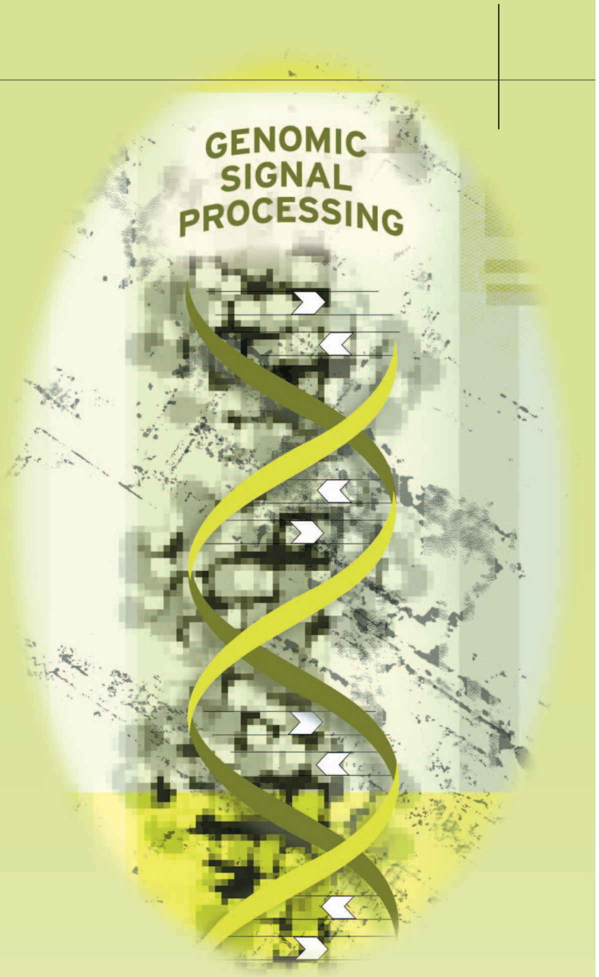
[Shih-Chieh Su, C.-C. Jay Kuo, and Ting Chen]

Single Nucleotide Polymorphism Data Analysis



© EYEWIRE

[State-of-the-art review on this emerging field from a signal processing viewpoint]

The basic structural units of the genome are nucleotides. There are four nucleobases: guanine (G), Adenine (A), Thymine (T), and Cytosine (C). A single nucleotide polymorphism (SNP) is a mutation at a single nucleotide position, where a possible nucleotide type is called an allele. For example, there are two nucleotides in the following two DNA fragments in the fourth position: CCACGTT and CCATGTT. In this case, we say that the SNP has two alleles, C and T. Although the polymorphism may consist of two, three, or four alleles, the triallelic and tetraallelic SNPs are extremely rare. Thus, the SNP is generally referred to as the bi-allelic polymorphism with the minor allele frequency (MAF) being larger than 1%. A locus is the location for a SNP (or gene) on a chromosome. It is called homozygous when two alleles are the same and heterozygous when they are different. The most frequent allele in a locus is called the wildtype while the second most frequent is the mutant allele.

Although more than 99% of human DNA sequences are the same across the population, the rest less-than-1% DNA variations can have a major impact on how humans respond to disease, environmental insults, drugs, and other therapies. SNPs make up 90% of all human genetic variations, and SNPs with a MAF at least 1% occur once every 100–300 bases along the human genome. SNPs may give clues as to why some subpopulations are more likely to have certain diseases than others, and why some drugs work in some subpopulations and not in others. This makes SNPs of great value for biomedical research and for developing pharmaceutical products or medical diagnostics. SNPs can also help identify multiple genes associated with complex diseases such as cancer and diabetes. Because a single altered gene may only contribute a little to a complex disease, these associations are difficult to establish with conventional gene-hunting methods.

We discuss several major problems in SNP data analysis and review some existing solutions in this work. Generally speaking, SNP data analysis is an emerging research field, and we foresee a rich set of SNP analysis problems to be cast in the signal processing framework. Our objective is to offer a state-of-the-art review on this topic from a signal processing viewpoint so that researchers in the signal processing field can grasp the important domain knowledge to overcome the barrier between the two fields.

DEFINITIONS AND BACKGROUND

GENOTYPE AND HAPLOTYPE

Human chromosomes appear in pairs. One of them is from the father, and the other is from the mother. The set of alleles that a person has on a pair of chromosomes is called a genotype. The term genotype can include the SNP alleles that a person has at a particular SNP location or many SNPs across the genome. A set of associated SNP alleles on the same chromosome is called a

1053-5888/07/$25.00©2007IEEE

haplotype. Genotyping is the method that discovers the genotype of a person.

The most accurate genotyping method is to apply the polymerase chain reaction (PCR) process, which is a technology that makes multiple copies of a DNA sequence, to the region where SNPs are located and then sequence the region directly. However, the PCR process cannot distinguish one chromosome from the other because they are almost identical (>99% similar). As a result, SNPs on both chromosomes will be amplified and sequenced, and their associations with the two chromosomes are lost in this process. In general, obtaining haplotype data through experiments is more difficult since it has to separate and purify a pair of almost identical chromosomes from human cells. In a population, most short chromosome regions have only a few common haplotypes, which account for most of the variations among people. Even though there may be many SNPs in such a chromosome region, the pattern of the haplotype variations in this region can be represented by a few selected SNPs, called tag SNPs.

A raw genotype data obtained from experiments is a sequence of SNP pairs, which is also called unphased SNPs. We use a simple example to illustrate the haplotype inference problem. Consider an individual A that has 00, 01, 01, 11, and 01 in five SNP locations in the genotype, where 0 represents the wildtype and 1 the mutant. Then, there are four possible configurations of haplotypes: (00010,01111), (00011,01110), (00110,01011) and (00111,01010). Additional information is needed to determine which one is more likely to happen. Suppose that we observe another genotype as 00, 00, 00, 11, and 00 in the corresponding five SNP locations, indicating two identical haplotypes of (00010). Thus, the haplotype configuration of individual A is more likely to be (00010,01111).

LINKAGE

In most cases, two SNP loci share a certain amount of correlation, that is, they are linked. Consider a pair of SNP loci (a, b) in the data set. There can be either allele X or x in location a and either allele Y or y in b. Sometimes, we are able to predict the allele at b based on the allele at a. This happens more frequently when positions a and b are close to each other. If the prediction accuracy is 100%, we say that these two SNPs are fully linked. There are also cases where a and b are independent, especially when they are far away from each other. Then, there is no linkage between them. More often, we have some partial information about the allele at b given the allele knowledge at a. The phenomenon is called linkage disequilibrium (LD). We address two popular LD metrics here, since they are closely related to the problem of disease association mapping, and the problem of tag SNP selections. The origin of LD metrics is the independent test

$$D = f(XY) - f(X)f(Y),$$

where (X,x) and (Y,y) are alleles of two SNPs. D is not a normalized measure. Lewontin [1] suggested to use the normalized one defined as

$$D = \frac{D}{D_{\max}},$$

where

$$\begin{cases} D_{\max} = \min\{f(X)f(Y),\ f(x)f(y)\}, & \text{if } D < 0, \\ D_{\max} = \min\{f(X)f(y),\ f(x)f(Y)\}, & \text{if } D \geq 0. \end{cases}$$

Please note that $|D| = 1$ denotes complete LD while $|D| = 0$ indicates that the two SNPs are independent. Historical recombinations between pairs of human chromosomes result in the decay of D toward zero. This metric has many extended versions and they have been widely applied, such as in [2] to partition the haplotype blocks.

Another LD metric is defined as

$$r^2 = \frac{D^2}{f(X)f(x)f(Y)f(y)},$$

which is also normalized. When $r^2 = 1$, the two SNPs are in complete LD, which means knowing one of them is directly predictive of the other. When $r^2 = 0$, the two SNPs are independent. This metric has been applied in [3] to select the tag SNPs. On the other hand, the value of r^2 is inversely related to the required sample size of association mapping, given a fixed genetic effect. That is, if one SNP was genotyped and it has $r^2 = 0.5$ to another ungenotyped SNP, then the sample size has to be doubled to provide the same statistical power for the ungenotyped SNP as the case with $r^2 = 1$.

RECOMBINATION

Before a single chromosome is produced, there is a process called meiosis that is essential to the generation of either sperms or eggs. During the meiosis, there is a small chance that a pair of paternal chromosomes exchange a segment of DNA with each other, called recombination. After the meiosis, these two paternal chromosomes are duplicated and separated into four different sperms. The same situation applies to the maternal chromosomes and eggs. Therefore, chromosomes are shuffled by recombinations, and so are haplotypes. When a recombination event occurs between two SNPs, it reduces the LD between them. Moreover, two SNPs close together are less likely to be affected by the recombination than two SNPs far away.

SNP DATA ANALYSIS: DATA SETS AND CHALLENGES

SNP DATA SETS

There are some data sets widely used in the literature. Daly et al. [2] reported the haplotype block structure in human chromosome 5p31, which denotes subband region 31 of the p-arm (i.e., the shorter arm) of chromosome 5. They released genotype data from 129 trios (father, mother, child; a total of 387 individuals). The amount of 103 SNPs were collected from each individual. There are 6,764 missing values and 3,868 heterozygous SNPs ($\sim 20\%$ unphased). The other data set is a set of haplotype data of human chromosome 21 by Patil et al. [4]. This collection gathered 24,047 SNPs from each of 20 individual chromosomes. This data set contains about 20% missing SNPs.

The International HapMap Project focuses on the construction of a haplotype map of the human genome, the HapMap, to reveal the common patterns of human DNA sequence variations. The first phase of the international HapMap project [5] genotyped more than 1 million SNPs in 269 individuals of European, Yoruba, Chinese, and Japanese ancestry. The density is around one SNP per 3 kilobase. It significantly increased the number and annotation of known SNPs in the public SNP map (dbSNP) from 2.6 million to 9.2 million. Motivated by the allele frequency distribution of variants in the human genome, SNPs with MAF $\geq 5\%$ was targeted in HapMap. The variants of these SNPs are referred to as common variants. The data show that the panels of Yoruba individuals are more diverse than other populations. This means that we need to select more tag SNPs to capture the same fraction of common variants for this population. The resource offered by HapMap has been applied to various research projects, especially the genome-wide association studies. The second phase of HapMap attempts to genotype an additional 4.6 million SNPs in each of the HapMap samples. This will increase the density of genotyped SNPs to one per kilobase.

CHALLENGES OF SNP DATA ANALYSIS

HAPLOTYPE INFERENCE

Genotype data can be less costly collected than haplotype data. However, haplotypes are still the desired data format ultimately. The problem of haplotype inference is to convert the genotype data into haplotypes. The phase of SNPs can be either homozygous, where both of the haplotypes have the same SNP, or heterozygous, where two haplotypes have different SNPs. Techniques in haplotype inference use the partial knowledge from the genotype source to infer the missing parts of the complete haplotype data. The inference processes are generally time consuming. As a result, precision and efficiency are both crucial in the context of haplotype inference.

HAPLOTYPE BLOCK PARTITIONING

Although the recombination process disturbs the genome, a large portion of DNA bases are still conserved. As a result, each SNP is somehow related to its neighboring SNPs. The linkage metric provides a measure to quantify the correlation between SNPs. When two SNPs have high correlation, they are closely linked. According to the recombination model, SNPs within a short genomic distance tend to be linked with each other. This phenomenon results in a block structure in the haplotype data. Haplotype blocks are also observed experimentally. The objective of haplotype block partitioning is to reduce the complexity of association mapping by using haplotypes rather than individual SNPs. On the other hand, recombination hot spots are likely to be located at the boundaries of haplotype blocks. These hot spots have significant impact on the population structure. Since every partitioning method gives a partition result, one challenging issue is to evaluate the performance of each block partitioning method in terms of power of association. Potential problems encountered in block partitioning are that sometimes boundaries between blocks are not clearly defined, and that SNPs in different partitioned blocks may still be linked to a certain degree.

TAG SNP SELECTION

The large number of SNPs makes SNP-based disease studies difficult, since we need to collect SNPs from hundreds to thousands of patients. It is however possible to collect a small set of representative SNPs, called tag SNPs, and use them to infer remaining SNPs. The intuition of the tag SNP selection comes from the LD among the SNP data. When a representative SNP is closely linked to a group of SNPs, it can be chosen to be a tag SNP for this group. Efforts have been made on selecting as few tag SNPs as possible for a data set. This set of tag SNPs can predict the remainder of the data set with a high precision. There are further issues on tag SNP selection. First, the selected tag SNPs cannot be better than the original set of SNPs in terms of representability. Thus, it is important to screen the SNP data and determine how well tag SNPs can catch the variation among individuals. Second, the LD observed from a sample set can be faulty if the sample size is too small. Finally, since most SNPs in the public databases have been discovered in a small sample of individuals, the ascertainment bias of SNPs in the initial set has to be considered. Despite the above challenges, the tag SNP selection problem still plays an essential role in SNP data analysis, since the cost and complexity of experiments can be significantly reduced. Recent studies on disease association are closely tied with tag SNP selection techniques.

MISSING DATA

Collected raw data often contain missing spots due to imperfect experiments. The portion of missing data indicates the quality of the data set. A missing SNP can be inferred using the information of its neighboring SNPs by exploiting the LD property. It is obvious that a properly partitioned block structure helps to infer missing SNPs more accurately. The distribution of missing locations is not uniform, i.e., some locations are more likely than others. Some missing data may be of the form of a short missing segment. These make the missing data inference problem more challenging.

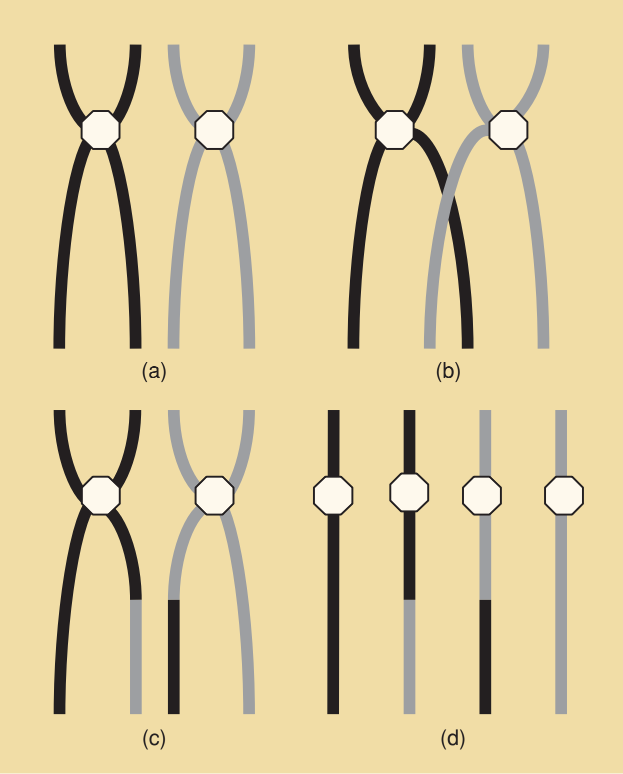
SOLUTION TECHNIQUES TO SNP DATA PROCESSING PROBLEMS

HAPLOTYPE INFERENCE

Although preliminary haplotype inference methods were developed in 1990s, rapid movements on SNP analysis have been driven by the accessibility to human SNP data sets recently. There have been several individual SNP data sets available since 2001. Furthermore, the International HapMap Project has also delivered the first phase data in late 2005. The real data facilitate the development of more efficient data processing methods, which in turn refines experimental techniques in data acquisition.

Haplotype inference is an essential step towards the processing of genotype data collected from experiments. It deals with incomplete data and attempts to infer missing data based on observed ones. The intuition of inference is to use homozygous locations in some individuals to predict SNPs at the same locations, but heterozygous in other individuals. Another useful information for haplotype inference is the neighboring SNPs due to the conservation of SNP patterns in a short range. The pioneering work of Clark [6] attempted to minimize the total number of different haplotypes. Clark's algorithm starts with the first haplotype and uses it to infer other genotypes. This process continues as more haplotypes are inferred from genotypes and stops until every genotype is resolved. The result of Clark's algorithm depends on the first haplotype to start with. Different starting haplotypes may result in different set of haplotypes. Sometimes this algorithm may not find any haplotype to start with. Moreover, genotype vectors that no compatible haplotypes are found in the solved set are left unresolved.

An expectation-maximization (EM) method for haplotype inference was proposed by Excoffier and Slatkin [7]. It is assumed that m different genotypes, with counts $n_1, n_2, \ldots n_m$, from n individuals are observed. The number of possible combinations of haplotype pairs leading to the jth genotype is



[FIG1] Illustration of the recombination process: (a) after meiotic replication and before cross over, (b) a cross over between two chromatids, (c) before meiotic divisions, and (d) after two meiotic divisions, which results in two recombinant and two non-recombinant chromatids.

$$c_j = \begin{cases} 2^{s_j - 1}, & \text{if } s_j > 0, \\ 1, & \text{if } s_j = 0, \end{cases}$$

where s_j is the number of heterozygous loci. Under the assumption of random mating, the probability P_j of the jth genotype is given by the sum of the probabilities of each of the possible c_j haplotype combinations as

$$P_j = \sum_{i=1}^{c_j} P(i\text{th haplotype combination}) = \sum_{i=1}^{c_j} P(h_k h_l),$$

where $P(h_k h_l)$ is the probability of the ith genotype made up of haplotypes k and l. $P(h_k h_l) = p_k^2$ if $k = l$ and $P(h_k h_l) = 2p_k p_l$ if $k \neq l$, where p_k and p_l are the population frequencies of the kth and the lth haplotypes, respectively.

The algorithm begins with the initial guess of haplotype frequencies. In the E-step, the haplotype frequencies are used to estimate the genotype frequencies. In the gth iteration, the haplotype frequencies in the previous iteration is used to calculate the probability of resolving each genotype into the different possible haplotype combinations

$$P_j\,(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{P(h_k h_l)^{(g)}}{P_j^{(g)}},$$

These expected genotype frequencies in turn estimate the haplotype frequencies in the M-step, where the haplotype frequencies are computed using a procedure equivalent to the gene-counting method

$$\hat{p}_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j(h_k h_l)^{(g)},$$

where δ_{it} is an indicator variable that gives the number of haplotype t is present in haplotype combination i. The E- and M-steps are performed iteratively until the haplotype frequencies converge, i.e., their change number is smaller than a threshold.

Nowadays, there are several commonly used tools for haplotype inference, including HAPLOTYPER [8], PHASE [9]–[11], SDPHapInfer [12], PPH [13], and HAP [14]. HAPLOTYPER adopts a Bayesian procedure that imposes no assumptions on the population evolution history using the Dirchlet prior. Their Gibbs sampling algorithm iterates between the following steps.

1) Conditioned on the Dirchlet prior, sample a pair of compatible haplotypes for each subject according to the candidate haplotypes' portion in current samples.
2) Conditioned on current samples, update the prior by a random draw from the posterior distribution modified by the sample counts.

Afterwards, two novel techniques, partition ligation (PL) and prior annealing, are used to solve the haplotype inference problem in a divide-and-conquer fashion. This procedure becomes a popular approach in haplotype inference research for lowering the computational cost. A direct derivative from HAPLOTYPER is the PL-EM algorithm [15], which uses EM instead of the Gibbs sampler for haplotype inference.

Another system using the Bayesian approach is PHASE, which appears to be the most accurate haplotype inference tool. The first version of PHASE [9] employs the Gibbs sampler with a prior approximating the coalescent (see [16] for a review). The approximation considers the distribution of the genealogy-related yet randomly sampled individuals as described by coalescence, which helps predict how similar a future-sampled chromosome and a previously sampled chromosome are likely to be. Moreover, future-sampled chromosomes tend to be more similar to previously sampled chromosomes as the sample size increases and as the mutation rate decreases. Coalescence theory is strongly supported by collected evolutional data. This is the main reason that PHASE infers the haplotype with higher accuracy than other methods.

The main shortcoming of the first version of PHASE is its long computational time, which has been significantly shortened in PHASE 2.0 [10] that applies the divide-and-conquer technique similar to PL in HAPLOTYPER. The fundamental difference between HAPLOTYPER and PHASE is their prior models. The Dirichlet prior used in HAPLOTYPER is suitable for the parent-independent mutation model. That is, the genetic sequence of a mutant offspring does not depend on the progenitor sequence. This model is however unrealistic for longer SNP sequences. We verify their accuracies on inferring the transmitted haplotypes of first 70 families in [2]. PHASE and HAPLOTYPER yield an error rate of 1.85% and 4.31%, respectively. However, the running time of PHASE is much longer than that of HAPLOTYPER.

PHH and HAP took a different approach from the Bayesian approach used in HAPLOTYPER and PHASE. They exploits ideas from a perfect phylogeny model. In such a model, no recombination is allowed and the site mutation is infinite. However, each SNP site can mutate only once, and the mutation lasts forever (namely, it cannot mutate back). Even though the perfect phylogeny model is not realistic in real data modeling, it does limit the diversity of haplotypes. A relaxed model, called imperfect phylogeny, is implemented in HAP. The system first picks "common" haplotypes from either perfect or imperfect phylogeny with enough evidence. Afterwards, it attempts to infer haplotypes from genotype data. Another program SDPHapInfer aimed to find the minimum subset of haplotypes that can resolve all the genotypes. It was shown in [12] that its result is within a factor of $O(\log n)$ away from the optimal solution, where n is the number of genotypes.

The running time of the HAP system is noticeably shorter than the PHASE system when the number of SNPs is small. The result obtained from HAP is comparable to that from PHASE. On the other hand, the running time for PHH is also linear in its later version. The performance of SDPHapInfer is similar to HAPLOTYPER for a small number of SNPs. We compare the popular PHASE, HAPLOTYPER and HAP in Table 1. It is worthwhile to point out that most of the haplotype inference algorithms produce different results in different runs. The final result is often the consensus of multiple runs of the same algorithm.

[TABLE 1] COMPARISON ON HAPLOTYPE INFERENCE PROGRAMS.

PROGRAM	PHASE	HAPLOTYPER	HAP
STRENGTH	ACCURATE	FAST	ACCURATE
WEAKNESS	SLOW	LESS ACCURATE	NO BLOCK PARTITIONING MECHANISM (SHORT BLOCKS ONLY)

HAPLOTYPE BLOCK PARTITIONING

The study of haplotype block partitioning was pioneered by Daly et al. [2] and Patil et al. [4], each offering a human SNP data set. The most important observation on these data sets is the existence of low-diversity regions, which are called haplotype blocks. Normally, there are two possible nucleotides in each SNP location. With a haplotype block spanning ten SNPs, it can have 2^{10} haplotypes. However, the actual number of haplotypes within this block is far less than that, say, ten. The problem to partition the data set into low-diversity haplotype blocks is called haplotype block partitioning.

Daly et al. [2] adopted a hidden Markov model (HMM) for haplotype block partitioning, where the transition probabilities between states at adjacent SNPs are related to the decay of LD and measured by a statistical value of D. Their method used several heuristics, including the window size to cover the local SNP information, the threshold of minority SNP frequencies, and the model itself. They developed a partitioning method and collected data of the trio format, which includes SNPs from fathers, mothers, and their children. They partitioned this data set into 11 inconsecutive blocks, which are generally referred to as the ground truth of this data set. Their partitioned haplotype blocks span up to 100 kb and contain five or more common SNPs. The blocks have only a few (two to four) major haplotype patterns. For example, the first haplotype block containing eight SNPs observed in two major patterns, GGACAACC and AATTCGGG. These patterns account for 95% of observed chromosomes. Another 3.8% of chromosomes match either of the two patterns at all alleles except one. This may be due to gene conversion or an undetected genotyping error.

Patil et al. [4] and Zhang et al. [17] partitioned haplotype blocks using the block diversity, which is measured by the number of haplotype tag SNPs (htSNPs) within a block. Patil et al. [4] determined block boundaries using a greedy algorithm while Zhang et al. [17] proposed a dynamic programming (DP) algorithm to find the block partition corresponding to a globally minimal number of htSNPs. Actually, the solution of Zhang et al. [17] was optimized for the haplotype block partitioning problem set up by Patil et al. [4].

Two haplotypes are compatible if the alleles are the same at the loci with no missing data. A haplotype in a block is ambiguous if it is compatible with two other incompatible haplotypes. Let r_i be a SNP locus whose value can be 0, 1, or 2, where 0 indicates missing data, 1 and 2 are the two alleles. For a block $(r_i, \ldots, r_j)$, the indicator function $b(r_i, \ldots, r_j) = 1$ if at least α percent of unambiguous haplotypes in the block are represented more than once. Here α

serve as a metric for diversity, where higher α implies lower diversity. Let $f(r_i, \ldots, r_j)$ be the minimum number of htSNPs required to uniquely distinguish at least α percent of unambiguous haplotypes within the block. The minimum total number of htSNPs needed for the first j SNPs, S_j, can be derived from the DP formula

$$S_j = \min\{S_{i-1} + f(r_i, \ldots, r_j); 1 \le i \le j, b(r_i, \ldots, r_j) = 1\}.$$

Although the objective function is optimized, there can be several partitions that yield the same optimum number of htSNPs. Then, the partition with the minimum number of blocks is chosen as the final partition.

SNP DATA ANALYSIS IS AN EMERGING RESEARCH FIELD, AND WE FORESEE A RICH SET OF SNP ANALYSIS PROBLEMS TO BE CAST IN THE SIGNAL PROCESSING FRAMEWORK.

Following the work of Zhang et al. [17], several researchers have studied a systematic way to achieve haplotype block partitioning based on the DP algorithm. For example, the objective function in the algorithm was replaced with the minimum description length (MDL) measurements by Koivisto et al. [18] and Anderson and Novembre [19]. Simply speaking, an MDL system selects the best model that yields the MDL for the whole data set. There exists an additional assumption in the MDL algorithms presented in [18] and [19]. That is, to derive the description length of a haplotype in a block, SNPs are assumed to be independent of each other within a block, which is however unrealistic. Anderson and Novembre [19] added another assumption, i.e., blocks follow the first-order Markovian relationship. These two methods yield different results, since different MDL models are applied to the same data set.

Low diversity is a common feature of haplotype blocks. It is convenient to use the entropy to measure the haplotype diversity within a block; namely, low entropy implies low diversity. Let (i, j) denote the set of consecutive SNPs from the ith SNP to the jth SNP. Let $\Phi(i, j)$ be the set of all haplotypes collected in this interval. The block entropy in the iterative partition-inference (IPI) system [20] is defined as

$$E(i, j) = \sum_{\phi \in \Phi(i,j)} P_\phi \log P_\phi^{-1}.$$

A DP formula similar to that in [17] can be used for haplotype block partitioning while minimizing the total block entropy. Let $B(j)$ denote the minimum total block entropy from the first SNP to the jth SNP. The DP structure is as

$$B(j) = \min_{1 \le i \le j} \{B(i-1) + E(i, j); \text{for } E(i, j) \le T\}. \qquad (1)$$

The condition $E(i, j) \le T$ in (1) defines a block, where T is a threshold on the maximum entropy allowed for a block. With proper choice of T, the partition algorithm yields similar partition result as in [2], which is considered as ground truth for their released data set.

TAG SNP SELECTION

Recently, several large genomic regions of around 500 kb have been comprehensively examined as part of the Encyclopedia of DNA Elements (ENCODE) project. This project resequenced 96 chromosomes to ascertain all common variants and genotyped all SNPs that are either in the dbSNP database or identified by resequencing. These studies strongly confirm the patterns of long segments revealed in [21]. Therefore, most of the common SNPs in the genome have groups of neighbors that are all in nearly perfect correlation with each other. One SNP, the tag SNP, can thereby serve as a proxy for many others in further studies.

Tag SNP selection provides feedback to further experiments in SNP collection. It screens currently available SNP data and yields representative SNPs. The LD metrics offer good measures for tag SNP selection. Moreover, they are viewed as the hinge to reconstruct ungenotyped SNPs from tag SNPs. In this section, we review several tag SNP selecting systems: HapBlock [22], ldSelect [3], and STAMPA [23]. Tag SNPs are selected to span a large portion of samples. The cases of the singleton, which is the single occurrence of a specific haplotype pattern and may come from genotyping errors, or SNPs with low MAF (usually below 10%) tend to be uncovered. Only selection algorithms were considered in the first two systems. Besides the selection algorithm, STAMPA also proposed a reconstruction method, which may further reduce the number of tag SNPs while achieving the same coverage.

The problem of finding the minimum number of representative SNPs within a block to uniquely distinguish all haplotypes is known as the minimum test set problem [17], which was proven to be NP-complete. HapBlock is a system built upon the extension of [17]. It uses DP for haplotype block partitioning and chooses the partition that has the fewest overall tag SNPs as the desired one. Tag SNPs are selected based on the haplotype data inferred from PL-EM. The program can run on the genotype data input.

ldSelect adopts the LD statistics measured in the r^2 metric. It runs a simple greedy algorithm that selects the SNP that is above a r^2-threshold with the maximum number of other SNPs until the selected set of tag SNPs can resolve the portion of all existing haplotypes. The fundamental idea in ldSelect is simple, and the greedy algorithm is considerably faster than the DP technique used in HapBlock and STAMPA. This attracts other researchers to follow this approach in their system to select tag SNPs in a larger scale data set.

In STAMPA, a stronger condition is imposed. That is, any unselected SNP should be restored using only two closest tag SNPs to its both sides as much as possible. Three auxiliary score functions on the prediction error are defined. DP is performed to minimize the score under a certain prediction function. The prediction process provides an algorithm to recover non-tag SNPs from selected tag SNPs, and phased haplotype data are needed in the prediction.

A different approach that targets at genome-wide tag SNP selection was proposed in [24]. It groups SNPs into segments of low haplotype diversity and selects a subset of SNPs that can discriminate all common haplotypes within blocks. While not relying on any predefined haplotype block structure, it is called a block-free selection. HapBlock and STAMPA involve systematic DP methods. Thus, the computation is very heavy in a genome-wide range. Furthermore, they may lose local selectivity due to the overall optimization.

The tag SNP selection method in [25] has the ability to distinguish haplotypes even when some tag SNPs are missing during the experiment. These selected tag SNPs are referred to as robust tag SNPs. Huang et al. [25] argued that finding minimum robust tag SNPs is equivalent to finding minimum tag SNPs in [17], and it is an NP-hard problem. They also proposed three algorithms to find approximate robust tag SNPs efficiently. Their solutions are close to the optimal solution while the genotyping cost can be saved by as high as 80%.

Similar to the block partitioning problem, evaluation of selected tag SNPs can be implicit. A system can always define a certain measure such as predictiveness, informativeness, prediction errors, or block diversity and then optimizes the chosen measure. In general, the more tag SNPs are selected, the higher percentages of haplotypes can be represented uniquely. Furthermore, more tag SNPs also means better protection against missing data and genotyping errors. The initial survey before tag SNP selection is also crucial step. If the survey is biased due to a small sample of individuals, the selection result will be biased as well.

MISSING DATA

Both missing SNP inference and haplotype inference are key problems in the processing of currently collected data. Some haplotype inference methods, e.g., [26], [10], [14], can handle missing SNPs. However, they are preliminary since only the local neighboring information within a fixed range is exploited.

The IPI method proposed in [20] partitions the haplotype globally and infers missing SNPs locally within a block. The partition algorithm was discussed earlier. For the inference, a single missing SNP is being updated under the assumption that the inference of other missing SNPs is true. An EM-like algorithm was employed to update the inference. The inference algorithm in [20] was proved to lower the entropy of the block in which the single missing SNP is located. This means that it helps organize the block content in a more desirable manner. The inference can further facilitate the block partitioning job for a lower total block entropy.

For the IPI method, every missing SNP is initialized to be the majority at its location, which is called “majority assignment”. This assignment uses only the single location to compute the likelihood without the help of its neighbors. Then, the data set is partitioned into haplotype blocks under this assignment result. After partitioning, the assignment is updated for every missing SNP. In the next round of iteration, the updated assignment is again employed for block partitioning. Both the error rate and the total block entropy are lowered during the iterative optimization process of IPI, until it converges. The results of various missing rates on data in [2] are shown in Table 2. The IPI has higher error rate when there are more missing SNPs.

[TABLE 2] ERROR RATES FOR THE INFERENCE OF MISSING SNPS ACCORDING TO DIFFERENT MISSING RATES, USING FULL DATA IN [2].

MISSING RATES	1%	5%	10%
MAJORITY ASSIGNMENT ERROR RATES	16.95%	19.01%	19.35%
FIRST ROUND	6.02%	8.51%	8.73%
SECOND ROUND	5.08%	7.75%	8.02%
THIRD ROUND	5.08%	7.38%	8.02%
FOURTH ROUND	5.08%	7.34%	7.98%
FIFTH ROUND	5.08%	7.34%	7.98%

CONCLUSION AND FUTURE PERSPECTIVES

Several SNP data processing problems and their solution techniques were reviewed in this work. Most of them involve statistical inference on SNP symbols (including haplotype and missing data inference) based on partial observations and their grouping (i.e., haplotype block partitioning) and representative selection (i.e., tag SNPs). A common framework of various solution methods consists of the selection of a proper signal model, a proper cost function, and an iterative optimization algorithm. This framework has been widely applied by researchers in the signal processing community to the analysis of speech, audio and communication signals. It is our belief that SNP data analysis will provide another excellent opportunity for signal processing researchers to contribute in the near future.

There are several interesting applications of SNP data analysis, which can lead to research problems in the future. One of them is the disease mapping problem, which is also known as the association study. The problem intends to identify the relationship between diseases and genomic data. In the association study, a genetic variant is genotyped in a population for which phenotypic information, such as disease occurrence or a range of trait values, is available. Recent developments set up a good platform for the genome-wide association study: the completion of the human sequence, the deposition of millions of SNPs into public databases, rapid improvement in genotyping techniques and the official release of the International HapMap Project.

Traditionally, genome-wide linkage analysis is the method used to identify disease genes. It has achieved great success in mapping genes underlying monogenic diseases. However, the linkage analysis is also much less powerful for identifying common genetic variants that have modest effects on disease (common diseases). Most common diseases and clinically important quantitative traits have complex architecture, for which the phenotype is determined by the sum of multiple genetic and environmental factors, and/or the interactions between them. The genome-wide association approach deals with common diseases by surveying most of the genomes for causal genetic variants. No assumptions are made about the genomic location of causal variants. Thus, it is an unbiased yet fairly comprehensive option, even in the absence of convincing evidence regarding the function or the location of causal genes. For more work on the genome-wide association, we refer to [27] and [28] for reviews and [29] and [30] for recent developments.

ACKNOWLEDGMENTS

This work is supported by NIH NIGMS PAR-02-021 (Center of Excellent in Genome Sciences: Implications of Haplotype Structure in the Human Genome) and NIH R01 LM008991-01 (Efficient Analysis of SNPs and Haplotypes with Applications in Gene Mapping).

AUTHORS

Shih-Chieh Su (shihchis@usc.edu) received his M.S. and Ph.D. degrees in electrical engineering from University of Southern California in 2001 and 2006, respectively. His research interests are mainly in genomic signal processing, bioinformatics and audio content analysis.

C.-C. Jay Kuo (cckuo@sipi.usc.edu) is director of the Signal and Image Processing Institute (SIPI) and professor of electrical engineering, computer science, and mathematics at the University of Southern California. His research interests are in the areas of digital signal/image/video analysis and modeling, multimedia data compression, communication and networking. He has guided 70 students to their Ph.D. degrees and served as coauthor of about 120 journal papers, 650 conference paper,s and seven books. He received the 1992 National Science Foundation Young Investigator Award and the 1993 Presidential Faculty Fellow Award. He is a Fellow of the IEEE and SPIE.

Ting Chen (tingchen@usc.edu) is an associate professor of biological sciences, computer science, and mathematics at the University of Southern California. His research interests are in the areas of computational biology/bioinformatics, algorithms, and pattern recognition. His current research topics include mass spectrometry data analysis, protein interactions and functions, and human genetics. He has published over 40 journal papers. He received the Alfred P. Sloan research fellowship in 2004.

REFERENCES

[1] R. Lewontin, "The interaction of selection and linkage. I. General considerations; heterotic models," *Genetics*, vol. 49, no. 1, pp. 49–67, 1964.

[2] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander, "High-resolution haplotype structure in the human genome," *Nature Genetics*, vol. 29, no. 2, pp. 229–232, 2001.

[3] C. Carlson, M. Eberle, M. Rieder, Q. Yi, L. Kruglyak, and D. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *Amer. J. Human Genetics*, vol. 74, no. 1, pp. 106–120, 2004.

[4] N. Patil, A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, C. Marjoribanks, D.P. McDonough, B.T. Nguyen, M.C. Norris, J.B. Sheehan, N. Shen, D. Stern, R.P. Stokowski, D.J. Thomas, M.O. Trulson, K.R. Vyas, K.A. Frazer, S.P. Fodor, and D.R. Cox, "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21," *Science*, vol. 294, no. 5547, pp. 1719–1723, 2001.

[5] The International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–1320, 2005.

[6] A. Clark, "Inference of haplotypes from pcr-amplified samples of diploid populations," *Mol. Biol. Evolution*, vol. 7, no. 2, pp. 111–122, 1990.

[7] L. Excoffier and M. Slatkin, "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population," *Mol. Biol. Evolution.*, vol. 12, no. 5, pp. 921–927, 1995.

[8] T. Niu, Z. Qin, X. Xu, and J. Liu, "Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms," *Amer. J. Human Genetics*, vol. 70, no. 1, pp. 157–169, 2002.

[9] M. Stephens, N. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *Amer. J. Human Genetics*, vol. 68, no. 4, pp. 978–989, 2001.

[10] M. Stephens and P. Donnelly, "A comparison of Bayesian methods for haplotype reconstruction from population genotype data," *Amer. J. Human Genetics*, vol. 73, no. 5, pp. 1162–1169, 2003.

[11] M. Stephens and P. Scheet, "Accounting for dacay of linkage disequilibrium in haplotype inference and missing-data imputation," *Amer. J. Human Genetics*, vol. 76, no. 3, pp. 449–462, 2005.

[12] Y.-T. Huang, K.-M. Chao, and T. Chen, "An approximation algorithm for haplotype inference by maximum parsimony," *J. Computat. Biol.*, vol. 10, no. 10, pp. 1261–1274, 2005.

[13] D. Gusfield, "Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions," in *Proc. RECOMB, 6th Annu. Conf. Research Computational Molecular Biology*, pp. 166-175, 2002.

[14] E. Halperin and E. Eskin, "Haplotype reconstruction from genotype data using imperfect phylogeny," *Bioinformatics*, vol. 20, no. 12, pp. 1842–1849, 2004.

[15] Z. Qin, T. Niu, and J. Liu, "Partitioning-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms," *Amer. J. Human Genetics*, vol. 71, no. 5, pp. 1242–1247, 2002.

[16] R. Hudson, "Gene genealogies and the coalescent process," in *Oxford Surveys in Evolutionary Biology*, vol. 7. London, U.K.: Oxford Univ. Press, 1990, pp. 1–44.

[17] K. Zhang, M. Deng, T. Chen, M. Waterman, and F. Sun, "A dynamic programming algorithm for haplotype block partitioning," *Proc. Natl. Acad. Sci.*, vol. 99, no. 11, pp. 7335–7339, 2002.

[18] M. Koivisto, M. Perola, R. Varilo, W. Hennah, J. Ekelund, M. Lukk, L. Peltonen, E. Ukkonen, and H. Mannila, "An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries," in *Proc. Pacific Symp. Biocomputing 8*, 2003, pp. 502–513.

[19] E. Anderson and J. Novembre, "Finding haplotype block boundaries by using the minimum-description-length principle," *Amer. J. Human Genetics*, vol. 73, no. 2, pp. 336–354, 2003.

[20] S.C. Su, C.C. Kuo, and T. Chen, "Inference of missing SNPs and information quantity measurements for haplotype blocks," *Bioinformatics*, vol. 21, no. 9, pp. 2001–2007, 2005.

[21] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler, "The structure of haplotype blocks in the human genome," *Science*, vol. 296, no. 5576, pp. 2225–2229, 2002.

[22] K. Zhang, Z. Qin, J. Liu, T. Chen, M. Waterman, and F. Sun, "Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies," *Genome Res.*, vol. 14, no. 5, pp. 908–916, 2004.

[23] E. Halperin, G. Kimmel, and R. Shamir, "Tag SNP selection in genotype data for maximizing SNP prediction accuracy," *Bioinformatics*, vol. 21, no. Suppl. 1, pp. i195–i203, 2005.

[24] B. Halldorsson, V. Bafna, R. Lippert, R. Schwartz, F. De La Vega, A. Clark, and S. Istrail, "Optimal haplotype block-free selecion of tagging SNPs for genome-wide association studies," *Genome Res.*, vol. 14, no. 8, pp. 1633–1640, 2004.

[25] Y.T. Huang, K. Zhang, T. Chen, and K.M. Chao, "Selecting additional tag SNPs for tolerating missing data in genotyping," *BMC Bioinformatics,* vol. 6, no. 263, 2005. [Online]. Available: http://www.biomedcentral.com/1471-2105/6/263

[26] S. Lin, D. Cutler, M. Zwick, and A. Chakravarti, "Haplotype inference in random population samples," *Amer. J. Human Genetics*, vol. 71, no. 5, pp. 1129–1137, 2002.

[27] J. Hirschhorn and M. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Rev. Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[28] W. Wang, B. Barratt, D. Clayton, and J. Todd, "Genome-wide association studies: Theoretical and practical concerns," *Nature Rev. Genetics*, vol. 6, no. 2, pp. 109–118, 2005.

[29] D.M. Maraganore, M. de Andrade, T.G. Lesnick, K.J. Strain, M.J. Farrer, W.A. Rocca, P.V. Pant, K.A. Frazer, D.R. Cox, and D.G. Ballinger, "High-resolution whole-genome association study of parkinson disease," *Amer. J. Human Genetics*, vol. 77, no. 5, pp. 685–693, 2005.

[30] R.J. Klein, C. Zeiss, E.Y. Chew, J.Y. Tsai, R.S. Sackler, C. Haynes, A.K. Henning, J.P. SanGiovanni, S.M. Mane, S.T. Mayne, M.B. Bracken, F.L. Ferris, J. Ott, C. Barnstable, J. Hoh, "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

SP