SNP and Haplotype Inference

Nikolas Meitanis
Physics Department, Massachusetts Institute of Technology
(Dated: May 11, 2005)

Single Nucleotide Polymorphisms (SNP) constitute the most prevalent form of variance in the human genome. Data indicate that certain complex, multi-gene conditions may depend on SNP structure. In particular, the complete sequencing of the SNP on both chromosome copies (haplotyping) in diploid organisms is often necessary. Experimental determination of the exact haplotype information is still difficult and expensive. Statistical methods to infer the information by genotype analysis have yielded promising results and merit further investigation.

I. INTRODUCTION: MUTATION AND THE HUMAN GENOME

Genetic polymorphism, DNA sequence variations in a population, takes many forms: deletions and insertions, single base substitutions, fragment repetitions etc. Among these, Single Nucleotide Polymorphism (SNP), the difference at a single location of two sequences, is statistically the most significant, comprising an estimated 90% of all variation. It is believed that around 10 million such SNPs exist in the human genome, with almost half of them identified and indexed to date. Such variations may be the result of an external agent (such as a virus) or due to chance.

For a variation to be categorized as an SNP, it has to be observed in at least 1% of the population, with the higher frequency variant called the *major allele* and the lower frequency the *minor allele*. SNPs occur through two diffferent substitution processes: transition, an exchange among pyramidines and purines independently; and transversion, a cross-exchange between the two groups. SNP distribution is non-uniform, with more of them found in the non-coding region of the DNA. High-density clusters have also been observed in areas of certain chromosomes and have attracted particular interest. In figures 1 and 2 we use the ENSEMBLE online search software to plot SNP information for chromosome 5.

II. SINGLE NUCLEOTIDE POLYMORPHISM

Among a variety of DNA sequence variations observed, *Single Nucleotide Polymorphisms (SNP)* are the most frequent. A specific variation is categorized as an SNP if it is found to occur in at least 1% of the population. The importance of SNP studies for medical research, diagnostics and pharmacogenomic development is paramount. SNPs are believed to affect propensity towards a disease or a virus, as well as the organism's repsonse to drug therapies. An accurate analysis of SNP locations may provide clues as to the gene-dependence of a number of multi-gene conditions, such as cancer. In addition, SNPs can be used as markers in navigating the human genome due to their dense coverage of both coding and non-coding sections of the DNA.

Several collaborations have worked to identify and catalogue the human genome's SNPs. The Human Genome Project (HGP) was established in 1998 as a joint effort funded by the Department of Energy and the National Institute of Health. The goal of the HGP was to identify at least 100000 SNP markers, develop the technology to study the data set and make it available to the public. An even more ambitious program, The SNP Consortium (TSC) [1], begun its research in 1999, largely funded by ten privately owned pharmaceutical companies. By the end of its discovery phase, the TSC had identified approximately 1.8 million SNP. The data can be accessed and searched online [19].

III. HAPLOTYPE ANALYSIS

While studying the effects of individual SNP on complex diseases, an association between multiple SNP at adjacent loci was observed. Patterns of SNP repetition spanning thousands of base-pairs emerged. Subsequent studies investigated the phenomenon, referred to as *Linkage Disequilibrium (LD)*, as a measure of observed allele frequency variations from expected equilibrium values. This motivated the separation of SNPs from genotype data, the combined information of both chromosome copies for human populations, into two individual sequences or *haplotypes*. While individual SNPs carry little information and obscure linkage associations which may be used for evolution studies etc, haplotypes may lead to identification of disease-susceptibility genes over long chromosomal segments. Studies by Daly et al. [2] and Rioux et al. [3] on Crohn's disease first demonstrated the importance of haplotype structure. The year 2002 saw the launch of the International Haplotype Map project to address the newfound significance of haplotypes.

The goal of the HapMap project is to determine common SNP patterns of the human genome, thus providing a more efficient means to analyze their functions. Haplotypes take advantage of two conclusions from human population genetics [4]. First, that common variants comprise the majority of sequence variations and second that the variants originated from single mutations and are thus linked to nearby variants on the ancestral chromosome. The HapMap collaboration completed its first

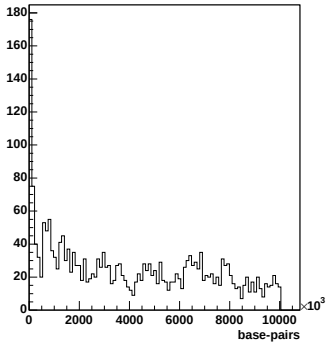

FIG. 1: SNP distribution versus base-pair for chromosome 5. The ENSEBLE search function was used with database Homo-Sapiens Genes (NCBI 35).

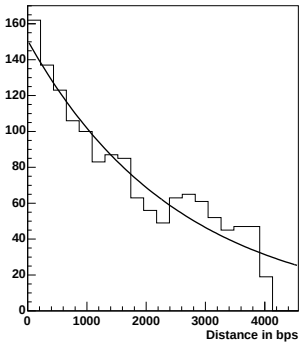

FIG. 2: Histogram of distance between adjacent SNPs for chromosome 5, fitted to an exponential decay.

draft of the map of the human genome and is now entering a second phase by adding several million more SNP [5].

Since full haplotype information cannot be achieved with currently available genotyping methods, additional experimental analyses are necessary. The most logical technique is the examination of family genetic material, especially parental DNA. However, this may prove unfeasible or at best expensive. In addition, it is complicated by *recombination*, a process that occurs during meiosis and tends to mix fragments of the maternal and paternal genetic material [6]. Alternatively, several experimental haplotyping techniques have been developed. *Polymerase Chain Reaction (PCR)* [20] is a molecular technique to achieve DNA amplification using a DNA polymerase. Most promising among the techniques that use PCR is the allele-specific long-range PCR which can amplify DNA regions tens of kbs long [7, 8] while related techniques include *Single-Molecule Dilution* [9] and *Polony PCR* [10] etc. *Conversion*, a strategy for the transformation of diploid to haploid cells [11] is yet another efficient method.

The success of several experimental derivations has to date failed to establish a fast and affordable method procedure. This fact has motivated the development of statistical techniques to infer the necessary information in order to fully reconstruct the haplotypes. These techniques are frequently referred to as *haplotype inference*. In the following section, four inference techniques are discussed.

IV. HAPLOTYPE INFERENCE

A. Clark's parsimony method

Clark [12] introduced a parsimony-based algorithm to infer the missing haplotype information. Clark's scheme requires an initial set of haplotypes known to exist in the population sample. Such a set is acquired by identifying all homozygotes and single heterozygotes to form an unambiguous set of haplotypes. All possible haplotypes for the remaining, ambiguous sequences are then compared to the initial set. If any of these can be combined with any haplotypes of the initial set to give the observed genotype, the corresponding "new" haplotypes are assumed correct and are added to the set of known sequences (figure 3). The algorithm proceeds until all haplotypes are resolved or until no more can be resolved.

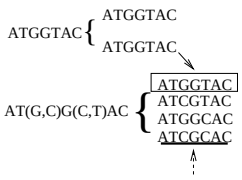

FIG. 3: Clark's algorithm identifies a known haplotype (ATGGTAC) in the sample and uses it for an unresolved subspace. The complementary haplotype (ATCGCAC in this case) is thus deduced and added to the resolved set.

Although the algorithm yielded good results, several problems remained with its implementation. In the event that an initial set cannot be constructed, the method will be unusable. Equation 1,

$$P = \left[1 - \frac{1}{\theta + 1} - \frac{\theta}{(\theta + 1)^2}\right]^n \qquad (1)$$

plotted in figure 4, shows the probability the algorithm will fail to start, which is small for expected values of θ. A more significant drawback is the possibility of not resolving all haplotypes in the sample correctly. Clark proposed that a maximum resolution principle guides the process: the number of correct inferences grows with the number of overall inferences, thus running the algorithm multiple times allows selection of the outcome with the best results. The hypothesis was verified by Gusfield [18]. Yet another disadvantage of the algorithm is that the

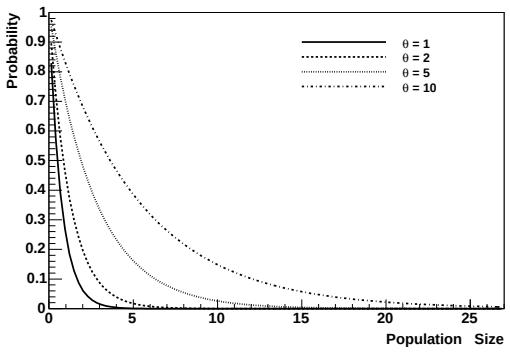

FIG. 4: The probability of failure to start as a function of the population sampled for Clark's algorithm

order in which the genotypes are analyzed will have an impact on the results.

B. The EM algorithm

The EM algorithm [13] approaches inference from a slightly different angle: given a set of genotype samples, the method identifies all possible haplotypes $(h_1, h_2, ..., h_h)$ and though multiple iterations estimates the frequency of occurrence for each member of the set. Formally, the algorithm is based on expressing the likellihood of a particular set of haplotype frequencies $(p_1, p_2, ..., p_h)$ as

$$L(p_1, p_2, ..., p_h) \;=\; \alpha_1 \prod_{j=1}^{m} \Big(\sum_{i=1}^{c_j} P(h_{ik} h_{il}) \Big)^{n_j} \qquad (2)$$

where c_j is the number of possible haplotype pairs, m is the number of different unresolved genotypes and n_j the number of times each such genotype is observed. The likelihood could then be maximized through partial derivatives with respect to the haplotype frequencies. However, the resulting system of equations can often be prohibitively complex to solve. Instead, the algorithm assumes an initial set of haplotype frequencies (in [13] the haplotypes are taken as equiprobable ad initio) as input. The probability of a particular genotype $h_k h_l$ at the gth iteration is expanded as a linear combination of all possible haplotype pairs. The haplotype frequencies for the next step of the iteration are derived from

$$p_t^{g+1} \;=\; \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P(h_k h_l)^{(g)} \qquad (3)$$

where δ_{it} is an integer for the number of times haplotype t is found in genotype i. Through successive iterations, the haplotype pair that maximizes the likelihood function is selected.

The main weakness of this method is a reduction in performance with sequence length. Another possible drawback is the assumption of Hardy-Weinberg equilibrium which may lead to bias in the resulting frequencies. Fallin and Schork [14] showed that the algorithm's sensitivity to a variety of simulated data sets and population structures is indeed small. Two additional improvements over Clark's original method are its success in yielding results even in the cases where the latter would fail to start and its independence of the order in which the samples are considered. However, frequency-estimation is an indirect method and still has to incorporated into actual haplotype analysis.

C. Gibbs Sampling

Using *Gibb's Sampling* , a type of Markov-chain Monte Carlo method (MCMC) [21], Stephens et al [15] constructed an algorithm based on some initial haplotype set $H^{(0)}$ to obtain a new set : $H^{(t+1)}$ for t=0,1,...,t. This forms a Markov chain by

1. Choosing an individual from the subset of all individuals whose genotype allows more than one possible haplotype pairs,
2. Expressing the chosen individual's genotype in terms of haplotypes taken from a random sampling of the distribution of all other individuals' haplotypes in the total sample $Pr(H_j \,|\, G, H_{-j})$, provided they could match the genotype under investigation.
3. Setting $H_j^{(t+1)} = H_j^{(t)}$ for the jth individual.

The complication arises in considering the probability $Pr(H_j \,|\, G, H_{-j})$ which is model dependent and often not well established for the model under consideration. The authors suggest that the next order haplotype can be obtained from a randomly chosen haplotype with a random number of mutations applied. Mathematically, this is expressed as $\pi(h \,|\, H)$ given by

$$\pi(h \,|\, H) \;=\; \sum_{\alpha \epsilon E} \sum_{s=0}^{\infty} \frac{r_\alpha}{r} \Big(\frac{\theta}{r+\theta} \Big)^s \frac{r}{r+\theta} (P^s)_{\alpha h} \qquad (4)$$

where r_α denotes the number of occurrences of the randomly selected haplotype in the set, r denotes the tital number of haplotypes s the number of mutations and θ the relevant mutation rate. In essence, the approximation above reflects the hypothesis that the next order haplotype is most likely to be identical or very similar to the observed haplotypes in the set.

This is the crucial difference with Clark's method and allows an improvement of the error rate by up to 50%. In addition, the algorithm performs well even with long sequences and neither fails to start nor provide a guess for every haplotype in the set. Finally, despite the method's reliance on Hardy-Weinberg equilibrium, similarly to the EM algorithm, Stephens et al. show that their results are not affected by variations within plausible population structures [22].

D. Bayesian Haplotype Inference

Niu et al.[16] introduced a more sophisticated method[23] which borrows several ideas from the aforementioned ones. Consider $G = (g_1, ..., g_n)$, $H = (h_1, ..., h_n)$ and $p = (p_1, ..., p_n)$ as particular genotype, haplotype and haplotype frequency sets respectively. An initial p set is selected from from a Dirichlet distribution with parameter set $\beta = (\beta_1, ..., \beta_M)$ where M is the number of possible haplotypes. To form the algorithm, a pair of allowed haplotypes for each individual from the distribution to form the likelihood function

$$L(H_i = (k,l) \mid p, G_i) = \frac{p_k p_l}{\sum'_{k'l'} p_{k'} p_{l'}} \quad (5)$$

is selected. The likelihood is then updated by picking a new set of haplotype frequencies from the Dirichlet distribution with the parameter set $\beta = (\beta_1 + N_{h_1}, ..., \beta_M + N_{h_M})$ where N is the number of occurrences of each haplotype from the sampling in equation 5.

Two novel techniques are also introduced: *Partition Ligation* refers to the process of segmenting the genotype under consideration into smaller parts, typically of size $K \leq 8$, resolving the appropriate haplotype segments and selecting the B most probable ones, with B an integer around 50. When all genotype segments are resolved, the various sets of B haplotypes are ligated in a pairwise process which results in the B most probable haplotypes each time until the entire set is exhausted. Yet another technique used is *Prior Annealing*, which refers to he progressive decrease of the β set for the Dirichlet distribution with each iteration. This is designed to reduce the possibility of local maxima in the results.

Niu et al. [16] performed several simulations and compared the performance of their method to that of the previous three, showing an edge for the Bayesian approach. In addition, their results cast doubts over some of the assumptions made by Stephens et al., in particular Hardy-Weinberg equilibrium dependence and genotyping order.

V. CONCLUSIONS

Several fast and accessible methods for statistical haplotype inference have yielded good results with small projected errors. However, models of recombination, population structures, resolution errors as well as to date unforeseen factors may undermine the efficiency and accuracy of the techniques. Additionally, the error tolerance for the methods to have any practical use in a field as sensitive as pharmacogenomics is not well established. A combination of statistical and inexpensive experimental techniques may be the ideal solution. Current public and private focus on haplotyping makes such a solution attainable within the near future.

[1] G.A.Thorisson and L.D.Stein, Nucleic Acids Res. 31 , 124 (2003)
[2] M.J.Daly et al., Nat Genet. 29, 229 (2001)
[3] J.D.Rioux et al., Nat Genet. 29, 223 (2001)
[4] The International HapMap Consortium. The International HapMap Project. Nature 426, 789 (2003)
[5] NIH press release , Feb.7, 2005.
[6] B.V.Halldorsson et al,. SNPs and Haplotype Inference S.Istrail et al. Eds., Springer-Verlag (2004)
[7] S.Michalatos-Beloin et al.,Nucleic Acids Res. 24, 4841 (1996)
[8] G.Pont-Kingdon et al., JMD 6, 264 (2004)
[9] G.Ruano et al., Proc.Natl.Acad.Sci.USA 87, 6296 (1990)
[10] R.D.Mitra, Proc.Natl.Acad.Sci.USA 100, 5296 (2003)
[11] J.A.Douglas et al., Nat Genet 26, 361 (2001)
[12] A.G.Clark, Mol.Biol.Evol. 7, 11 (1990)
[13] L.Excoffier and M.Slatkin, Mol.Biol.Evol. 12, 921 (1995)
[14] D.Fallin and N.J.Schork, Am.J.Hum.Genet. 67, 947 (2000)
[15] M.Stephens et al., Am.J.Hum.Gen. 68, 978 (2001)
[16] T.Niu et al., Am.J.Hum.Genet. 70, 157 (2002)
[17] L.Wang and Y.Xu, Bioinformatics 19, 1773 (2003)
[18] D.Gusfield, J.Comp.Biol. 8(3) (2001)
[19] The TSC data can be found at http://snp.cshl.org/. The site provides interactive map search tools and the options of selecting specific chromosomes, locations and actual variations. SNP data are also provided by the NCBI at http://www.ncbi.nlm.nih.gov/SNP/ and by the HGVbase at http://hgvbase.cgb.ki.se/
[20] PCR consists of 3 main steps: 1.Denaturation of the target DNA segment through heating. 2. Hybridization, the process in which primer molecules with sequences identical to parts of the target DNA bind on the now single-stranded target. 3. Copying through polymerase reaction with the primers.
[21] Markov chains are sets of random variables in which the future state depends only on the present one and not on past values. Gibbs sampling MCMC methods use iterative conditional sampling for the phase space.
[22] The software package PHASE was developed by the authors and may be obtained upon request
[23] The software program HAPLOTYPER was developed and used by Niu et al. for their analysis.