基于枚举策略的三倍体个体单体型重建算法

张 倩[1] 吴璟莉[1,2,3]

(广西师范大学计算机科学与信息工程学院 桂林 541004)[1]

(广西师范大学广西多源信息挖掘与安全重点实验室 桂林 541004)[2]

(广西区域多源信息集成与智能处理协同创新中心 桂林 541004)[3]

摘 要 求解三倍体个体单体型对于探索三倍体物种的遗传特性和表型差异等方面的研究具有重要的推动作用。针对带基因型信息的最少错误更正(MEC/GI)模型,提出了一种基于枚举策略的三倍体个体单体型重建算法EHTR。该算法依次重建3条单体型上的每一个单核苷酸多态性位点取值,对于给定位点,首先根据其基因型取值枚举该位点的3种单体型取值情况,然后选择片段支持度最高的取值作为该位点的重建值,算法的总时间复杂度为$O(mn+m\log m+cnl)$。采用CELSIM和MetaSim两种测序片段模拟生成器生成实验测试数据,在片段覆盖率、错误率、单片段长度、单体型长度和单体型海明距离等参数的不同设置下,对算法EHTR,GTIHR,W-GA和Q-PSO的重建率和运行时间进行对比分析。实验结果显示,算法EHTR在不同的参数设置下均能以更短的运行时间获得更高的重建率。

关键词 序列分析,三倍体,单体型,基因型,最少错误更正,枚举

中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.014

Triploid Individual Haplotype Reconstruction Algorithm Based on Enumeration Strategy

ZHANG Qian[1] WU Jing-li[1,2,3]

(College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China)[1]

(Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, China)[2]

(Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, Guilin 541004, China)[3]

Abstract Determining triploid individual haplotype plays an important role in promoting the study of exploring genetic traits and phenotypic differences of triploid species. In this paper, based on the minimum error correction with genotype information (MEC/GI) model, an enumeration strategy based enumeration haplotyping triploid (EHTR) algorithm was proposed for solving triploid individual haplotype reconstruction problem. The EHTR algorithm reconstructs the SNP sites of the three haplotypes one after another. When reconstructing a given SNP site, it enumerates three kinds of SNP values in terms of the genotype of the site, and chooses one with the most high support degree coming from the SNP fragments that are covering the corresponding SNP site. The total time complexity is $O(mn+m\log m+cnl)$. In the experiments, two kinds of simulators CELSIM and MetaSim were invoked to generate SNP fragments. The reconstruction rate and running time were compared and analyzed among algorithms EHTR, GTIHR, W-GA and Q-PSO with different parameters setting, such as fragment coverage, error rate, single fragment length, haplotype length and haplotype hamming distance. Under different parameter setting, the EHTR algorithm can obtain higher reconstruction rate in shorter running time, which is proved by a number of experiments.

Keywords Sequence analysis, Triploid, Haplotype, Genotype, Minimum error correction, Enumeration

1 引言

基因测序技术的飞速发展产生了大量的基因数据,使后基因组研究工作不仅在人类等二倍体生物上展开,还过渡到动植物等三倍体生物。在生物界中,三倍体普遍存在于鱼类、贝类和高等植物中,通常三倍体鱼、贝类的生长周期更短,成熟鱼寿命更长、存活率更高、抗病力更强,而三倍体植物也普遍具有生长旺盛、果实大且少籽或无籽、产量高、适应性和抗逆性强等优点。因此,研究三倍体生物基因组对于探索三倍体物种的遗传特性起到了重要的推动作用。

到稿日期:2015-08-24 返修日期:2015-10-26 本文受国家自然科学基金项目(61363035,61502111),广西自然科学基金项目(2015GXNSFAA139288,2013GXNSFBA019263,2012GXNSFAA053219),"八桂学者"工程专项,广西多源信息挖掘与安全重点实验室系统性研究基金项目(14-A-03-02,15-A-03-02)资助。

张 倩(1991-),女,硕士生,主要研究领域为生物信息学;吴璟莉(1978-),女,博士,教授,硕士生导师,CCF会员,主要研究领域为生物信息学、算法和复杂度,E-mail:wjlhappy@mailbox.gxnu.edu.cn。

单核苷酸多态性（Single Nucleotide Polymorphism，SNP）是基因组中最常见的一种遗传多态性，它是由染色体基因组在单个核苷酸碱基尺度上的变化而引起的 DNA 序列多态性。研究 SNP 对于探索三倍体生物的遗传性状和表型差异[1-3]以及植物分子育种[4,5]等发挥着着重要的作用。由于连锁不平衡现象以及缺乏遗传重组事件，染色体上某一区域的一组 SNP 位点趋于共同遗传，这组位于一条染色体上的相关 SNP 位点序列称为单体型（Haplotype）。研究表明，单体型数据比单个 SNP 位点携带更多的遗传信息，其在生物个体的表型差异、基因表达和疾病预测等方面发挥更大的作用。但在当前的实验技术水平下，获取完整单体型数据的成本很高，而获取基因型数据和 SNP 片段数据则较为容易，因此通常采用计算的方式，由测序片段数据推测单体型，由此产生了个体单体型重建问题，也称为个体单体型组装问题。

目前，对三倍体个体单体型重建问题的研究较少，有文献基于最少错误更正模型（the Minimum Error Correction，MEC）[6]、带基因型最少错误更正模型（the Minimum Error Correction with Genotype Information，MEC/GI）[6]及最少片段删除模型（the Minimum Fragment Removal，MFR）[7]提出求解 K-个体单体型重建问题的算法。Wang 等基于 MEC 和 MEC/GI 模型，提出重建二倍体单体型的遗传算法，并指出对其稍加修改后可用于重建 K-个体单体型[6]（本文称为 W-GA）。该算法的编码长度与片段数目相等，这使得算法 W-GA 的解空间庞大，制约了算法的求解性能[8]。Li 等[7]基于 MFR 模型，针对无空隙片段的问题实例，提出时间复杂度为 $O(m^2n+m^{K+1})$ 的多项式算法，其中 m 表示片段个数，n 表示 SNP 位点数；针对片段中洞的最大个数不超过 t 的问题实例，提出时间复杂度为 $O(2^{2t}m^2n+2^{(K+1)t}m^{(K+1)})$ 的动态规划算法，该算法仍然只适用于 m 值（片段数目）较小的情况，随着 m 值的增大，算法扩展性不好，性能受到很大的影响。Qian 等基于 MEC 和 MEC/GI 模型，提出求解二倍体重建问题的粒子群优化算法，并指出对其稍加修改后可用于求解 K-个体单体型重建问题[9]（本文称为 Q-PSO），该算法的的编码方式与算法 W-GA 相同，故性能仍受到片段数目的制约。吴璟莉等[10]通过引入新颖的染色体编码方法及有效的爬山算子，提出了求解 MEC 模型的三倍体个体单体型重建算法 GTIHR，该算法编码长度等于单体型中杂合位点数，当单体型长度增长或杂合率增大时，算法性能受到较大的影响。

本文针对 MEC/GI 模型研究了三倍体个体单体型重建问题，提出基于枚举策略的三倍体个体单体型重建算法 EHTR。对于每个待重建的 SNP 位点，算法 EHTR 根据其基因型枚举出 3 种取值情况，并选择片段支持度最高的取值作为该位点的取值，以此依次重建出 3 条单体型上的每个 SNP 位点。由于三倍体测序数据很难得到，本文采用模拟数据对算法 EHTR，W-GA，Q-PSO 和 GTIHR 进行实验测试。实验结果显示，在不同的参数设置下，算法 EHTR 均能有效地求解问题，与其它 3 种算法相比，算法 EHTR 能在更短的时间内获得更高的重建率，有较强的实用价值。

2 问题及符号定义

三倍体生物体细胞中含有 3 个染色体组，其上一段连续区域内相关的 SNP 序列即构成 3 条单体型序列。由于 SNP 位点的二态性，即每个 SNP 位点仅有两种不同的变异形式，可以采用定义在字符集{0,1}上的二进制序列表示单体型，而无需采用真实的{A,T,C,G} 4 个碱基字符。在 3 条染色体上共同表现出来的 SNP 复合序列构成了基因型序列，若 3 条染色体在某 SNP 位点的碱基取值相同，则这个位点上的基因型称为纯合子，否则称为杂合子[11]。例如，$(000)^T$ 和 $(111)^T$ 表示同合 SNP 位点的基因型取值，$(001)^T$ 和 $(011)^T$ 均表示杂合 SNP 位点的基因型取值。假设给定某号染色体上 3 条长度为 n 的单体型，经测序产生一组 m 条 SNP 测序片段，记为 SNP 矩阵 $\boldsymbol{M}_{m\times n}$，其中每行表示一条 SNP 片段，每列表示一个 SNP 位点，每个元素 $m_{ij}\in\{0,1,-\}(i=1,2,\cdots,m,j=1,2,\cdots,n)$（－表示片段 i 未覆盖第 j 个 SNP 位点，或者片段 i 在第 j 个位点的取值未知）。令 $\boldsymbol{g}=(\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_n)$ 表示对应的基因型矩阵，其中 $\boldsymbol{g}_j=(g_{j1},g_{j2},g_{j3})^T(g_{jk}\in\{0,1\},k=1,2,3,j=1,2,\cdots,n)$ 表示第 j 个 SNP 位点的基因型。

下面介绍本文采用的符号：

给定矩阵 $\boldsymbol{M}$ 的列 $\boldsymbol{M}_{\cdot j}(j=1,\cdots,n)$，令 $r(j)$ 表示覆盖第 j 个位点的片段集。给定矩阵 $\boldsymbol{M}$ 的行 $\boldsymbol{M}_{i-}(i=1,\cdots,m)$，令 $l(i)$ 表示第 i 行中元素 $m_{ij}\neq-$ $(j=1,2,\cdots,n)$ 的最小列号。

给定两条字符序列 $\boldsymbol{U}=\langle u_1,\cdots,u_n\rangle$ 和 $\boldsymbol{V}=\langle v_1,\cdots,v_n\rangle$，且 $u_j,v_j\in\{0,1,-\}(j=1,\cdots,n)$，海明距离 $HD(\boldsymbol{U},\boldsymbol{V})$ 定义为 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 中对应位取值不同的位数，如式(1)所示：

$$HD(\boldsymbol{U},\boldsymbol{V})=\sum_{j=1}^{n}d(u_j,v_j) \tag{1}$$

其中，

$$d(u_j,v_j)=\begin{cases}1, & \text{若 } u_j\neq-,v_j\neq-,\text{且 } u_j\neq v_j\\ 0, & \text{否则}\end{cases} \tag{2}$$

$HD(\boldsymbol{U},\boldsymbol{V})$ 可用于求解两条 SNP 片段之间的距离、SNP 片段和单体型之间的距离以及两条单体型之间的距离。当 $HD(\boldsymbol{U},\boldsymbol{V})$ 表示两条 SNP 片段之间的距离时，若 $HD(\boldsymbol{U},\boldsymbol{V})$ 大于 0，表示片段 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 冲突，否则表示它们相容。冲突的片段一般来自于不同的单体型或是含有测序错误。如果所有片段均没有测序错误或者错误已被更正，矩阵 $\boldsymbol{M}$ 中行可被划分成 3 个彼此不相交且均包含相容片段的子集，每个子集决定一条单体型，则称矩阵 $\boldsymbol{M}$ 可行[11]。

假设给定单体型 $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3)(\boldsymbol{h}_k=(h_{k1},h_{k2},\cdots,h_{kn}),k=1,2,3)$，基因型 $\boldsymbol{g}=(\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_n)(\boldsymbol{g}_j=(g_{j1},g_{j2},g_{j3}),j=1,2,\cdots,n)$，当 $\sum_{k=1}^{3}h_{kj}=\sum_{k=1}^{3}g_{jk}(j=1,2,\cdots,n)$ 时，称单体型 $\boldsymbol{h}$ 与基因型 $\boldsymbol{g}$ 相容。

2002 年，Lippert 等[12]提出重建二倍体单体型的最少错误更正模型（the Minimum Error Correction，MEC）。由于基因型数据较单体型数据更容易测定，Wang 等[6]引入带基因型信息的最少错误更正模型（the Minimum Error Correction with Genotype Information，MEC/GI），本文将 MEC/GI 模型延伸到三倍体个体单体型重建问题。

带基因型信息的最少错误更正模型：给定 SNP 矩阵 $\boldsymbol{M}$ 及基因型矩阵 $\boldsymbol{g}$，修改最少数目的矩阵元素（0 改为 1 或将 1 改为 0）以使得 SNP 矩阵可行，且重建的 3 条单体型与给定基因型相容。

本文用重建率[6,10,13](Reconstruction Rate，RR)衡量重建单体型的精度。令 $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3)$ 为原始单体型，$\boldsymbol{h}'=(\boldsymbol{h}_1',\boldsymbol{h}_2',\boldsymbol{h}_3')$ 为重建单体型，RR 定义为单体型 $\boldsymbol{h}'$ 中正确构建的核苷酸比例，如式(3)所示：

$$RR(h,h')=1-\frac{\min\{\sum_{k=1}^{3}r_{i_kj_k}\mid i_k,j_k\in\{1,2,3\},\sum_{k=1}^{3}i_k=\sum_{k=1}^{3}j_k=6\}}{3n} \tag{3}$$

其中，$r_{i_kj_k}=HD(h_{i_k},h'_{j_k})(i_k,j_k=1,2,3)$。

3 EHTR 算法

本节提出一种基于枚举策略重建三倍体个体单体型的启发式算法 EHTR (Enumeration Haplotyping Triploid)。算法的输入为片段矩阵 $\boldsymbol{M}$ 和基因型矩阵 $\boldsymbol{g}$，输出为单体型 $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3)$。EHTR 首先预处理基因型矩阵 $\boldsymbol{g}$ 和 SNP 矩阵 $\boldsymbol{M}$，删除基因型中的纯合子及矩阵中的纯合 SNP 位点；然后根据某个 SNP 位点的基因型枚举其 3 种可能取值，选择片段支持度最高的取值作为该 SNP 位点的取值，并以此循环生成 3 条只含杂合位点的单体型 $\boldsymbol{h}'=(\boldsymbol{h}_1',\boldsymbol{h}_2',\boldsymbol{h}_3')$；最后将单体型 $\boldsymbol{h}'$ 扩展成为 $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3)$。下面详细介绍算法 EHTR 的主要步骤。

3.1 预处理

由于纯合位点对单体型重建工作没有作用，首先删除基因型矩阵 $\boldsymbol{g}$ 和片段矩阵 $\boldsymbol{M}$ 中的纯合位点，以提高求解问题的效率。删除 $\boldsymbol{g}$ 中满足条件 $g_{j1}=g_{j2}=g_{j3}(j=1,2,\cdots,n)$ 的项，并将片段矩阵 $\boldsymbol{M}$ 中的对应列删除，记该列取值为 g_{j1}。经过列删除后，矩阵 $\boldsymbol{M}$ 中会产生某些空行(元素值全为－)，将这些行也删除。预处理后基因型全为杂合子，SNP 矩阵中保留的 SNP 位点也均为杂合位点。为描述方便，仍用 $\boldsymbol{M}_{m\times n}$ 和 $\boldsymbol{g}$ 表示新的片段矩阵和基因型矩阵。将片段矩阵的行 $\boldsymbol{M}[i,-](i=1,\cdots,m)$ 按照 $l(i)$ 值非降序排列，令 $r_1,\cdots,r_m$ 为排序后的行，则 $l(r_p)\leqslant l(r_q)(p<q,p,q=1,\cdots,m)$。对于给定列 $j(j=1,\cdots,n)$，计算覆盖该列的行集 $r(j)$。

3.2 枚举计算单体型

如前所述，算法 EHTR 按列构建 3 条只含杂合位点的单体型 $\boldsymbol{h}'=(\boldsymbol{h}_1',\boldsymbol{h}_2',\boldsymbol{h}_3')$。假设 3 条单体型 h' 的前 $j-1$ 列即 $(h'_{k1},\cdots,h'_{kj-1})(k=1,2,3,j=2,\cdots,n)$ 已经构建好，当前重建列为第 j 列 h'_{kj}。首先根据 $\boldsymbol{g}_j$ 取值枚举该 SNP 位点的 3 种可能取值：1)若 $\sum_{k=1}^{3}g_{jk}=1$，则 3 种取值可能为 $(h'_{1j}=0,h'_{2j}=0,h'_{3j}=1)$，$(h'_{1j}=0,h'_{2j}=1,h'_{3j}=0)$ 和 $(h'_{1j}=1,h'_{2j}=0,h'_{3j}=0)$。2)若 $\sum_{k=1}^{3}g_{jk}=2$，则 3 种取值可能为 $(h'_{1j}=0,h'_{2j}=1,h'_{3j}=1)$，$(h'_{1j}=1,h'_{2j}=0,h'_{3j}=1)$ 和 $(h'_{1j}=1,h'_{2j}=1,h'_{3j}=0)$；然后选择片段支持度最大的取值作为第 j 列的重建值。令 $S(h'_{1j},h'_{2j},h'_{3j})$ 记录第 j 列取值为 $(h'_{1j},h'_{2j},h'_{3j})$ 的片段支持度，其定义如式(4)所示。

$$S(h'_{1j},h'_{2j},h'_{3j})=\sum_{i=1}^{m}\max\{\sum_{k=1}^{j-1}c(m_{ik},m_{ij},h'_{lk},h'_{lj})\mid l=1,2,3\},j=2,\cdots,n \tag{4}$$

其中：

$$c(x_1,x_2,y_1,y_2)=\begin{cases}1, & \text{若 } x_1\neq-,x_2\neq-,\\ & x_1=y_1 \text{ 且 } x_2=y_2\\ 0, & \text{否则}\end{cases} \tag{5}$$

3.3 扩展结果

最后，需要将预处理时删除的纯合 SNP 位点重新加回。对于重建的仅含杂合位点的单体型 $\boldsymbol{h}'=(\boldsymbol{h}_1',\boldsymbol{h}_2',\boldsymbol{h}_3')$，若某个已删除位点的取值记为 g_{j1}，则将单体型 $\boldsymbol{h}_1'$，$\boldsymbol{h}_2'$ 和 $\boldsymbol{h}_3'$ 的相应 SNP 位点处插入 g_{j1}，由此扩展成最终的单体 $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3)$。根据上述算法设计，算法 EHTR 的详细描述如算法 1 所示。

算法 1　算法 EHTR

输入：SNP 矩阵 $\mathbf{M}_{m\times n}$，基因型矩阵 $\mathbf{g}$

输出：单体型 $\mathbf{h}=(\mathbf{h}_1,\mathbf{h}_2,\mathbf{h}_3)$

1. 对 M 和 g 进行预处理，得到新的片段矩阵 $\mathbf{M}_{m\times n}$ 和基因型矩阵 g
2. 令 $h'_{i1}=g_{1i}(i=1,2,3)$
3. for j=2,…,n do
4. support=0
5. if ($\sum_{k=1}^{3}g_{jk}=1$) then
6. if (S(0,0,1)> support) then
7. $h'_{1j}=h'_{2j}=0,h'_{3j}=1$,support=S(0,0,1)
8. if (S(0,1,0)>support) then
9. $h'_{1j}=h'_{3j}=0,h'_{2j}=1$,support=S(0,1,0)
10. if (S(1,0,0)>support) then
11. $h'_{2j}=h'_{3j}=0,h'_{1j}=1$,support=S(1,0,0)
12. else if ($\sum_{k=1}^{3}g_{jk}=2$) then
13. if (S(0,1,1)>support) then
14. $h'_{2j}=h'_{3j}=1,h'_{1j}=0$,support=S(0,1,1)
15. if (S(1,0,1)>support) then
16. $h'_{1j}=h'_{3j}=1,h'_{2j}=0$,support=S(1,0,1)
17. if (S(1,1,0)>support) then
18. $h'_{1j}=h'_{2j}=1,h'_{3j}=0$,support=S(1,1,0)
19. 扩展单体型 $\mathbf{h}'=(\mathbf{h}_1',\mathbf{h}_2',\mathbf{h}_3')$，输出 $\mathbf{h}=(\mathbf{h}_1,\mathbf{h}_2,\mathbf{h}_3)$

3.4 算法复杂性

本节对算法 EHTR 的时间复杂性进行分析。算法主要分为 3 个阶段：1)预处理阶段，处理基因型矩阵的时间复杂度为 $O(n)$，处理 SNP 片段矩阵的时间复杂度为 $O(mn)$，片段排序的时间复杂度为 $O(m\log m)$，计算覆盖列的行集的时间复杂度为 $O(mn)$；2)重建阶段，重建仅含杂合位点单体型的时间复杂度为 $O(cnl)$，c 表示片段的平均覆盖率，l 为片段的平均长度；3)扩展阶段，其时间复杂度为 $O(n)$。因此，算法的总时间复杂度为 $O(mn+m\log m+cnl)$。

4 实验结果

由于很难得到真实的生物数据，本文利用具有真实测序数据特征的模拟数据集对算法 EHTR，W-GA[6]，Q-PSO[9] 和 GTIHR[10] 的重建率和运行时间进行比较分析。由于文献[6]和文献[9]没有介绍算法 W-GA 和 Q-PSO 求解 MEC/GI 模型时杂合位点的修正方法，因此本文仅与其求解 MEC 的结果进行了实验测试。本文实验在一台索尼工作站 (Inter Core i5 2.50GHz，内存为 6GB) 上进行，操作系统为 Windows 7，程序编译器为 Microsoft Visual C# 2012。

4.1 实验数据

本文实验采用的模拟单体型由如下方法产生[10]：随机生

成长度为 n 的单体型 $\boldsymbol{h}_1$，根据单体型的海明距离 d 随机生成单体型 $\boldsymbol{h}_2$，最后生成单体型 $\boldsymbol{h}_3$，且 h_{3j} 按均匀概率随机取值为 h_{1j} 或 h_{2j} $(j=1,2,\cdots,n)$。

片段数据采用 CELSIM[14] 和 MetaSim[15] 两种模拟片段生成器生成，分别称为 CELSIM 实例和 MetaSim 实例。CELSIM 实例的生成方法如下[14,16]：CELSIM 模拟鸟枪法测序，对于每组测试数据，生成 m_1 条长度范围为[f_min，f_max]的单片段和 m_2 条长度为 $n/10$ 的 mate-pair 片段，两种片段的覆盖率均为 $c/2$，总片段覆盖率为 c。根据错误率 p_s 在片段中置入测序错误。实际测序中，[f_min，f_max]取值为[3，7]，c 的取值范围为 5～10，p_s 的取值范围为 2%～5%[16]。MetaSim 实例的生成方法如下：MetaSim 模拟 454 测序，对于每组测试数据，生成 m 条片段，包括 $m_1=(1-p_m)\times m$ 条单片段和 $m_2=p_m\times m$ 条 mate-pair 片段，其中 p_m 表示 mate-pair 片段的比率，实验中取值为 0.25。单片段的期望长度为 l，mate-pair 片段的期望长度为 $3l$。由于一条 mate-pair 片段由同条单体型的一对单片段构成，因此片段覆盖率 $c=[(m_1+2m_2)\times l]/3n$。

4.2 性能评价

本节在不同的测序参数取值下对算法 EHTR，W-GA，Q-PSO 和 GTIHR 的性能进行测试分析，每组测试参数生成 100 个数据集，实验结果取 100 次重复测试的平均值。算法 GTIHR，W-GA 和 Q-PSO 的参数设置分别与文献[6，9，10]一致。

表 1－表 5 列出了对 CELSIM 实例进行测试的结果。表 1 针对错误率 p_s 设置了 12 组参数，其中 $c=10$，$f\_min=3$，$f\_max=7$，$n=100$，$d=0.3$。表 1 显示，在各错误率取值下，算法 EHTR 均获得较 GTIHR，W-GA 和 Q-PSO 算法更高的重建率。错误率 p_s 为 0 时，算法 EHTR 的重建率为 0.971，较算法 GTIHR，W-GA 和 Q-PSO 的重建率分别高出 4.0%，7.2%和 15.5%。当 p_s 增加到 0.2 时，EHTR 算法仍获得 0.926的重建率，较 GTIHR，W-GA 和 Q-PSO 的重建率分别高出5.3%，4.9%和 5.5%。此外，算法 EHTR 的运行速度很快，较算法 GTIHR，W-GA 和 Q-PSO 的平均运行速度分别提高了约 1133 倍、952 倍和 341 倍。

表 1 不同错误率下的比较(CELSIM 实例)

p_s	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
0	0.971	0.934	0.906	0.841	0.02	21.77	19.25	6.92
0.01	0.964	0.923	0.899	0.851	0.02	22.30	19.13	6.88
0.02	0.965	0.934	0.902	0.864	0.02	21.48	19.30	6.83
0.03	0.968	0.927	0.901	0.858	0.02	22.93	19.30	6.94
0.04	0.969	0.927	0.903	0.854	0.02	22.02	19.53	7.00
0.05	0.964	0.935	0.904	0.849	0.02	20.78	18.77	6.59
0.06	0.960	0.924	0.902	0.853	0.02	20.99	17.95	6.46
0.07	0.961	0.924	0.902	0.868	0.02	21.36	18.33	6.64
0.08	0.958	0.920	0.903	0.861	0.02	21.95	19.28	7.00
0.09	0.951	0.925	0.901	0.866	0.02	22.49	19.27	6.92
0.1	0.952	0.926	0.898	0.863	0.02	22.91	19.52	7.04
0.2	0.926	0.879	0.883	0.878	0.02	31.09	19.07	6.91

表 2 针对片段覆盖率 c 设置了 9 组参数，其变化范围为 2～10。在这 9 组实例中，$n=100$，$f\_min=3$，$f\_max=7$，$p_s=0.05$，$d=0.3$。表 2 中数据显示，在不同的覆盖率下，算法 EHTR 均能获得较算法 GTIHR，W-GA 和 Q-PSO 更高的重建率。当覆盖率 c 为 2 时，算法 EHTR 的重建率为 0.940，较算法 GTIHR，W-GA 和 Q-PSO 的重建率分别高出 7.1%，13.8%和 16.0%。随着覆盖率的增加，算法能运用的片段原始信息越多，重建率有所提高。当 c 增加到 10 时，算法 EHTR，GTIHR，W-GA 和 Q-PSO 的重建率分别达到0.964，0.935，0.904 和 0.849。算法 EHTR 的运行速度很快且受片段覆盖率的影响不大，当 c 由 2 提高到 10 时，算法 EHTR 的运行时间增加了 1 倍，由 0.01s 增加到0.02s，而算法 GTIHR，W-GA 和 Q-PSO 的运行时间分别增加了 3.63 倍、3.61 倍和 3.61 倍。

表 2 不同片段覆盖率下的比较(CELSIM 实例)

c	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
2	0.940	0.878	0.826	0.810	0.01	4.49	4.07	1.43
3	0.941	0.900	0.856	0.841	0.01	6.45	5.85	2.11
4	0.942	0.907	0.876	0.856	0.01	9.18	8.25	2.96
5	0.955	0.913	0.887	0.862	0.01	10.93	9.99	3.63
6	0.954	0.925	0.893	0.866	0.01	13.03	11.80	4.19
7	0.959	0.923	0.897	0.871	0.01	18.95	17.49	6.25
8	0.958	0.927	0.898	0.858	0.02	18.52	16.26	5.95
9	0.958	0.925	0.899	0.854	0.02	20.35	18.03	6.47
10	0.964	0.935	0.904	0.849	0.02	20.78	18.77	6.59

表 3 针对单体型长度 n 设置了 6 组测试实例，其中 $c=10$，$f\_min=3$，$f\_max=7$，$p_s=0.05$，$d=0.3$。从表 3 中可以看出，在各组参数设置下，算法 EHTR 均获得最高的重建率。随着 n 值的增大，算法 EHTR，GTIHR 和 W-GA 的重建率都有所下降。当 n 从 100 增加到 1000 时，算法 EHTR 的重建率从0.964下降到 0.924，算法 GTIHR 的重建率从 0.935 下降到0.876，算法 W-GA 的重建率从 0.904 下降到 0.874，而 Q-PSO 的重建率在 0.849 到 0.886 之间波动。n 值的增加对 4 种算法的运行时间都有显著影响。当 $n=100$ 时，算法 EHTR，GTIHR，W-GA 和 Q-PSO 的运行时间分别为 0.02s，20.78s，18.77s 和 6.59s；而当 $n=1000$ 时，它们的运行时间分别增加到 11.08s，1207.61s，1051.55s 和326.50s。

表 3 不同单体型长度下的比较(CELSIM 实例)

n	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
100	0.964	0.935	0.904	0.849	0.02	20.78	18.77	6.59
200	0.963	0.904	0.886	0.886	0.13	57.78	56.37	18.71
300	0.943	0.890	0.885	0.885	0.34	105.95	114.47	36.80
500	0.932	0.879	0.877	0.880	1.28	273.68	291.79	91.59
800	0.925	0.877	0.877	0.877	5.37	750.68	711.03	218.23
1000	0.924	0.876	0.874	0.875	11.08	1207.61	1051.55	326.50

表 4 针对单片段长度范围[f_min，f_max]设置了 3 组参数，其中 $c=10$，$p_s=0.05$，$n=100$，$d=0.3$。表 4 中数据显示，在不同的单片段取值范围下，算法 EHTR 都能获得较算法 GTIHR，W-GA 和 Q-PSO 更高的重建率。同时，算法 EHTR 的运行时间的增长速度很缓慢，当单片段长度范围从[3，7]减小到[1，2]时，算法 EHTR，GTIHR，W-GA 和 Q-PSO 分别增加了 0.01s，12.12s，45.55s 和 16.65s。

表 4 不同单片段取值范围下的比较(CELSIM 实例)

[f_min, f_max]	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
[3,7]	0.964	0.935	0.904	0.849	0.02	20.78	18.77	6.59
[2,4]	0.943	0.920	0.897	0.879	0.02	23.50	27.20	9.75
[1,2]	0.928	0.893	0.890	0.880	0.03	32.90	64.32	23.24

表 5 针对海明距离 d 设置了 10 组参数。在这 10 组参数中，$f\_min=3$，$f\_max=7$，$c=10$，$p_s=0.05$，$n=100$。从表 5 中可以看出，随着 d 值的增大，算法 GTIHR，W-GA 和 Q-PSO

的重建率均有所降低，而算法 EHTR 的重建率在0.964～0.992之间波动。当 $d=0.1$ 时，算法 EHTR，GTIHR，W-GA 和 Q-PSO 的重建率分别为 0.986，0.972，0.966 和0.897；当 d 增加到 1 时，它们的重建率分别为 0.992，0.738，0.672和 0.654，算法 GTIHR，W-GA 和 Q-PSO 重建率的下降率分别为 24.1%，30.4%和 27.1%。此外，d 值的增加对算法 EHTR 和 GTIHR 的运行时间有显著影响，而对 W-GA 和 Q-PSO算法的影响不显著。当 d 由 0.1 增加到 1 时，算法 EHTR 和 GTIHR 的运行时间分别增加了 0.24s 和 61.31s，而 W-GA 和 Q-PSO 算法的运行时间基本保持不变。

表 5　不同海明距离下的比较(CELSIM 实例)

d	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
0.1	0.986	0.972	0.966	0.897	0.01	6.60	20.44	7.45
0.2	0.971	0.954	0.939	0.870	0.01	15.86	20.62	7.46
0.3	0.964	0.935	0.904	0.849	0.02	20.78	18.77	6.59
0.4	0.965	0.909	0.870	0.836	0.04	31.12	19.88	7.11
0.5	0.970	0.880	0.838	0.813	0.04	36.55	19.32	6.91
0.6	0.968	0.841	0.810	0.791	0.06	45.60	20.25	7.29
0.7	0.974	0.827	0.769	0.756	0.09	49.27	18.70	6.72
0.8	0.985	0.796	0.747	0.725	0.13	56.70	19.21	7.05
0.9	0.988	0.780	0.708	0.697	0.18	66.20	20.52	7.32
1	0.992	0.738	0.672	0.654	0.25	67.91	19.40	6.83

表 6　不同覆盖率下的比较(MetaSim 实例)

c	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
5	0.933	0.880	0.859	0.849	0.01	12.20	10.21	3.65
10	0.938	0.901	0.893	0.842	0.02	23.05	20.19	7.08
15	0.943	0.896	0.892	0.857	0.03	35.23	29.94	11.04
20	0.943	0.908	0.900	0.840	0.04	42.68	38.88	13.62
25	0.941	0.904	0.896	0.832	0.05	55.48	48.47	17.35
30	0.943	0.913	0.903	0.828	0.06	71.02	62.56	21.80
35	0.942	0.912	0.901	0.823	0.07	80.21	68.60	24.34
40	0.945	0.911	0.900	0.819	0.08	95.97	82.84	29.00
45	0.945	0.912	0.901	0.816	0.09	109.71	94.49	33.36
50	0.945	0.912	0.902	0.816	0.11	120.01	104.23	36.72

表 7　不同单体型长度下的比较(MetaSim 实例)

n	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
100	0.943	0.908	0.900	0.840	0.04	42.68	38.88	13.62
200	0.932	0.894	0.891	0.880	0.31	166.05	159.13	51.86
300	0.928	0.889	0.890	0.891	1.03	356.57	328.83	105.45
500	0.921	0.884	0.887	0.889	4.15	1193.31	837.29	262.75
800	0.920	0.879	0.885	0.885	18.14	4224.58	2427.49	748.90
1000	0.915	0.875	0.883	0.886	35.04	6610.63	3249.52	1014.95

表 8　不同单片段长度下的比较(MetaSim 实例)

c	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
10	0.951	0.908	0.899	0.858	0.02	27.76	19.96	7.17
5	0.943	0.908	0.900	0.840	0.04	42.68	38.88	13.62
3	0.915	0.877	0.880	0.831	0.04	32.61	63.71	23.10

表 9　不同海明距离下的比较(MetaSim 实例)

d	RR				running time(s)			
	EHTR	GTIHR	W-GA	Q-PSO	EHTR	GTIHR	W-GA	Q-PSO
0.1	0.970	0.946	0.957	0.847	0.01	20.49	47.18	16.58
0.2	0.950	0.930	0.930	0.846	0.02	31.00	39.25	13.98
0.3	0.943	0.908	0.900	0.840	0.04	42.68	38.88	13.62
0.4	0.934	0.880	0.865	0.813	0.07	64.73	41.49	14.73
0.5	0.927	0.865	0.832	0.803	0.12	89.16	46.62	16.85
0.6	0.929	0.838	0.799	0.774	0.17	108.64	49.28	17.71
0.7	0.924	0.799	0.763	0.752	0.25	126.22	48.47	17.27
0.8	0.924	0.768	0.725	0.709	0.29	139.22	48.09	16.89
0.9	0.942	0.750	0.703	0.688	0.37	149.43	46.89	16.18
1	0.948	0.718	0.655	0.657	0.43	155.48	45.27	15.75

表 6－表 9 对 MetaSim 实例进行了测试。表 6 针对覆盖率设置了 10 组测试实例，其中 $n=100$，$p_s=0.05$，$l=5$，$d=0.3$。表 7 针对单体型长度设置了 6 组测试实例，其中 $c=20$，$p_s=0.05$，$l=5$，$d=0.3$。表 8 针对单片段取值设置了 3 组测试实例，其中 $n=100$，$c=20$，$p_s=0.05$，$d=0.3$。表 9 针对海明距离设置了 10 组测试实例，其中 $n=100$，$c=20$，$p_s=0.05$，$l=5$。从表中数据可以看出，在各种参数设置下，相较于算法 GTIHR，W-GA 和 Q-PSO，算法 EHTR 均能获得更高的重建率，且运行速度更快。

结束语　单体型数据在研究三倍体物种的基因表达和遗传特性等方面发挥着重要的作用，个体单体型重建问题是获得单体型数据的有效手段。本文针对带基因型信息的最少错误更正模型，对三倍体个体单体型重建问题进行研究，提出基于枚举策略的重建算法 EHTR。算法 EHTR 根据 SNP 位点的基因型取值枚举出单体型在该位点上的 3 种取值情况，并选择获最高片段支持度的取值作为位点的重建值。大量实验结果显示，在各种参数设置下，算法 EHTR 都能以更快的运行速度获得比算法 GTIHR，W-GA 和 Q-PSO 更高的重建率，具有较好的实用价值。该枚举策略还可扩展用于四倍体或者更高倍体的单体型重建，今后将对其展开进一步的研究。

参 考 文 献

[1] AYEH K O. Expressed sequence tags(ESTs) and single nucleotide polymorphisms(SNPs): Emerging molecular marker tools for improving agronomic traits in plant biotechnology[J]. African Journal of Biotechnology, 2008, 7(4): 331-341.

[2] OLLITRAULT P, TEROL J, GARCIA-LOR A, et al. SNP mining in C. clementina BAC and sequences: transferability in the Citrus genus(Rutaceae), phylogenetic inferences and perspectives for genetic mapping [J]. BMC Genomics, 2012, 13(13): 2-9.

[3] CUENCA J, ALEZA P, NAVARRO L, et al. Assignment of SNP allelic configuration in polyploids using competitive allele-specific PCR: application to citrus triploid progeny [J]. Annals of Botany, 2013, 111(4): 731-742.

[4] HAYWARD A, MASON A S, DALTON-MORGAN J, et al. SNP discovery and applications in Brassica napus [J]. Journal of Plant Biotechnology, 2012, 39(1): 49-61.

[5] ADESOYE A, MMEKA E, VROH B. Single Nucleotide Polymorphism Markers Discovery in Musa Spp(Plantain Landraces, AAB Genome) and its Potentials for Use in Gibberellic Acid and Parthenocarpy Trait Mapping [J]. Journal of Plant Molecular Biology & Biotechnology, 2012, 3(1): 9-21.

[6] WANG R S, WU L Y, LI Z P, et al. Haplotype reconstruction from SNP fragments by minimum error correction [J]. Bioinformatics, 2005, 21(10): 2456-2462.

[7] LI Z P, WU L Y, ZHAO Y Y, et al. A dynamic programming algorithm for the k- haplotyping problem [J]. Acta Mathematicae Applicate Sinica, English Series, 2006, 22(3): 405-412.

[8] WU J L, WANG J X, CHEN J E. A parthenogenetic algorithm for single individual SNP haplotyping [J]. Engineering Applications of Artificial Intelligence, 2009, 22(3): 401-406.

(下转第 112 页)

图 5 显示了分组的平均端到端时延随 cbr 流数目的变化情况。从图中可知，随着负载的增加，分组的平均端到端时延不断增加，主要是由分组在节点处的排队时延引起的。改进后的协议具有预测负载的功能，使用负载较轻的节点作为中间节点转发分组，降低了端到端的时延；相比原来的协议，采用本协议后平均时延降低了约 16%。

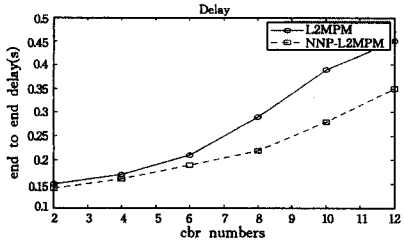

图 5 端到端平均时延随数据流数目的变化情况

结束语 本文提出了一种基于流量预测机制的负载均衡协议 NNP-L2MPM，用于进一步改善路由选路的性能。协议利用 RBF 神经网络预测模型，实现了对节点流量负载的预测，进而实现了负载均衡。实验表明，NNP-L2MPM 协议具有较好的性能。

参 考 文 献

[1] SONG J H, WONG V, LEUNG V C M. Load-aware on-demand routing (LAOR) protocol for mobilead hoc networks[C]// The 57th IEEE Semiannual Conference on Vehicular Technology, 2003(VTC 2003-Spring). IEEE, 2003: 1753-1757.

[2] LI Y, MAN H. Three load metrics for routing in ad hoc networks[C]// 2004 IEEE 60th Vehicular Technology Conference, 2004(VTC2004-Fall). IEEE, 2004: 2764-2768.

[3] TON C K, LE A N, CHO Y Z. Load balanced routing protocols for ad hoc mobile wireless networks[J]. Communications Magazine, IEEE, 2009, 47(8): 78-84.

[4] WANG Sha-sha, ZHU Guo-hui, WANG Xin. Research on load balancing routing algorithm for Ad Hoc networks[J]. Modern Electronics Technique, 2013, 36(3): 40-42. (in Chinese)
王莎莎，朱国晖，王鑫. Ad Hoc 网络负载均衡路由协议研究[J]. 现代电子技术，2013，36(3)：40-42.

[5] SHA Yi, YAN Xing-xing, TANG Xun. Load Balancing Routing Protocol Based on Traffic Prediction for Ad Hoc Network[J]. Journal of Northeastern University(Natural Science), 2012, 33(10): 1403-1406. (in Chinese)
沙毅，闫星星，唐逊. 基于流量预测的 Ad Hoc 网络负载均衡协议[J]. 东北大学学报(自然科学版)，2012，33(10)：1403-1406.

[6] SHA Yi, LI Gui-rong, ZHANG Li-li, et al. Prediction of node traffic routing protocol based on neural network in Ad hoc networks[J]. Computer Engineering and Appfications, 2011, 47(36): 118-122. (in Chinese)
沙毅，李贵荣，张立立，等. 神经网络预测 Ad hoc 节点流量的路由协议[J]. 计算机工程与应用，2011，47(36)：118-122.

[7] LEE S J, GERLA M. Dynamic load-aware routing in ad hoc networks[C]// IEEE International Conference on Communications, 2001(ICC 2001). IEEE, 2001: 3206-3210.

[8] YIN Shou-yi, LIN Xiao-kang. Traffic Self-similarity in Wireless Mesh Network[J]. Telecommunications Science, 2005, 21(4): 53-55. (in Chinese)
尹首一，林孝康. 无线 Mesh 网络流量自相似性研究[J]. 电信科学，2005，21(4)：53-55.

[9] LIANG Q. Ad hoc wireless network traffic-self-similarity and forecasting[J]. Communications Letters, IEEE, 2002, 6(7): 297-299.

[10] DONG Meng-li, YANG Geng, CAO Xiao-mei. Methods of Network Traffic Prediction[J]. Computer Engineering, 2011, 37(16): 98-100. (in Chinese)
董梦丽，杨庚，曹晓梅. 网络流量预测方法[J]. 计算机工程，2011，37(16)：98-100.

[11] WANG Jun-song, GAO Zhi-wei. Network traffic modeling and prediction based on RBF neural network[J]. Computer Engineering and Applications, 2008, 44(13): 6-7. (in Chinese)
王俊松，高志伟. 基于 RBF 神经网络的网络流量建模及预测[J]. 计算机工程与应用，2008，44(13)：6-7.

[12] ZHOU Wei-hua. Optimization Study of the Hidden Structure and Parameters in the RBF Neural Networks [D]. Shanghai: East China University of Science and Technology, 2014. (in Chinese)
周维华. RBF 神经网络隐层结构与参数优化研究[D]. 上海：华东理工大学，2014.

(上接第 79 页)

[9] QIAN W Y, YANG Y J, YANG N N, et al. Particle swarm optimization for SNP haplotype reconstruction problem [J]. Applied Mathematics and Computation, 2008, 196(1): 266-272.

[10] WU Jing-li, WANG Zhao-can. Genetic Algorithm for Solving Triploid Individual Haplotype Reconstruction Problem [J]. Journal of Chinese Computer Systems, 2014, 35(4): 840-844. (in Chinese)
吴璟莉，王兆灿. 求解三倍体个体单体型重建问题的遗传算法[J]. 小型微型计算机系统，2014，35(4)：840-844.

[11] WU Jing-li. Research on the combinatorial optimization problem in detection of genetic diversities [D]. Changsha: Central South University, 2008. (in Chinese)
吴璟莉. 遗传多态性检测中组合优化问题的研究[D]. 长沙：中南大学，2008.

[12] LIPPERT R, SCHWARTZA R, LANCIA G, et al. Algorithmic strategies for the SNPs haplotype assembly problem [J]. Briefings in Bioinformatics, 2002(3): 23-31.

[13] FILIPPO G. A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem [J]. Bioinformatics, 2010, 26(18): 2217-2225.

[14] MYERS G. A dataset generator for whole genome shotgun sequencing [C]// Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. California: AAAI Press, 1999: 202-210.

[15] RICHTER D C, OTT F, et al. MetaSim-A Sequencing Simulator for Genomics and Metagenomics [J]. PLOS ONE, 2008, 3(10): e3373.

[16] PANCONESI A, SOZIO M. Fast hare: a fast heuristic for single individual SNP haplotype reconstruction [C]// Proceedings of 4th Workshop on Algorithms in Bioinformatics. Heiderberg: Springer-Verlag, 2004: 266-277.