An Effective Algorithm for SNP Haplotype Block Inference

Chia-Ling Sun[†], Chang-Biau Yang[†],
Yow-Ling Shiue[‡] and Hsing-Yen Ann[†]

[†] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan. cbyang@cse.nsysu.edu.tw.

[‡] Institute of Biomedical Science, National Sun Yat-sen University, Kaohsiung, Taiwan.

Abstract

Recently, it has been shown that there should exist a block-like structure in human genome, and thus only limited haplotype diversity can be obtained. In this paper, we propose a fixed-diversity strategy to find out the suitable block diversity and block boundaries. The diversity value in one block is defined as $d = 1 - \sum_{i=1}^{n}(x_i)^2$, where x_i denotes the frequency ratio of the ith kind of haplotype within the block and n denotes the number of distinct types of haplotype. We figure out that once a putative block stretches across the primary block boundaries, the diversity will increase rapidly. And the secondary block boundary effects occur when two types are merged or split into different types. The threshold in our algorithm is decided by the two detections of the primary and secondary block boundary effects. We obtain a reasonable diversity of chromosome 21 SNP data with our algorithm. Our partition result shows highly concurrence property to the haplotype data downloaded from NCBI website.

Key words: *SNP, haplotype, diversity*

1 Introduction

Since Human Genome Project (HGP) finished, we are eager to know what makes us different rather than what we are in common. As a genetic marker, *SNP* (single nucleotide polymorphism) data can be used to capture human disease traits because of its abundance and low diversity. In recent research results, it has been shown that there is a block-like structure in human genome, and only limited haplotype diversity can be observed.

SNP occurs about every thousands of base pairs, and there is a special phenomenon at these loci. Only two kinds of possible nucleotides may appear at each SNP site. The one with higher frequency is called *major allele* and another is called *minor allele*. The frequency of major allele is usually higher than 90%. Because there are only two possible nucleotides, SNP data can be easily described as binary numbers, e.g. {*true, false*}, {*on, off*} or {*0, 1*}. Figure 1 shows an example of binary number representation.

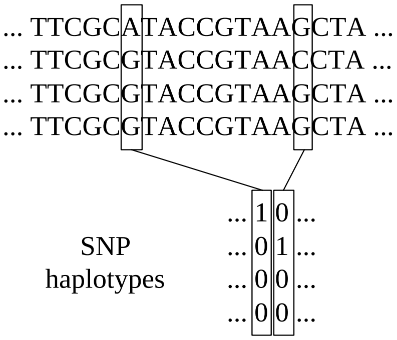

... TTCGCATACCGTAAGCTA ...
... TTCGCGTACCGTAACCTA ...
... TTCGCGTACCGTAAGCTA ...
... TTCGCGTACCGTAAGCTA ...
SNP
haplotypes
... 1 0 ...
... 0 1 ...
... 0 0 ...
... 0 0 ...

Figure 1: SNP haplotype expression.

The numerous number of SNP data make them useful in association study and also lead the hardness to retrieve the information inside. Mendelian genetic model demonstrates the principle of segregation and the principle of independent assortment. That phenomenon of continuous allele tends to be inherited together is different from classical genetics. Alleles of close SNP sites tend to be inherited together, and thus *linkage disequilibrium* (LD) occurrs. A set of associated SNP alleles in a region of a chromosome that always inherit together is called a *haplotype*. The International HapMap Project started in October 2002 and its aim is to develop a haplotype map of the human genome [1]. Recent reports have shown that SNP data tends to be a key to human genome disease association study and evolution [7, 10, 13, 17].

Hot and cold spots of recombination phenomenon have been discovered in several studies via different ways of observation or calculation [4–6]. Because of the non-random crossover in meiosis process, the linkage disequilibrium is formed so that there exists block-like structure in our genome, as shown in Figure 2. Be-

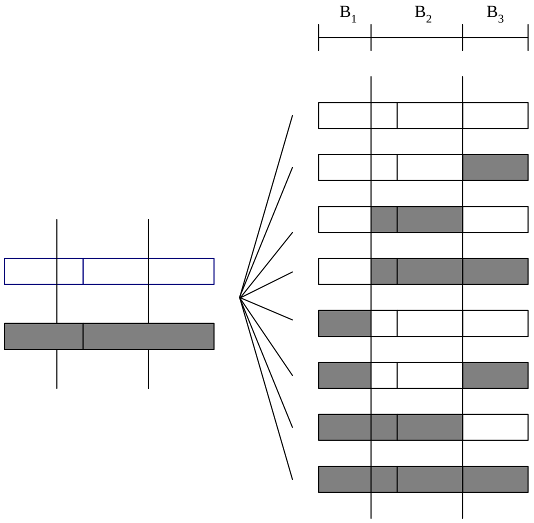


Figure 2: Limited diversity phenomenon in recombination.

sides, only limited haplotype diversity can be observed within each block [12, 16, 18].

The property of limited diversity in DNA sequence within one race is also stated in many researches [12, 16, 18]. Patil et al. found that there exists limited haplotype diversity in human chromosome 21, and four most frequently haplotypes occupy about 80% haplotypes in population [14]. Defining the threshold of diversity needs a great knowledge, and it is somehow subjective. Different haplotype block definitions make it difficult to study the mechanism based on the formation of haplotype block structure [4]. It is hard to judge whether it is right or wrong in using different criterion of a block, thus we are going to use a more objective view to define the threshold of acceptable haplotype diversity.

Since we know that there should be limited haplotype diversity in each block, only a few kinds of common haplotypes can account for a bulk percentage of population [8, 9, 14, 15]. The measurement of the haplotype diversity becomes an important subject.

In Section 2, we will provide an overview on type classification process and diversity calculation. Our approach to inference the haplotype blocks will be described in Section 3. In Section 4, we will present some experimental results. We finish this paper with discussions of what we have learned about the haplotype block inference problem so far and what the remainder issues are in Section 5.

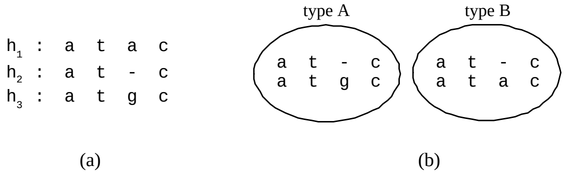


Figure 3: Example for compatible haplotypes. (a)Three haplotypes h_1, h_2 and h_3. (b)The haplotypes that are compatible to the types.

2 Preliminaries

In this section, we first briefly describe how to classify the haplotypes into several different types. By using the type classification process, we can calculate the diversity values of the given haplotypes. After all, we will give the definition of the haplotype block ingerence problem.

Before calculating the diversity within a block region, we have to classify the haplotypes into several different types, where within a type, all haplotypes are compatible to each other. Two haplotypes are said to be *compatible* if each allele position of the two haplotypes are the same, except the positions of missing data (denoted by ’$-$’). In Figure 3(a), pair (h_1, h_2) is compatible and pair (h_2, h_3) is also compatible. It is clear that pair (h_1, h_3) is not compatible due to the third SNP allele.

When we say a haplotype is compatible with a type, it means that the haplotype is compatible with all haplotypes in this type. See Figure 3. For example, h_1 is compatible with type B, but it is not compatible with type A. On the other hand, h_2 is compatible with both type A and type B. If a haplotype belongs to a type, it must be compatible with this type.

According to the definition of a type, we can now classify the haplotypes into different types. Figure 4 illustrates the concept of classification in the block. It is clear that there are three distinct types, $\{10110, 01110, 10011\}$, where the darker square denotes the major allele and it is represented as 1. Thus we can obtain the frequency of each type from the classification result. For example, the frequency ratio of $\{10110, 01110, 10011\}$ is $\{0.7, 0.2, 0.1\}$, respectively.

After the frequency of each type is obtained, the entire diversity then can be calculated from these frequencies. The diversity value is defined as $d = 1 - \sum_{i=1}^{n}(x_i)^2$, where x_i represents the frequency ratio of the ith kind of haplotype within the block, n represents the number of distinct types of haplotype. The value of d is the probability that two haplotypes chosen at random from a block are different from each other. This diversity calculation formula applies the idea of gene

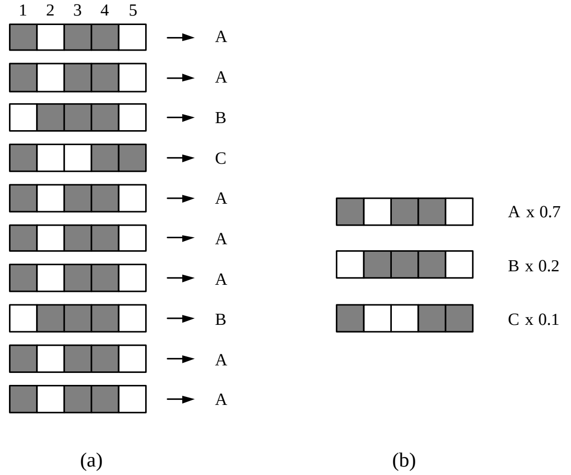


Figure 4: Classification of haplotypes. (a)The classified types of haplotypes. (b)The frequency ratio of each type.

diversity formula proposed by Li [11]. Take Figure 4 as an example, the diversity of the block is calculated as $d = 1 - 0.7^2 - 0.2^2 - 0.1^2 = 0.46$.

The problem we are going to solve is to find the suitable block diversity value in SNP data, and to inference the block boundaries. Because there is only limited diversity in each block, when we are given the SNP data shown in Figure 5(a), we try to partition the sequence into blocks such that the diversities of these blocks are limited. Figure 5(b) shows one of the possible partitions of the given input data in Figure 5(a).

According to the previous researches and analysis on SNP data, we have following assumptions:

Assumption 1. When we compute the diversities of the putative blocks inside a haplotype, the diversity should increase slowly. Once a putative block crosses the haplotype boundaries, the diversity will increase rapidly.

Assumption 2. The diversities of the different haplotypes should be very close. To simplify the design of our algorithm, we assume that all haplotypes are of the same diversity value.

3 A Fixed-diversity Strategy to Inference the Haplotype Blocks

The fundamental procedure of our fixed-diversity approach is to partition blocks with different diversities. We increase the diversity gradually and observe how the number of blocks changes. For example, when the diversity is fixed at 0.55 for partitioning blocks, we try to enlarge the block size as large as we could but the diversity of the entire block can not exceed 0.55. Figure 6 describes an example of partitioning block with diversity fixed at 0.55.

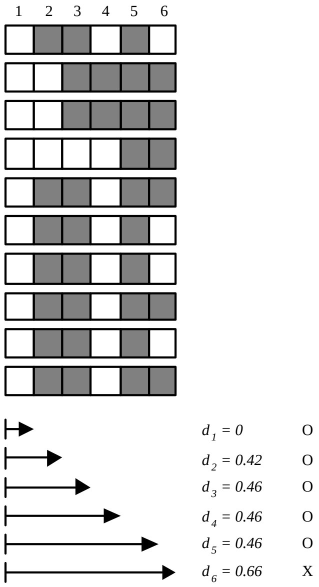
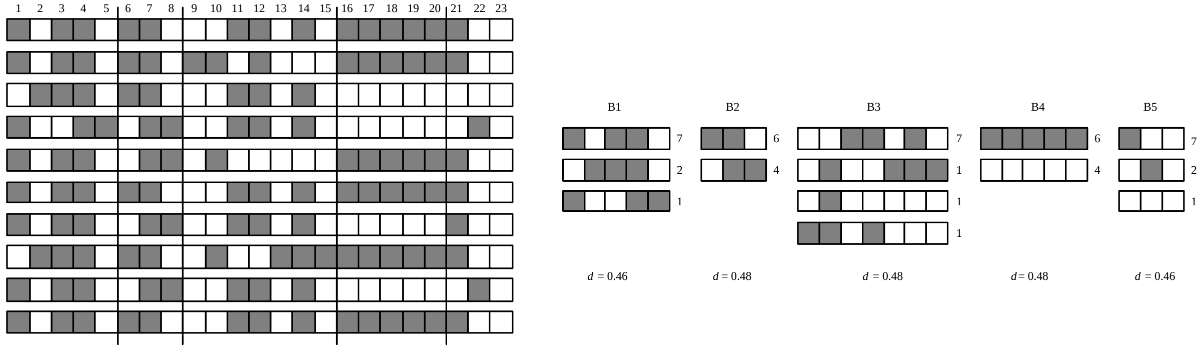


Figure 6: The partition result with diversity fixed at 0.55. d_i represents the diversity of the block from position 1 to i. With diversity fixed at 0.55, the block size can be enlarged until position 5.

3.1 Primary Block Boundary Effects

The boundaries are called *primary boundaries* if the diversity between two adjacent boundaries is bounded by the given threshold. Take Figure 5 as an example. Solid lines partition the SNP sequences into several blocks, and the diversity of each block is limited. Thus, these solid lines are the primary boundaries. Because there is no relationship between two adjacent blocks, once a putative block crosses the primary block boundaries, the diversity will increase rapidly. In Figure 7, b_i stretches within two adjacent primary block boundaries. Thus the diversity calculated from b_i is less than or equal to the primary diversity in this block. Once the block length extends to b_k, the diversity in b_k will be much larger than that in b_i. Enlarge the block length, the types in block will branch from larger groups into smaller ones, the behavior can be represented by a tree relation in Figure 8.

Once a putative block stretches across the primary block boundary, it will cause a rapid diversity increase. This is the most important property of block structure in DNA sequence, and it is why we can use diversity-based algorithm to capture the appropriate diversity



(a) (b)

Figure 5: An example of haplotype block ingerence result. (a)The origional input SNP data. (b)One possible ingerence result (with fixed diversity 0.5).

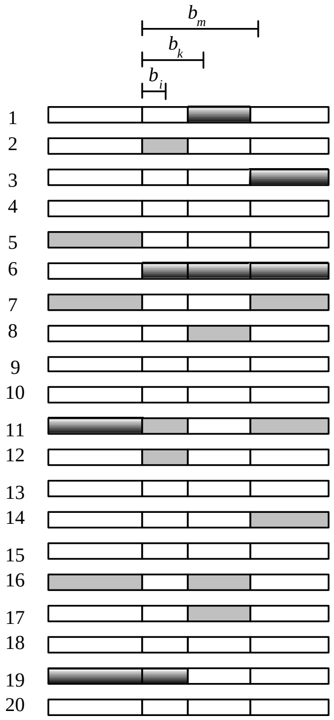


Figure 7: The concept of SNP data with diversity 0.405. Different color represents different types in each block. Vertical solid lines are the primary block boundaries. The putative blocks are b_i, b_k and b_m.

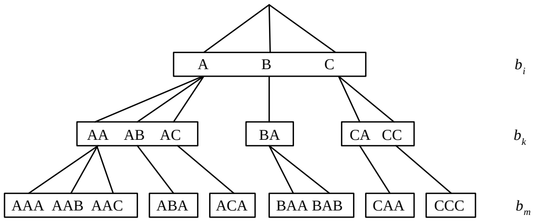


Figure 8: Tree relation representation for the putative blocks.

with SNP data.

Our fixed-diversity approach is as follows. We try to partition the blocks with some fixed diversity values 0.05, 0.1, 0.15, ..., 0.95. In other words, each time we increase 0.05 in diversity value and observe the change of total number of blocks. With a fixed diversity value μ, we scan the input sequences linearly and calculate the accumulated diversity η at each position. When $\eta \geq \mu$, we find a possible boundary, which is the line between the preceding and current positions. The change of η is rapid when one block stretches across the block boundary. Then, we can calculated the number of blocks partitioned with μ, which is denoted as Φ_μ. If $|\Phi_{\mu_{i-1}} - \Phi_{\mu_i}|$ is large, the possibility that μ_i is a good diversity value is high. We can use the first order differentiation chart to represent the degree of difference. Large difference will trigger a peak in the chart.

See Figure 9. Assume that there is a set of SNP data with block diversity $\mu_{opt} = 0.63$. There will not trigger a peak at other diversities, such as 0.55 and 0.6. Since our algorthm linearly scans the input sequences from the left to the right, either 0.55 or 0.6 possibly partitions the block when the primary block boundary is encountered. Otherwise, one block may increase its diversity value and may exceed 0.55 or 0.6 if the block

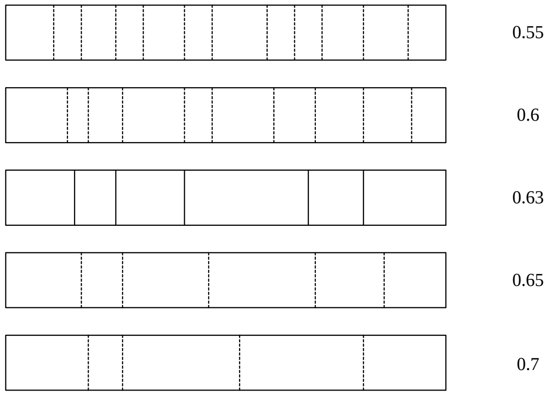


Figure 9: Concept of partition results when fixed at different diversities.

includes the boundary. Thus, the difference of $\Phi_{0.55}$ and $\Phi_{0.6}$ is not large.

Similarly, the difference of $\Phi_{0.65}$ and $\Phi_{0.7}$ will not trigger a peak, because 0.65 is only a little larger than μ_{opt}, and the boundaries found by 0.65 are close to primary boundaries. When we set the fixed diversity at 0.7, the slight diversity difference will not make the block length stretch to next primary block boundary. The block boundaries created by 0.7 will locate near them created by 0.65, thus the difference of $\Phi_{0.65}$ and $\Phi_{0.7}$ is not large. The concept of the adjacent diversity results is shown in Figure 9. We obtain an important conclusion: Where there is a block, there is a rapid diversity change.

3.2 Secondary Block Boundary Effects

When the fixed diversity is across the primary block boundary diversity, there will be a force in diversity increasing. The force formed by diversity gap will trigger a peak in the first order differentiation chart. Diversity gap not only occurs when the entire block diversity meets, but also occurs when two types of haplotype are mergeed together. This kind of diversity gaps caused by two types merging are called *secondary block boundary*. Take merged diversity gap as an example. If each block only contains three different types, the possible secondary block structures are pointed out in Figure 10. The dotted lines indicate that if we extend block from the first position to the next position, there will occur a diversity gap at those positions.

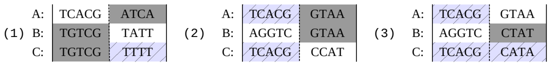


Figure 10: An example of secondary block boundaries. Solid lines represent primary block boundaries, and the dotted lines represent the secondary block structure. Different texture of haplotype points out different types of haplotype within each secondary blocks.

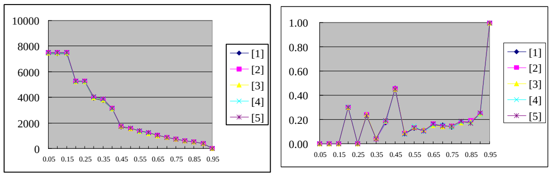


Figure 11: The block number variance and the first order differentiation chart on artificial data with diversity 0.405.

According to the concept stated above, the first order differentiation chart will arise at the positions when the diversity meets the special diversity value where two or more haplotypes tend to merge together or split from one another. For instance, suppose that in each block, 15 individuals are with type A, 3 individuals are with type B and 2 individuals are with type C. Then the diversity in the block is $1-(\frac{15}{20})^2-(\frac{3}{20})^2-(\frac{2}{20})^2=0.405$. There should be an rapid arise in the first order differentiation chart at diversity 0.45, since 0.405 is the primary diversity and it can be detected from 0.4 to 0.45. If we merge types A and B, we will get diversity $1-(\frac{15+3}{20})^2-(\frac{2}{20})^2=0.18$, which will cause a secondary block boundary. Thus, it will trigger large variance on the number of blocks at diversity 0.2. If types A and C are merged together, the diversity is 0.255 and it will be detected when diversity reaches 0.3. Merging of type B and C, we can obtain diversity 0.375 and we can detect it when diversity reaches 0.4. Our experiments also reveal these secondary block structure effects in Figure 11.

The secondary block boundaries are formed by different combinations of type merging. It is possible that two combinations of merged types of haplotypes share the same diversity of the secondary block structure, or it may equals to the diversity of the primary block, thus the number of peaks in first order differentiation chart may be less than the number of different combinations.

When we enlarge the block length to cross the primary block boundaries, it will generate a diversity gap, while secondary block structure boundaries only generate small peaks. Thus, there are some small peaks in the first order differentiation chart.

3.3 Dealing with Missing Data

Without missing data, we can calculate the diversity within each block easily. The SNP data we use for analysis was downloaded from Perlegen Sciences,

		n
P_1	a a t g	0
P_2	a a - g	1
P_3	a t a c	0
P_4	a t a c	0
P_5	- - a c	2
P_6	a a t -	1
P_7	a t - c	1
P_8	c t a -	1
P_9	c t a c	0
P_{10}	a t a c	0

(a)

		n	Type
P_1	a a t g	0	A
P_3	a t a c	0	B
P_4	a t a c	0	B
P_9	c t a c	0	C
P_{10}	a t a c	0	B
P_2	a a - g	1	A
P_6	a a t -	1	A
P_7	a t - c	1	B
P_8	c t a -	1	C
P_5	- - a c	2	B/C

(b)

Figure 12: An example of input SNP data with ten persons, where n represents the amount of missing data in each sequence. (a)The origional ten input SNP sequences. (b)The reordered SNP data with the classification procedure.

Inc. [2]. We reconstruct the data and obtain the SNP sequences for twenty persons. There are many missing data in the SNP sequences, thus we need to deal with them. We tried several ways to deal with missing data and encounter different problems. The method we adopt can get rid of these side effects that occurred in other methods. When we find one haplotype is compatible with more than one types in the classification process, we say this haplotype is *ambiguous*.

In our algorithm, when we execute the classification procedure, we need to reorder the input SNP sequences according to the amount of the missing data. The one with fewer missing data will be processed by the classification procedure first. For example, Figure 12, illustrates the reordering of ten input SNP sequences.

After input sequences have been reordered, we perform the classification procedure, which assigns each haplotype (sequence) a *type*. Because P_1 is the first sequence, so that there is no type at the beginning. Classification procedure assigns P_1 to one new type, say type A. Next, the type of P_3 is to be decided. Since no type is compatible with P_3, we assign this haplotype to a new type, say type B. When a haplotype h is to be determined, we examine all existing types to check if there exists any type compatible with h. If no type is compatible with h, we build up a new type.

We use a frequency table to store the number of haplotypes compatible with each type. The frequency table is shown in Table 1.

Obviously, P_5 is an ambiguous haplotype because it is compatible with two types. Our algorithm allocate this haplotype to type B because type B has the highest frequency in all types compatible with P_5. The allocation idea is based on the report that few common haplotypes construct 80% of the genetic diversity among the samples [14].

Table 1: The frequency table after performing the classification procedure on the input SNP data shown in Figure 12.

Type	Compatible individuals	Number of individuals
Type A	P_1,P_2,P_6	3
Type B	P_3,P_4,P_{10},P_7,P_5	5
Type C	P_9,P_8,P_5	3

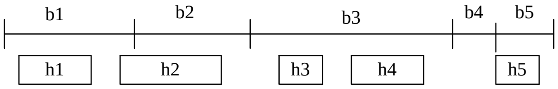


Figure 13: Inter-haplotype and intra-haplotype. h_i represents one haplotype and b_i represents one block. Intra-haplotypes regions are the regions inside h_1, h_2, h_3, h_4 and h_5. Inter-haplotypes are the regions between (h_1, h_2), (h_2, h_3), (h_3, h_4) and (h_4, h_5).

4 Experimental Results

In this section, we first define the penalty functions to evaluate the simulation results of our algorithm. Then, we brifly describe the testing data and show what the experimental results are.

4.1 Evaluation Functions

SNPs that lie close to each other along the DNA molecule construct a haplotype block and they tend to be inherited together. SNP variants that are far from each other along the DNA molecule tend to be in different haplotype blocks and are less likely to be inherited together. According to this definition of haplotype, we can use the properties to evaluate our partition result with respect to other random partition results. More clearly, we can say that two adjacent haplotypes tend to locate in different blocks, while one whole haplotype tends to locate within one block.

We define *inter-haplotype* and *intra-haplotype* first. Inter-haplotype is the relation between two adjacent haplotypes, and intra-haplotype is the region within each haplotype itself, as illustrated in Figure 13.

We would like to know whether one intra-haplotype locates within one block, and whether two adjacent haplotypes or two inter-haplotypes locate at different blocks. If blocks are set to be larger, two inter-haplotypes will locate in the same block. To evaluate

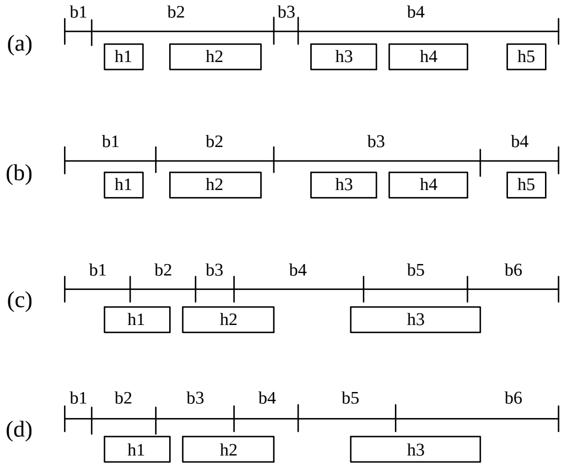


Figure 14: Some block partition results. (a)Haplotypes locateing within larger blocks. (b)Haplotypes locating within smaller blocks. (c)Haplotypes locating across blocks with higher penalty. (d)Haplotypes locating across blocks with lower penalty.

this phenomenon, we use adjacent penalty, denoted as *adj* to evaluate it. The function *adj* is defined as

$$\sum_{(h_i,h_{i+1}) \in \text{ haplotypes locating in one block}} adj(h_i,h_{i+1})$$

$$adj(h_i,h_{i+1}) = \begin{cases} 0\text{: } h_i \text{ and } h_{i+1} \text{ locate in different blocks.} \\ 1\text{: } h_i \text{ and } h_{i+1} \text{ locate in the same block.} \end{cases}$$

For example, in Figure 14(a) and 14(b), each intra-haplotype locates in only one block. It is obviously that we prefer the result of Figure 14(b), because that haplotypes locating in one block is not due to larger block size. When we use adjacent penalty to evaluate them, we check the adjacent haplotypes, (h_1,h_2), (h_2,h_3), (h_3,h_4) and (h_4,h_5). The adjacent penalty of Figure 14(a) is equal to 3 while in Figure 14(b), only one inter-haplotype locates in the same block. The adjacent penalty ratio is equal to $\frac{adjacent\ penalty}{number\ of\ haplotypes\ locate\ in\ one\ block}$ which can be regarded as the degree of larger block size. Larger adjacent penalty ratio means that the blocks are partitioned too large, so that two adjacent haplotypes usually locate in the same block.

If blocks are set to be smaller, the testing haplotypes will not locate within only one block. Cross penalty is defined as follows:

$$\sum_{h_i \in \text{ haplotypes not locating in one block}} cross(h_i).$$

$$cross(h_i) = \text{ The number of blocks } h_i \text{ crosses}$$

The cross penalty of Figure 14(c) is equal to 8, and Figure 14(d) has corss penalty 6. We define cross penalty ratio to represents the degree of partition the block in a small one. The cross penalty ratio is defined as $\frac{cross\ penalty}{number\ of\ haplotypes\ not\ locating\ within\ one\ block}$. The cross penalty ratio is used to find out how small the blocks are partitioned. With the two evaluation fuctions, it is easy to judge whether it is a suitable block result.

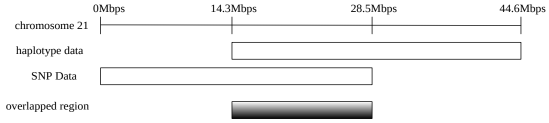


Figure 15: The overlapped region of the haplotype and SNP data in chromosome 21.

4.2 Testing Data

We use the haplotype data downloaded from NCBI website [3] to evaluate the performance of our partition results. Chromosome 21 spans 44.6Mbps long, the SNP data we adopt locates at the region from the beginning of chromosome 21 to 28.5Mbps, and the haplotype we use to evaluate partition results is in the range from 14.3Mbps to the end of this chromosome. Figure 15 represents these different ranges clearly. We focus on the overlapped region and evaluate the performance for different partition results. The overlapped region in Figure 15 contains 9911 SNPs and 840 haplotypes. Dealing with missing data by our fixed-diversity approach, we obtained 2593 blocks in the overlapped region. We will check the relationship between partitioned boundaries and the haplotype data.

4.3 Statistics of Partition Result

We scanned the SNPs of chromosome 21 and obtained the entire diversity located at 0.5. Table 2 shows the statistics of our partition result versus random partition results. We can see that both adjacent penalty ratio and cross penalty ratio of random partition with 2593 blocks are higher than our method. Statistics shows that while haplotype locates in one block, blocks with random partition are too large so that 22.7% inter-haplotypes are crowded in one block. The cross penalty ratio shows that blocks with random partition has worse result than ours, haplotypes locating over a blocks tends to span 4.41 blocks, while our partition method only 4.298.

We also present the random partition results with different block numbers. When we partition the entire sequence into 1000∼2000 or 2000∼3000 blocks randomly, the adjacent penalty ratio are higher than

Table 2: Statistics of our result and random partition results.

penalty functions	our result	random partition blocks			
		2593	1000∼2000	2000∼3000	3000∼4000
haplotypes locate in one block	344	366.8	460.6	377.0	314.1
adjacent penalty	61	83.20	169.85	91.15	49.45
adjacent penalty ratio	0.177	0.227	0.364	0.239	0.157
haplotypes not in one block	496	473.2	379.4	463.0	526.0
cross penalty	2132	1149.1	1298.4	2024.4	2729.4
cross penalty ratio	4.298	4.410	3.402	4.356	5.185

that of our method. Because the blocks may be too large, adjacent haplotypes tend to locate in one block. The cross penalty ratio of random partition with 2000∼3000 and 3000∼4000 blocks are higher than that of our method. Because the blocks may be too small, one haplotype spans more than one block. According to the statistics shown in Table 2, we have more confidence that our method is reasonable.

We also calculate the number of SNPs in each haplotype. The statistics shows that when we partition the 2593 blocks randomly, almost no block contains more than 30 SNPs. Adopting our fixed-diversity approach, there can be 54 SNPs in one block at most. The statistics of numbers of SNPs in one block is shown in Table 3.

5 Conclusion

Most of the existing SNP partition methods are based on compressing the storage size. However, in this paper, we focus on the entire primary block boundary structure. The result statistics shows that our main idea is feasible and suitable for our block structure assumption. The positions of boundaries are influenced by the missing data, and the missing data is the main nondeterministic factor in solving the boundary partition problem.

Several factors may affect the partition result, and we point out the main reasons that we may not partition the block in the right way from the point of view of SNP data and haplotype data. The missing SNP data is surely the most difficulty of the problem. From the aspect of SNP data, the block structure is stronger within one race. But, the SNP data we downloaded from NCBI website may contain more than one race, thus the block structure may not be very clear. Furthermore, the SNP data is gathered by a large number of biological experiments, there may exist experiment errors while acquiring these SNP data. From the haplotype data point of view, the haplotype we used in verification may contains different races, or the reliability of haplotype data is suspected. There still needs a lot of efforts to make it progress of this problem. We believe, after the uncertainty of haplotype property is dissolved, there will be a better way in solving this problem.

References

[1] http://www.hapmap.org/.

[2] http://www.perlegen.com/haplotype/.

[3] http://www.ncbi.nlm.nih.gov/.

[4] N. W. J. Akey, K. Zhung, R. Chakraborty, and L. Jin, “Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination and mutation,” *The American Journal of Human Genetics*, Vol. 71, pp. 1227–1234, 2002.

[5] E. Anderson and M. Slatkin, “Population-genetic basis of haplotype block in the 5q31 region,” *The American Journal of Human Genetics*, Vol. 74, pp. 40–49, 2004.

[6] N. Arnheim, P. Calabrese, and M. Nordborg, “Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved,” *The American Journal of Human Genetics*, Vol. 73, pp. 5–16, 2003.

[7] A. J. Brookes, “The essence of SNPs,” *Gene*, Vol. 234, pp. 177–186, 1999.

[8] M. J. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander, “High-resolution haplotype structure in the human genome,” *Nature Genetics*, Vol. 29, pp. 229–232, 2001.

Table 3: Statistics of our result and random partition results.

number of SNPs	NCBI	our	random partition blocks			
in one block	haplotype	method	2593	1000∼2000	2000∼3000	3000∼4000
1 ∼ 10	634	2486	2470.65	1176.8	2372.35	3499.35
11 ∼ 20	118	88	117	226.9	126.9	43.4
21 ∼ 30	50	13	5.05	50.1	10.05	0.65
31 ∼ 40	20	3	0.3	11.2	0.75	0
41 ∼ 50	6	2	0	2.6	0.05	0
51 ∼ 60	6	1	0	0.9	0	0
61 ∼ 70	2	0	0	0.2	0	0
71 ∼ 80	2	0	0	0	0	0
81 ∼ 90	1	0	0	0	0	0
91 ∼ 100	1	0	0	0	0	0

[9] S. Gabriel, S. Schaffner, H. Ngyen, J. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. Deflice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. Lander, M. Daly, and D. Altshuler, “The structure of haplotype blocks in the human genome,” *Science*, Vol. 296, No. 21, pp. 2225–2229, 2002.

[10] I. Gray, D. Campbell, and B. Spurr, “Single nucleotide polymorphisms as tools in human genetics,” *Human Molecular Genetics*, Vol. 9, No. 16, pp. 2403–2408, 2000.

[11] W. H. Li and D. Graur, *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., first ed., 1990.

[12] W. H. Li and L. A. Sadler, “Low nucleotide diversity in man,” *Genetics*, Vol. 129, pp. 513–523, 1991.

[13] P. Nowotnyand, J. Kwon, and A. Goate, “SNP analysis to dissect human traits,” *Current Opinion in Neurobiology*, Vol. 11, pp. 637–641, 2001.

[14] N. Patil, A. Berno, D. Hinds, W. Barrett, J. Doshi, C. Hacker, C. Kautzer, D. Lee, C. Marjoribanks, C. Kautzer, B. Nguyen, M. Norris, J. Sheehan, N. Shen, D. Stern, R. Stokowski, D. Thomas, M. Trulson, K. Vyas, K. Frazer, S. Fodor, and D. Cox, “Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21,” *Science*, Vol. 294, No. 23, pp. 1719–1723, 2001.

[15] D. Reich, S. Schaffner, M. Daly, G. McVean, J. Mllikin, J. Higgins, D. Richter, E. Lander, and D. Altshuler, “Human genome sequence variation and the influence of gene history, mutation and recombination,” *Nature Genetics*, Vol. 32, pp. 135–142, 2002.

[16] J. A. Schneider, M. Pungliya, J. Choi, R. Jiang, X. J. Sun, B. Salisbury, and C. Stephens, “DNA variability of human genes,” *Mechanisms of Ageing and Development*, Vol. 124, pp. 17–25, 2003.

[17] B. S. Shastry, “SNP alleles in human disease and evolution,” *Journal of Human Genetics*, Vol. 47, pp. 561–566, 2002.

[18] D. Wang, J. Fan, C. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, vHubbell, E. Robinson, M. Mittmann, M. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. Hudson, and E. Lander, “Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome,” *Science*, Vol. 280, No. 5366, pp. 1077–1082, 1998.