枚举单体型组装问题多个最优解的遗传算法设计

谢民主[1],刘新求[2,3]
XIE Min-zhu[1],LIU Xin-qiu[2,3]

1.湖南师范大学 物理信息与科学学院,长沙 410081
2.湖南师范大学 数学与计算机科学学院,长沙 410081
3.湖南工程职业技术学院 基础科学系,长沙 410151
1.College of Physics and Information Science,Hunan Normal University,Changsha 410081,China
2.College of Mathematics and Computer Science,Hunan Normal University,Changsha 410081,China
3.Department of Foundation Science,Hunan Engineering Vocational Technical College,Changsha 410151,China
E-mail:xieminzhu@sina.com

XIE Min-zhu,LIU Xin-qiu.Design of Genetic Algorithm to enumerate multiple optimal solutions to haplotype assembly problem.Computer Engineering and Applications,2010,46(11):7-9.

Abstract: The haplotype assembly problem aims to reconstruct a pair of haplotype of an individual from its DNA sequencing fragment data.There are some heuristic algorithms and parameterized algorithms for the various computational optimal models.However,these algorithms work out with only one optimal solution,i.e.a pair of haplotypes.However,the optimal solution to a biological problem is usually not unique,or the real solution may be suboptimal.The paper proposes a new genetic algorithm to enumerate multiple optimal solutions to the haplotype assembly problem.Experimental results show that this algorithm is more accurate in haplotype reconstruction and provides the chance for the biologists to choose one from these multiple solutions based on some biological knowledge.

Key words: Single-Nucleotide Polymorphisms(SNPs);haplotype;heuristic algorithm;bioinformatics

摘 要:单体型组装问题就是根据个体基因组测序获得的DNA序列数据重构出该个体的一对单体型。目前单体型组装问题的各种优化计算模型已有相关的启发式算法和参数化精确算法,但是这些算法只能得出一个最优解,即一对单体型。可是生物问题的最优解往往不是唯一的,或者真实解可能只是接近最优的。该文设计了一个新的能枚举出最优的多个解的遗传算法。实验结果表明该算法具有较高的单体型重建精度,并为生物学家根据领域知识在算法获得的多个解的基础进一步选择提供了可能。

关键词:单核苷酸多态性;单体型;启发式算法;生物信息学
DOI:10.3778/j.issn.1002-8331.2010.11.003 **文章编号:**1002-8331(2010)11-0007-03 **文献标识码:**A **中图分类号:**TP301

1 引言

Levy等[1]的研究成果表明两个不同个人基因组的相似度为99.5%，而剩下0.5%的基因组差异有短的DNA序列的插入、删除、重复等,但主要体现在基因组某个位点上单个碱基的变化。在人类中,基因组存在于细胞核的24对染色体中。单核苷酸多态性(Single-Nucleotide Polymorphisms,SNPs)为在人群中1%上个体中出现的染色体某个位点上的碱基变异。在人类全基因组中有几百万个SNPs,SNPs已成为确定人类个体差异的重要分子标记。单体型(Haplotype)为一条染色体某个区域相关的SNP序列。单体型比单个SNP在复杂疾病的全基因组关联分析中能提供更多的信息[2],因而具有更加重要的应用价值。

由于技术的限制,当前直接使用生物实验测定个体单体型在时间和金钱上代价过大,因而计算技术是目前获取单体型的主要手段。确定单体型的计算技术主要分为两大类[3]:第一类是单体型推断(haplotype inference),第二类是单体型组装(haplotypes assembly)。该文研究单体型组装问题。

2 单体型组装问题

人类等双倍体生物的染色体是成对出现的,由于一对染色体很难被分开，因此当今DNA测序得到的片断数据都是来自一对染色体的。基于直接测定单体型的技术困难性和单体型在遗传分析上的重要性,Lancia等[4]首次提出利用个体联配的

基金项目:湖南省自然科学基金(the Natural Science Foundation of Hunan Province of China under Grant No.09JJ3116);中国博士后科学基金一等资助(the Postdoctoral Science Foundation of China Under Grant No.20090450189)。
作者简介:谢民主(1969-),男,博士,副教授,主要研究领域为生物信息学,计算机算法;刘新求(1971-),女,博士研究生,讲师,主要研究领域为图论及其应用。
收稿日期:2009-12-31 **修回日期:**2010-02-01

DNA 片断数据,根据某些计算优化准则重建出两条单体型,即单体型组装问题。单体型组装问题也叫个体单体型问题(single individual haplotyping)。

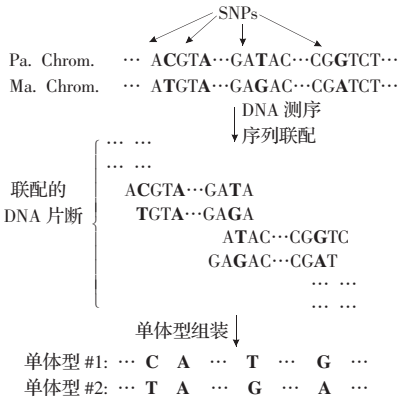

图 1　单体型组装示意图

注:Pa.Chrom:表示来自于父亲的染色体,Ma.Chrom 表示来自于母亲的染色体。

图 1 是一个单体型组装问题的示例。当 DNA 测序过程没有错误时,根据联配的 DNA 片断数据在对应的 SNP 位点上的取值,所有的 DNA 片段很容易划分为两个集合,使得每个集合中的片断在对应的 SNP 位点上具有相同的值。这样两个单体型就可以由每个集合中的片断在各 SNP 位点上的取值决定。如图中联配的 DNA 片断数据可以通过单体型组装获得一对单体型"…CA…T…G…"和"…TA…G…A…"。

可是由于测序过程中的错误是不可避免的,因此实验室测得的 DNA 片段数据常常不可能划分为两个集合,使得在同一个集合里的片段在对应的 SNP 位点上的取值均相等,这样单体型组装问题在计算上就变得很困难。国内外众多学者对单体型组装问题都有过深入研究,提出了众多的计算模型,其中主要的计算模型是最少错误更正模型。

最少错误更正(Minimum Error Correction,MEC):MEC 模型也叫做最少字符翻转(Minimum Letter Flips,MLF),该模型要求修改 DNA 片段上最少的 SNP 值,使得修改后的 DNA 片段可以分成两个子集,使得每个子集的片段能决定一个单体型。

对于 MEC 模型,Cilibrasi 等[5]证明了其是 NP-难和 APX-难的。Wang 等[6]曾设计了时间复杂度为 $O(2^m)$的分支限界算法,其中 m 为 DNA 片段数,由于该算法是指数时间复杂度的,只适合片段数不多的场合。对于片段数较多的情况,Wang 等[6]设计了遗传算法求其近似解。当片段覆盖的 SNP 最大位点数 k_1 和覆盖任意 SNP 位点的最大片段数 k_2 较小时,谢民主等[7]提出了时间复杂度为 $O(nk_2 2^{k_2}+m\log m+mk_1)$的算法,其中 n 为SNP 位点数。上述目前已有的算法只能得出一个最优解,即一对单体型。考虑到生物问题的最优解往往不是唯一的,加之由于生物现象本身的复杂性,即使在最优解只有一个的情况下,生物学家对一些接近最优的解也很感兴趣。快速的能提供最优多个解的算法,为生物学家根据领域知识做出进一步选择提供了可能,因而更能满足生物学家的需求。

3　枚举出单体型组装问题的多个解

3.1　k-最小距离模型

一对条染色体在一个 SNP 位点上的碱基值可以是相同,这叫纯合(homozygous);也可不同,这叫杂合(heterozygous)。这样单体型就可以用{0,1}上的字符序列来表示,不必用真正的碱基字符,其中'0'表示人群在该位点上常见的 SNP 值,'1'表示另一个。

在单体型组装问题中,DNA 片断中只有在 SNP 位点上的值才会被考虑。n 个 SNP 位点按在染色体上的次序从左到右记作$\{s_1,s_2,\cdots,s_n\}$,m 个片断记作$\{r_1,r_2,\cdots,r_m\}$。任意片断在某个 SNP 位点的取值为'0'、'1'或'-',其中'-'为空值,表示片断在该位点的取值未知。这样联配的 DNA 片断在各 SNP 位点上的取值就可以表示为在$\{0,1,-\}$上的一个 $m\times n$ 的矩阵,叫做 SNP 矩阵 $\boldsymbol{M}$。图 2 是一个 5×4 的 SNP 矩阵。SNP 矩阵的列表示 SNP 位点,行表示片段在对应的 SNP 位点上的取值。

片段(fragment) \ SNP				
	0	1	0	0
	0	-	0	-
	1	0	1	-
	-	0	1	0
	1	-	1	0

图 2　5×4 SNP 矩阵

两个来自于字母表$\{0,1,-\}$的字符 a 和 b 的距离定义为:

$$d(a,b)=\begin{cases}1, \text{if } a='0' \text{ and } b='1', \text{or } a='1' \text{ and } b='0'\\ 0, \text{otherwise}\end{cases}$$

给定一个'0'、'1'或'-'组成的长为 n 的 DNA 片断 r 和一对由字母'0','1'组成的长为 n 的单体型 $H=(h_1,h_2)$,r 和 H 的距离定义为:

$$d(r,H)=\min\left(\sum_{i=1}^{n} d(r[i],h_1[i]),\sum_{i=1}^{n} d(r[i],h_2[i])\right)$$

其中 $r[i]$和 $h[i]$分别表示 DNA 片断 r 和单体型 h 的第 i 个字符。

r 和 H 的距离表示如果 r 来自于 H 代表的一对染色体,r 这个片断中最少的错误 SNP 值的个数。片段 r 和 H 的距离为 0 时称 r 和 H 兼容。

给定一个 $m\times n$ 的 SNP 矩阵 $\boldsymbol{M}$ 和一对长为 n 的单体型 $H=(h_1,h_2)$,$\boldsymbol{M}$ 和 H 的距离定义为:

$$dist(\boldsymbol{M},H)=\sum_{i=1}^{m} d(r_i,H)$$

其中 r_i 为 $\boldsymbol{M}$ 的第 i 行。

$\boldsymbol{M}$ 和 H 的距离表示使得 $\boldsymbol{M}$ 中的任意片断和 H 兼容,必须修改的 SNP 值的最少个数。

基于上述定义,提出以下模型,使其能够为单体型组装问题提供最优的多个解。

k-最小距离模型(k-Minimum Distance):给定一个 $m\times n$ 的 SNP 矩阵 $\boldsymbol{M}$ 和一个正整数 k,在所有可能的长为 n 的单体型对中,找出与 $\boldsymbol{M}$ 距离最小的 k 对单体型。

k-最小距离模型是对最少错误更正模型的扩展,一个最少错误更正模型的实例就是一个 1-最小距离模型的一个实例,由此可以证明 k-最小距离模型是 NP-难和 APX-难的。

3.2　求解 k-最小距离模型

3.2.1　预处理

由于 DNA 测序的错误较小,对于一个纯合的 SNP 位点,绝大部分片断在该位点的非空值应该相同,因此对 SNP 矩阵进行如下预处理:对 SNP 矩阵的每一列 j 统计取值为'0'和'1'的行数,分别记作 N_0 和 N_1。对于事先给定的一个阈值 t(设定为

20%),如果 $N_0/(N_0+N_1)<t$(或$N_1/(N_0+N_1)<t$),则该列被认为是纯合的,该个体的一对单体型在该位点上取值 0(或 1)。去掉纯合的列和由此带来的空行。

3.2.2 遗传算法

经过预处理后,SNP 矩阵 $\boldsymbol{M}$ 所有的列被认为是杂合的,即所有的单体型对在经过预处理后剩下的列上取值均不相同。假设 $\boldsymbol{M}$ 的列数为 n,对应的一对杂合单体型可用一个长为 n 的在{0,1}上字符串表示。如'0101'表示一对杂合单体型'0101'和'1010'。这样整个解空间为 2^n,当 n 比较大时,穷尽搜索是不可行的,因此下面采用遗传算法。

同上所述,给定一个经过预处理后的 SNP 矩阵 $\boldsymbol{M}$,遗传算法中遗传个体采用长为 n 的在{0,1}上字符串,该字符串编码一对单体型 H。遗传个体的适应度函数为 $1-dist(\boldsymbol{M},H)/N$,其中 N 为 $\boldsymbol{M}$ 中非空值的个数。在生成下一代个体时,采用 Wang 等[6]的相同方法,具体算法如下:

k-GA 算法:

输入:经过预处理后的 $m\times n$ SNP 矩阵 $\boldsymbol{M}$,要求最优的单体型对数 k;遗传算法本身参数:群体中个体个数 s,交叉率 p_c,变异率 p_m,进化代数最大值 g。

输出 k 对单体型。

步骤 1 令 l=0,随机初始化群体 $P(0)=\{H_1,H_2,\cdots,H_s\}$,即 H_i 为长 n 的在{0,1}上的随机字符串,i=1,2,…,s。

利用适应度函数计算 $P(l)$中的每个个体 H_i 的适应度,记录其中适应度最大的 k 个个体在解集 S 中。

步骤 2 l=l+1,用下述方法创建 $P(l)$:

(1)使用锦标赛选择算子从群体 $P(l-1)$中选择$(1-p_c)s$ 个个体加入群体 $P(l)$中;

(2)使用轮盘赌选择算子从群体 $P(l-1)$中选择 $p_c s/2$ 对个体随机进行单点交叉或均匀交叉操作,把由此获得的新个体加入群体 $P(l)$中;

(3)从群体 $P(l)$中随机选择 $p_m s$ 个个体。对每个被选中的个体等概率进行如下操作:随机选择一个位点修改该位点的值,或者交换两个随机选择位点的值。

步骤 3 利用适应度函数计算 $P(l)$中每个个体 H_i 的适应度,如果其适应度大于 S 中个体适应度的最小值,则用 H_i 替换 S 中的具有该最小适应度的一个个体。

步骤 4 如果 $l>g$,则算法结束,根据适应度从大到小依次返回 S 中的 k 个个体对应的 k 对杂合单体型;否则重复步骤 2 和步骤 3。

显然,把预处理中去掉的纯合列的值插入到 k-MDGA 算法返回的 k 对杂合单体型中,即可获得预处理前 SNP 矩阵的 k-最小距离模型的近似解。称由预处理、遗传算法和后续处理组成的整个算法为 k-MD 算法。

4 实验结果

对 k-MD 算法求出的第一个解和来自于 Wang 等[6]的遗传算法(GAMEC)求出的一个解进行比较测试,并对 k-MD 算法求出的多个解进行分析。k-MD 算法用 C++语言实现。实验测试在一台 Linux 服务器(4 个 Intel Xeon 3.6 G CPU,4 G RAM)上进行。

测试数据生成方法同文献[7-8],单体型采用两种方式得到:第一种采用真实的单体型数据,实验采用来自于国际人类基因组单体型图计划真实单体型数据,随机选择一个体长度 n=100 的一对单体型。第二种用计算机模拟生成,即首先随机生成长度 n=100 的单体型,然后根据 20%的差异率来随机生成另一个单体型。与文献[10]一样,在得到一对单体型的基础上采用著名的 shotgun 测序模拟数据生成器 Celsim[11]根据指定的参数随机生成片段数据集。在实验中,测序误差设置为 5%,片段数据集包含两类片断:第一类片断由 3 到7 个连续的非空值构成;第二类片断由 7 个连续的非空值跟 10 个空值再跟 7 个连续的非空值组成。第一类片断数 m_1 由片断覆盖率 c_1 决定:$m_1=nc_1$/片断的平均长度;同样第二类片断数 m_2 由片断覆盖率 c_2 决定。模拟数据生成器的详细情况请参照文献[11]。

算法参数的设置:两个算法中群体的个体个数 s,交叉率 p_c,变异率 p_m,进化代数的最大值 g 均相同,分别为 400、0.8、0.2 和 1 500。对于 k-MD,k 设置为 10。

实验主要测试指标为单体型重建率 R、算法运行时间 T。单体型重建率为算法重建出的单体型对与真实单体型对具有相同值的 SNP 位点数与总的 SNP 位点数的比值。表 1 中的测试数据是相同参数下对算法重复测试 100 次的结果平均值,其中 R 为百分比,T 的单位为秒(s)。由于 k-MD 每次测试返回 10 对单体型,表 1 中 R_0 为其每次运行返回第一对单体型重建率的平均值,R_b 为其返回的 10 对单体型中最高单体型重建率的平均值。N_f 为 100 次重复测试中,k-MD 返回的第一对单体型在 10 对中具有最高单体型重建率的测试次数。

从表 1 的实验结果可以看出,当考虑其返回的第一对单体型,k-MD 比 GAMEC 在单体型重建精度上高出约 3%。这可能是因为这两个遗传算法的遗传个体编码方法不同,GAMEC 的遗传个体编码空间为 2^m,而 k-MD 遗传个体编码空间最大为 2^n。当覆盖度增大时,两个算法的单体型重建精度和运行时间均有所提高。对于 k-MD 而言,返回的 10 对单体型中,第一次返回的取得最高单体型重建率的概率约为 50%,返回的 10 对单体型中最好的单体型对比第一对在单体型重建率上高出约 1%。

表 1 GAMEC 和 k-MDGA 性能比较

		GAMEC			k-MD		
		R/(%)	T/s	R_0/(%)	T/s	N_f	R_b/(%)
真实的单	c_1=c_2=10	90.6	0.013 0	94.3	0.004 1	49	95.6
体型数据	c_1=c_2=20	92.4	0.022 0	95.8	0.007 5	46	96.8
模拟的单	c_1=c_2=10	89.8	0.009 9	93.1	0.003 7	54	94.2
体型数据	c_1=c_2=20	91.8	0.020 8	94.6	0.007 4	53	96.0

注:表中单体型重建率 R、R_0、R_b 和运行时间 T 为 100 次重复运行结果的平均值,c_1 和 c_2 分别是普通片段和 mate-pair 的覆盖率

5 结论与展望

同一物种不同个体基因组的多态性主要体现为 SNP,而同一染色体上相关的 SNP 序列构成单体型。单体型对复杂遗传疾病相关基因的定位、个性化治疗等相关应用有重要的作用。由于直接测定单体型的生物实验代价过分昂贵,在高通量、大规模的实验数据上利用计算技术重建单体型具有重要的现实意义。目前国内外学者已提出了各种优化模型和相应的计算机算法,但是已有的算法是以一个最优解(即一对单体型)为最终目的。可一个问题很可能不止一个最优解,且对于复杂的生物问题,正确解也可能不是最优的。基于上述考虑,提出了能提供最优多个解的 k-最小距离模型,并设计出了对应的近似算法。

(下转 17 页)