

家系数据紧密连锁位点的单体型频率估计^{*}

李彩霞¹, 黎培兴², 方积乾¹

(1. 中山大学医学统计与流行病学系, 广东 广州 510080;
2. 中山大学数学系, 广东 广州 510275)

摘 要: 在对单体型进行重构或进行连锁不平衡分析时, 往往首先要估计单体型频率。针对家系数据, 提出了紧密连锁位点的单体型频率估计方法。利用子代信息减少亲代单体型的不确定性, 构建家系数据的似然函数, 把家系中的个体潜在的单体型看成缺失数据, 采用 EM 迭代算法, 给出了家系数据单体型频率的极大似然估计。其结果表明, 家系数据紧密连锁位点的单体型频率估计可通过简单的递归迭代进行, 后一步的单体型的频率估计即为前一步的加权单体型频率。

关键词: 单体型; 似然函数; EM 算法; 极大似然估计

中图分类号: Q-332 **文献标识码:** A **文章编号:** 0529-6579 (2005) 03-0009-04

所谓单体型 (haplotype), 或称单倍型, 就是一条染色体上的 2 个或者 2 个以上的多态位点组合。我们可以通过分析某些单体型多态性和某种复杂性疾病是否存在关联, 如病例与对照的单体型多态性分布比较^[1]。由于实验数据通常只能提供各个位点的基因型, 而不能直接得到单体型数据, 单体型通常需要依赖基因型的组合情况进行推测与重构。个体的各个位点的基因型虽然已知, 但是当只要有两个位点是杂合的, 其两条单体型存在不确定性, 即观测到的多位点基因型可能有好几种单体型对与之匹配。

1 家系数据单体型的确定

由于高密度的 SNP 位点以及单体型块

(haplotype block) 的存在, 用紧密连锁 (没有重组) 位点进行连锁分析或连锁不平衡分析方法能更精确定位疾病基因。对于家系中的子代, 亲代的单体型对中一条传递, 一条未传递, 根据亲代向子代单体型传递过程, 家系数据能为紧密连锁位点的单体型的确定提供信息, 减少其不确定性, 从而家系数据理论上更能准确估计单体型频率估计与重构单体型。

例如, 考虑表 1 中的核心家系 1, 由上面可知双杂合子父亲的单体型存在不确定性, 但是根据亲代向子代单体型传递过程, 此家系潜在的单体型对可以惟一确定, 但是表 1 中的核心家系 2, 由观测的基因型推测潜在的单体型对, 存在 2 种可能性, 仍然不能惟一确定。

表 1 连锁相能确定与不能确定的核心家系
Tab 1 Nuclear families with certain phase and uncertain phase

家系编号	家系中个体	位点 1	位点 2	可能的单体型对 1	可能的单体型对 2
1	父	$A_1 A_2$	$B_1 B_2$	$(A_1 B_2, A_2 B_1)$	—
1	母	$A_1 A_1$	$B_1 B_2$	$(A_1 B_1, A_1 B_2)$	—
1	子代	$A_1 A_2$	$B_1 B_1$	$(A_1 B_1, A_2 B_1)$	—
2	父	$A_1 A_2$	$B_1 B_2$	$(A_1 B_1, A_2 B_2)$	$(A_2 B_1, A_1 B_2)$
2	母	$A_1 A_1$	$B_1 B_2$	$(A_1 B_1, A_1 B_2)$	$(A_1 B_1, A_1 B_2)$
2	子代	$A_1 A_2$	$B_1 B_2$	$(A_1 B_1, A_2 B_2)$	$(A_1 B_2, A_2 B_1)$

收稿日期: 2004-11-18
基金项目: 国家自然科学基金资助项目 (30170523)
作者简介: 李彩霞 (1970 年生), 女, 讲师, 博士生; E-mail: lpx@gzsums.edu.cn
(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 核心家系数据单体型频率估计

关于群体数据的单体型频率估计与重构最常用的便是 EM (expectation - maximization) 算法^[2-4], 另外还有基于 Bayes 方法的 MCMC 算法^[5-9]。

当位点之间紧密连锁时, 常把这些位点看成一个整体位点, 而由此多个位点产生的各种单体型便相当于这个整体位点的等位基因, 基于这点, 针对紧密连锁位点的单体型, 相应已有很多统计分析方法, 如已被广泛应用的核家系数据的单体型传递不平衡检验^[7,8], 但是由于单体型的不确定性, 常常需要通过估计单体型频率从而重构个体的单体型。而对家系数据单体型频率估计, 往往把家系中的父母看成来自一般群体中独立的个体, 即把家系数据当成群体数据, 没有利用子代信息^[8]。下面考虑家系数据的信息, 估计家系数据的单体型频率。

设有 l 个紧密连锁的位点 (两两位点之间的重组可忽略不计), 其等位基因数目分别为 $a_1, a_2, \dots, a_l$, 则各位点可能基因型数目分别 $\frac{1}{2}a_i(a_i + 1), i=1, 2, \dots, l$, 这些位点所形成的可能单体型个数为 $M=\prod_{i=1}^l a_i$, 将所有单体型从 1 到 M 编号, 设单体型频率 $F=(F_1, F_2, \dots, F_M)$ 。为了讨论方便, 下面不妨只考虑单个子代的情形。

设有 n 个核心家系, $G_i=(G_i^f, G_i^m, G_i^c)$ 表示家系 i 中父亲, 母亲与小孩的基因型。 $G_i=(G_i^f, G_i^m, G_i^c)$ 表示家系 i 中父母, 母亲与小孩的单体型对。令 $G=(G_1, G_2, \dots, G_n)$ 表示 n 个家系的 (观测的) 基因型, $G=(G_1, G_2, \dots, G_n)$ 表示相应 (未知的) 单体型对。

假定 $2n$ 个亲代是来自某群体的一份随机样本, 此群体满足 Hardy-Weinberg 平衡与随机婚配, 由于小孩能为亲代的单体型确定提供信息, 下面利用家系的基因型数据以及子代的信息重构亲代的单体型。对于家系中的子代, 亲代的单体型对中一条传递, 一条未传递, 从而可设 $G_i^f=(h_{1i}, h_{2i}), G_i^m=(h_{3i}, h_{4i}) G_i^c=(h_{1i}, h_{3i}), (i=1, 2, \dots, n)$ 。根据 Hardy-Weinberg 平衡与随机婚配条件,

$$P(G_i^f, G_i^m, G_i^c | F) = \sum_{h'_{1i}h'_{2i}h'_{3i}h'_{4i} \in H_i} F_{h'_{1i}}F_{h'_{2i}}F_{h'_{3i}}F_{h'_{4i}}$$
 这里, H_i 为所有满足如下条件的 $\{h'_{1i}h'_{2i}h'_{3i}h'_{4i}\}$ 的集合: 单体型对 $\{h'_{1i}h'_{2i}\}$ 与父亲的观测基因型 G_i^f 匹配, 单体型对 $\{h'_{3i}h'_{4i}\}$ 与母亲的观测基因型 G_i^m 匹配, 单体型对 $\{h'_{1i}h'_{3i}\}$ 与小孩的观测基因型 G_i^c 匹配。则对数似然函数

$$\ln L(F | G) = \sum_{i=1}^n \ln P(G_i^f, G_i^m, G_i^c | F) = \sum_{i=1}^n \ln \left\{ \sum_{h'_{1i}h'_{2i}h'_{3i}h'_{4i} \in H_i} F_{h'_{1i}}F_{h'_{2i}}F_{h'_{3i}}F_{h'_{4i}} \right\}$$

下面考虑把 G 或各 $(h_{1i}, h_{2i}, h_{3i}, h_{4i})$ 看成缺失值, 采用 EM 算法。添加 G 后的完全数据的对数似然函数为

$$\ln L(F | G, G) = \sum_{i=1}^n \ln P(G_i^f, G_i^m, G_i^c, G_i | F) = \sum_{i=1}^n \ln (F_{h_{1i}}F_{h_{2i}}F_{h_{3i}}F_{h_{4i}}) = \sum_{r,s,u,v=1}^M \sum_{i=1}^n \hat{q}(r,s,u,v) \ln (F_rF_sF_uF_v)$$

这里

$$\hat{q}(r,s,u,v) = \begin{cases} 1, & \text{若 } (h_{1i}, h_{2i}, h_{3i}, h_{4i}) = (r,s,u,v) \\ 0, & \text{否则} \end{cases}$$

E 步: 计算完全数据的对数似然函数的期望。

给出 F 的初始值 $F^{(0)}$, $F^{(t)}$ 为 t 步后的 F 的迭代值。虽然缺失值的取值未知, 但我们可以利用它的分布计算对数似然函数的期望。由于

$$P(G | G, F^{(t)}) = \prod_{i=1}^n \frac{P(G_i | G_i, F^{(t)})}{P(G_i | F^{(t)})} = \prod_{i=1}^n \frac{F_{h_{1i}}F_{h_{2i}}F_{h_{3i}}F_{h_{4i}}}{\sum_{h'_{1i}h'_{2i}h'_{3i}h'_{4i} \in H_i} F_{h'_{1i}}F_{h'_{2i}}F_{h'_{3i}}F_{h'_{4i}}}$$

从而当给定 $F=F^{(t)}$ 时, 可以求出分布 $P(G | G, F^{(t)})$ 的具体取值, 利用此分布, 求对数似然函数的期望。令

$$P_i^{(t)}(r,s,u,v) = \begin{cases} \frac{F_r^{(t)}F_s^{(t)}F_u^{(t)}F_v^{(t)}}{\sum_{h'_{1i}h'_{2i}h'_{3i}h'_{4i} \in H_i} F_{h'_{1i}}F_{h'_{2i}}F_{h'_{3i}}F_{h'_{4i}}}, & \text{若 } (r,s,u,v) \in H_i \\ 0, & \text{否则} \end{cases}$$

则 $P_i^{(t)}(r,s,u,v)$ 是 $\{1, 2, \dots, M\}$ 或 H_i 上的概率分布, 有 $\sum_{r,s,u,v \in H_i} P_i^{(t)}(r,s,u,v) = 1$ 。

完全数据的对数似然函数的期望

$$Q(F | G, F^{(t)}) = E[\ln L(F | G, G, F^{(t)})] = \sum_G \ln L(F | G, G, P(G | G, F^{(t)})) = \sum_{h_{11}h_{21}h_{31}h_{41} \in H_1} \dots \sum_{h_{n1}h_{n2}h_{n3}h_{n4} \in H_n} \left\{ \left[\sum_{r,s,u,v=1}^M \sum_{j=1}^n \hat{q}(r,s,u,v) \ln (F_rF_sF_uF_v) \right] \cdot \prod_{i=1}^n P_i^{(t)}(h_{1i}, h_{2i}, h_{3i}, h_{4i}) \right\} = \sum_{r,s,u,v=1}^M \ln (F_rF_sF_uF_v) \cdot$$

$$\left\{ \sum_{j=1}^n \left[\left(\sum_{(h_{1j}, h_{2j}, h_{3j}, h_{4j}) \in H_j} \hat{g}(r, s, u, v) p_j^{(t)}(h_{1j}, h_{2j}, h_{3j}, h_{4j}) \right) \right] \right\}$$

注意到 $\sum_{(h_{1i}, h_{2i}, h_{3i}, h_{4i}) \in H_i} P_i^{(t)}(h_{1i}, h_{2i}, h_{3i}, h_{4i}) = 1$, 上式可以简化为

$$Q(F \mid G, F^{(t)}) = \sum_{r, s, u, v=1}^M \ln(F_r F_s F_u F_v) \left\{ \sum_{j=1}^n P_j^{(t)}(r, s, u, v) \right\}$$

M 步: 求参数集 F 的估计 F , 使得 $Q(F \mid G, F^{(t)})$ 达到最大, 然后令 $F^{(t+1)} = F$ 。

在 $\sum_{i=1}^M F_i = 1$ 的条件下, $F_k^{(t+1)}$ 应满足方程组

$$\begin{cases} \frac{\partial}{\partial F_k} \left\{ Q(F \mid G, F^{(t)}) - \lambda \left(\sum_{i=1}^M F_i - 1 \right) \right\} = 0 \\ (k = 1, 2, \dots, M) \\ \sum_{k=1}^M F_k^{(t+1)} = 1 \end{cases}$$

其中 λ 为待定参数, 由于

$$\frac{\partial Q(F \mid G, F^{(t)})}{\partial F_k} =$$

$$\frac{1}{F_k} \left[\sum_{i=1}^n \sum_{r, s, u, v=1}^M n_k(r, s, u, v) p_i^{(t)}(r, s, u, v) \right],$$

这里 $n_k(r, s, u, v)$ 为 (r, s, u, v) 中含 k 的个数, 例如, 当 $(r, s, u, v) = (k, k, k, v)$ 时, $n_k(r, s, u, v) = 3$ 。可解得

$$F_k^{(t+1)} = \frac{\sum_{i=1}^n \sum_{r, s, u, v=1}^M n_k(r, s, u, v) p_i^{(t)}(r, s, u, v)}{4n},$$

$k = 1, 2, \dots, M$

实际上, $\sum_{i=1}^n \sum_{r, s, u, v=1}^M n_k(r, s, u, v) p_i^{(t)}(r, s, u, v)$ 为第 t 步迭代后, $2n$ 个亲代中具有单体型 k 的加权频数, 故上式可以理解为, 第 $(t+1)$ 步的单体型 k 的频率为 $2n$ 个亲代 (共 $4n$ 条单体型) 中具有单体型 k 的加权频率, 此权重依赖第 (t) 步的各单体型频率。

3 讨论

本文针对家系数据, 提供了单体型频率估计的 EM 算法的迭代过程。上面针对的家系虽然是只有

一个子代的情形, 但事实上, 对于家系中有多个小孩情形, 只要将每个 H_i 定义为满足如下条件的 $\{h_{1i}, h_{2i}, h_{3i}, h_{4i}\}$ 集合: 单体型对 $\{h_{1i}, h_{2i}\}$ 与父亲的观测基因型 G_i^f 匹配, 单体型对 $\{h_{3i}, h_{4i}\}$ 与母亲的观测基因型 G_i^m 匹配, 单体型对 $\{h_{1i}, h_{3i}\}$ 与某一指定小孩 (如第一个小孩) 的观测基因型 G_i^c 匹配, 其他小孩的基因型与单体型对 $\{h_{1i}, h_{3i}\}$ 或 $\{h_{1i}, h_{4i}\}$ 或 $\{h_{2i}, h_{3i}\}$ 或 $\{h_{2i}, h_{4i}\}$ 其中任何一个匹配。

另外, 上面的计算过程仍然适合家系中个体的基因型有缺失的情形, 只是家系的 H_i 的满足条件更宽松。特别地, 当家系中无小孩的基因型, 即无小孩信息时, 此方法即等价于把父母看成相互独立的个体, 在无其他信息条件下的单体型估计。

参考文献:

[1] FALLIN D, COHEN A, ESSIUX L, et al. Genetic analysis of case control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease [J]. Genome Res 2001, 11: 143—151.

[2] EXCOFFIER L, SLATKIN M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population [J]. Mol Biol Evol 1995, 12: 921—927.

[3] LONG J C, WILLIAMS R C, USBANEK M. An EM algorithm and testing strategy for multiple-locus haplotypes [J]. Am J Hum Genet 1995, 56: 799—810.

[4] HAWLEY M E, KIDD K K. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes [J]. J Hered 1995, 86: 409—411.

[5] STEPHENS M, SMITH N J, Donnelly P. A new statistical method for haplotype reconstruction from population data [J]. Am J Hum Genet. 2001, 68: 978—989

[6] NIU T, QIN Z S, XU X, et al. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms [J]. Am J Hum Genet, 2002, 70: 157—169.

[7] CLAYTON D G. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission [J]. Am J Hum Genet 1999, 65: 1170—1177.

[8] ZHAO H, ZHANG S, MERIKANGAS K R, et al. Transmission/disequilibrium test using multiple tightly linked markers [J]. Am J Hum Genet 2000, 67: 936—946.

(下转第 24 页)

[9] PIEGL L, TILLER W. The NURBS Book[M] . NewYork :
Springer, 1995.

[10] ERIC J S, TONY D D DAVID H S. Wavelets for Computer
Graphics; A Primer, Part 1[J] . IEEE Computer Graphics
and Applications, 1995, 15(3): 76— 84.

[11] ERIC J S, TONY D D DAVID H S. Wavelets for Computer
Graphics; A Primer, Part 2[J] . IEEE Computer Graphics
and Applications, 1995, 15(4): 75— 84.

[12] 孙延奎, 朱心雄. B 样条曲线的小波光顺法[J] . 工程
图学学报, 1998 (4): 51— 58.

[13] 刘 建, 关右江, 秦开怀. 任意 NUBS 曲线的小波分析
和造型技术[J] . 中国图像图形学报, 2002, 7(9): 894
— 900.

Multi-scale Modeling Based B-Spline Wavelet on DEM Topography Surface

TANG Xin jian^{1, 2}, ZHU Tong lin^{2, 3}, LIU Jian³, WANG Chuan ying¹

(1. Institute of Rock and Soil Mechanics, Chinese Academy of Sciences, Wuhan 430071, China;
2. Institute of Patters Recognition & Artificial Intelligence,
Huazhong University of Science and Technology, Wuhan 430074, China;
3. Information College, South China Agricultural University, Guangzhou 510642, China)

Abstract: Multi-resolution real time modeling and rambling on topography surface has an important application in GIS, but it is a difficult problem in algorithm design. The modeling features of B Spline curve and surface fitting are studied, and B Spline wavelet also. This paper realizes automated multi-resolution modeling on DEM (Digital Elevation Model) topography surface by combining non-uniform B Spline surface fitting with quasi-uniform B Spline” wavelet multi-scale effectively. The proposed approach in the article is applied to the slope DEM simulation of one constructing expressway located in mountains. The resulting simulation shows that the modeling is realistic and resolution is visible.

Key words: multi-resolution modeling on DEM; topography surface construction; non-uniform B-spline; wavelet

(上接第 11 页)

Haplotype Frequency Estimation for Tightly Linked Loci from Family Data

LI Cai xia¹, LI Pei xing², FANG Ji qian¹

(1. Department of Medicine Statistics, Sun Yat Sen University, Guangzhou, 510080, China
2. Department of Mathematics, Sun Yat Sen University, Guangzhou 510275, China)

Abstract: Haplotype frequencies must be estimated first when haplotypes are reconstructed or linkage disequilibrium is tested. From nuclear family data, Haplotype frequency estimation for tightly linked loci is given. The information supported by offspring is used to reduce the uncertainty of parents’ haplotype. The likelihood function is given for family data. The underlying haplotypes are regarded as missing data, and then the MLE (Maximum Likelihood Estimate) of the haplotype frequencies for the family data via an EM algorithm is given. The results show that haplotype frequency for tightly linked loci from family data could be estimated by a simple recursion. The successive haplotype frequency estimation is the weighted haplotype frequency of the preceding step.

Key words: haplotype; likelihood function; EM algorithm; maximum likelihood estimation