.51018
6. J. Zheng, W. Zhang, J. Luo, W. Zhou, V. Liesaputra (2014) Variant Map Construction to Detect Symmetric Properties of Genomes on 2D Distributions. *J Data Mining Genomics Proteomics* 5:150. doi: 10.4172/2153-0602.1000150

Thanks