PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2007, 34(7): 724~731
www.pibb.ac.cn

基于支持向量机的细菌基因组水平转移基因预测 *

吴建盛 ** 谢建明 ** 周 童 翁建洪 孙 啸 ***
(东南大学生物电子学国家重点实验室，南京 210096)

摘要 随着各种生物基因组序列测定工作的完成，大量的 DNA 序列数据涌现出来，为研究在基因组中寻找水平转移基因提供了极大的便利. 将基因序列特征分析和支持向量机技术结合起来，通过分析基因序列的特征差异发现水平转移基因. 依据以前研究工作的基础，选取了绝对密码子使用频率(*FCU*)作为序列特征，主要因为它既包含了基因密码子使用偏性的信息，也包含了基因所编码蛋白的氨基酸组成信息，支持向量机利用这些信息进行水平转移基因分析和预测，可以提高预测的准确性. 另外，提出了基于分链的水平转移基因预测新方法，即将细菌基因组前导链和滞后链上的基因区别对待，分别进行水平转移基因预测. 结果显示，基本预测方法要优于目前预测结果最好的 Tsirigos 等提出的基于八联核苷酸频率的打分算法，命中率的相对提高率最高达 31.47%，而基于分链的方法对水平转移基因的预测取得了更好的结果.

关键词 细菌基因组，水平转移基因，支持向量机，绝对密码子使用频率
学科分类号 S718.83

随着人类及其他生物基因组测序工作相继完成，人们发现不同物种之间，甚至亲缘关系很远的生物之间基因组上大量存在水平基因转移(horizontal gene transfer，HGT)现象. 水平基因转移，又称横向基因转移(lateral gene transfer，LGT)，是指在不同物种的生物个体之间，或单个细胞内部细胞器之间所进行的遗传物质的交流.

随着对各种生物基因组序列的研究，大量的基因组序列数据涌现出来，为研究人员在基因组中寻找水平转移基因提供了便利. 目前，在基因组序列中发现水平转移基因的方法有很多种，比较典型的就是利用系统发生学的方法构建进化树，当基因组数据足够多时，这种方法比较有效[1]. 另外还有一类方法是基于基因序列特征的，这些方法基本上都基于这样一个假设：某个基因组的某个组成特征模式是这个基因组特有的，即这个基因组的特征标志(genomic signature)，而该基因组中与这个“signature”背离的区域可能就是水平转移区域. 基因组的 GC 含量[2]和密码子自适应系数(codon adaptation index，CAI)值[3]便是有关“geno- mic signature”经典的例子. Karlin 等[4]利用了密码子使用频率设计了一个打分系统，通过设定阈值来识别水平转移基因，而与密码子使用频率类似的算法还有利用核酸序列的双联碱基频率或 N 联碱基频率来发现水平转移基因[5,6]. Tsirigos 等[7]提出了一种基于八联核苷酸频率的打分法来进行水平转移基因的预测，这种方法对于不同的基因组可以自动设定阈值，并且相比于以往的那些算法在识别水平转移基因的命中率上有了显著的提高. 本文中，我们将基因序列特征分析和支持向量机技术结合起来，通过分析基因序列的特征差异发现水平转移基因. 依据我们以前研究工作的基础[8]，选取了绝对密码子使用频率(*FCU*)作为序列特征，主要因为它既包含了基因密码子使用偏性的信息，也包含了基因所编码蛋白的氨基酸组成信息，利用这些信息并借助支持向量机进行水平转移基因分析和预测，并将实验结果和 Tsirigos 等[7]的结果进行比较.

1 数据及方法

1.1 人工模拟的细菌基因组水平基因转移事件

我们所选用的 3 个细菌物种分别是大肠杆菌

*国家自然科学基金资助项目(60671018, 60121101).
** 共同第一作者. *** 通讯联系人.
Tel: 025-83795174, E-mail: xsun@seu.edu.cn
收稿日期：2006-12-01，接受日期：2007-02-01

(*Escherichia coli* K12)、包氏螺旋体(*Borrelia burgdorferi*)以及蜡状芽孢杆菌(*Bacillus cereus* ZK)，原因是：三者都是常见的病原性细菌，*Escherichia coli* K12 是一种模式生物，其密码子使用偏好性不明显，*Borrelia burgdorferi* 密码子使用具有很强的偏好性，而 *Bacillus cereus* ZK 的介于前两者之间. 它们的基因组序列都来自于 GenBank 数据库，登记号分别是 NC_000913、NC_001318 和 NC_006274 .

由于在细菌基因组中已知的水平转移基因数据很少，所以在本文中，我们采用人工的办法模拟在细菌基因组中插入水平转移基因. 在先前的研究中，一般采用人造基因序列作为给体基因，这些基因序列的某些特征符合一些统计规则，然而在构造这些人造基因序列时，总不可避免地加入了更多的人为因素，人们不可能构造出完全符合自然界规则的序列来. 所以在最近的研究中，一般选用噬菌体基因或细菌基因作为给体基因，这种方法的好处在于所有的水平转移基因都是自然界客观存在的,并且噬菌体基因或细菌基因作为给体基因，水平转移到细菌基因组中的事件在自然界中是真实存在的[7,9]. 本文中，采用了 Tsirigos 等[7]提出来的方法——选取了 27 种噬菌体基因组中共 1 485 个基因作为给体基因数据集. 然后随机从这个数据集中挑出给体基因插入对象细菌基因组中作为人工模拟的水平转移基因，其中挑出给体基因的个数是根据对象细菌基因组的大小来决定的(本文中，我们挑出的给体基因数量是对象细菌基因组基因总数的 2%).

1.2 序列特征提取

在本文中，我们提取细菌及噬菌体基因组中每种密码子的绝对密码子使用频率(*FCU*)作为密码子使用偏性的衡量指标，它的计算方法如式(1).

$$FCU_i=\frac{obs_i}{total} \quad (1)$$

其中，obs_i 指某一特定的密码子 i 在基因中出现的次数，*total* 指整段基因中密码子的个数.

这种衡量方法的优势在于，它含有较多的序列信息. 首先，它包含了基因的密码子使用偏性信息. 不同物种、不同生物体的基因密码子使用存在着很大的差异，其偏爱的密码子和偏爱程度即密码子使用频率是不同的，不同的基因组具有不同的密码子使用偏性. 其次，它还含有基因所编码蛋白的氨基酸组成信息. 细菌通过获得外来基因是使之能更好地适应环境甚至产生新的物种，其功能是通过获得的外来基因所编码蛋白来实现的，而不同蛋白质的氨基酸组成存在差异.

1.3 支持向量分类机

支持向量分类机(C-Support vector classification，C-SVC)是 Vapnik[10]提出的一类新型机器学习方法. 由于其出色的学习性能，在高维小训练样本情况下有着很好泛化能力，该技术已成为机器学习界的研究热点，并在很多领域都得到了成功的应用. 它是以结构化风险最小化(SRM) 代替常用的经验风险最小化(ERM) 作为优化准则，其基本思想是对于非线性可分样本，将其输入向量经非线性变换映射到另一个高维空间 Z 中，在变换后的空间中寻找一个最优的分界面(超平面)，使其推广能力最好. 具体应用 C-SVC 的步骤为：选择适当的核函数→求解优化方程，获得支持向量及相应的 Lagrange算子→写出最优分界面方程.

本文中，我们通过 C-SVC 对 HGT 的预测分为两步来操作. 第一步是人工插入水平转移基因到对象细菌基因组中，人工插入的基因作为正类集，对象细菌基因组中的基因作为负类集，然后通过提取每个基因的绝对密码子使用频率(*FCU*)即 64 维的向量作为 C-SVC 的输入文件. 第二步，我们利用 SVMlight 6.01 (http://svmlight.joachims.org/)对样本进行训练与分类，采用 10 倍交叉验证方法和不同的核函数进行水平转移基因预测. 对于人工插入水平转移基因这个事件来说，偶然性肯定很大. 所以对于每个细菌基因组，我们都采用了重复 100 次插入实验，然后取所有实验的预测结果平均值来消除人工模拟基因水平转移事件的偶然性.

1.4 One-class SVM

1.3 节中介绍了利用 C-SVC 来识别水平转移基因，这种方法在实际运用中需要一些已知的水平转移基因来训练支持向量机，然而对于很多细菌基因组来说，水平转移基因方面的信息还是很不足的，所以需要一种无监督的机器学习方法来预测细菌基因组中的水平转移基因，可以用 One-class 支持向量机来完成这个任务.

Schölkopf 等[11]于 1999 年提出了 One-class 支持向量机，用于解决一类问题. One-class 支持向量机的基本思想是把要描述的对象作为一个整体, 建立一个封闭而紧凑的区域 Ω，使被描述的对象全部或尽可能多地包容在 Ω 内，而非该类对象没有或尽可能少地包含在 Ω 内. 给定一个包含 N 个数据对象的数据集 $\{X_i, i=1,\cdots, n\}$，即 One-class 分类器的学

习样本，然后试图找到一个最小体积的超球体(圆心为 A，半径为 R)，使尽可能多的 X_i 都包含在该球体内. 本文中，通过计算全部或尽可能多地包含基因组中所有基因数据的最小超球形边界来对该组数据进行描述，以此最小超球作为 One-class 问题的分类器，然后找出此时位于超球之外的奇异点作为可能的水平转移基因，然后通过与我们人工模拟插入的水平转移基因进行比较，可以看出我们分类器的预测效果.

我们利用 One-class SVM 对 HGT 进行预测的具体方法与 **1.3** 中所描述的类似，只是将 C-SVC 换成了 One-class 支持向量机.

1.5 算法性能的衡量标准

对于识别水平转移基因的算法，最理想的结果便是把所有的水平转移基因都预测出来. 然而，大家要注意一点，我们现在所要识别出来的基因都是人为地插入细菌基因组的，而细菌基因组本身还拥有着自己原有的水平转移基因. 如果算法合理的话，那些细菌基因组原本拥有的水平转移基因也应该被预测出来一部分，然而我们是没有办法对这部分基因的识别准确率作出判断的. 因此，当人们在设计识别水平转移基因的算法时，一般采用命中率(*Hit ratio, HT*)这个参数来衡量算法的好坏. 所谓命中率就是考察我们人为插入的基因中到底有几个能被算法识别出来，如果全部的人为插入的基因都预测出来了，那么命中率就是 100% . 在本文中，我们对每个细菌基因组的 100 次基因插入实验的命中率求平均值，如式(2)所示.

$$HT=\frac{1}{100}\sum_{i=1}^{100}HT_i(G) \tag{2}$$

其中，G 代表某个细菌基因组.

同时，我们也关注我们的算法与 Tsirigos 等[7]算法比较所取得的相对提高率(relative improvement，*RI*)，计算公式如式(3)所示.

$$RI=\frac{HT_{\text{our\_method}}-HT_{\text{Tsirigos'}}}{HT_{\text{Tsirigos'}}} \tag{3}$$

其中，$HT_{\text{Tsirigos'}}$ 为 Tsirigos 等算法的 *HT* 值，$HT_{\text{our\_method}}$ 为我们算法的 *HT* 值.

当我们关心命中率的同时，也要考察一下算法的特异性(*Specificity*). 如果特异性过低，这个算法也是不成功的. 特异性的计算公式如式(4)所示.

$$Specificity=\frac{TN}{TN+FP} \tag{4}$$

这里，*TN*、*FP* 分别对应于识别结果中的真阴性和假阳性样本的数目.

2 实验结果

2.1 利用 C-SVC 对水平转移基因的预测

我们采取 **1.1** 中所描述的人工插入水平转移基因的方法，利用 C-SVC，并结合基因序列的 *FCU* 特征，对 3 种细菌基因组分别进行了水平转移基因的识别，当我们选用径向基核作为核函数时效果最好(γ=100)，通过 10 倍交叉验证后的结果如表 1 所示. 表 1 中第 2 列和第 3 列分别表示 Tsirigos 等的算法和我们算法所得到的命中率，第 4 列表示我们的算法与 Tsirigos 等的算法比较在命中率上所取得的相对提高率. 从表 1 中，我们可以清楚地看出，我们的这种基于密码子使用偏性的 C-SVC 方法要比 Tsirigos 等最近提出的基于八联核苷酸频率的打分法要更敏感，对于水平转移基因的识别率要更高，特别是对 *Escherichia coli* K12 的命中率要明显高出 Tsirigos 等的方法，相对提高率高达 31.47%. 另外，在对 3 个细菌基因组的水平转移基因识别时，特异性均高于 95%.

Table 1 Comparison of our method and Tsirigos' method performed on prediction of horizontal gene transfers

Species	*HT* of Tsirigos' method	*HT* of our method	*RI* on *HT* of our method over Tsirigos' method / %
Escherichia coli K12	0.375	0.493	31.47
Bacillus cereus ZK	0.541	0.571	5.55
Borrelia burgdorferi	0.758	0.767	1.19

2.2 利用 C-SVC 对水平转移基因的分链预测

在细菌基因组中，复制的前导链和滞后链上的基因在密码子使用偏性上存在着一定的差异，这主要是由于前导链上的 G 和 T 要比滞后链的多一些，而滞后链上的 C 和 A 要比前导链的多一些. 我们原本的预测方法是假设细菌基因组上的基因在密码子使用偏性上基本具有一致的 “Signature”，然而事实上，我们在以前的研究中发现，原核生物中同一个基因组中前导链和滞后链上的基因还是有差别的[12]. 那么，我们在预测细菌基因组的水平转移基

因时，能不能将两条链上的基因分开对待？

基于这种想法，我们对预测的方法进行了一些修改. 我们首先将每个细菌基因组中前导链和滞后链上的基因分开(用 GC skew 正负性的变化找出基因组中的复制起始点和终止点[13]，然后再判断基因复制时的方向)，然后分别对前导链上的基因集合与滞后链上的基因集合人工插入水平转移基因，每个基因集合中插入给体基因的数量决定于该集合本身的基因数量(插入给体基因的数量为该集合中基因总数的 2%)，最后，我们利用 C-SVC 分别对前导链与滞后链上的基因集合中的水平转移基因进行预测. 通过 10 倍交叉验证后的结果如表 2 所示. 表中第 2 列为 **2.1** 节描述的方法的预测结果，后面的 3列 Leading strand，Lagging strand 和 Mean 分别表示对水平转移基因进行分链预测时前导链、后滞链的预测结果和两者的平均值. 从表 2 中我们可以看出，对水平转移基因进行分链预测的结果要明显好于不分链的结果(特异性均高于 95%). 这说明，前导链和滞后链上基因在密码子使用偏性上的差异，确实会影响我们对水平转移基因的预测.

Table 2 Comparison of non-strand method and new method considering strand asymmetry performed on prediction of horizontal gene transfers

Species	*HT* of original method	*HT* of new method considering strand asymmetry		
		Leading strand	Lagging strand	Mean
Escherichia coli K12	0.493	0.539	0.496	0.519
Bacillus cereus ZK	0.571	0.647	0.619	0.640
Borrelia burgdorferi	0.767	0.849	0.860	0.852

2.3 利用 One-class 支持向量机对水平转移基因的预测

这里我们还是利用基因序列的 *FCU* 特征，对 3 种细菌基因组分别进行水平转移基因的识别，人工插入水平转移基因的方法都相同，只是 C-SVC 换成了 One-class 支持向量机，此时我们选用三阶多项式核作为核函数效果最好. 通过 100 次模拟水平转移实验后的平均结果如表 3 所示. 表 3 中第 2 列和第 3 列分别表示 Tsirigos 等的算法和我们的算法所得到的命中率，第 4 列表示我们的算法与Tsirigos等的算法比较在命中率上所取得的相对提高率(relative improvement，*RI*). 从表 3 中，我们可以清楚地看出，在 *Borrelia burgdorferi* 和 *Bacillus cereus* ZK 中，我们的这种基于密码子使用偏性的 One-class 支持向量机的方法，要比 Tsirigos 等的方法具有更高的命中率，*Borrelia burgdorferi* 的 *RI* 值达到 11.61%，而在 *Escherichia coli* K12 中的表现要略逊于 Tsirigos 等的方法. 另外，我们这种基于 One-class 支持向量机的方法在对 3 个细菌基因组的水平转移基因识别时特异性也都高于 95%.

Table 3 Comparison of our new method based on one-class SVM and Tsirigos' method performed on prediction of horizontal gene transfers

Species	*HT* of Tsirigos' method	*HT* of our method	*RI* on *HT* of our method over Tsirigos' method /%
Escherichia coli K12	0.375	0.346	−7.73
Bacillus cereus ZK	0.541	0.566	4.62
Borrelia burgdorferi	0.758	0.846	11.61

2.4 利用 One-class 支持向量机对水平转移基因的分链预测

与 **2.2** 节类似，这里我们在预测细菌基因组的水平转移基因时，也将前导链和滞后链上的基因分开对待，具体方法与 **2.2** 节中所描述的类似，只是将 C-SVC 换成了 One-class 支持向量机. 具体的计算结果如表 4 所示. 表中第 2 列为 **2.3** 节描述的方法的预测结果，后面的 3 列 Leading strand，Lagging strand 和 Mean 分别表示对水平转移基因进行分链预测时前导链、后滞链的预测结果和两者的平均值. 从表 4 中我们可以看出，在 *Escherichia coli* K12 和 *Borrelia burgdorferi* 中，利用 One-class

支持向量机对水平转移基因进行分链预测的结果要明显好于不分链的结果，只是在 *Bacillus cereus* ZK 中分链预测的表现略逊于不分链的结果. 但是不管怎么说，在 3 个细菌基因组中，我们这种分链并利用基于密码子使用偏性的 One-class 支持向量机方法在识别水平转移基因的表现，确实都要优于 Tsirigos 等提出的方法.

Table 4 Comparison of non-strand method and new method considering strand asymmetry both by one-class SVM performed on prediction of horizontal gene transfers

Species	*HT* of non-strand method	*HT* of new method considering strand asymmetry		
		Leading strand	Lagging strand	Mean
Escherichia coli K12	0.346	0.527	0.436	0.482
Bacillus cereus ZK	0.566	0.596	0.528	0.562
Borrelia burgdorferi	0.846	0.940	0.945	0.942

2.5 算法在粪肠球菌基因组中的实际检验

为了检验我们以上的算法在实际应用中的可靠性，我们对粪肠球菌(*Enterococcus faecalis* V583)基因组中的水平转移基因进行一次实际的预测. 目前已经知道，在 *Enterococcus faecalis* V583 中确实存在着一些通过水平转移而得到的耐万古霉素(Vancomycin-resistance)基因，一共有 7 条，分布在同一个基因簇(Cluster)中——EF2293～EF2299[14] .

我们采用了 **2.4** 节中描述的方法，即利用 One-class 支持向量机对不同链上的水平转移基因进行基于密码子使用偏性的识别. 采用 **2.2** 节中提到的 GC skew 方法，我们通过计算确定这 7 条 vancomycin-resistance 基因都位于前导链上，因此我们单独对 *Enterococcus faecalis* V583 基因组中前导链上的水平转移基因进行识别，但是这次我们不再人为地向 *Enterococcus faecalis* V583 基因组中插入给体基因，而是将 EF2293～EF2299 这 7 条基因作为正类集，前导链上的其余 2 500 个基因作为负类集. 我们的算法将这 7 条基因全部识别了出来，另外，我们还发现在基因 EF2293 上游 7 kb 与 EF2299 基因下游 38 kb 的区域内存在 28 个基因被预测水平转移呈阳性. Paulsen 等[14]对 *Enterococcus faecalis* V583 耐万古霉素基因簇上游 9 kb 与下游 51 kb 的区域内的共 64 个基因(EF2282～EF2347 及 EF1983～EF1987)进行了 Blast 分析，发现，其中的 8 个基因与 Tn916 转座子中的基因相似(*P*-value $<1\times10^{-5}$)及 11 个基因与 Tn1549 转座子中的基因相似(*P*-value $<1\times10^{-5}$)，这有可能说明基因 EF2293 上游 7 kb 与 EF2299 基因下游 38 kb 这整段区域的序列都是与这 7 条耐万古霉素基因同时转移到 *Enterococcus faecalis* V583 基因组中的.

3 讨 论

本文中，我们介绍了一种新的识别水平转移基因的方法. 这种方法一方面继承了前人利用密码子使用偏性进行水平转移基因预测的思想，在 Genomic signature 的选取上我们采用了密码子使用频率(*FCU*)作为序列特征，因为它不仅包含了基因密码子使用偏性的信息，还包含了基因所编码蛋白的氨基酸组成信息，另一方面又采用了在小样本分类上性能好的支持向量机算法，特别是我们提出了要把基因组前导链和滞后链上的基因区别对待的新思路. Tsirigos 等[7]提出了一种基于八联核苷酸频率的打分法来进行水平转移基因的预测，并且相比于以往的算法在识别水平转移基因的命中率上有了显著的提高，因此将我们的方法和 Tsirigos 等的结果进行了比较，发现在相同的数据集上我们的结果确实优于 Tsirigos 等的结果.

本文用支持向量机方法来预测噬菌体基因到细菌基因组的水平转移，采用人工的办法在细菌的基因组中插入噬菌体基因来模拟基因的水平转移，主要因为噬菌体基因作为给体基因水平转移到细菌基因组中的事件在自然界中是真实存在的，且我们的算法在粪肠球菌基因组中的实际检验中也得到了满意的结果. 一个细菌基因组中的密码子使用频率一般是不均匀的，噬菌体人工插入事件必须充分多，结果才有意义. 本文中，我们利用了 27 种噬菌体基因组中共 1 485 个基因作为给体基因数据集. 对于每个细菌基因组，我们都采用了重复 100 次插入实验，然后计算所有实验预测结果的平均值. 每次随机从给体基因数据集中挑出对象细菌基因组基因总数 2%的给体基因插入对象细菌基因组中，这样做

可以使插入的给体基因尽量涵盖到全部的给体基因数据集，保证噬菌体人工插入事件充分多，确保我们的模型有意义.

基于序列特征预测基因的水平转移这类方法，基本上都是假设所研究的对象基因组具有特有的Genomic signature，通过发现该基因组中与这个“Signature”不一致的区域来作为可能的水平转移区域. 为了检验我们研究的 3 种细菌基因组和人工模拟插入的噬菌体基因组是否具有特有的 *FCU* 组成特征模式，我们不仅利用 k-means 聚类方法分别把所研究的 3 种细菌基因组与噬菌体给体基因数据集放在一起进行了聚类分析(表 5)，还应用主成分分析(principal component analysis，PCA)方法选取了贡献率最大的前两维进行分析(图 1). 表 5 中的第 2 列中括号内的数值为对应细菌基因组的基因总数，括号外的数值为错误聚类到噬菌体类的基因个数，第 3 列中括号内的数值为噬菌体的基因总数，括号外的数值为错误聚类到对应细菌基因组类的基因个数，第 4 列为对应细菌基因组的我们基于支持向量分类机的算法与 Tsirigos 等的算法比较在命中率上所取得的相对提高率(relative improvement，*RI*)，第 5 列为对应细菌基因组的我们基于 One-class SVM 的算法与 Tsirigos 等的算法，比较在命中率上所取得的 *RI* 值(第 4、5 列的值来自

Table 5　Number of error cases in each cluster by genes' *FCU* motif between a specified bacterium genome and phage genomes[1)]

Species	Error cases[2)] (Total gene number)	Error cases for phages[3)] (Total gene number of phages)	*RI* on *HT* of our C-SVC method/%[4)]	*RI* on *HT* of our method by one-class SVM /%[5)]
Escherichia coli K12	540 (4254)	696 (1485)	31.47	−7.73
Bacillus cereus ZK	6 (5134)	809 (1485)	5.55	4.62
Borrelia burgdorferi	0 (851)	823 (1485)	1.19	11.61

[1)]Application of k-means clustering means for three bacterium genomes being considered and phage genomes; [2)]Number of bacterium genes being mis-clustered into phage group; [3)]Number of phage genes being mis-judged into the specified bacterium group; [4)]Source from **Table 1**; [5)]Source from **Table 3**.

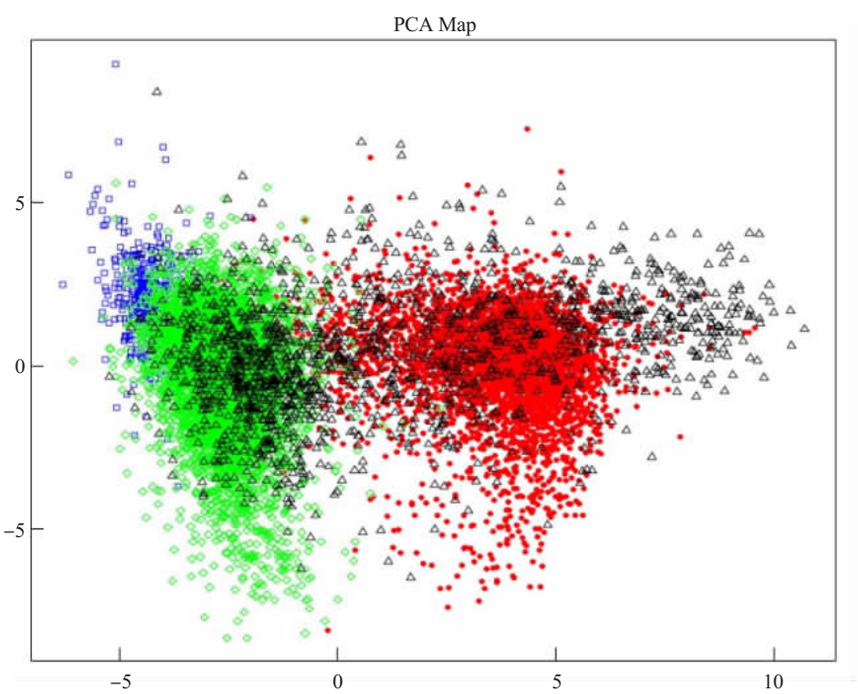

Fig. 1　Plot of the two most dominant axes generated with PCA analysis method for three bacteria genomes and phage genomes by genes' *FCU* motif

The result shows the exact difference between *Escherichia coli* K12 cases and *Bacillus cereus* cases, also between *Escherichia coli* K12 cases and *Borrelia burgdorferi* cases , and shows a proportion of overlap occurring between *Bacillus cereus* and *Borrelia burgdorferi* cases. The result indicates that these three bacteria genomes have their own special *FCU* motif usage. The main reason for wide distribution of Phage cases is that its cases came from 27 kinds of phage genomes. ●: *Escherichia coli* K12; □: *Borrelia burgdorferi*; ◇: *Bacillus cereus*; △: Phage.

表 1 和表 3). 图 1 中，*Escherichia coli* K12 与 *Bacillus cereus*，*Borrelia burgdorferi* 的样本通过前两维的 PCA 分析，区分得非常明显，*Bacillus cereus* 与 *Borrelia burgdorferi* 之间有部分重叠，说明这 3 种细菌基因组基因的 *FCU* 模式是对应基因组所特有的，而 Phage 样本分布广泛，主要是因为这些样本来自 27 种噬菌体基因组. 表 5 和图 1 的结果显示，我们研究的 3 种细菌基因组确实都具有特有的 *FCU* 模式，*Borrelia burgdorferi* 最明显，*Bacillus cereus ZK* 其次. 图 1 显示，噬菌体给体基因数据集中的样本分布非常广泛，*FCU* 特征不明显，主要是因为这些样本来自 27 种噬菌体基因组，这样可确保人工插入的噬菌体基因的 *FCU* 特征不具有偏向性，使噬菌体人工插入事件充分得多，保证我们的模型是有意义的. 对基于 C-SVC 的方法而言，*Escherichia coli* K12 命中率的相对提高率(*relative improvement*，*RI*)最高(31.47%)，*Borrelia burgdorferi* 最低(1.19%)，表明这种方法更适用于发现 *FCU* 特征不明显的细菌基因组的水平转移基因(表 5).

对于我们所研究的 3 个细菌基因组来说，各自的水平转移基因的预测命中率各不相同，对 *Escherichia coli* K12 和 *Borrelia burgdorferi* 的命中率差异达到了 30%，为什么会产生如此大的差距呢？我们认为这与各个基因组经历的选择与突变压力的不同有关. 在 *Escherichia coli* K12 基因组中，选择压力对基因的影响相当大，密码子使用偏性主要与基因的翻译选择有关[15]，因此 *Escherichia coli* K12 基因组中的基因密码子使用偏性或者碱基组成不是很一致，基因的个体差异大，这样就导致了 *Escherichia coli K12* 基因组的“Genomic signature”不是很显著. 当使用计算的方法识别水平转移基因时，效果相对较差也就不难理解了. 从这个角度来讲，当利用计算的方法预测水平转移基因时，对象基因组如果经历的突变压力大于选择压力，那么它的基因碱基组成应当比较一致，这样识别效果就会比较好，*Borrelia burgdorferi* 基因组便是一例. 另外，*Borrelia burgdorferi* 基因的密码子使用偏性主要是由于前导链和滞后链之间碱基组成上的差异[16]，所以，当我们对 *Borrelia burgdorferi* 基因组中的水平转移基因进行分链预测时，取得了非常好的效果，命中率接近 95%.

需要指出的是，任何一种预测水平转移基因的算法都有可能预测出本身并不是水平转移的基因. 这些基因主要集中在高表达基因中，比如核糖体蛋白基因、生长因子基因等. 这类基因由于选择压力的影响可变性较小，比较稳定，不容易发生水平转移[17]，相对于基因组的其他基因来说，它们的碱基组成、密码子使用偏性等序列特征是比较特别的，因此才会被误测出来. 然而，并不是所有的高表达基因都不会发生水平转移，有文献表明，高表达的核糖体蛋白基因 S14 就可能发生水平转移[18]，另外还有研究表明，氨酰 tRNA 合成酶(aminoacyl-tRNA synthetases)基因这种与翻译过程密切相关的高表达基因也可能经历水平转移[19,20].

参 考 文 献

1 Syvanen M. Horizontal gene transfer: evidence and possible consequences. Annu Rev Genet, 1994, **28**: 237～261

2 Ochman H, Lawrence J G, Groisman E A. Lateral gene transfer and the nature of bacterial innovation. Nature, 2000, **405** (6784): 299～304

3 Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol, 1985, **2** (1): 13～34

4 Karlin S, Mrazek J, Campbell A M. Codon usages in different gene classes of the *Escherichia coli* genome. Mol Microbiol, 1998, **29** (6): 1341～1355

5 Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet, 1995, **11** (7): 283～290

6 Sandberg R, Winberg G, Branden C, *et al*. Capturing whole-genome characteristics in short sequences using a naive Bayesian Classifier. Genome Res, 2001, **11** (8): 1404～1409

7 Tsirigos A, Rigoutsos I. A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res, 2005, **33** (3): 922～933

8 Zhou T, Weng J H, Sun X, *et al*. Support vector machine for classification of recombination hotspots and coldspots in *Saccharomyces cerevisiae* based on codon composition. BMC Bioinformatics, 2006, **7**: 223

9 Tsirigos A, Rigoutsos I. A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. Nucleic Acids Res, 2005, **33** (12): 3699～3707

10 Vapnik V. The Nature of Statistical Learning Theory. New York : Springer-Verlag, 1995. 1～188

11 Schölkopf B, Platt J C, Shawe-Taylor J, *et al*. Estimating the support of a high-dimensional distribution. Neural Comput, 2001, **13** (7): 1443～1471

12 Zhou T, Sun X, Lu Z H. Synonymous codon usage in *environmental chlamydia UWE25* reflects and evolutional divergence from *Pathogenic chlamydiae*. Gene, 2006, **368**: 117～125

13 Frank A C, Lobry J R. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics, 2000, **16** (6): 560～561

14 Paulsen I T, Banerjei L, Myers G S, *et al*. Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science, 2003, **299** (5615): 2071～2074

15 Ikemura T. Correlation between the abundance of *Escherichia coli*

transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol, 1981, **151** (3): 389～409

16 McInerney J O. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci USA, 1998, **95** (18): 10698～10703

17 Jain R, Rivera M C, Lake J A. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA, 1999, **96** (7): 3801～3806

18 Brochier C, Philippe H, Moreira D. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet, 2000, **16** (12): 529～533

19 Doolittle R F, Handy J. Evolutionary anomalies among the aminoacyl-tRNA synthetases. Curr Opin Genet Dev, 1998, **8** (6): 630～636

20 Woese C R, Olsen G J, Ibba M, *et al*. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev, 2000, **64** (1): 202～236

Support Vector Machine for Prediction of Horizontal Gene Transfers in Bacteria Genomes*

Wu Jian-Sheng**, Xie Jian-Ming**, Zhou Tong, Weng Jian-Hong, Sun Xiao***

(*State Key Laboratory of Bioelectronics Southeast University, Nanjing* 210096, China)

Abstract Horizontal gene transfer (HGT), also Lateral gene transfer (LGT), is any process in which an organism transfers genetic material to another species that is not its offspring. With the increase of available genomic data, it has become more convenient to study the way to detect the genes, which are products of horizontal transfers among a given genome. There are few data about known horizontal gene transfers in three bacterium genomes under consideration, so the experiments, which simulated gene transfer by artificially inserting phage genes, were carried out. Combining the feature analysis methods of gene sequences with support vector machine (SVM), a novel method was developed for identifying horizontal gene transfers (HGT) in 3 fully sequenced bacterium genomes (*Escherichia coli* K12, *Borrelia burgdorferi*, *Bacillus cereus* ZK). According to our previous work, codon use frequency (FCU) was selected as the sequence feature, in respect that it is inherently the fusion of both codon usage bias and amino acid composition signals. In addition, another computational method was proposed considering strand asymmetry and predicting horizontal gene transfers of leading strand and lagging strand of genomes under consideration, respectively. To avoid the occasionality of simulating gene transfer through artificially inserting phage genes, 100 times of the transfer-and-recover experiment were repeated and arithmetic average of measurement for each genome being considered were reported to evaluate algorithm's performance. Ten-fold cross-validation was used for both parameter and accuracy estimation. The best results were obtained for C-Support Vector Classification (C-SVC) type by using the radial basis function kernel with γ=100, while for one-class SVM type the best performance was obtained using the polynomial kernel of three degree. The performance of the approach was compared with that of Tsirigos' method ,which is one of the best predictive approachs to date in detecting of horizontal transfer genes. Firstly, for the original method that did not consider the strand asymmetry, the C-SVC type has a high relative improvement(*RI*) of 31.47% on hit ratio for *Escherichia coli* K12, while the one-class SVM type has *RI* of 11.61% for *Borrelia burgdorferi*. Moreover, as theoretically expected, the method considering the strand asymmetry resulted in higher *RI* than the original method. In order to examine the approach's performance in detecting factual gene transfer events, the approach was applied in genome of *Enterococcus faecalis* V583. It is not only succeed in recovering all the seven factual horizontally transferred genes, also found that the whole segment from 7 kb upstream of gene *EF2293* to 38 kb downstream of gene *EF2299* was probably transferred into *E. faecalis* V583 genome simultaneously with the above seven genes.

Key words bacteria genomes, horizontal gene transfer (HGT), support vector machine (SVM), codon use frequency (*FCU*)

*This work was supported by a grant from The National Natural Science Foundation of China (60671018,60121101).

**Contributed to this paper equally.

***Corresponding author . Tel: 86-25-83795174, E-mail: xsun@seu.edu.cn

Received: Decmeber 1, 2006 Accepted: February 1, 2007