Journal of
Data Mining in Genomics & Proteomics
ISSN: 2153-0602

Zheng et al., J Data Mining Genomics Proteomics 2014, 5:1
DOI: 10.4172/2153-0602.1000150

Research Article Open Access

Variant Map Construction to Detect Symmetric Properties of Genomes on 2D Distributions

Jeffrey Zheng[1]*, Weiqiong Zhang[2], Jin Luo[3], Wei Zhou[1] and Veronica Liesaputra[4]

[1]School of Software, Yunnan University, Kunming Yunnan, China
[2]School of Software and Microelectronics, Peking University, Beijing, China
[3]School of Life Sciences, Yunnan University, Kunming Yunnan, China
[4]School of Computing, Unitec, Auckland, New Zealand

Abstract

Visualization Methods have played a key role in the Human Genome Project. After further development in other international projects such as ENCODE, larger numbers of Genome Databases are established and mass Genome-wide gene expression measurements are developed. In current situation, it is necessary to shift targets in computational cell biology from collecting sequential data to making higher-level interpretation and exploring efficient content-based retrieval mechanism for genomes. Mammalian genomes encode thousands of large non-coding RNAs (lncRNAs), many of which regulate gene expression, interact with chromatin regulatory complexes, and are thought to play a role in localizing these complexes to target loci across the genome. Using higher dimensional visualization tools, their complex interactive properties could be organized as different visual maps. The Variant Map Construction VMC as an emerging scheme is systematically proposed in this paper to apply multiple maps that uses four Meta symbols as same as DNA or RNA representations. System architecture of key components and core mechanism on the VMC are described. Key modules, relevant equations and their I/O parameters are discussed. Using the VMC system, two DNA data sets of multiple real sequences are tested to show their visible properties in higher levels of intrinsic relationships among relevant DNA sequences in 2D maps. Visual results are briefly analyzed to explore their intrinsic properties and symmetric characteristics on relevant genome sequences under 2D maps of the Variant Map Construction. A set of sample 2D maps are included and their characteristics are illustrated under various controllable environment.

Keywords: Symmetric property; Genomic sequence; Partition; Segment; Measurement; 2D maps; Visual distribution; Variant map construction

Introduction

Visualization Methods have played a key role in the Human Genome Project (HGP) [1,2]. After HGP completed successfully, a public research consortium-the Encyclopedia of DNA Elements (ENCODE) were launched by the National Human Genome Research Institute (NHGRI) in 2003 to find all functional elements in the human genome as one of the most critical projects by NHGRI to explore genomes after HGP.

In 2012, ENCODE released a coordinated set of 30 papers published in key Journals of Nature, Genome Biology and Genome Research. These publications show that approximately 20% of noncoding DNA in the human genome is functional while an additional 60% is transcribed with no known function [3]. Much of this functional non-coding DNA is involved in the regulation of the expression of coding genes [4]. Furthermore the expression of each coding gene is controlled by multiple regulatory sites located both near and distant from the gene. These results demonstrate that gene regulation is far more complex than was previously believed [5]. Mammalian genomes encode thousands of large non-coding RNAs (lncRNAs), many of which regulate gene expression, interact with chromatin regulatory complexes, and are thought to play a role in localizing these complexes to target loci across the genome [6]. Associated with different international projects, larger numbers of Genome Databases are established and mass Genome-wide gene expression measurements are developed. Due to huge amount of DNA sample collections and extremely difficulties to determine their variation properties in wider applications [7], it is essential for us to extend advanced DNA analysis models, methods and tools in further extensions to explore emerging models and concepts to interpret complex interactions among complicated sets of DNA sequences in real environments.

In current situation, it is necessary for advanced researchers to shift targets in computational cell biology from directly collecting sequential data to making higher-level interpretation and exploring efficient content-based retrieval mechanism for genomes. Using higher dimensional visualization tools, their complex interactive properties could be organized as different visual maps systematically.

DNA Analysis

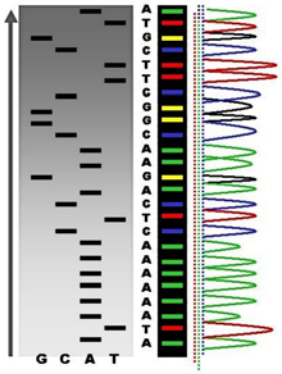
DNA analysis plays a key role in modern genomic application [2]. The HGP is heavily relevant to advanced DNA sequencing and analysis techniques. DNA sequences are composed of four Meta symbols on {A, T, G, C} as basic structure. Classical DNA double helix structure makes the first level of pair construction of DNA sequences with A & T and G & C complementary structures as the first level of symmetric relationships. A typical DNA sequencing result is shown in Figure 1a. Four Meta symbols could be separated as four projective sequences.

In ENCODE, recent Genomic analysis results are indicated that encoded sequences have only 20 percent in human genomes and around 80 percent genomes look like useless sequences. Under further assumptions, it seems that additional symmetric properties are required to satisfy the second, third and higher levels of structural constructions

***Corresponding author:** Jeffrey Zheng, School of Software, Yunnan University, Kunming Yunnan, China, Tel: 86-13108839090; E-mail: conjugatesys@gmail.com

Received July 24, 2013; **Accepted** January 26, 2014; **Published** January 29, 2014

Citation: Zheng J, Zhang W, Luo J, Zhou W, Liesaputra V (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



(a)

DNA Sequences	Variant Logic
G	0-0 : ⊥
A	0-1: +
T	1-0: −
C	1-1: T

(b)

Figure 1: Modern DNA sequencing & Their correspondences on Variant Logic; (a) a sample DNA sequencing result and its four projection sequences; (b) Four Meta DNA Symbols and relevant linkages to Variant Logic.

to explore complex interactive properties [1-12].

Variant Construction and DNA

Variant construction is a new structure composed of logic, measurement and visualization models to analyze 0-1 sequences under variant conditions. The further details of this construction can be checked on variant logic [13,14], 2D maps [15,16] and variant phase spaces [17]. Since the variant system uses another set of four Meta symbols $\{\perp, +, -, T\}$ to describe system, this typical correspondence shown in Figure 1b may provides a natural mapping between DNA and variant data sequences.

Since DNA sequences are played an essential role to explore different symmetric properties based on analysis approaches, in this paper, measurement and visual models are proposed systematically to use a fixed segment structure to measure four meta symbols distributions in their spectrum construction. Under this construction, refined symmetric features can be identified from their polarized distributions and further symmetric properties are visualized.

Target of this Paper

The target of this paper is to establish the Variant Map Construction (VMC) as an emerging scheme systematically based on variant logic schemes [13-17] to apply multiple maps that uses four Meta symbols as same as DNA or RNA representations. System architecture of key components and core mechanism on the VMC are described. Key modules, relevant equations and their I/O parameters are discussed. Applying the VMC system, two DNA data sets of multiple real sequences are collected to show their intrinsic properties in higher levels of intrinsic relationships among relevant DNA sequences on various 2D maps. Further detailed descriptions and discussions are contained.

System Architecture

In this section, system architecture and their core components are briefly discussed with the use of diagrams. The refined definitions and equations of this system are described in the next section–Variant Map Construction.

Architecture

The three components of a variant map construction are the Binary Probability Measurement (BPM), Mapping Position (MP), and Visual Map (VM) as shown in Figure 2. The architecture is shown in Figure 2a with the key modules of the three core components being shown in (Figures 2b-2d) respectively.

In the first part of the system, four vectors of probability measurements are created from the t-th selected DNA sequence with N_t elements as an input. Multiple segments are partitioned by a fixed number of n elements for each segment; at least m_t segments can be identified by the BPM component. Next component uses the four vectors of probability measurements and a given k value as input data, a pair of position values are created for each Meta symbol. Four pairs of values are generated by the MP component. Then, in order to process multiple selected DNA sequences, all selected sequences are processed by the VM component and each sequence may provide a set of pair values to generate relevant variant maps to indicate their distribution properties respectively. With six parameters in an input group, there are two sets of parameters in the intermediate group and one set of parameters in the output group. The three groups of parameters may be listed as follows.

1. Inputgroup:

t An integer indicates the t-th DNA sequence selected, $0 \le t<T$

n An integer indicates the number of elements in a segment, n>0

N_t An integer indicates the number of elements in the t-th DNA sequence, $N_t >> n$

k An integer indicates the control parameter for mapping, K>0

VA symbol is selected from four DNA symbols {A, G, T, C}=D, $V \in D$

X^t An input DNA vector with elements, $X^t \in D^{N_t}$

2. Intermediate group:

Four sets of probability measurements with $0 \le 1<m_t$, $V \in D$ {(x)}

Four paired values, k>0, $V \in D$

3. Output group:

$\{Map_v\}$ Four 2D maps for distributions, $V \in D$

BPM-Binary probability measurement

The BPM component as shown in Figure 2b is composed of two modules: BM Binary Measure and PM Probability Measurement. Five parameters are shown as input signals; four vectors of binary measures are outputted from the BM component as an intermediate group and four sets of probability measurements are outputted as output group.

Input group: tan integer indicates the t-th DNA sequence selected, $0 \le t<T$

n An integer indicates the number of elements in a segment, n > 0

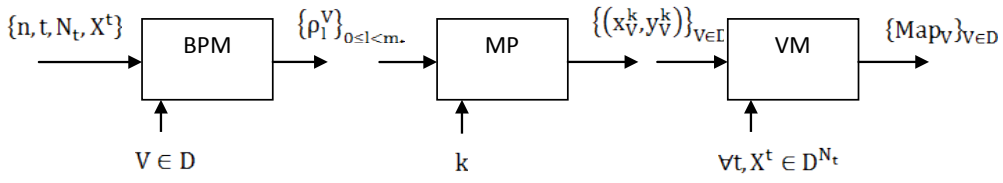
N_t An integer indicates the number of elements in the t-the DNA sequence, $N_t >> n$

VA symbol is selected from four DNA symbols {A,G,T,C}=D, $V \in D$

X^t An input DNA vector with N_t elements, $X^t \in D^{N_t}$

Intermediate group: $\{M_v^t\}$ Four 0-1 vectors with N_t elements, $M_v^t(\mathbf{I}) \in \{0,1\} = B, M_v^t \in B^{N_t}, V \in D$

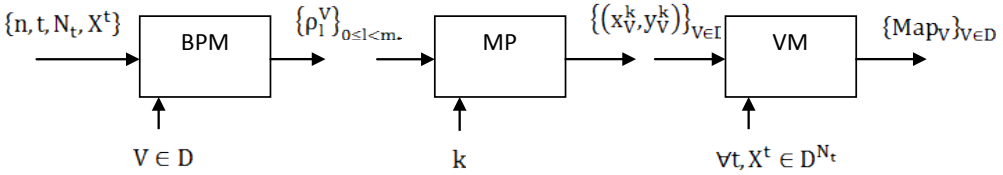
Output group: $\{\rho_l^V\}$ Four sets of probability measurements with $0 \le 1<m_t$, $V \in D$



$0 \le t < T,\ 0 < n \ll N_t,\ X^t \in D^{N_t},\quad m_t = \left\lfloor N_t / n \right\rfloor$

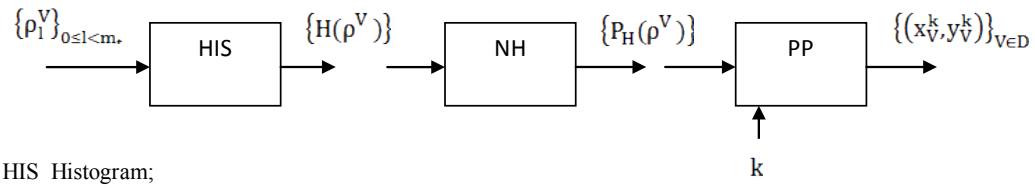
BPM Binary Probability Measurement;
MP Mapping Position;
VM Visual Map

(a) Architecture of VMC Variant Map Construction composed of three components: BPM, MP and VM



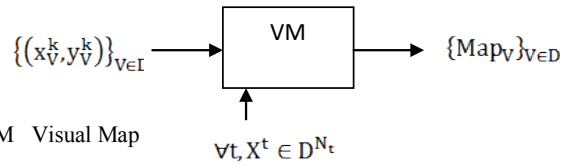
BM Binary Measure;
PM Probability Measurement.

(b) BPM Binary Probability Measurement composed of two components: BM and MP



HIS Histogram;
NH Normalized Histogram;
PP Pair Position

(c) MP module composed of three components: HIS, NH and PP



VM Visual Map

(d) VM module is itself: VM

Figure 2: Variant Map Construction VMC and key components (a) Architecture; (b) BPM component; (c) MP component; (d) VM component.

The BPM component transforms a selected DNA sequence to generate four 0-1 vectors by BM module for the input DNA sequence. Then four probability vectors are generated by the PM module as the output of the BPM under a fixed length of segment condition.

MP-Mapping position

The MP component as shown in Figure 2c is composed of three modules: HIS Histogram, NH Normalized Histogram and PP Pair Position. Two parameters are shown as input signals; four histograms and four normalized histograms are generated from the HIS component and the NH component as intermediate groups respectively. Then four paired values are generated by the PP component as the output group.

Input group: $\{\rho_l^V\}$ Four sets of probability measurements with $0 \le 1<m_t$, $V \in D$

An integer indicates the control parameter for mapping, $k > 0$

Intermediate group: {H(ρV)}Four histograms for relevant probability measurements, $V \in D$

{PH(ρV)} Four normalized histograms for relevant probability measurement, $V \in D$

Output group: $\{(x_V^k, y_V^k)\}$ Four paired values, $k > 0, V \in D$

The MP component uses probability measurements as input, under a given k condition to generate each relevant histogram and its normalized distribution. The output of the MP component is composed of four paired values controlled in a given condition

VM-Visual map

The VM component shown in Figure 2d is composed of one module: VM Visual Map. Three parameters are input signals. Collected all selected DNA sequences, four 2D maps are generated by the VM component as the output result.

Inputgroup: ∀t all possible DNA sequences are selected, $0 \le t < T$

X^t An input DNA vector with N_t elements, $X^t \in D^{N_t}$

$\{(x_V^k, y_V^k)\}^t$ Four paired values for the t-th DNA sequence, $k > 0, V \in D$

Output group: $\{Map_V\}$ Four 2D maps for distributions, $V \in D$

The VM component process all selected DNA sequences as input to generate relevant paired values for each sequence. The output of the VM component is composed of four 2D maps to show the final visual distribution for the system.

Variant Map Construction

In this section, brief definitions and equations are provided to describe the proposed VMP system. In addition to the initial preparation, six core modules are involved in the BM, PM, HIS, NH, PP and VM components respectively.

Initial preparation

Let X denotes a DNA sequence with N elements, D denotes a symbol set with four elements i.e. D={A,G,T,C}. This type of a DNA sequence can be described as a four valued vector follows:

$X=(X(0),\ldots,X(I),\ldots,X(N-1)),\ 0 \le I<N, X(I) \in D=\{A,G,T,C\}, X \in D^N$.

BM module

For a given I-th element, four projective operators can be defined and denoted as

$\{M\_A\ (I), M\_G\ (I), M\_T\ (I), M\_C\ (I)\}$

$M_A(I)=\begin{cases}1, \text{if } X(I)=A;\\ 0, \text{Otherwise;}\end{cases} M_G(I)=\begin{cases}1, \text{if } X(I)=G;\\ 0, \text{Otherwise;}\end{cases} M_T(I)=\begin{cases}1, \text{if } X(I)=T;\\ 0, \text{Otherwise;}\end{cases}$ and $M_T(I)\begin{cases}1, \text{if } X(I)=C;\\ 0, \text{Otherwise;}\end{cases}$ respectively

Applying the four operators to all elements, the DNA sequence X can be reorganized as four binary sequences in 0-1 values. i.e.

$$M_V:\{X(I)\}_{I=0}^{N-1} \to \{M_A(I), M_G(I), M_T(I), M_C(I),\}_{I=0}^{N-1}, M_V(I) \in B=\{0,1\}, V \in D$$

E.g. Let a DNA sequence X=CTGATTAGCCAT, N = 12, its four binary sequences can be represented as follows.

X=CTGATTAGCCAT

M_A=000100100010

M_G=001000010000

M_T=010011000001

M_C=100000101100

It is interesting to notice that the basic relationship between a DNA sequence X and its four M_V sequences are exactly same as in a modern DNA sequencing procedure to separate a selected DNA sequence into the four Meta symbol sequences shown in Figure 1a. This correspondence could be the key feature to apply the proposed scheme naturally in simulating complex behaviors for any DNA sequence. The projection M_V provides the required operation in the BM component as the first module shown in Figure 2b.

PM module

For this set of the four binary sequences, it is convenient to partition them into m segments and each segment contained a fixed number of n elements. For the l-th segment, let $0 \le l<m$, $0 \le j<n$, the I-th position will be I=1*n+J, four probability measurements $\{\rho_A, \rho_G, \rho_T, \rho_C,\}$ can be defined.

$$\rho_l^v = \frac{\sum_{I=1*n}^{1*(n+1)-1} M_v(I)}{n}, V \in D,\ 0 \le I<N=n*m$$

Under this construction, four sets of probability measurements established.

$$\rho_l^v:\{M_A(I), M_G(I), M_T(I), M_C(I),\}_{I=0}^{N-1} \circledR \{\rho_l^A, \rho_l^G, \rho_l^T, \rho_l^C\}_{l=0}^{m-1}$$

The probability operator ρ^V generates four probability measurement vectors in the PM component as the second module shown in Figure 2b. After the BM and PM processes, the whole procedure of the BPM component is complete in Figure 2b.

HIS module

Since the BPM generates four sets of probability measurement, it is necessary to perform further operations in the MP component shown in Figure 2c as follows.

In the HIS component as the first module in Figure 2c, each probability sequence $\{\rho_l^V,\}_{l=0}^{m-1}, V \in D$ can be calculated from n positions, at most n+1 distinguished value identified in a vector. Under this organization, a histogram distribution can be established.

Let H (.) be a histogram operator, for each position, it satisfies following relation,

$$H(\rho_{0,}^{1},)=\begin{cases}1, \text{if} & \rho_l^V=\frac{i}{n} V \in D;\\ 0, \text{ Otherwise;} & 0 \le i \le n.\end{cases}$$

$$H(\rho_1^V)=\begin{cases}1, \text{ if } & \rho_1^V=\frac{i}{n}V\in D;\\ 0, \text{ Otherwise;} & 0\le i\le n.\end{cases}$$

Collecting all possible values, a histogram distribution can be established,

$$H(\rho^V)=\sum_{1=0}^{m-1}H(\rho_1^V)$$

The histogram $H(\rho^V)$ is the output of the HIS module. Four histograms are generated after HIS process. Further normalized process will be performed in the NH component as the second module in Figure 2c.

NH module

Under this construction, a normalized histogram can be defined as $P(\rho^V)=H(\rho^V)/m$

After the NH component processed, its output provides the PP component for further operations as the third module in Figure 2c.

PP Module

Relevant probability vectors have (n+1) distinguished values; four sets of normalized vectors can be organized as a linear order as follows,

$$p_i^V=\sum_{1=0}^{m-1}H(\rho_1^V=\frac{i}{n})/m,\ 0\le i\le n$$

Under this condition, four linear sets of probability vectors are established,

$$P_H(\rho^V)=\{p_i^A,p_i^G,p_i^T,p_i^C\}_{i=0}^n,\qquad p_i^V\in[0,1],\quad V\in D,\quad 0\le i\le n$$

For four vectors, their components can be normalized respectively,

$$\sum_{i=0}^{m-1}p_i^V=1,\qquad V\in D$$

Four sets of probability vectors are composed of a complete partition on their measurements.

Using this set of measurements, two mapping functions can be established to calculate a pair of values to map analyzed DNA sequence into a 2D map as follows.

Let y=F (P,V,k) and x= F (P,V,1/k) or (x_v^k, y_v^k) be a pair of values defined by following equations,

$$y_v^k=F(P,V,k)=\left(\sum_{i=0}^n\sqrt[k]{p_i^V}\right)^k\ \&\ x_v^k=F(P,V,1/k)=\sqrt[k]{\sum_{i=0}^n(p_i^V)^k},\ V\in D$$

In the PP component, four paired values are generated and each pair indicates a specific position on a 2D map relevant to the selected DNA sequence. The core operations of two key components: BPM and MP for a selected DNA sequence are performed in Figures 2b and 2c.

VM module

Since only one point of a 2D map is determined for a selected DNA sequence, it is essential to apply larger number of DNA sequences as inputs to generate visible distributions. This type of operations will be performed in the VM component shown in Figure 2d.

In a general condition, the VM component processes a selected DNA set $\{X^t\}_{t=0}^{T-1}$ composed of T sequences, the t-th sequence with N_t elements can be expressed by $X^t=(X^t(0),...,X^t(I),...,X^t(N_t\text{-}1),X^t\in D^{N_t}$ each sequence can be processed to apply the same procedures of the BPM and MP components. Since for each segment, its length n will be fixed for all selected sequences, it is essential to make number of segments be $m^t=\lceil N_t/n\rceil$ in convention to match each sequence. Under this expression, the last module VM collects all T pairs of positions on relevant 2D visual maps as follows,

$$VM:\{X^t\}_{t=0}^{T-1}\rightarrow\{x_V^k,y_V^k\}_{t=0}^{T-1}\rightarrow\{MAP_V\},\quad V\in D$$

A sample 2D map of VM is shown in Figure 3; this provides an assistant illustration for this type of visual maps on a case of multiple sequences.

Under this construction, T DNA sequences are transformed as T visual points on four 2D visual maps that would be help analyzers to explore their intrinsic symmetry properties among four binary sequences.

Sample Results on 2D Maps

Data resources

It is important to use some real DNA sequences to illustrate various test results of the VMC system. Two sets of DNA sequences are selected and relevant resource features are briefly described as follows.

The first data set originally comes from the human genome assembly version 37 and was taken from the reference sequences of 13 anonymous volunteers from Buffalo, New York. Hi-C technology [1] used to analyze chromatin interaction role in genome. From a genomic analysis viewpoint, this set of data may contain more complex secondary or higher level structures. A special structure nearly the GRCh37 DNA sequence has been identified to explore their spatial characteristics. After positive and negative sequencing, each data file contain 2700 DNA sequences and each sequence has around 500 elements stored in two files left and right respectively.

The second data set are selected from some plant gene database for comparison. Two DNA sequence sets are stored in file201-500 and random_201-500. Each data file contains 2700 DNA sequences and each sequence has around 200-600 elements. Both of them may be ordinary single sequences without complex secondary structures.

Sample results

Using the two sets of DNA sequences, four groups of 2D maps are listed in Figures 4-7 under different conditions to illustrate their spatial distributions using Variant Map Construction in a controllable environment. In Figure 4, two groups of ten 2D maps are shown in the range of $n=3, k=7, N\cong 200\sim 600, T=2700$ for comparison; (a1-

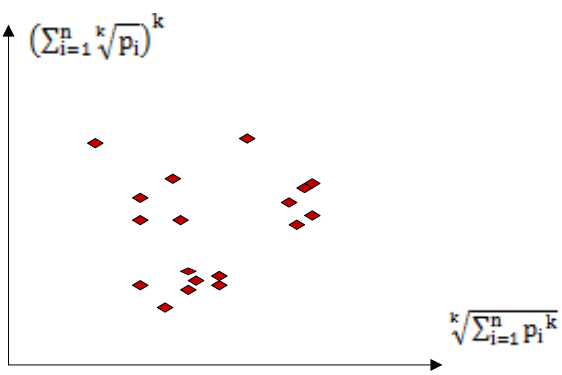
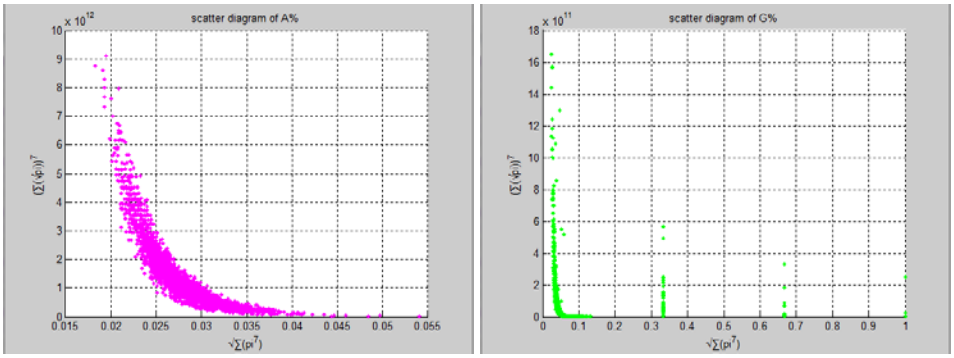
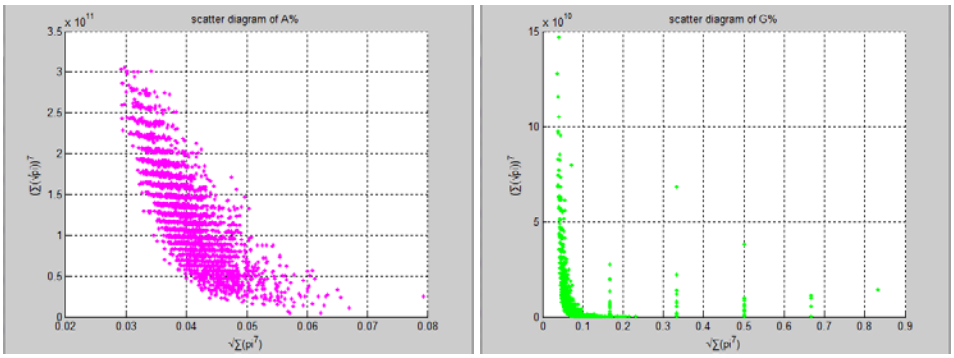


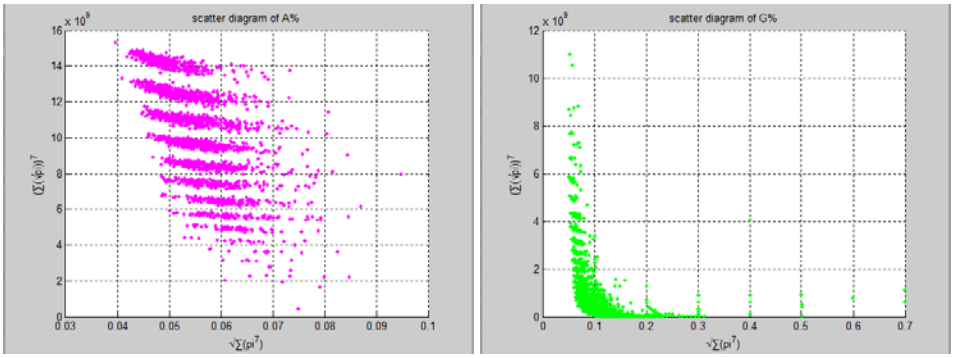
Figure 3: A sample 2D map of VM on multiple sequences.



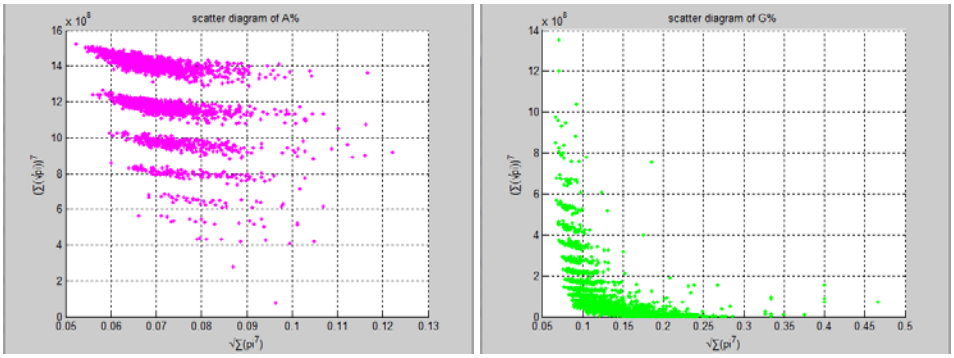
(a1) n=3 (b1)



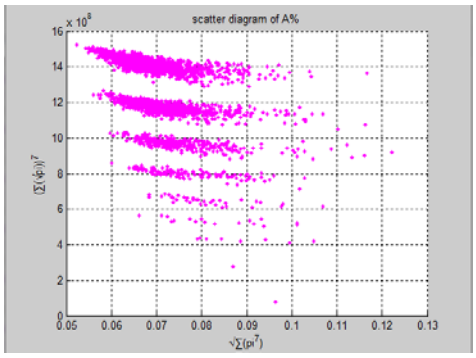
(a2) n=6 (b2)

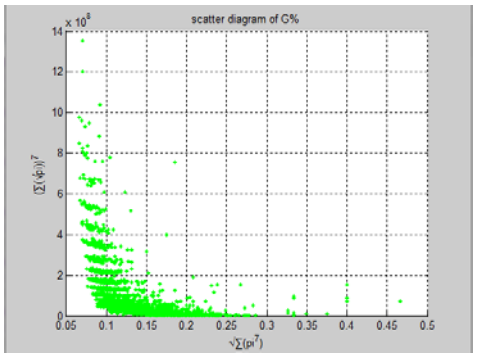


(a3) n=10 (b3)

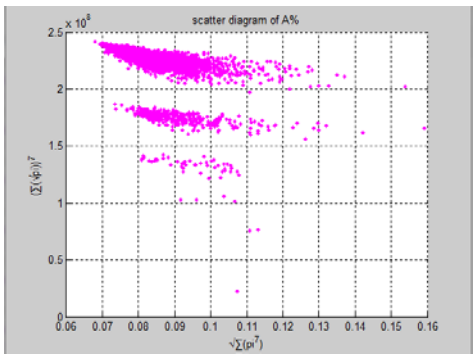


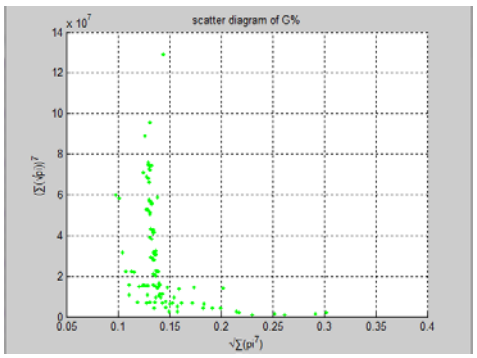
(a4) n=15 (b4)



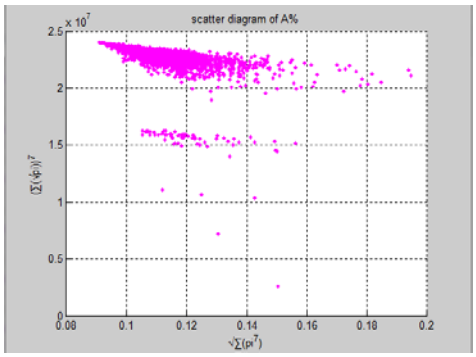


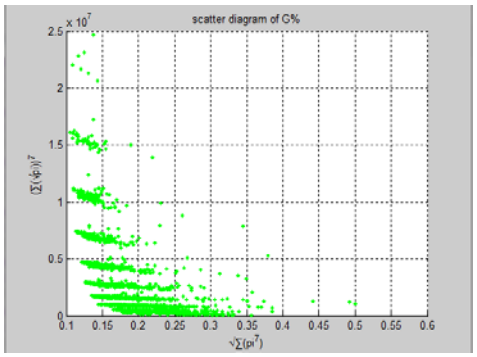
(a4) n=15 (b4)



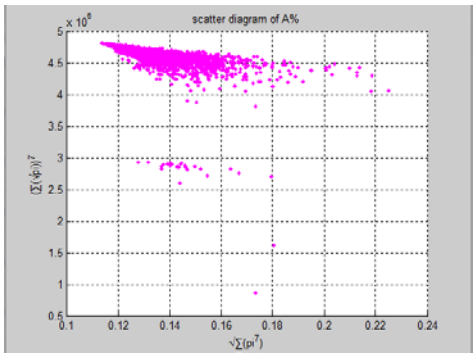


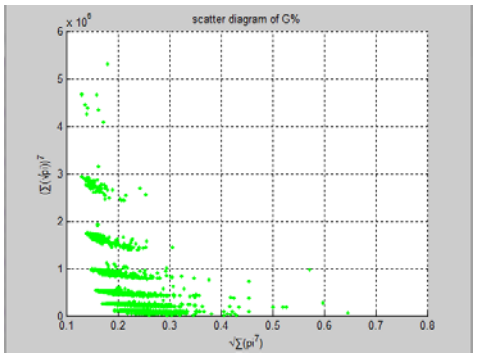
(a5) n=20 (b5)



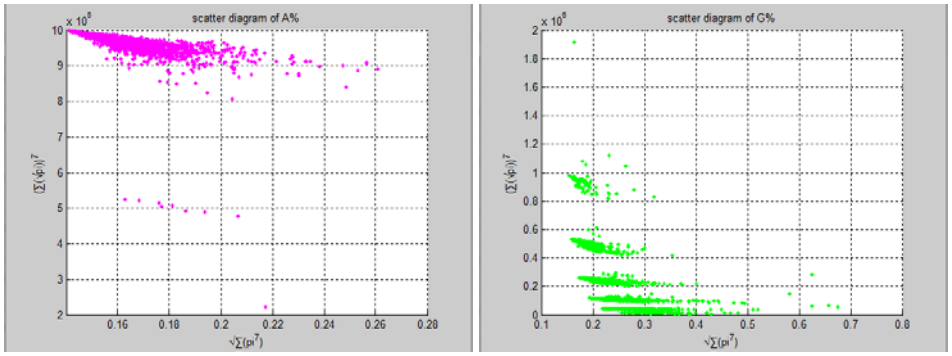


(a6) n=30 (b6)

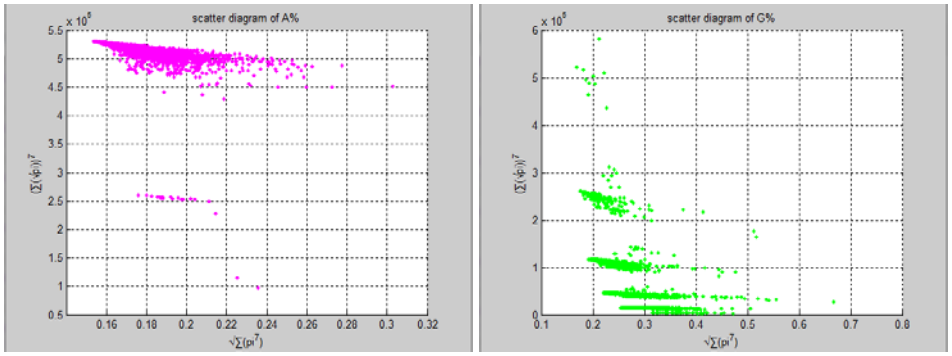




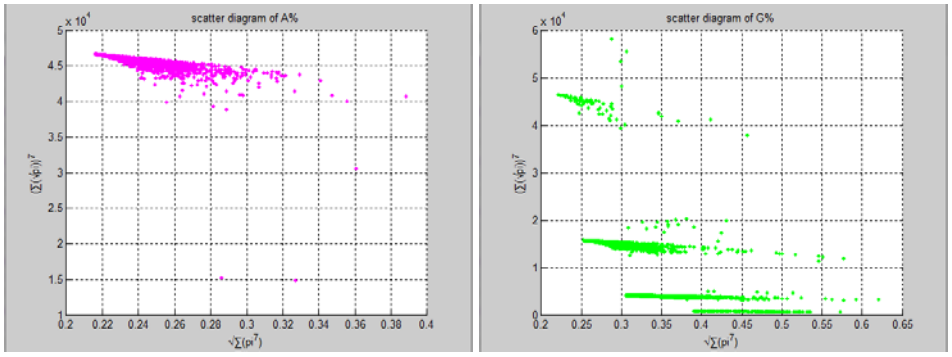
(a7) n=40 (b7)



(a8) n=50 (b8)

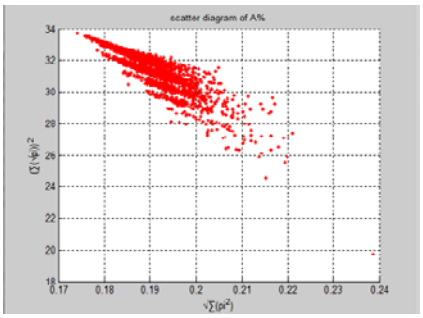


(a9) n=60 (b9)

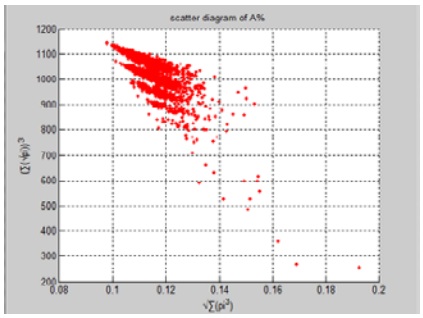


(a10) n=90 (b10)

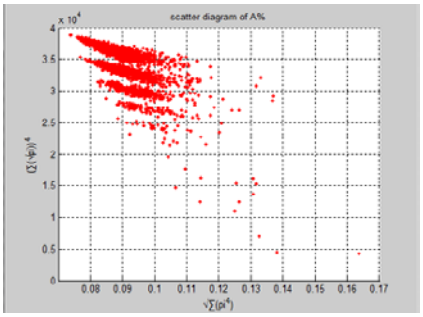
Figure 4: Two groups of ten 2D maps in the range of $n=3 \sim 90$, $k=7$, $N \cong 200 \sim 600$, $T=200 \sim 600$, $T=2700$; (a1-10) Map_V for the file Right; (b1-10) Map_G for the file 201-500.



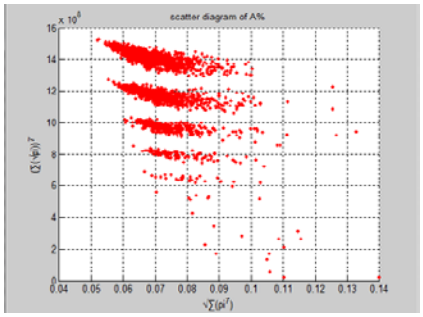
(a1) Map_A k=2



(a2) Map_A k=3

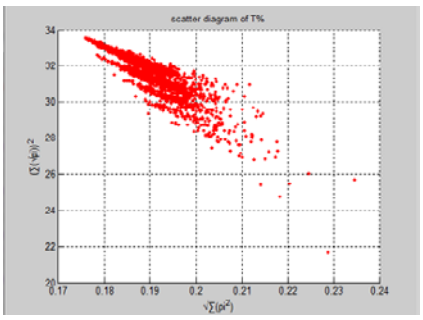


(a3) Map_A k=4

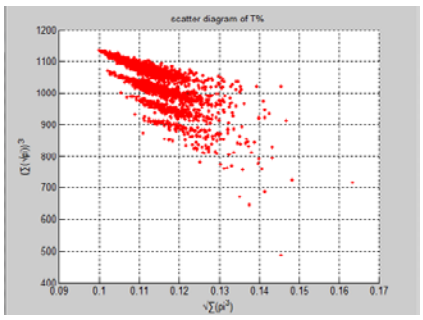


(a4) Map_A k=7

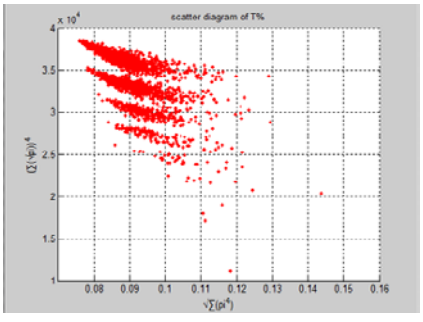
(a)



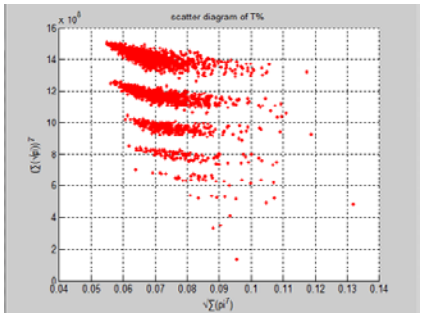
(b1) Map_T k=2



(b2) Map_T k=3

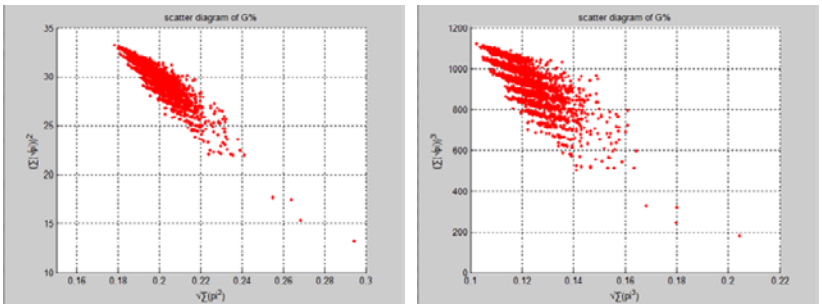


(b3) Map_T k=4

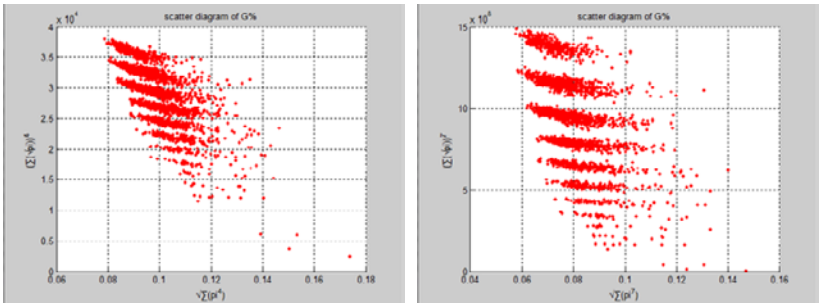


(b4) Map_T k=7

(b)

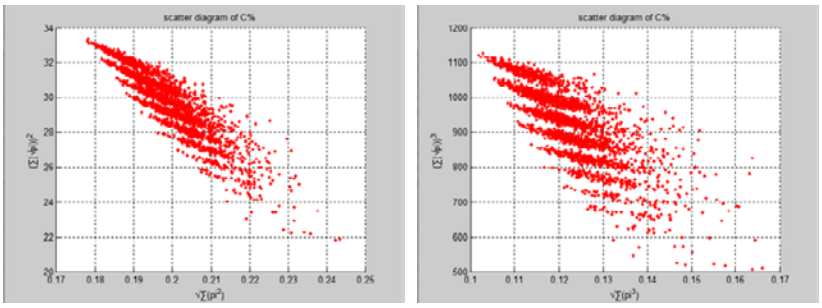


(c1) Map_G k=2 (c2) Map_G k=3

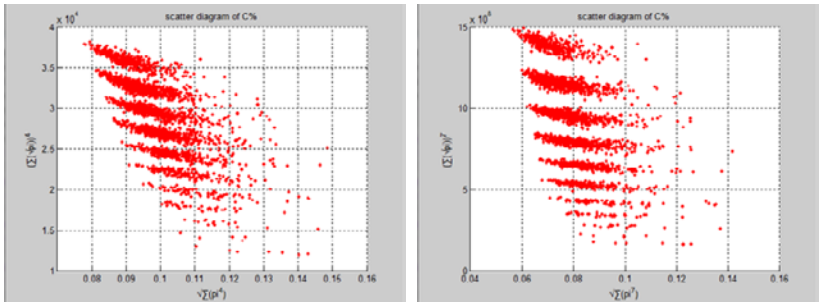


(c3) Map_G k=4 (c4) Map_G k=7

(c)



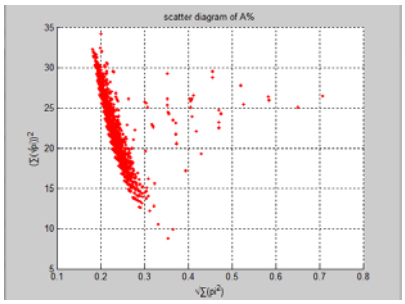
(d1) Map_C k=2 (d2) Map_C k=3



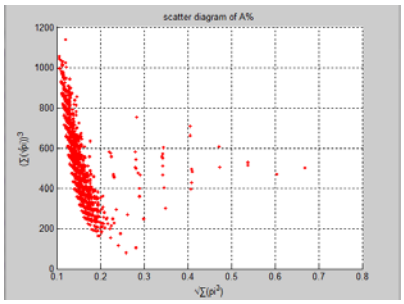
(d3) Map_C k=4 (d4) Map_C k=7

(d)

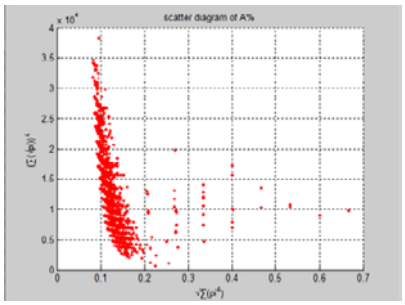
Figure 5: Four groups of sixteen 2D maps in the range of $n=15, k=\{2,3,4,7\}, N \cong 200 \sim 600, T=2700$ (a) group (a1-a4) four Map_A maps; (b) group (b1-b4) four maps; (c) (c1-c4) four Map_G maps; (d) (d1-d4) four Map_C maps for the file *right*.



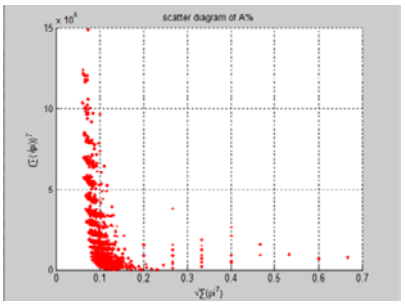
(a1) Map_A k=2



(a2) Map_A k=3

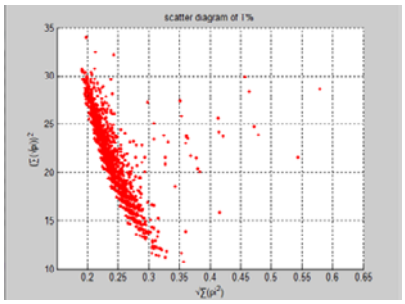


(a3) Map_A k=4

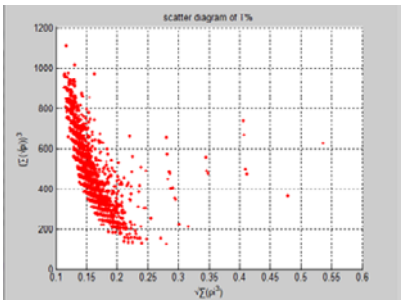


(a4) Map_A k=7

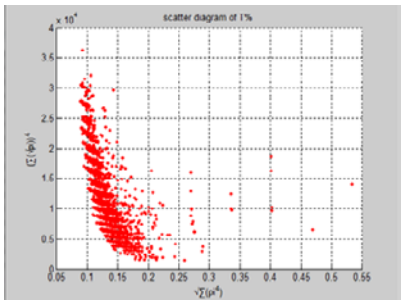
(a)



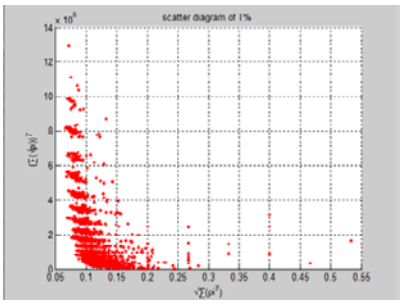
(b1) Map_T k=2



(b2) Map_T k=3

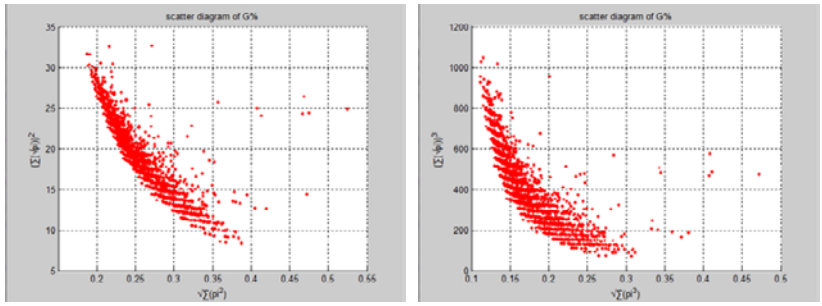


(b3) Map_T k=4

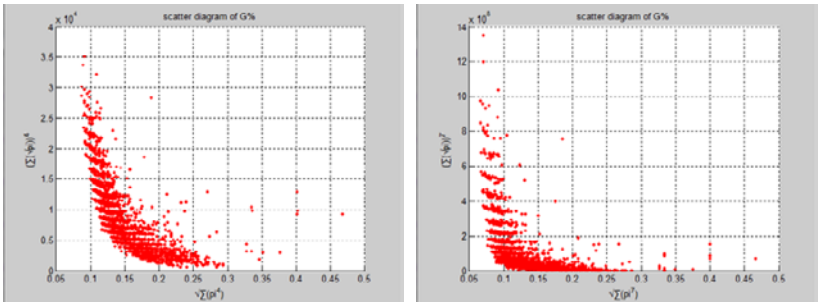


(b4) Map_T k=7

(b)

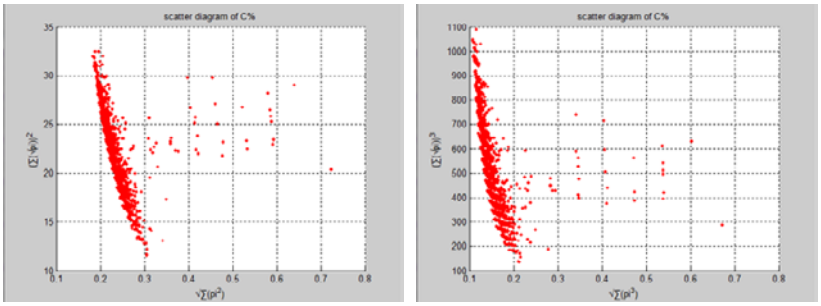


(c1) Map_G k=2 (c2) Map_G k=3

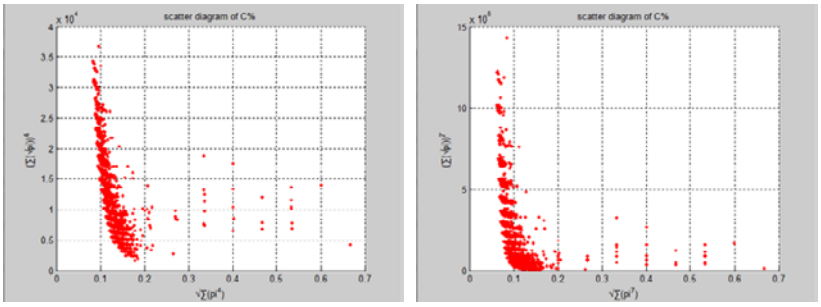


(c3) Map_G k=4 (c4) Map_G k=7

(c)



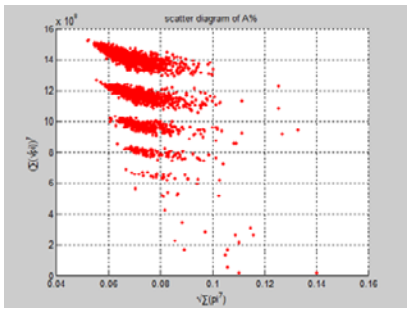
(d1) Map_C k=2 (d2) Map_C k=3



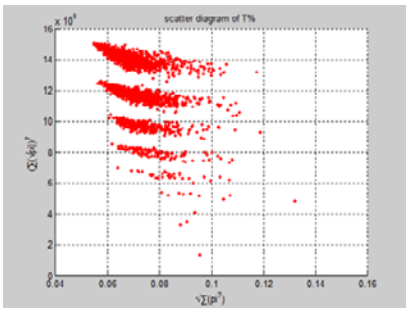
(d3) Map_C k=4 (d4) Map_C k=7

(d)

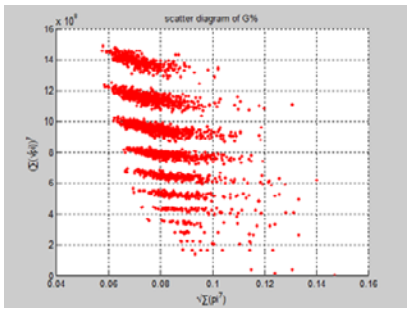
Figure 6: Four groups of sixteen 2D maps in the range of $n=15$, $k=\{2,3,4,7\}$, $N \cong 200 \sim 600$, $T=200 \sim 600$, $T=2700$; (a) group (a1-a4) four Map_A maps; (b) group (b1-b4) four Map_T maps; (c) (c1-c4) four Map_G maps; (d) (d1-d4) four Map_C maps for the file 201-500.



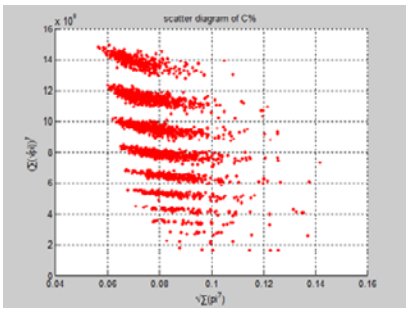
(a1) $\mathrm{Map_A}$



(a2) $\mathrm{Map_T}$

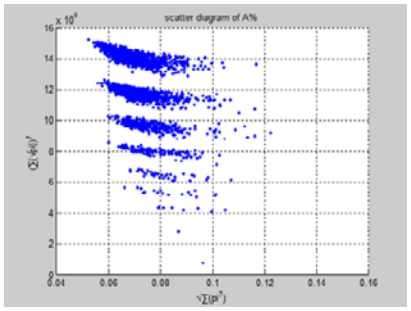


(a3) $\mathrm{Map_G}$

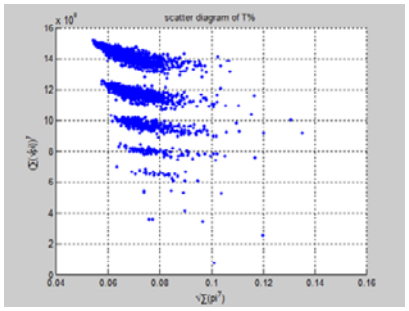


(a4) $\mathrm{Map_C}$

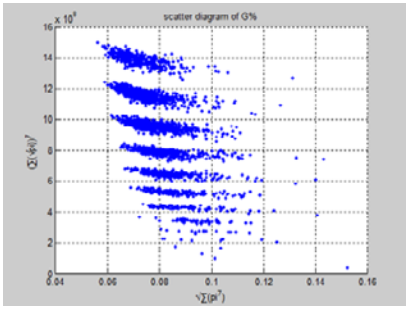
(a) Four maps for the file *left*



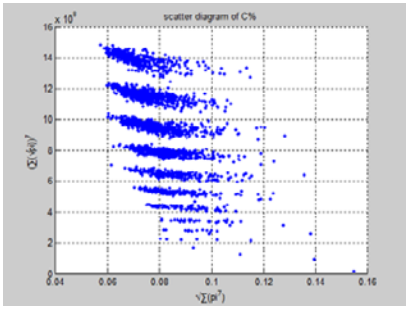
(b1) $\mathrm{Map_A}$



(b2) $\mathrm{Map_T}$

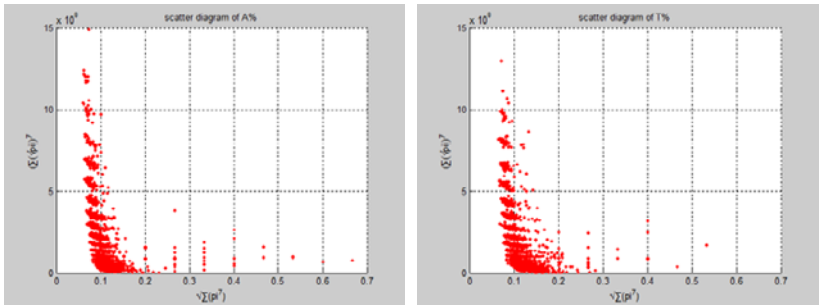


(b3) $\mathrm{Map_G}$

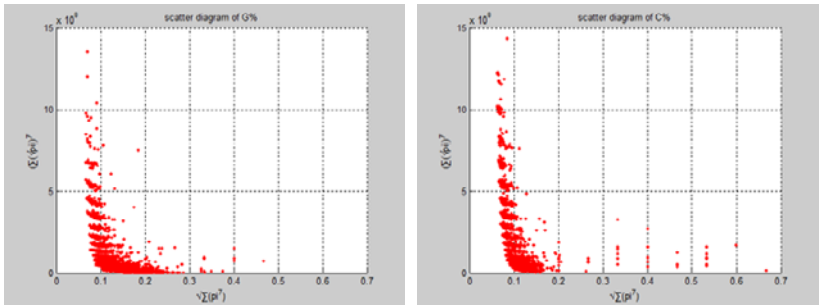


(b4) $\mathrm{Map_C}$

(b) Four maps for the file *right*

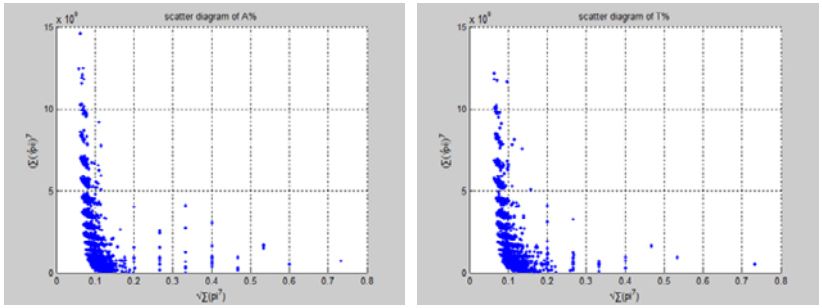


(c1) Map_A (c2) Map_T

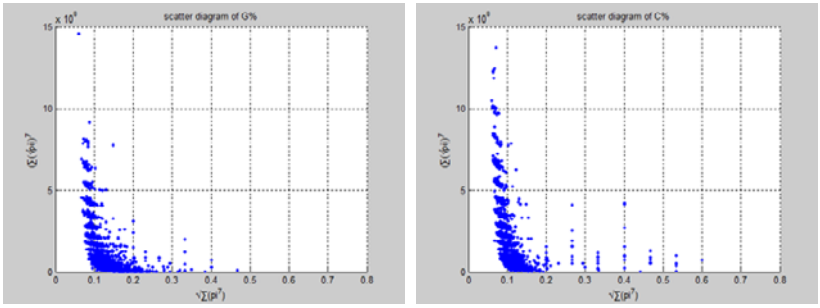


(c3) Map_G (c4) Map_C

(c) Four maps for the file *201-500*



(d1) Map_A (d2) Map_T



(d3) Map_G (d4) Map_C

(d) Four maps for the file random_201-500

Figure 7: Four groups of sixteen 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; (c) group (c1-c4) four Map_v maps for the file 201-500; (d) group (d1-d4) four Map_v maps for the file random_201-500.

10) ten Map_A maps for the file Right; (b1-10) ten Map_G maps for the file 201-500 respectively. In Figure 5, four groups of sixteen 2D maps are listed 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) group (c1-c4) four Ma_{pG} maps; (d) group (d1-d4) four Map_C maps for the file right. In Figure 6, four groups of sixteen 2D maps are selected in the range of $n=15, k=\{2,3,4,7\}, N \cong 200 \sim 600, T=2700$; (a) group (a1-a4) four Map_A maps; (b) group (b1-b4) four Map_T maps; (c) group (c1-c4) four Map_G maps; (d) group (d1-d4) four Map_C maps for the file 201-500. In Figure 7, Four groups of sixteen 2D maps are compared in the range of $n=15, k=\{2,3,4,7\}, N \cong 200 \sim 600, T=2700$; (a) group (a1-a4) four Map_v maps for the file left; (b) group (b1-b4) four Map_v maps for the file right; (c) group (c1-c4) four Map_v maps for the file 201-500; (d) group (d1-d4) four Map_v maps for the file random_201-500.

Symmetric Analysis Based on 2D Maps

Four groups of 2D maps contain different information, it is necessary to make a brief discussion on their important issues as follows. The first group of results shown in Figure 4 presents two sets of ten 2D maps from two data files: right and 201-500 undertaken various lengths of basic segment from 3-90 to illustrate their variations respectively. Ten 2D maps of Figure 4 (a1-a10) for the file right show significant trace on their visual distributions; the numbers of main visible clusters identified are decreased when the length of segment has being increased e.g. (a3-a10). However lesser length of segment does not provide refined visual distinctions with larger region in fuzzy areas e.g. (a1-a2). From a structural viewpoint, middle ranged numbers of length provide better clustering results e.g. (a3-a5) for further analysis targets. To check another ten 2D maps of Figure 4 (b1-b10) for the file 201-500, different visual distributions are observed; the numbers of main visible clusters identified are decreased when the length of segment has being increased less significantly e.g. (b4-b10). However lesser length of segment does not provide refined visual distinctions with wider regions in fuzzy areas e.g. (b1-b3). In general, middle ranged numbers of length still provide better clustering effects e.g. (b4-b7) for further analysis purpose. It is interesting to observe different maps when control parameter k changed. Four groups of sixteen 2D maps for the file right are shown in Figure 5 on the range of $n=15, k=\{2,3,4,7\}, N \cong 200 \sim 600, T=2700$; four groups in (a)–(d) provide four maps to share the same other parameters with different k values. Checking visible clusters in different maps, it is important to notice nearly same numbers of clusters identified in the same group, but different groups may contain significantly different numbers. Lesser k value (e.g. k=2) makes a tighter distribution and larger k value (e.g. k=7) takes better separation on the maps. Through k=7 maps provide better separation effects, it is easy to observe their y axis values already in 10^8 range. Four groups of sixteen 2D maps for the file 201-500 are shown in Figure 6 in the range of $n=15, k=\{2,3,4,7\}, N \cong 200 \sim 600, T=2700$. This group of 2D maps can be compared with 2D maps in Figure 5. Under the same parameters, similar visible effects and feature clustering properties could be observed if various k values are selected. Using a set of selected parameters, four groups of sixteen 2D maps are compared in Figure 7 for four files: left, right, 201-500 and random_201-500 to explore higher levels of symmetric properties for secondary or higher levels of structures potentially contained in DNA sequences. Selected parameters are in the range of $n=15, k=\{2,3,4,7\}, N \cong 200 \sim 600, T=2700$. Group (a) provides four Map_v maps (a1-a4) for the file left; group (b) uses four Map_v maps (b1-b4) for the file right; group (c) gives four Map_v maps (c1-c4) for the file 201-500; group (d) lists four Map_v maps (d1-d4) for the file random_201-500. In convenient description, let ~ be a similar operator, for groups (a) & (b), four pairs of {(a1) ~ (b1), (a2) ~ (b2), (a3) ~ (b3), (a4) ~ (b4)} maps i.e. (left-A ~ right-A, left-T ~ right-T, left-G ~ right-G, left-C ~ right-C) have a stronger similar distribution between left & right. In addition, only two clustering classes could be significantly identified as {(a1) ~ (a2) ~ (b1) ~ (b2), (a3) ~ (a4) ~ (b3) ~ (b4)} i.e. (left-A ~ right-A ~ left-T ~ right-T, left-G ~ right-G ~ left-C ~ right-C) respectively. This type of similar clustering distributions indicates eight maps with intrinsically higher levels of DNA sequences with extra A-T & G-C pairs of symmetric relationships between two files: left & right. However, for groups (c) & (d), let be a weak similar operator, four pairs of {(c1) ≃ (d1), (c2) ≃ (d2), (c3) ≃ (d3), (c4) ≃ (d4)} maps i.e. (201-500-A ≃ random_201-500-A, 201-500-T ≃ random_201-500-T, 201-500-G ≃ random_201-500-G, 201-500-C ≃ random_201-500-C) have a stronger similar distribution between 201-500 & random_201-500. In addition, two clustering classes may be identified as {(c1) ≃ (c4) ≃ (d1) ≃ (d4), (d2) ≃ (d3) ≃ (d2) ≃ (d3)} i.e. (201-500-A ≃ random_201-500-A ≃ 201-500-T ≃ random_201-500-T, 201-500-G ≃ random_201-500-G ≃ 201-500-C ≃ random_201-500-C) respectively. Such clustering distributions may indicate eight maps with A-C & T-G pairs of similar relationships between two files: 201-500 & random_201-500. Such types of pairs could be less important than A-T & G-C relationships, further investigations may be required. From a comparison viewpoint, (a)-(b) groups have significantly much stronger similar relationship than (c)-(d) groups.

Conclusion

This paper proposes architecture to support the Variant Map Construction. Using a DNA sequence, variant measures, probability measurement and normalized histogram, a pair of values can be determined by a series of controlled parameters. Collecting relevant pairs on multiple DNA sequences, four 2D maps can be generated. The main results of this paper provide a brief architecture description in diagrams, core components, modules, expressions and important equations for the construction.

In addition to core models and diagrams, sample results are illustrated to apply two sets of selected DNA sequences for testing. After proper set of parameters selected, suitable visual distributions could be observed using the VMC system. Results in Figures 4-7 provide useful evidences systematically to support proposed VMC system useful in checking higher levels of symmetric properties among complex DNA sequences. This construction could provide useful spatial information on complex DNA sequences via 2D maps to explore higher dimensional environments in near future.

Acknowledgement

Thanks to the school of software Yunnan University, to the key laboratory of Yunnan software engineering and the key laboratory for Conservation and Utilization of Bio-resource for excellent working environment, to the Yunnan Advanced Overseas Scholar Project (W8110305), the Key R&D project of Yunnan Higher Education Bureau (K1059178) and National Science Foundation of China (61362014) for financial supports to this project.

References

1. Sakamoto K (2000) Molecular Computation by DNA Hairpin Formation. Science 283:1223-1227.
2. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-293.
3. Sakamoto K, Gouzu H, Komiya K, Kiga D, Yokoyama S, et al. (2011) Multi-scale coding of genomic information: From DNA sequence to genome structure and function. Physics Reports 498: 45-188.
4. Engela S, Alemany A, Forns N (2011) Folding and unfolding of a triple-branch

DNA molecule with four conformational states. Philosophical Magazine 91: 2049-2065.

5. Urquiza JM, Rojas I, Pomares H, Herrera LJ , Ortega J, et al. (2011) Method for prediction of protein–protein interactions in yeast using genomics/proteomics information and feature selection, Neurocomputing 74: 2683-2690.
6. Zhang H, Liu X (2011) A CLIQUE algorithm using DNA computing techniques based on closed-circle DNA sequences. Biosystems 105: 73-82.
7. Bánfai B, Jia H, Khatun J, Wood E, Risk B, et al. (2012) Long noncoding RNAs are rarely translated in two human cell lines. Genome Res 22: 1646-1657.
8. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489: 91-100.
9. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.
10. Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA. Science 337: 1159, 1161.
11. Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A 110: 5294-5300.
12. Engreitz JM, Pandya-Jones A, McDonel P (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.
13. Jeffrey Zhi J. Zheng, Christian H. Zheng (2010) A framework to express variant and invariant functional spaces for binary logic. Frontiers of Electrical and Electronic Engineering in China 5: 163-172.
14. Jeffrey Zheng, Christian Zheng and Tosiyasu Kunii (2011) A Framework of Variant Logic Construction for Cellular Automata, in Cellular Automata – Innovative Modelling for Science and Engineering, Edited by A. Salcido, In Tech Press, 325-352.
15. Qingping Li, Jeffrey Zhi Zheng (2010) 2D Spatial Distributions for Measures of Random Sequences Using Conjugate Maps, in the Proceedings of the 11th Australian Information Warfare and Security Conference, Perth 1-9.
16. Zhang WQ, Zheng J (2012) Randomness Measurement of Pseudorandom Sequence Using different Generation Mechanisms and DNA Sequence. Journal of Chengdu University of Information Technology 27: 548-555.
17. Jeffrey Zheng, Christian Zheng, Tosiyasu Kunii (2013) Interactive Maps on Variant Phase Spaces-From Measurements-Micro Ensembles to Ensemble Matrices on Statistical Mechanics of Particle Models, in Emerging Application of Cellular Automata, Edited A. Salcido, InTech Press, 113-196.