For a haplotype (where is a set of strings of length of ), let denote the allele of SNP at the *k*-th locus of . The Hamming distance between two haplotypes and is defined as

where if and otherwise. As the Hamming distance is length-dependent, we define the following as a length-independent distance between haplotype and :

.

Next, let and be haplotype pairs to be compared, where . We defined the distance between haplotype pairs and as

For unphased-diplotypes , let and be the *i*-th and the *j*-th candidate haplotype-diplotypes for and respectively. and are the numbers of haplotype-diplotype candidates for and respectively. If we were given a population model , we can compute probability that haplotype-diplotype candidate is correct for the haplotype-diplotype data . Let and be the conditional probabilities of the candidate haplotype-diplotype and under the population model , respectively. We considered hidden Markov model (HMM) defined in HIT algorithm [1] as the population model. Then we defined HIT HMM-based distance (HHD) between two haplotype-diplotypes and is defined as follows:

where and representing the normalized predicted frequencies of the candidate haplotype-diplotypes and , respectively.

[1] Rastas P, Koivisto M, Mannila H, Ukkonen E. A Hidden Markov Technique for Haplotype Reconstruction. Lect. Notes. Bioinform. 3692. 2005;140-151. doi: 10.1007/11557067\_12.