Available online at www.sciencedirect.com

ScienceDirect

Information Sciences 177 (2007) 804–814

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

New methods for imputation of missing genotype using linkage disequilibrium and haplotype information

Ho-Youl Jung [a,*], Yun-Ju Park [b], Young-Jin Kim [b], Jung-Sun Park [b], Kuchan Kimm [b], InSong Koh [b]

[a] *Bioinformatics Team, IT-BT Group, IT Convergence Technology Research Division, Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Republic of Korea*

[b] *Division of Epidemiology and Bioinformatics, National Genome Research Institute, National Institute of Health, 5 Nokbun-dong, Eunpyung-gu, Seoul 122–701, Republic of Korea*

Received 13 January 2005; received in revised form 30 March 2006; accepted 7 July 2006

Abstract

In this paper, we propose new missing imputation methods for the missing genotype data of single nucleotide polymorphism (SNP). The common objective of imputation methods is to minimize the loss of information caused by experimental missing elements. In general, imputation of missing genotype data has used a major allele method, but this approach is not far from the objective of the imputation – minimizing the loss of information. This method generally produces high error rates of missing value estimation, since the characteristics of the genotype data are not considered over the structure of given genotype data. In our methods, we use the linkage disequilibrium and haplotype information for the missing SNP genotype. As a result, we provide the results of the comparative evaluation of our methods and major allele imputation method according to the various randomized missing rates.

Keywords: Bioinformatics; SNP; Genotype; Haplotype; Missing imputation

1. Introduction

The most significant development to date in the molecular study of disease genetics has emerged from the availability of the human genome sequence. This data provided the template from which to generate extensive amounts of information on single nucleotide variants. Several important advantages emerge from the availability of such single nucleotide polymorphisms (SNPs). These are by far the commonest form of polymorphism within the genome. These variants will account for the vast majority of polymorphism responsible for human disease. The variation occurs in both coding and non-coding sequences at a frequency of approximately 1 per 1000 base pairs [2].

[*] Corresponding author. Tel.: +82 42 860 1502; fax: +82 42 860 1208.
E-mail address: hoyoul.jung@etri.re.kr (H.-Y. Jung).

0020-0255/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2006.07.017

SNPs are of interest for a variety of reasons. First, a SNP, particularly when found in a functional gene region, may itself encode differences in protein form and expression, which in turn lead to disease and other, often subtler, phenotypic differences. Second, SNPs may mark or track the presence of other, perhaps less easily detected and processed, genetic differences that cause phenotypes of interest. Last, they are useful in studying mutation rates and evolutionary history [13].

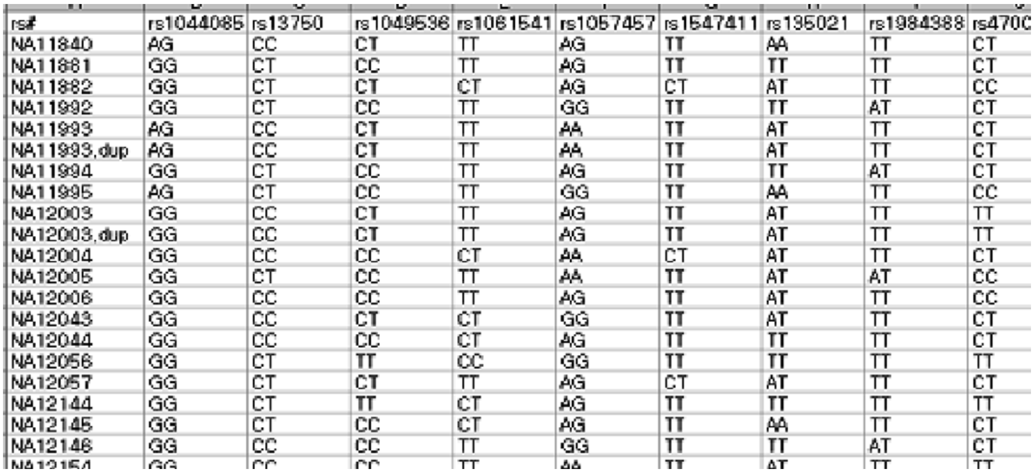
Almost all of experimental data contain missing elements. When the data containing missing elements are analyzed, we may eliminate the elements which have missing values, or estimate them using other given information. The former is called as ''filtering'' and the latter is called as ''imputation''. This kind of missing values in experimental data is a common phenomenon in not only medicine and health which manipulate clinical data but also molecular biology [10]. Genotype data for SNPs which are kind of biological or clinical experimental data have also many experimental missing values. This hinders that we cannot rely on the result of further SNP analysis, like as disease association.

Table 1 shows the missing rate for the genotype data of the international HapMap project from chromosome I to XXII. Fig. 1 shows an example of the genotype data of the international HapMap Project. First column represents the sample ID, and first row denotes SNP site ID [15]. In Fig. 1, allele pair like as GG or AA is called homozygote and allele pair like as AG is called heterozygote.

Table 2 shows the missing rate in duplicate experiments. Surprisingly the missing rate in duplicate is 23.11%, which is very high probability even though these are duplicate experiments. This means that the

Table 1
The missing rate for genotype data from the international HapMap project (only chromosome from I to XXII)

Number of genotypes	33,151,320
Number of missing genotype	382,428
Missing rate (%)	1.15



rs#	rs1044085	rs13750	rs1049536	rs1061541	rs1057457	rs1547411	rs195021	rs1984388	rs470C
NA11840	AG	CC	CT	TT	AG	TT	AA	TT	CT
NA11881	GG	CT	CC	TT	AG	TT	TT	TT	CT
NA11882	GG	CT	CT	CT	AG	CT	AT	TT	CC
NA11992	GG	CT	CC	TT	GG	TT	TT	AT	CT
NA11993	AG	CC	CT	TT	AA	TT	AT	TT	CT
NA11993,dup	AG	CC	CT	TT	AA	TT	AT	TT	CT
NA11994	GG	CT	CC	TT	AG	TT	TT	AT	CT
NA11995	AG	CT	CC	TT	GG	TT	AA	TT	CC
NA12003	GG	CC	CT	TT	AG	TT	AT	TT	TT
NA12003,dup	GG	CC	CT	TT	AG	TT	AT	TT	TT
NA12004	GG	CC	CC	CT	AA	CT	AT	TT	CT
NA12005	GG	CT	CC	TT	AA	TT	AT	AT	CC
NA12006	GG	CC	CC	TT	AG	TT	AT	TT	CC
NA12043	GG	CC	CT	CT	GG	TT	AT	TT	CT
NA12044	GG	CC	CC	CT	AG	TT	TT	TT	CT
NA12056	GG	CT	TT	CC	GG	TT	TT	TT	TT
NA12057	GG	CT	CT	TT	AG	CT	AT	TT	CT
NA12144	GG	CT	TT	CT	AG	TT	TT	TT	TT
NA12145	GG	CT	CC	CT	AG	TT	AA	TT	CT
NA12146	GG	CC	CC	TT	GG	TT	TT	AT	CT
NA12154	GG	CC	CC	TT	AA	TT	AT	TT	TT

Fig. 1. An example for the genotype data of the international HapMap Project [15].

Table 2
The missing rate for duplicate experiments from the international HapMap project (only chromosome from I to XXII)

Original experiment	Duplicate	Number of genotypes
Missing	Missing	7344
Missing	Success	24,434
Missing rate (%)		23.11
Success rate (%)		76.89