Chapter 5

A Novel Approach to Mine for Genetic Markers via Comparing Class Frequency Distributions of Maximal Repeats Extracted from Tagged Whole Genomic Sequences

Jing-Doo Wang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75113

Abstract

The cost to extract one new biomarker within genomic sequences is very huge. This chapter adopts a scalable approach, developed previously and based on MapReduce programming model, to extract maximal repeats from a huge amount of tagged whole genomic sequences and meanwhile computing the similarities of sequences within the same class and the differences among the other classes, where the types of classes are derived from those tags. The work can be extended to any kind of genomic sequential data if one can have the organisms into several disjoint classes according to one specific phenotype, and then collect the whole genomes of those organisms. Those patterns, for example, biomarkers, if exist in only one class, with distinctive class frequency distribution can provide hints to biologists to dig out the relationship between that phenotype and those genomic patterns. It is expected that this approach may provide a novel direction in the research of biomarker extraction via whole genomic sequence comparison in the era of post genomics.

Keywords: biomarker, comparative genomics, class frequency distribution, maximal repeat, MapReduce programming

1. Introduction

It is very attractive and challenging to discover markers [1] from genomic sequences and then to use these markers for genetic tests [2] to diagnose diseases and for personalized medicine to adverse drug responses [3, 4]. Nowadays, genome-wide association studies (GWASs) [5] have

already examined single-nucleotide polymorphisms (SNPs) across human genomes to identify specific SNPs related to some diseases, for example, diabetes, heart abnormalities, Parkinson disease, and Crohn disease [6]. Furthermore, GWAS is also used to predict cancer [7] and to influence human intelligence [8].

Most of GWASs are achieved with SNP arrays [9]. The "Illumina" [10] uses the "whole-genome genotyping" to interrogate SNPs across the entire genome to obtain the most comprehensive view of genomic variation; the Affymetrix Genome-Wide Human SNP Array 6.0 features 1.8 million genetic markers which includes more than 906,600 single-nucleotide polymorphisms (SNPs) [11]. The majority of these SNPs are designed to investigate the coding regions of genes in genomic sequences. However, some of the non-coding regions, once being mistaken as "junk DNA," are believed to contain functions to regulate gene transcription and to account for the genetic differences between individuals [12]. Although, on the one hand, the number of SNPs on one chip may be several hundreds of thousands, on the other hand, its coverage is still not enough [13] to figure out the relationship between genotypes and phenotypes in humans as given in the database of "dbGaP" [14].

As the era of post genomics with Next-Generation Sequencing (NGS) is coming, it is expected that the cost of genomic sequencing is decreasing and the availability of complete whole genomes of individual creatures is becoming popular. After using NGS for DNA sequencing [15], as shown on the right side in **Figure 1**, for example, one creature, for example, a virus, is supposed to contain three chromosomes with eight genotypes. On the other side of **Figure 1**, there are three phenotypes, for example, "Drug Resistance" "Envelope," and "Contents," inspected and detected by three domain experts, respectively. Under the assumption that these three phenotypes are totally dominated by those eight genotypes, represented as different icons, without considering the epigenetics [16], as shown in **Figure 2**, it is difficult for biologists in wet laboratory to analyze aimlessly the relationships among these phenotypes and those

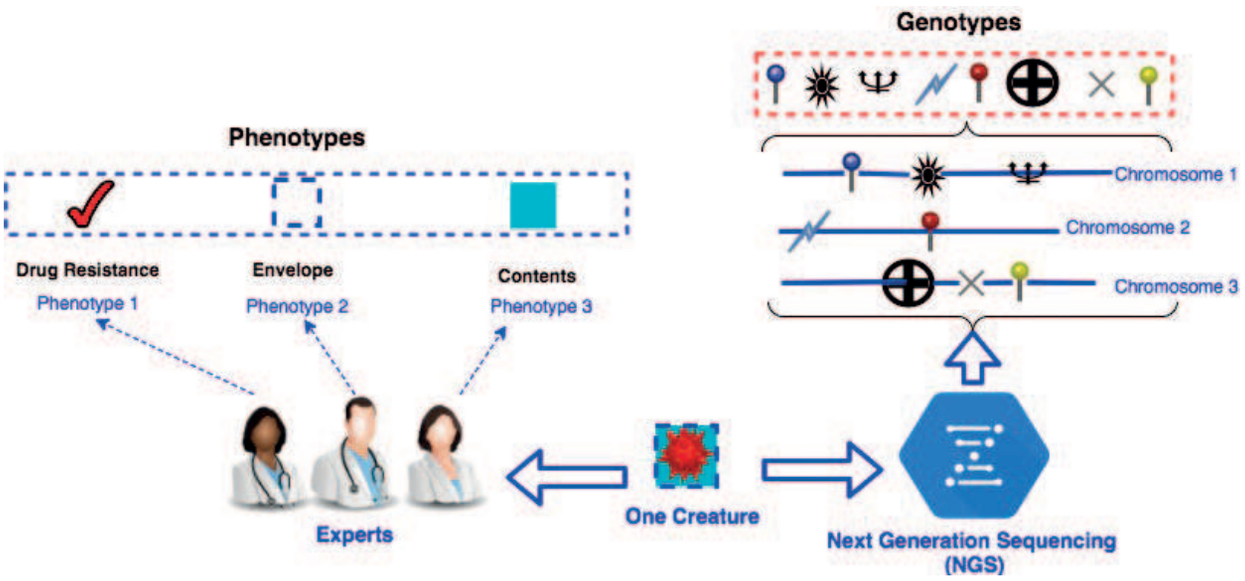

Figure 1. An example of one creature with three phenotypes and eight genotypes.

http://dx.doi.org/10.5772/intechopen.75113

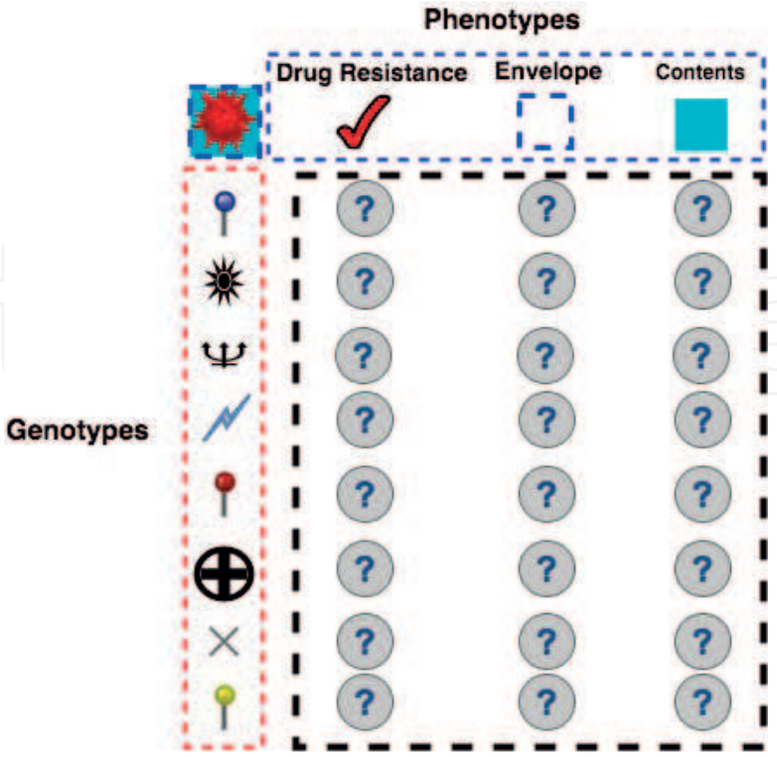

Figure 2. Example: how to identify the relationships among genotypes and phenotypes as described in **Figure 1**.

genotypes without further bioinformatics information or techniques such as comparative genomics [17].

With more and more complete whole genomes of distinctive creatures being available and popular in the coming days, it is very interesting and desired to extract common significant subsequences as candidate genomic markers as genotypes via comparing these creatures' whole DNA sequences according to the classes (or types) of their phenotypes observed and specified by domain experts. **Figure 3** shows the conceptual diagram of the corresponding classes for each of these three phenotypes given in **Figure 1**. With precise observations or experiments (phenotypes), biologists or experts can divide these creatures with complete whole genomes into disjoint classes if possible. Then, it is highly expected for biologists that some distinctive patterns (genotypes) hidden within their DNA sequences can be extracted as the candidates of class markers (phenotypes) if the frequency distributions of these patterns among classes are extremely biased, or some patterns are just in one class solely and appear in all instances belonging to that class ideally. To achieve the earlier-mentioned goal, one needs to extract repeats and to compute class frequency distributions of these repeats from a huge amount of tagged genomic sequences, where the types of classes are derived from the tags.

Due to the availability of genomic sequences in National Center for Biotechnology Information (NCBI) [18], The Cancer Genome Altas (TCGA) [19], it is interesting to have class frequency distribution of maximal repeats from these tagged genomic sequences for mining the biomarker or specific patterns. As the age of Next-Generation Sequencing (NGS) is going to be introduced for the project "Cancer Moonshot" in the National Cancer Institute [20], it is very

attractive to identify specific biomarkers from these genomic sequences with tags, such as cancer types or distinctive genotypes. **Figure 4** gives the conceptual diagram of how to reduce the gap between phenotypes and genotypes by using the phenotypes as classes to identify those subsequences that appear in unique class only as biomarkers.

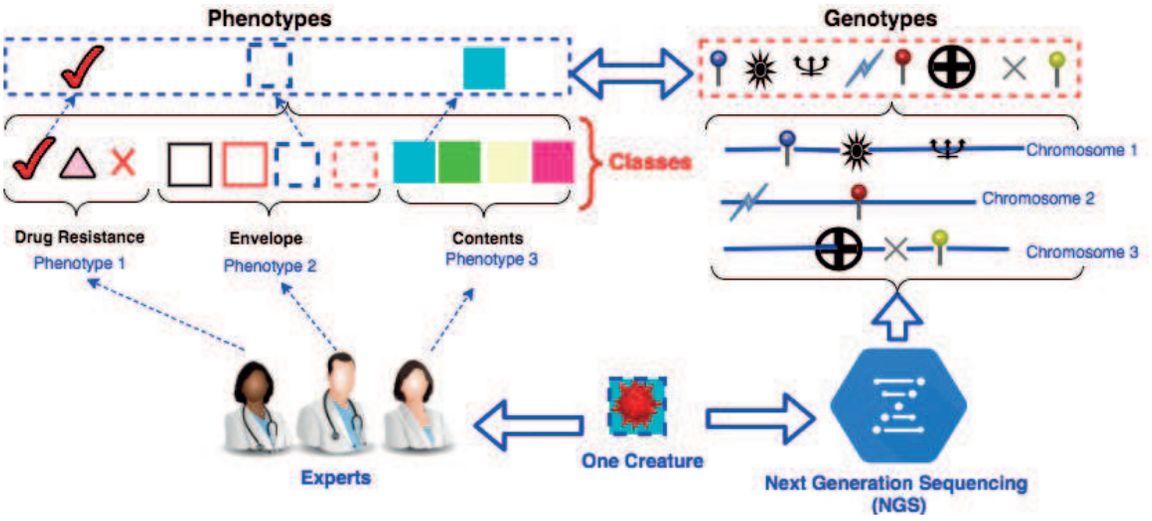

Figure 3. Mining the relationship of phenotypes and genotypes via classes comparison.

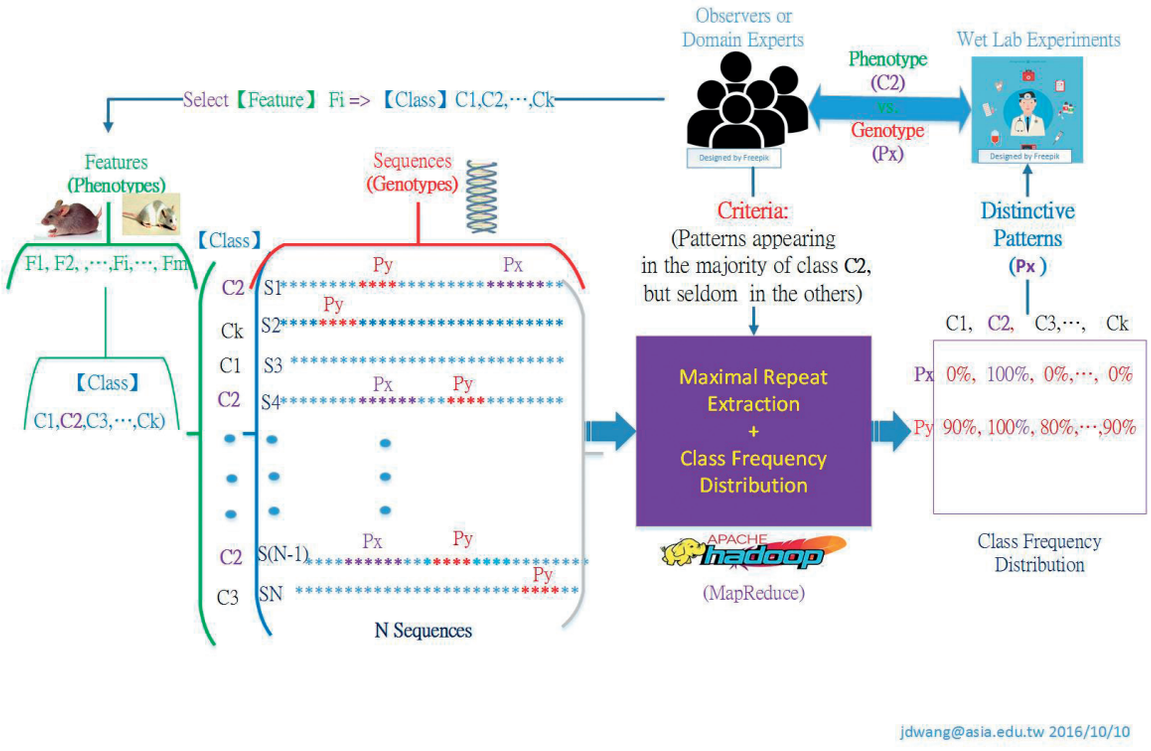

Figure 4. The conceptual diagram of reducing the gap between phenotypes and genotypes.

http://dx.doi.org/10.5772/intechopen.75113

The remainder of this chapter is organized as follows. Section 2 gives the review of potential applications with class frequency distributions of maximal repeats. Section 3 shows the scalable approach to extract maximal repeat from tagged sequential data. Section 4 describes the most recent work [21] that compute co-occurrences of DNA maximal repeat patterns appearing in both humans and viruses. Section 5 concludes and discusses on future works.

2. Potential applications with class frequency distribution of maximal repeats extracted from tagged sequential data

The previous work in [22] was a scalable approach based on Hadoop MapReduce programming model to overcome the computational bottleneck of using single computer with external memory [23, 24]. Furthermore, it had been applied for a USA patent (US-2017-0255634-A1) [25] whose publication data is as "Sep. 7, 2017" [25]. Recently, in these 2 years, many novel and potential applications, derived from that work, were launched in diverse fields successfully, due to its scalability being able to handle a huge amount of sequential data. There were many experiments in diverse applications with a huge amount of tagged sequential data, such as textual data for trend analysis [26–28], genomic sequences for biomarker identification [21, 29, 30], time-stamped gantry sequences for significant travel time intervals [31] and, most recently, the sequences of product traceability for quality control [32].

3. Methods

The scalable approach of maximal repeat extraction adopted in this chapter is based on Hadoop MapReduce programming model, and the details can be found in [22]. To illustrate the concept of the earlier approach clearly, as shown in **Figure 5**, there are 20 creatures generated manually. Each of them is with three phenotypes, "Drug Resistance," "Envelope," and "Contents," as given in **Figure 3**, and all of its chromosomes are concatenated into one line which may contain genotypes including motifs, domains, or unknown DNA segments that are represented as icons for simplicity. Even though with the conceptual diagram as shown in **Figure 5**, it is still very difficult for users to catch the hidden connection (or relationship) among these three phenotypes and those icons (genotypes) at first glance, let alone each of these icons (genotypes) presents one continuous subsequence whose length is not fixed and its location is unknown within chromosomes.

To reveal the possible mapping of phenotype "Drug Resistance," for example, to genotypes on purpose, **Figure 6** presents the rearrangement in the order of these 20 chromosomes which may contain icons as hidden or unknown DNA segments. The mapping of different types of phenotype "Drug Resistance" to the corresponding genotypes (icons) can be observed. Similarly, one can have the mapping of different types of phenotype "Envelope"

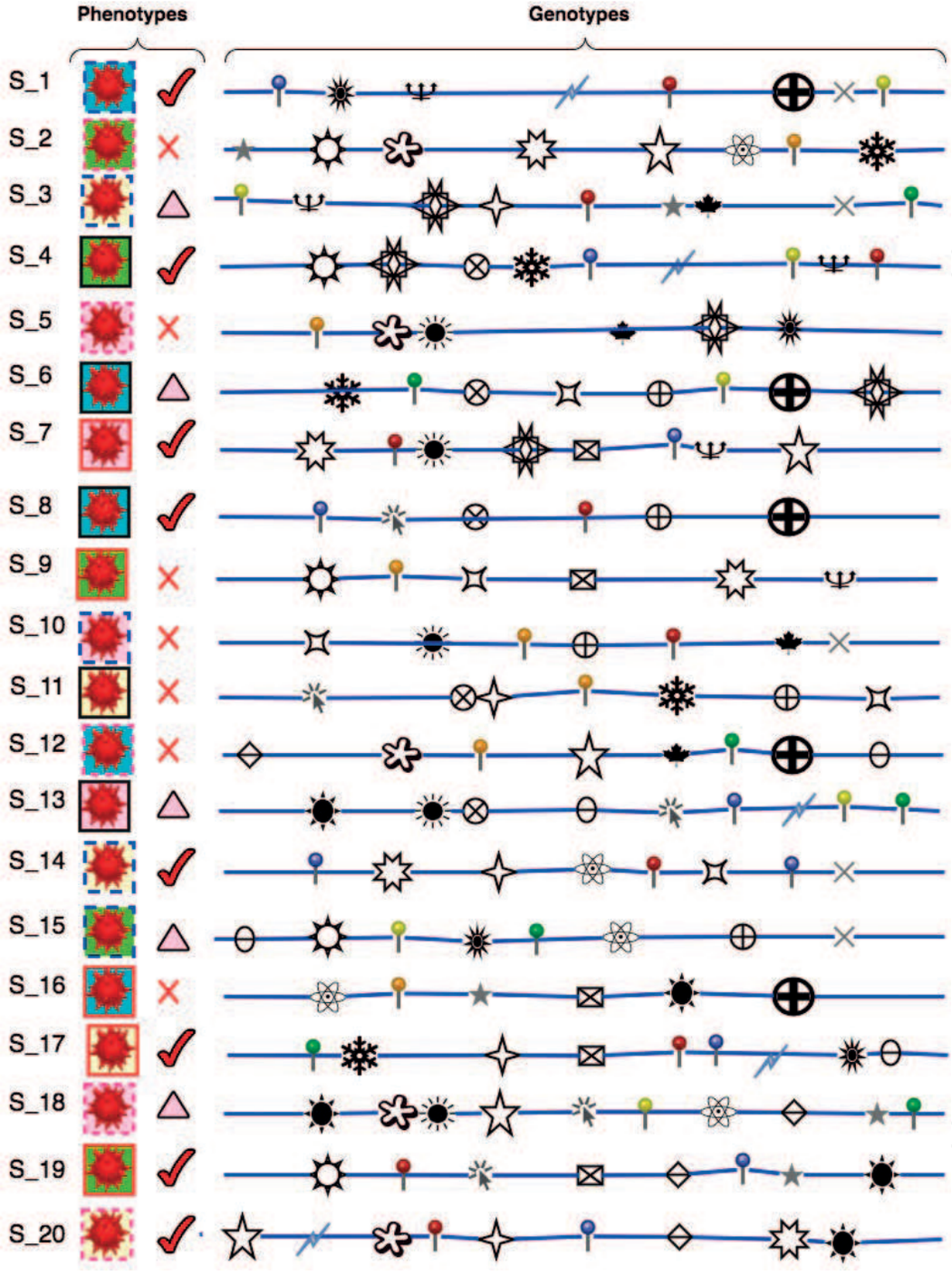

Figure 5. Each of 20 creatures is with three kinds of phenotypes as given in **Figure 3** and all of its chromosomes are concatenated as one line containing several icons as motif, domain, or unknown patterns.

and "Contents" to the corresponding genotypes (icons). Due to the page limitation, the corresponding mapping of figures for "Envelope" and "Contents" are given in the supplements. Focusing on the repeats whose class frequency distributions are biased, as shown in

http://dx.doi.org/10.5772/intechopen.75113

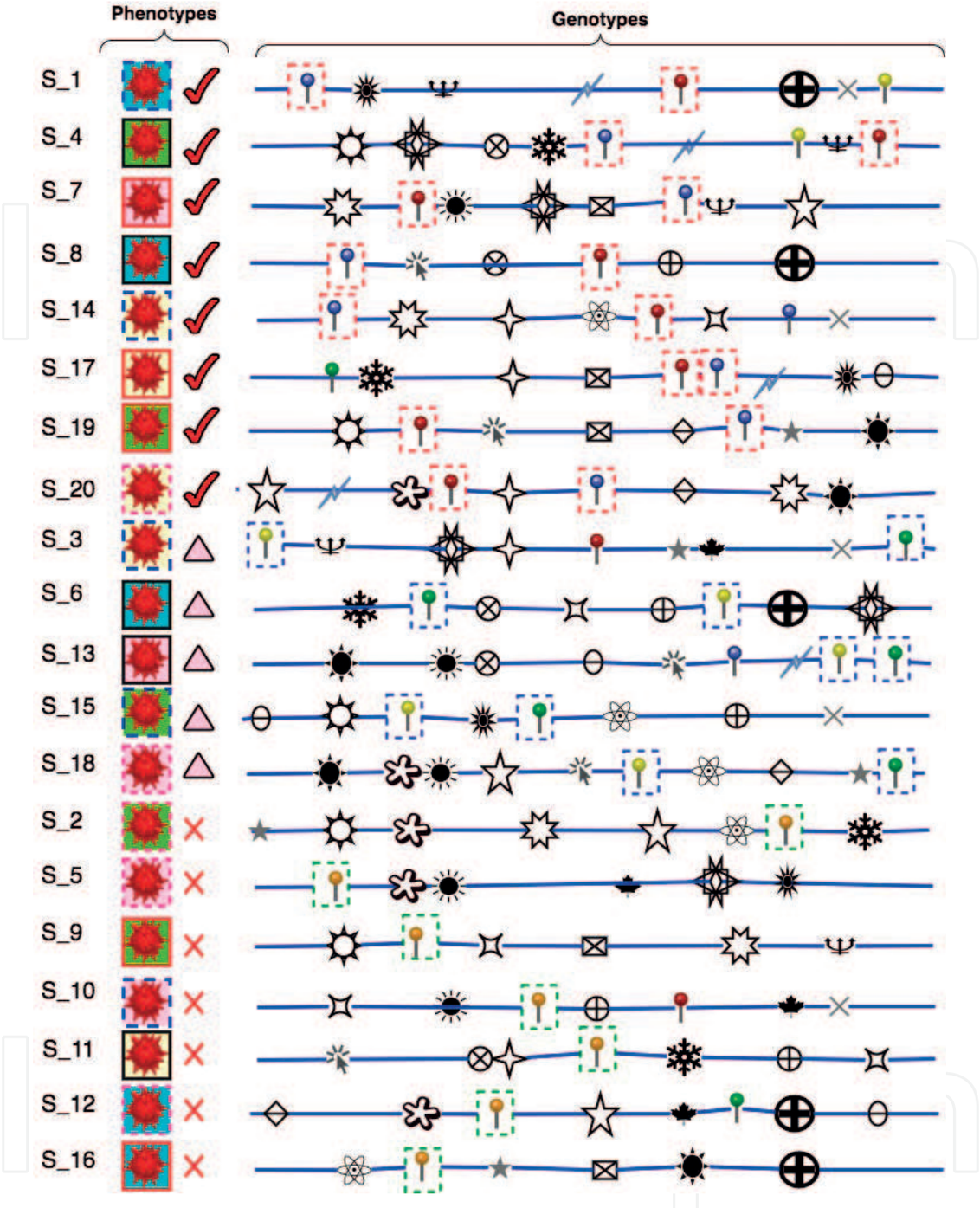

Figure 6. The mapping of different types of phenotype "Drug Resistance" to the corresponding genotypes (icons).

Figure 7, one can estimate these repeats as candidate class markers which can be the clues for further experiments of analyzing the mapping of phenotypes and genotypes derived from 20 creatures in **Figure 5**.

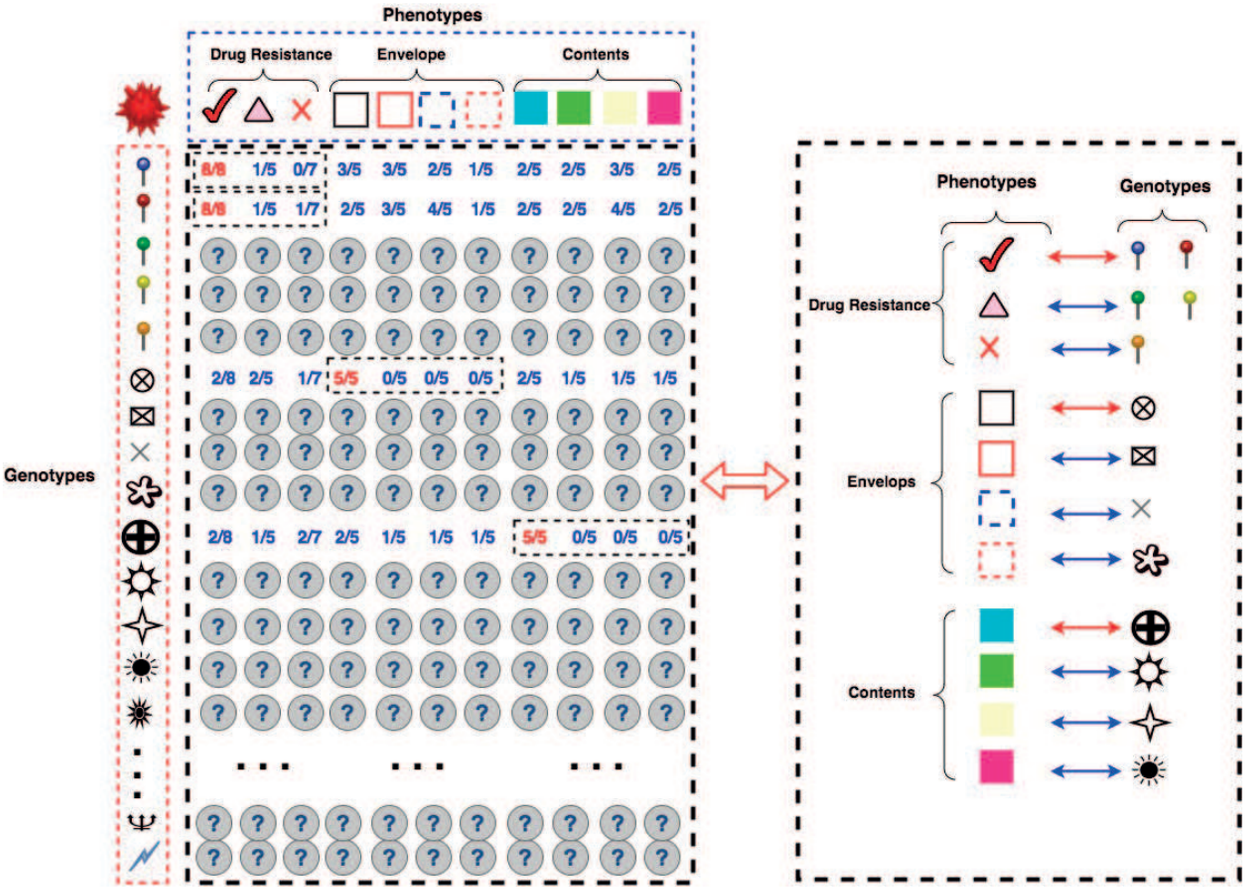

Figure 7. The mapping of phenotypes and genotypes derived from 20 creatures in **Figure 5**.

4. Case study: mining for the co-occurrences of DNA maximal repeat patterns in both human and viruses

There were three studies with a huge amount of genomic sequences [21, 29, 30] based on the scalable approach of maximal repeat extraction with class frequency distribution mentioned in this chapter. This chapter only describes the most recent work [21] that the co-occurrences of DNA maximal repeat patterns appearing in both humans and viruses are extracted via a scalable approach that is based on Hadoop distributed computing [22]; that work aimed to mine for specific DNA patterns within human genomes via observing class frequency distribution of DNA maximal repeats extracted from the whole genomic DNA sequences of humans and 559 virus genuses. The detail in [21] is described for reference in the following.

4.1. Genome resources

In [21], Wang et al. extracted significant DNA sequences appearing in both the genomes of humans and viruses. In this study, the taxonomic level of viruses is "genus" and is selected as

http://dx.doi.org/10.5772/intechopen.75113

the classes (tags) for further experiments. Experimental resources included the complete whole genomes of humans (GRCh38.p7 Primary Assembly) downloaded from the NCBI FTP [33] and that of 559 virus genuses, including 2712 viruses that had genus name and were selected from the total of 4388 viruses download from in NCBI FTP [34] on January 14, 2017. **Table 1** shows the partial statistics of 560 classes, including 559 virus genuses and the humans as "C248." Note that each of the 24 human chromosomes is estimated as one individual instance for observing the frequency distribution among human chromosomes. This chapter, for

Class ID	Human and virus genuses	No of Instances
C1	Alfamovirus	1
C2	Allexivirus	6
C3	Allolevivirus	3
C4	Alpha3microvirus	2
C5	Alphabaculovirus	40
C6	Alphacarmotetravirus	1
C7	Alphabaculovirus	7
…	…	…
C247	Human mastadenovirus E	1
C248	HumanGenomes_23_Assembled	24
C249	Hunnivirus	1
C250	Hypovirus	4
C478	Sobemovirus	15
C479	Solendovirus	1
C480	Soymovirus	4
…	…	…
C553	Xipapillomavirus	1
C554	Xp10virus	5
C555	Yatapoxvirus	2
C556	Yatapoxvirus	3
C557	Zeavirus	1
C558	Zetapapillomavirus	1
C559	Zetatorqueviurs	1
C560	primate papillomaviruses	1

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 1. The partial statistics of 559 virus genuses and human genomes (C248).

simplicity, only takes the positive-strand DNA sequences of humans and viruses for further experiments.

4.2. Computational time and environment

To show the scalability of this approach from a practical view of point, as shown in **Figure 8**, the computational platform was the Hadoop cluster with eight computing nodes, two name (master) nodes, and six data (slave) nodes; **Table 2** showed the specifications of hardware and software of one computing node; the computational time was about 37.5 h when the maximum length of maximal repeat patterns was limited to 500 bp (base pair).

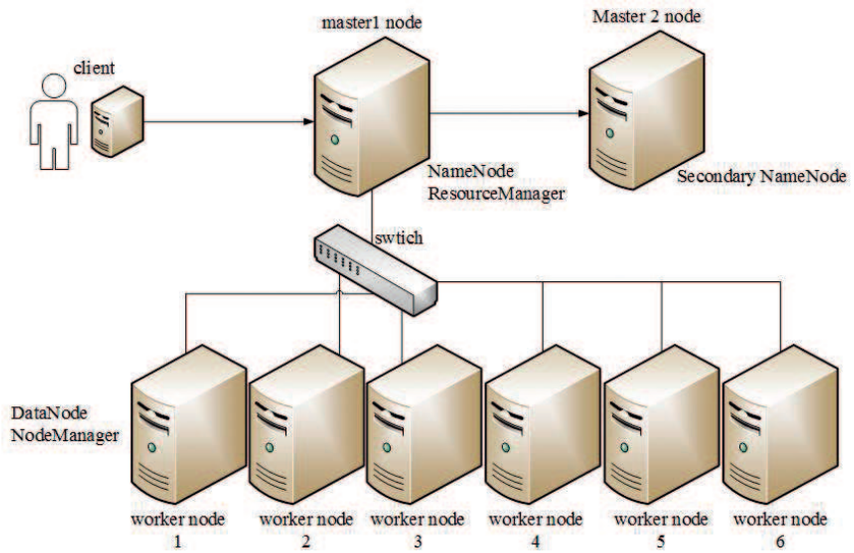

Figure 8. The conceptual diagram of a Hadoop cluster with two name (master) nodes, and six data (worker) nodes; "Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017" [41].

Hardware	CPU	Intel® Xeon® Processor E5-2630 v3 (8 cores)
	RAM	128 GB (16GB*8, ECC/REG DDR4 2133)
	Hard Disk	6 TB (SATA3 2 TB*3, 7200 rpm 3.5 inch)
	Network Card	Intel Ethernet X540 10GBASE-T RJ45 DualPort *4
Software	OS	CentOS 6.7
	Hadoop	Hadoop 2.6 ("Cloudera Express 5.4.5")

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 2. The hardware and software of one computing node in Hadoop cluster.

http://dx.doi.org/10.5772/intechopen.75113

Length	Virus (only)	Human (only)	Human and virus
5	426	127	1341
6	245	102	4234
7	84	48	16,454
8	29	26	65,556
9	5	11	262,154
10	1	9	1,048,579
11	956	4093	4,189,216
12	95,386	1,198,404	15,310,125
13	547,437	23,069,913	34,360,563
14	788,030	110,159,534	42,567,207
15	547,766	273,869,697	36,497,761
16	305,641	322,333,237	22,317,495
17	206,969	209,993,387	10,170,128
18	86,585	103,569,439	3,920,359
19	47,417	48,474,700	1,407,005
20	66,719	25,284,157	493,326
21	25,068	15,882,880	175,934
22	18,507	11,902,168	67,700
23	39,947	9,921,624	29,793
24	14,802	8,649,670	14,795
25	12,227	7,794,361	8749
…	…	…	…
98	165	107,159	15
99	710	102,830	15
100	707	99,579	13
101	1607	96,326	13
102	608	93,630	12
103	638	92,129	11
…	…	…	…
460	19	1933	1
461	27	2000	1
462	22	1812	1
463	26	1936	1
464	23	1993	
465	19	1817	
…	…	…	…

Length	Virus (only)	Human (only)	Human and virus
495	7	1542	
496	16	1564	
497	27	1408	
498	14	1451	
499	16	1494	
500	22	1542	

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 3. The partial of frequency distribution of DNA maximal repeats (length 5–500 bp).

4.3. The length distribution of DNA maximal repeats in both the genomes of human and 559 virus genuses

Comparing the maximal repeats that appear only in virus (Virus only), only in humans (Human only) or in both human and virus (Human and virus), **Table 3** shows the partial frequency distribution of maximal repeats whose lengths are from 5 to 500 bp. It is observed that the majority of those maximal repeats whose length range from 7 to 11 almost belong to the "Human and Virus." Note that there may exist extra nucleic acid codes, for example, "N," within these DNA sequences such that the number of maximal repeat (length = 5) appearing in both humans and viruses in **Table 3** is 1,341 and that is great than 4^5 (= 1024).

4.4. The longest DNA maximal repeat (length = 463 bp) appearing in both the genomes of human and 559 virus genuses

Table 3 shows the length of the longest maximal repeat extracted in both the genomes of humans and selected viruses of 559 virus genuses is 463 bp. In [21], the result of blasting two sequences, "*Homo sapiens* chromosome 5" (NC_000005.10) and "Human herpesvirus 6B" (NC_000898.1), as shown in **Table 4**, that longest repeat appears 109 times within human chromosome 5 and two times within virus "Human herpesvirus 6B." To further inspect the longest maximal repeat, as show in **Figure 9**, one can find that the longest one is a tandem repeat (TAACCC) and appears within virus "Human herpesvirus 6B" at two intervals, the front (8249–8711 bp) and tail (161570–162,032 bp), that are located within the regions of direct repeats (DR) [35]. **Figure 10** gives one of two longest patterns aligned within "Human herpesvirus 6B" (8249–8711 bp) in **Figure 9**.

4.5. The statistics of DNA maximal repeat patterns (length = 100 bp) appearing in both human and 559 virus genuses

Table 5, for example, shows the statistics of 13 DNA maximal repeat patterns (length = 100 bp) appearing in both human and 559 virus genuses. It is observed that the three repeats as the

http://dx.doi.org/10.5772/intechopen.75113

Maximal repeat patterns	DF	TF	Length	Class frequency distribution (ClassID#DF#TF)	Regular expression	Human chromosome (GRCh38.p7 Primary assembly)	Viruses
ctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaacccta accctaaccctaaccctaaccctaaccct aaccctaaccctaaccctaaccctaaccc taaccctaaccctaaccctaaccctaacc ctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaacccta accctaaccctaaccctaaccctaaccct aaccctaaccctaaccctaaccctaaccc taaccctaaccctaaccctaaccctaacc Ctaaccctaaccctaaccctaaccctaac Cctaaccctaaccctaaccctaaccctaa Ccctaaccctaaccctaaccctaacccta Accctaaccctaaccctaaccctaaccc	2	111	463	(C248#1#109) (C442#1#2)	(TAACCC)n	5	Human herpesvirus 6B

Table 4. The longest DNA maximal repeat patterns (Length = 463 bp) appearing in both humans and viruses.

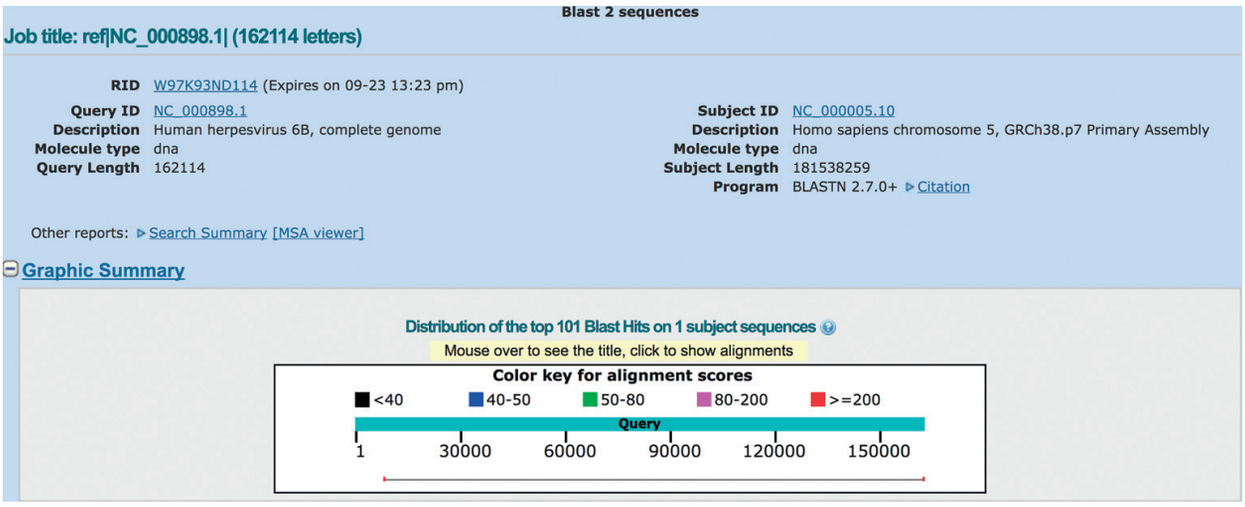

Figure 9. BLAST: "*Homo sapiens* chromosome 5" versus "human herpesvirus 6B"; "Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017" [21].

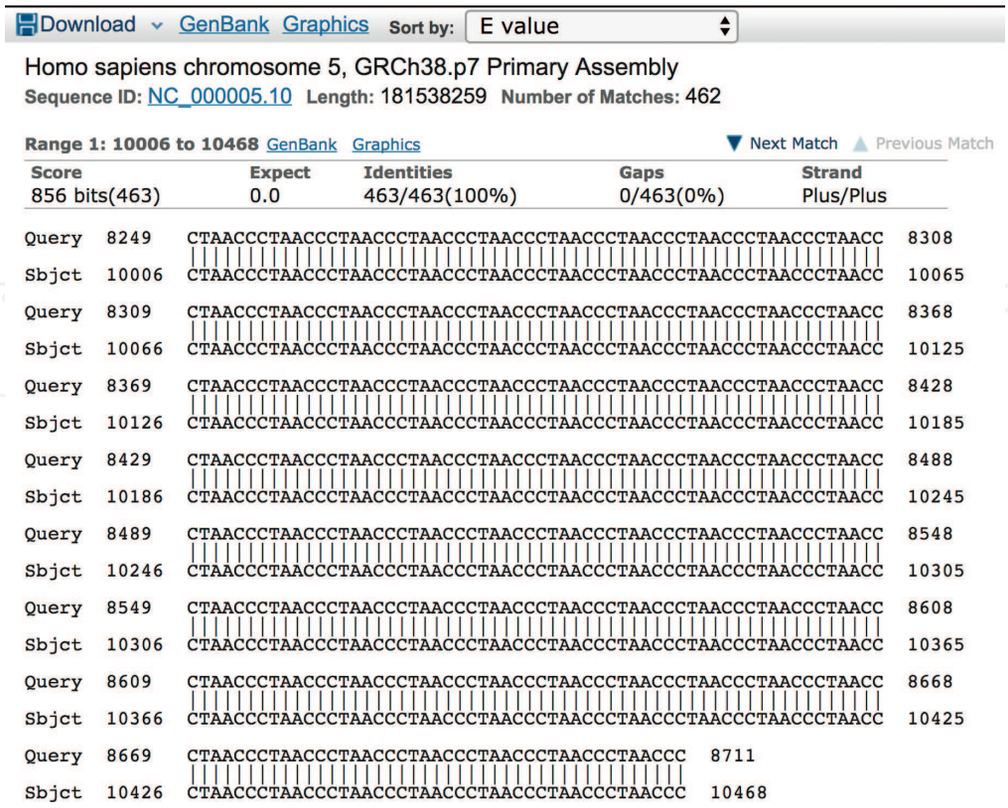

Figure 10. One of two aligned patterns (8249–8711 bp) in **Figure 9** "Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017" [41].

1st, the 6th, and the11th, for example, have the similar regular expression as "(AACCCT)n", "(CTAACC)n," and "(TAACCC)n", respectively, and all of them appear in human chromosomes "1," "5," "10" and "12"; all of these three repeats appear in these viruses, "Cyprinid herpesvirus 1," "Falconid herpesvirus 1," "Gallid herpesvirus 2," "Human herpesvirus 6A," "Human herpesvirus 6B," and "Equid herpesvirus 3." It is very interesting to investigate the relationship between these human chromosomes and those viruses for further research. On the other hand, from the biological viewpoint, furthermore, (AACCCT)n, (CCCTAA)n, and (CTAACC)n may comprise the same maximal repeat pattern with different repeat frame; (GGGTTA)n, and (AGGGTT)n can also comprise the same maximal repeat pattern in complementary sequence. Moreover, the (GGGTTA)n is expected to be targeted by cisplatin [36].

4.6. Phenotypes: "Group I(dsDNS)" in Baltimore virus classification

It is observed that all of these viruses in **Table 5** belong to the "Group I(dsDNS)" of Baltimore classification [37], as shown in **Table 6**, and most of them are from the family "Herpesviridae" and order "Herpesvirales." Indeed, it is very interesting and attractive to have all the viruses compared with human whole genome and to inspect these co-occurrences of repeats for virus prevention from the genomic point of view in the future.

		Total		Length	Class frequency distribution (ClassID#DF#TF)	Regular expression	Human chromosome (GRCh38.p7 primary assembly)	Viruses
	Maximal Repeat Patterns	DF	TF				C248	C5,C14,C149,C284,C305,C357, C442C541
1	aaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacc	10	591	100	(Cl 49# 1 # 18) (C248#4#299) (C305#2#41) (C442#2#208) (C541#l#25)	(AACCCT)n	1, 5, 10, 12	Cyprinid herpesvirus 1, Falconid herpesvirus 1, Gallid herpesvirus 2, Human herpesvirus 6A, Human herpesvirus 6B, Equid herpesvirus 3
2	aatagaatagaatagaatagaatagaatagaatagaatagaatagaatagaa atagaatagaatagaatagaatagaatagaatagaatagaatagaatag	3	42	100	(C248#1#18) (C284#1#11) (C357#1#13)	(AATAG)n	X	Rabbit fibroma virus, Taterapox virus,
3	agggttagggttagggttagggttagggttagggttagggttagggttagg gttagggttagggttagggttagggttagggttagggttagggttaggg	9	188	100	(C248#7#152) (C305#2#36)	(AGGGTT)n	2,12,13,18,22, X, Y	Falconid herpesvirus 1, Gallid herpesvirus 2
4	atatatatatatatatatatatatatatatatatatatatatatatatatata tatatatatatatatatatatatatatatatatatatat	6	533	100	(C14#1#34) (C248#5#499)	(AT)n	2, 3, 7, 19, X	Gryllus bimaculatus nudivirus
5	ccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccct aaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccc	3	4	100	(C248#1#1) (C305#1#2) (C541#1#1)	(CCCTAA)n	5	Meleleagrid herpesvirus 1, Equid herpesvirus 3
6	ctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaa	11	591	100	(C149#1#18) (C248#4#298) (C305#2#41) (C442#3#210) (C541#l#24)	(CCCTAA)n	1, 5,10,12	Cyprinid herpesvirus 1, Falconid herpesvirus 1, Gallid herpesvirus 2, Human herpesvirus 6A, Human herpesvirus 6B, Human herpesvirus 7, Equid herpesvirus 3
7	gagagagagagagagagagagagagagagagagagagagagagagag agagagagagagagagagagagagagagagagagagagagagagaga gagagagaga	3	55	100	(C149#1#32) (C248#2#23)	(GA)n	6, 11	Cyprinid herpesvirus 3
8	ggggttagggttagggttagggttagggttagggttagggttagggttagg gttagggttagggttagggttagggtagggttagggttagggttaggg	3	3	100	(C248#2#2) (C305#1#1)	G(GGGTTA)n	13, 18	Meleleagrid herpesvirus 1
9	gggttagggttagggttagggttagggttagggttagggttagggttaggg ttagggttagggttagggttagggttagggttagggttagggttagggt	9	188	100	(C248#7#152) (C305#2#36)	(GGGTTA)n	2, 12,13,18,22, X, Y	Falconid herpesvirus 1, Gallid herpesvirus 2

		Total	Length	Class frequency distribution (ClassID#DF#TF)	Regular expression	Human chromosome (GRCh38.p7 primary assembly)	Viruses
10	gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtg tgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtg	5	78	100	(C248#4#38) (C5#1#40)	(GT)n	2,10, 16, 19	Orgyia pseudotsugata MNPV
11	taaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac	10	588	100	(C149#1#16) (C248#4#298) (C305#2#41) (C442#2#208) (C541#1#25)	(TAACCC)n	1, 5,10,12	Cyprinid herpesvirus 1, Gallid herpesvirus 2, Falconid herpesvirus 1 Human herpesvirus 6A, Human herpesvirus 6B, Rquid herpesvirus 3
12	tgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt	5	76	100	(C248#4#35) (C5#1#41)	(GT)n	2,10, 16, 19	Orgyia pseudotsugata MNPV
13	ttagggttagggttagggttagggttagggttagggttagggttagggttag ggttagggttagggttagggttagggttagggttagggttagggttag	9	186	100	(C248#7#150) (C305#2#36)	(AGGGTT)n	2,12,13,18,22, X, Y	Falconid herpesvirus 1, Gallid herpesvirus 2

Table 5. The statistics of 13 DNA maximal repeat patterns (length = 100 bp) appearing in both humans and viruses.

Viruses	Class ID	The International Committee on Taxonomy of Viruses (ICTV)			Baltimore classification
		Genus	Family	Order	
Orgvia pseudotsugata MNPV	C5	Alphabaculovirus	Baculoviridae	N	Group I(dsDNA)
Gryllus bimaculatus nudiviras	C14	Alphanudivirus	Nudiviridae	N	Group I(dsDNA)
Cyprinid herpesvirus 1	C149	Cyprinivirus	Alloherpesviridae	Herpesvirales	Group I (dsDNA)
Rabbit fibroma virys	C284	Leporipoxvirus	Poxviridae	N	Group I (dsDNA)
Falconid herpesvirus 1	C305	Mardivirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Gallid herpesvirus 2	C305	Mardivirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Meleagrid herpesvirus 1	C305	Mardivirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Taterapox virus	C357	Orthopoxvirus	Poxviridae	N	Group I(dsDNA)
Human herpesvirus 6A	C442	Roseolovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Human herpesvirus 6B	C442	Roseolovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Human herpesvirus 7	C442	Roseolovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Equid herpesvirus 3	C541	Varicellovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 6. The taxonomy of 12 viruses selected in **Table 5**.

5. Conclusions and future works

Except considering the phenotypes that result from the epigenetics [38], it is believed that some of the phenotypes of creatures (or organisms) are determined by their genotypes as they are born in the beginning. This chapter proposes a novel approach to mine for genetic markers via comparing class frequency distributions of maximal repeats extracted from tagged genomic sequences of creatures, where the classes are derived from the tags given by domain experts. Once domain experts can divide the creatures into disjoint classes as precisely as possible according to their features (phenotypes), then they can adopt the scalable approach developed in [22] to extract the maximal repeats and compute class frequency distributions of these repeats via comparing the whole genomic sequences of these creatures. The repeats or the combination of some repeats that are with extremely biased class frequency distribution can be seen as class markers (genotypes) and can provide clues to biologists to analyze the relationship among these class markers (genotypes) and their corresponding features (phenotypes).

Due to the availability of cloud computing with flexible infrastructure, nowadays, it becomes possible to compute class frequency distributions of maximal repeats from a huge amount of tagged whole genomic sequences of many creatures across species via the scalable maximal repeat extraction approach [22] with Hadoop MapReduce programming model. The function mentioned in this chapter is somewhat like "Archimedes' Law of the Lever," as shown in

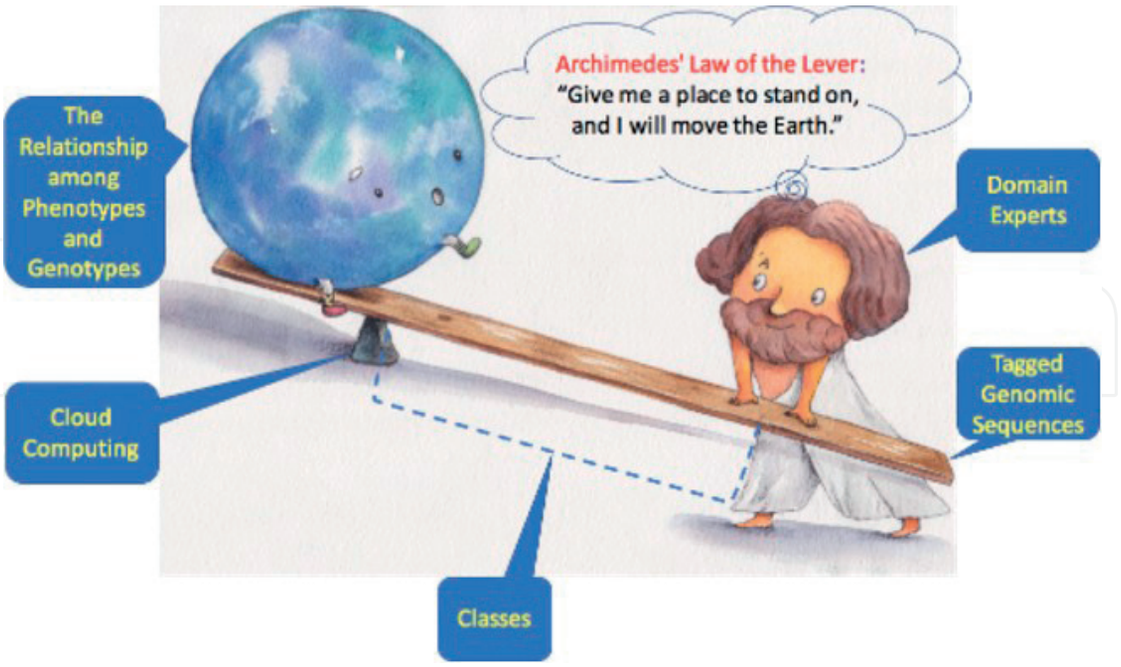

Figure 11. "Mining for biomarkers via observing class frequency distributions of maximal repeats from tagged genomic sequences" is somewhat like "Archimedes' Law of the Lever".

Figure 11, the Archimedes, an outstanding ancient Greek scientist, said that "Give me a place to stand on, and I will move the Earth." With scalable computing power and enough tagged genomic sequences, in other words, a domain expert can figure out the relationship among phenotypes and genotypes if the classes are properly and precisely defined. It is desired to have further cooperation with domain experts, especially who have collected the whole genomes of diverse organisms and desire to find or identify the relationship between genomic signatures and the features they concern in the future.

From a practical point of view, it is inconvenient for general users to have experiments of maximal repeat extraction by themselves in the beginning because there are a lot of preprocessing works and need considerable hardware infrastructure to support such a big-data computing. Furthermore, it might be a bottleneck or nightmare for general users, for example, biologists, to implement Hadoop MapReduce programming as described in [22]. Therefore, it is highly desired if maximal repeat extraction can be provided in public cloud services, such as Amazon Elastic Container Service (AWS ECS) [39], Google Cloud Platform [40], and Azure Container Service (AKS) [41]. It is highly expected that one will develop novel comparative genome with tagged genomic sequences and bring users with novel cloud services of computing class frequency distribution of maximal repeats in the future.

Acknowledgements

This study is supported by Ministry of Science and Technology, Taiwan, under project *MOST 106-2632-E-468-002*. I thank Prof. Jeffrey J.P. Tsai who provided computational environments

and finical support. I also thank Prof. Yi-Chun Wang for collecting the human chromosomes and inspecting experimental results about viruses and humans. I also thank Prof. Rouh-Mei Hu who helped in explaining the SNPs and the relationship between genotypes and phenotypes; Prof. Jan-Gowth Chang and Charles C.N. Wang had valuable discussions; I thank Jazz Wang for providing valuable Hadoop programming discussions. Finally, I thank Ling-Yu Ji for offering her drawing to show the conceptual diagram of "Archimedes' Law of the Lever."

Author details

Jing-Doo Wang[1,2]*

*Address all correspondence to: jdwang@asia.edu.tw

1 Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan

2 Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan

References

[1] Azuaje F. Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine. Wiley; 2011

[2] Novelli G, Ciccacci C, Borgiani P, Amati MP, Abadie E. Genetic tests and genomic biomarkers: Regulation, qualification and validation. Clinical Cases in Mineral and Bone Metabolism. 2008;**5**(2):149154

[3] Glauser TA. Biomarkers for antiepileptic drug response. Biomarkers in Medicine. 2011; **5**(5):635641

[4] Sun W et al. Common genetic polymorphisms influence blood biomarker measurements in COPD. PLoS Genetics. 2016;**12**(8):e1006011

[5] What are genome-wide association studies? https://ghr.nlm.nih.gov/primer/genomicresearch/gwastudies.

[6] Genome-wide association studies. https://www.yourgenome.org/stories/genome-wide--association-studies.

[7] Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: Current insights and future perspectives. Nature Reviews Cancer. 2017;**17**:692704

[8] Sniekers S, Stringer S, Watanabe K, Jansen PR, Coleman JRI, Krapohl E, Taskesen E, Hammerschlag AR, Okbay A, Zabaneh D, Amin N, Breen G, Cesarini D, Chabris CF, Iacono WG, Arfan Ikram M, Johannesson M, Koellinger P, Lee JJ, Magnusson PKE,

McGue M, Miller MB, Ollier WER, Payton A, Pendleton N, Plomin R, Rietveld CA, Tiemeier H, van Duijn CM, Posthuma D. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nature Genetics. 2017;**49**:1107. EP –, 05

[9] Louhelainen J. SNP arrays. Microarrays. 2016;**5**(4):27

[10] Illumina genotyping solutions. https://www.illumina.com/techniques/popular-applications/genotyping.html.

[11] Genome-Wide Human SNP Array 6.0. https://www.thermofisher.com/order/catalog/product/901182

[12] Clark DP, Pazdernik NJ. Chapter e9 - genomics and systems biology. In: Clark DP, Pazdernik NJ, editors. Molecular Biology. 2nd ed. Boston: Academic Press; 2013. p. e110, e117

[13] Ha N-T, Freytag S, Bickeboeller H. Coverage and efficiency in current snp chips. European Journal of Human Genetics. 2014;**22**:11241130

[14] The database of Genotypes and Phenotypes (dbGaP). https://www.ncbi.nlm.nih.gov/gap.

[15] Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;**107**(1):1-8

[16] Han Y, He X. Integrating epigenomics into the understanding of biomedical insight. Bioinformatics and Biology Insights. 2016;**10**(267289)

[17] Brown JR. Comparative Genomics: Basic and Applied Research. CRC Press; 2007

[18] NCBI Whole Genomes FTP Site. ftp://ftp.ncbi.nih.gov/genomes.

[19] The Cancer Genome Altas (TCGA). https://cancergenome.nih.gov/

[20] Cancer Moonshot. https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative

[21] Wang J-D, Wang Y-C, Hu R-M, Tsai J. Extracting the co-occurrences of dna maximal repeats in both human and viruses. In: The 17th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE2017); 2017

[22] Wang J-D. Extracting significant pattern histories from timestamped texts using mapreduce. The Journal of Supercomputing. 2016:1-25

[23] Wang J-D. An external memory approach to compute the statistics of maximal repeats across classes from whole genome sequences. In: 2005 National Computer Symposium, Taiwan, R.O.C. p. BIC1–2, 2005

[24] Wang J-D. External memory approach to compute the maximal repeats across classes from DNA sequences. Asian Journal of Health and Information Sciences. 2006;**1**(2):276-295

[25] Wang C-T. Method for extracting maximal repeat patterns and computing frequency distribution tables, Sep 2017. US Patent App. 15/208,994

[26] Wang J-D. A novel approach to compute pattern history for trend analysis. In: The 8th International Conference on Fuzzy Systems and Knowledge Discovery; 2011. pp. 1796-1800

[27] Wang J-D, Heri W. Extracting retrospective patterns from time-stamped texts according to variable query time interval. In: The International Multi-Conference on Engineering and Technology Innovation 2015 (IMETI2015); 2015

[28] Wang J-D, Jiang A-K, Chen J-C. Shape query for pattern history in PubMed literatures via Haar wavelet. International Journal of Advanced Information Technologies. 2015;**9**(6):67-76

[29] Chan W-L, Wang J-D, Chang J-G, Tsai J. Genome-wide functional identification of maximal consensus patterns derived from multiple species pirnas. In: The 16th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE2016); 2016

[30] Wang J-D, Chan W-L, Wang CCN, Chang J-G, Tsai JJP. Mining distinctive DNA patterns from the upstream of human coding and non-coding genes via class frequency distribution. In 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2016); 2016

[31] Wang J-D, Hwang M-C. A novel approach to extract significant patterns of travel time intervals of vehicles from freeway gantry timestamp sequences. Applied Sciences. 2017;**7**(9)

[32] Wang J-D. A novel approach to improve quality control by comparing the tagged sequences of product traceability. In: The 3rd International Conference on Inventions; 2017

[33] NCBI Whole Genomes FTP Site *Homo Sapiens* Assembled Chromosomes. ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/seq/

[34] NCBI Whole Genomes FTP Site Virus Whole Genomes. ftp://ftp.ncbi.nih.gov/genomes/Viruses/all.gbk.tar.gz

[35] Dominguez G, Dambaugh TR, Stamey FR, Dewhurst S, Inoue N, Pellett PE. Human Herpesvirus 6B genome sequence: Coding content and comparison with human Herpesvirus 6A. Journal of Virology. 1999;**73**(10):8040-8052

[36] Nguyen HTQ, Galea AM, Murray V. The interaction of cisplatin with a human telomeric DNA sequence containing seventeen tandem repeats. Bioorganic & Medicinal Chemistry Letters. 2013;**23**(4):1041-1045

[37] Baltimore D. Animal Virology. Number 4. Elsevier Science; 1976

[38] Felsenfeld G. A brief history of epigenetics. Cold Spring Harbor Perspectives in Biology. 2014;**6**(1)

[39] Amazon Elastic Container Service (AWS ECS). https://aws.amazon.com/tw/documentation/ecs/

[40] Google Cloud Platform : CONTAINER ENGINE. https://cloud.google.com/container-engine/

[41] Introduction to Azure Container Service (AKS). https://docs.microsoft.com/en-us/azure/aks/intro-kubernetes

IntechOpen