2016 年全国研究生数学建模竞赛 B 题

具有遗传性疾病和性状的遗传位点分析

人体的每条染色体携带一个 DNA 分子，人的遗传密码由人体中的 DNA 携带。DNA 是由分别带有 A,T,C,G 四种碱基的脱氧核苷酸链接组成的双螺旋长链分子。在这条双螺旋的长链中，共有约 30 亿个碱基对，而基因则是 DNA 长链中有遗传效应的一些片段。在组成 DNA 的数量浩瀚的碱基对（或对应的脱氧核苷酸）中，有一些特定位置的单个核苷酸经常发生变异引起 DNA 的多态性，我们称之为位点。染色体、基因和位点的结构关系见图 1.

在 DNA 长链中，位点个数约为碱基对个数的 1/1000。由于位点在 DNA 长链中出现频繁，多态性丰富，近年来成为人们研究 DNA 遗传信息的重要载体，被称为人类研究遗传学的第三类遗传标记。

大量研究表明，人体的许多表型性状差异以及对药物和疾病的易感性等都可能与某些位点相关联，或和包含有多个位点的基因相关联。因此，定位与性状或疾病相关联的位点在染色体或基因中的位置，能帮助研究人员了解性状和一些疾病的遗传机理，也能使人们对致病位点加以干预，防止一些遗传病的发生。

近年来，研究人员大都采用全基因组的方法来确定致病位点或致病基因，具体做法是：招募大量志愿者（样本），包括具有某种遗传病的人和健康的人，通常用 1 表示病人，0 表示健康者。 对每个样本，采用碱基(A,T,C,G)的编码方式来获取每个位点的信息(因为染色体具有双螺旋结构，所以用两个碱基的组合表示一个位点的信息)；如表 1 中，在位点 rs100015 位置，不同样本的编码都是 T 和 C 的组合，有三种不同编码方式 *TT*,*TC* 和 *CC*。类似地其他的位点虽然碱基的组合不同，但也只有三种不同编码。研究人员可以通过对样本的健康状况和位点编码的对比分析来确定致病位点，从而发现遗传病或性状的遗传机理。

表 1. 在对每个样本采集完全基因组信息后，一般有以下的数据信息
(以 6 个样本为例，其中 3 个病人，3 个健康者)：

样本编号	样本健康状况	染色体片段位点名称和位点等位基因信息			
		rs100015	rs56341	...	rs21132
1	1	TT	CA	...	GT
2	0	TT	CC	...	GG
3	1	TC	CC	...	GG
4	1	TC	CA	...	GG
5	0	CC	CC	...	GG
6	0	TT	CC	...	GG

注：位点名称通常以 rs 开头。

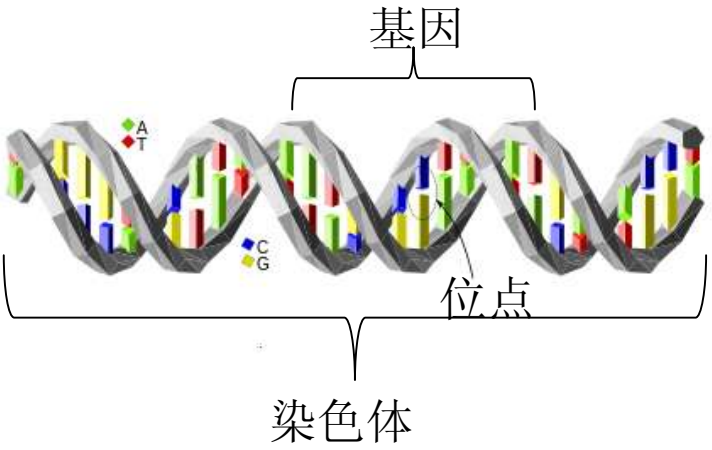

图 1. 染色体、基因和位点的结构关系.

本题目针对某种遗传疾病(简称疾病 A)提供 1000 个样本的信息，这些信息包括这 1000 个样本的疾病信息、样本的 9445 个位点编码信息，以及包含这些位点的基因信息。这些信息包含在附录中的 2 个文件(phenotype.txt ，genotype.dat)和 1 个文件夹 gene_info(包含 300 个文件)中。

phenotype.txt 文件中包含了样本具有遗传疾病 A 的信息，即一列 0 和 1 组成的数据，其中共有 500 个 0，500 个 1，表示我们现在共有 1000 个样本，其中 500 个 0 就是 500 个没患有疾病 A 的人，500 个 1 就是有 500 个患有遗传病 A 的人。如同表一中的第二列。

genotype.dat 文件中包含了上述 1000 个样本在某条染色体片段上所有的位点信息。该文件总共有 1001 行，9445 列。如同上表 1 中第三列到第六列的编码信息。具体来说，第一行表示 9445 个位点的名称，都是以字母 rs 开头的；接下来，有 1000 行，每一行表示一个样本在该条染色体片段上所有位点(9445 个位点)的编码信息。例如，该文件中第 2 行，就表示 1 号样本在该条染色体片段上 9445 个位点的编码信息。

文件夹 gene_info 中包含了 300 个 dat 文件，表示 300 个基因的信息；每个 dat 文件中包含了若干个位点的名称，表示该基因包含的位点信息，事实上，可以把基因理解为若干个位点组成的集合。注意到在 genotype.dat 文件中已包含所有位点的编码信息，所以我们可以得到每一个基因所包含位点的编码信息。例如 gene_1.dat，表示基因 gene_1 包含了 rs3094315, rs3131972,..., rs4040617，共 7 个位点。

另外，人体的许多遗传疾病和性状是有关联的，如高血压，心脏病、脂肪肝和酒精依赖等。科研人员往往把相关的性状或疾病放在一起研究，这样能提高发现致病位点或基因的能力；附录中的 multi_phenos.txt 文件中提供了上述 1000 个样本的 10 种相关性状的信息。文件中的每一列表示一个性状，每一行对应一个样本。文件中的 0 和 1 信息同 phenotype.txt 文件。

所有这些文件都可以利用 Notepad++软件打开。装好 notepad++后，当需要打开某个数据文件时，先点击该文件，然后点击右键，屏幕出现菜单，其中一栏是“edit with notepad++”,点击这一栏即可。许多软件也可以将文件中的数据直接读入内存。（如 matlab 可用 importdata 函数读入）

本题包含以下问题：
问题一、请用适当的方法，把 genotype.dat 中每个位点的碱基(A,T,C,G) 编码方式转化成数值编码方式，便于进行数据分析。

问题二、根据附录中 1000 个样本在某条有可能致病的染色体片段上的 9445 个位点的编码信息(见 genotype.dat)和样本患有遗传疾病 A 的信息（见 phenotype.txt 文件）。设计或采用一个方法，找出某种疾病最有可能的一个或几个致病位点，并给出相关的理论依据。

问题三、同上题中的样本患有遗传疾病 A 的信息（phenotype.txt 文件）。现有 300 个基因，每个基因所包含的位点名称见文件夹 gene_info 中的 300 个 dat 文件，每个 dat 文件列出了对应基因所包含的位点(位点信息见文件 genotype.dat)。由于可以把基因理解为若干个位点组成的集合，遗传疾病与基因的关联性可以由基因中包含的位点的全集或其子集合表现出来请找出与疾病最有可能相关的一个或几个基因，并说明理由。

问题四、在问题二中，已知 9445 个位点，其编码信息见 genotype.dat 文件。在实际的研究中，科研人员往往把相关的性状或疾病看成一个整体，然后来探寻与它们相关的位点或基因。试根据 multi_phenos.txt 文件给出的 1000 个样本的 10 个相关联性状的信息及其 9445 个位点的编码信息(见 genotype.dat)，找出与 multi_phenos.txt 中 10 个性状有关联的位点。

对你得到的结果都应该进行适当的统计分析和检验，从而从理论上说明你所发现的致病位点和基因的合理性。

关键词：遗传统计学， 全基因组关联性分析(GWAS)，位点(SNPs)