PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2011, 38(7): 642~651
www.pibb.ac.cn

基于迭代自学习的操纵子结构预测 *

吴文琪 ** 郑晓斌 ** 刘永初 汤 凯 朱怀球 ***
(北京大学，湍流与复杂系统国家重点实验室，工学院生物医学工程系，理论生物中心，蛋白质科学中心，北京 100871)

摘要 原核生物操纵子结构的准确注释对基因功能和基因调控网络的研究具有重要意义，通过生物信息学方法计算预测是当前基因组操纵子结构注释的最主要来源．当前的预测算法大都需要实验确认的操纵子作为训练集，但实验确认的操纵子数据的缺乏一直成为发展算法的瓶颈．基于对操纵子结构的认识，从基因间距离、转录翻译相关的调控信号以及 COG 功能注释等特征出发，建立了描述操纵子复杂结构的概率模型，并提出了不依赖于特定物种操纵子数据作为训练集的迭代自学习算法．通过对实验验证的操纵子数据集的测试比较，结果表明算法对于预测操纵子结构非常有效．在不依赖于任何已知操纵子信息的情况下，算法在总体预测水平上超过了目前最好的操纵子预测方法，而且这种自学习的预测算法要优于依赖特定物种进行训练的算法．这些特点使得该算法能够适用于新测序的物种，有别于当前常用的操纵子预测方法．对细菌和古细菌的基因组进行大规模比较分析，进一步提高了对基因组操纵子结构的普遍特征和物种特异性的认识．

关键词 基因组分析，操纵子，迭代自学习，预测评价
学科分类号 Q61，Q93
DOI: 10.3724/SP.J.1206.2010.00686

操纵子(operon)是原核生物基因组特有的组织结构，它由多个毗邻的结构基因加上相关的调控序列构成，以转录单元(transcript unit)的形式成为基因表达的基本单元．研究发现，原核生物的许多基因在功能上的关联及表达调控都可以通过操纵子机制来实现，构成同一操纵子的基因往往功能相关，或处于同一代谢通路[1-2]．因此，操纵子结构的研究对于基因功能和调控网络的理解具有十分重要的意义．随着实验技术的快速发展，近年来 RNA-seq 及 tiling arrays 两种实验手段已经开始运用于大规模的转录组测定．但是，迄今为止人们也只是获得了约 10 个基因组全转录组数据[3]，通过转录组数据全面获得操纵子信息的方式目前还受到很大的限制．显然，鉴于当前原核基因组测序的速度远快于转录组测定的速度，计算预测方法在未来相当长的时期内仍然是获得原核基因组操纵子结构的重要途径．

概括地说，当前操纵子结构的计算预测方法主要通过对三类特征信息进行建模来实现．第一类是操纵子结构相关的序列特征信息，包括基因间距离[4]、转录调控信号及其他信号[5-7]、密码子的使用偏好[6, 8]等，其中基因间距离被认为是最有效的特征[4, 7]，在当前的预测方法中被广泛使用．第二类是基于比较基因组学获得的操纵子结构信息，其基本思想是构成同一操纵子的相邻基因在进化压力下表现出相同的保守性，由此可以确定一部分操纵子结构[9]．但是，研究表明大多数操纵子结构在进化上是崭新的[7]，通过这种同源性设计的识别方法只适合于进化过程中相对保守的那部分操纵子[9]．第三类主要依赖基因产物功能注释的信息，包括 Riley 的功能分类[4]、GO(gene ontologies)[7, 10-11]、蛋白质 COG(clusters of orthologous groups of proteins)[8]、代谢通路[2, 11]等．由于功能注释方法的复杂性，对新

* 国家自然科学基金资助项目(30970667, 30770499, 10721403)，“重大新药创制”科技重大专项资助项目(2009ZX09501-002)，北京市优秀博士学位论文指导教师科技项目(YB20101000102)，国家重点基础研究发展计划(973)资助项目(2011CB707500).
** 共同第一作者.
*** 通讯联系人.
Tel: 010-62767261, E-mail: hqzhu@pku.edu.cn
收稿日期：2010-12-28，接受日期：2011-04-22

测序物种的注释可靠性还十分有限，因此功能注释信息对于新测序物种的操纵子预测并不十分理想.

基于上述三类信息，人们已发展了一系列的操纵子预测算法，包括隐马氏模型方法(HMM)[12]、简单统计方法[4, 13]、贝叶斯决策[6, 8, 14]、神经网络[11]、支持向量机[5]及其他模式识别方法[7]. 这些算法大多数都需要依赖训练集来确定数学模型的参数. 但是，它们使用的训练集都来自少数的模式生物[8, 14]，如大肠杆菌 *Escherichia coli* 和枯草芽孢杆菌 *Bacillus subtilis* 等基因组. 虽然 RNA-seq 和 tiling arrays 等技术已经提供了若干基因组的转录组试验数据，但这些数据尚未被广泛应用于当前的操纵子预测算法的模型训练和检测. 应该指出的是，对训练数据的依赖和数据集的缺乏使得这些预测方法难以应用于大多数原核生物. 例如，有人通过 *E. coli* 和 *B. subtilis* 等物种的已知操纵子结构数据作为训练集得到基因间距特征，试图作为普适的参数来应用于其他基因组[7-8]，但是有研究表明不同基因组的基因间距分布并不完全相同，基因间距具有一定的物种特异性[8, 15]. 近年来，人们发现，基于比较基因组学的操纵子预测方法可以减少对训练数据集的依赖，这一类方法虽然可以推广到任意物种，却无法预测全基因组的操纵子结构[9, 14]. 综上所述，当前操纵子结构预测方法的研究需要克服两个难题：一是已知操纵子结构数据的相对缺乏；二是操纵子结构特征的物种特异性. 随着原核生物基因组测序计划的不断加快，至 2010 年已有超过 1 000 种细菌和古细菌的全基因组序列和注释被发布，而人们对它们的操纵子结构的注释和认识还非常滞后. 因此，发展有效的操纵子结构预测方法是当前原核生物基因组研究中亟待解决的问题.

本文基于对操纵子结构的认识，从转录相关的调控信号、基因间距离以及 COG 注释等特征出发，建立了描述操纵子复杂结构的概率模型，并提出了不依赖于特定物种操纵子数据作为训练集的迭代自学习算法. 从算法设计而言，这种方法可以很好地克服上述两个难题. 我们还通过比较证明，自学习的预测算法要优于依赖特定物种操纵子数据进行训练的预测方法. 通过对实验验证的操纵子数据集的测试，结果表明本文的算法对于预测操纵子结构非常有效. 在不依赖于任何已知操纵子信息的情况下，我们的算法在总体预测水平上超过了目前最好的操纵子预测方法，从而表现出有别于当前主要预测方法的优势.

1　材料与方法

1.1　数据集

所有基因组的全基因组序列及基因注释从 RefSeq 数据库下载得到(版本号 30)，共 762 个基因组.

操纵子的预测通常是基于预测基因对(gene pair，即 2 个相邻基因)是否属于同一操纵子来实现[4]，本文也采用相同的策略. 这里，基因对包括两类：操纵子基因对(operon gene pair)和边界基因对(boundary gene pair)，前者是操纵子内的 2 个相邻基因，后者为同链上位于 2 个不同操纵子的相邻基因对(图 1). 为了刻画整个基因组的调控网络，我们把同链上被异链基因隔开的基因对也纳入预测范畴(尽管它们基本上不可能属于同一操纵子).

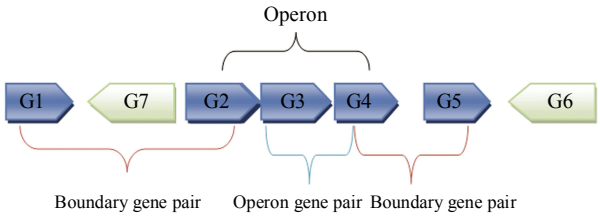

Fig. 1　Two kinds of gene pair

G1, G2, G3, G4, G5 are genes in the same direction, while G6, G7 are in the opposite direction, and G2, G3, G4 constitute one operon. Thus, (G2, G3), (G3, G4) are operon gene pairs and (G1, G2), (G4, G5) are boundary gene pairs.

为测试预测的精度，本文根据相关文献和数据库的注释构造已知操纵子结构的数据作为测试集，共包含 7 个物种：大肠杆菌 *Escherichia coli*、枯草芽孢杆菌 *Bacillus subtilis*、天蓝色链霉菌 *Streptomyces coelicolor*、铜绿假单胞菌 *Pseudomonas aeruginosa*、肺炎支原体 *Mycoplasma pneumonia*、硫磺矿硫化叶菌 *Sulfolobus solfataricus*、硫还原泥土杆菌 *Geobacter sulfurreducens*. 其中，模式细菌 *E. coli* 与 *B. subtilis* 已被多数方法作为衡量操纵子预测算法有效性的参考物种. *E. coli* 操纵子数据来源于 RegulonDB 数据库(本文剔除了可变转录单元和单个基因的转录单元)[16]；*B. subtilis* 操纵子数据则来源于 DBTBs 数据库中通过 RNA 印迹注释的操纵子[17]. 根据基因对的定义，共获得 *E. coli* 基因组的 1 622 个操纵子基因对和 1 127 个边界基因对，以及 *B. subtilis* 基因组的 530 个操纵子基因对和

385 个边界基因对．*S. coelicolor* 广泛分布于土壤中，有着复杂的生命周期，并能产生大量的抗生素，该基因组的操纵子数据为来源于 Charaniya 等[5]获得的 146 个操纵子基因对和 61 个边界基因对．*P. aeruginosa* 在自然界分布广泛，是人体手术后易感染的重要病菌之一．通过整理 ODB 数据库[18]，共获得 32 个实验确认操纵子，由此得到 67 个操纵子基因对和 39 个边界基因对．通过大规模测定原核生物转录组，共获得 *M. pneumonia*、*S. solfataricus*、*G. sulfurreducens* 三个基因组的全转录组数据．*M. pneumonia* 为最小的自我复制生物，同时也是人类支原体肺炎的病原体，共获得 366 个操纵子基因对和 199 个边界基因对[19]．*S. solfataricus* 为代谢硫磺的需氧型古细菌，其最适宜生长环境为 80℃ 和 pH 2～3，是被广泛研究的古细菌模式生物，共获得 856 个操纵子基因对和 1 028 个边界基因对[20]．*G. sulfurreducens* 能通过氧化有机污染物获取电能，共获得 1 904 个操纵子基因对和 725 个边界基因对[21]．

1.2 操纵子结构特征及数学模型

根据当前绝大多数预测方法的做法，对操纵子结构作如下 2 点假设：a．一个操纵子只对应一个转录物．虽然研究证实原核生物的少部分操纵子具有不同的转录单元[5, 22]，但对它们的复杂调控机制还有待于研究，当前的大多数预测方法并不考虑这种情况．b．为便于算法的设计，本文把单个基因的转录单元也看成是一种操纵子结构．

概括地说，本文的操纵子结构模型就是计算每一个基因对(gene pair)属于同一个操纵子的概率．结合基因之间的距离 d、上游基因的下游序列(即转录终止子区域 S_{tr})和下游基因的上游序列(即转录启动子区域 S_{pr}、翻译信号区域 S_{sd})、以及基因对是否属于同一 COG 功能分类这些特征，建立一个概率模型．若用 Ω 代表全部模型参数(或事件)构成的集合，则 Ω 的似然函数可以写成：

$$\begin{aligned}L(\Omega)&=P(d, S_{sd}, S_{pr}, S_{tr}, cog|\Omega)\\&=P(opr, d, S_{sd}, S_{pr}, S_{tr}, cog|\Omega)+P(nop, d, S_{sd}, S_{pr}, S_{tr}, cog|\Omega)\\&=P(d)\cdot P(opr|d)\cdot P(S_{sd}|opr, d)\cdot P(S_{pr}|opr, d)\cdot P(S_{tr}|opr, d)\cdot P(cog|opr)+P(d)\cdot P(nop|d)\cdot P(S_{sd}|nop, d)\cdot P(S_{pr}|nop, d)\cdot P(S_{tr}|nop, d)\cdot P(cog|nop)\end{aligned} \quad (1)$$

这里，*opr* 代表操纵子基因对，*nop* 代表边界基因对．$P(d)$代表所有相邻基因对间距的概率分布，是基因组本身特有的性质．我们还假定 S_{sd}、S_{pr}、S_{tr} 和 *cog* 是相互独立的，*cog* 和 d 也相互独立．对于各个部分分别建立模型如下：

a．基因间距离．基因间距离一直作为识别操纵子的重要特征[4, 7]．这里，定义相邻基因中上游基因的终止位点到下游基因的起始位点之间的距离为基因间距离．若两个基因发生重叠，则间距为负．分析表明，操纵子基因对与边界基因对的基因间距离差异显著，前者的基因间距离较短且分布紧凑，在–4、–1、–8 bp 等处出现明显峰值，而后者的基因间距离较长且分布比较平缓．这一特征与目前的研究认识是一致的[4]．

当一个基因对间距为 d 时，这对基因属于操纵子基因对和边界基因对的概率符合 Logistic 模型，即

$$P(opr|d)=\frac{e^{\alpha+\beta d}}{1+e^{\alpha+\beta d}},\ P(nop|d)=\frac{1}{1+e^{\alpha+\beta d}} \quad (2)$$

其中，α 和 β 反映基因间距离与属于操纵子基因对概率的线性关系，当 $\beta<0$ 时，表示概率与距离成反比．显然，当 $d=-\alpha/\beta$ 时上述两个概率相等，我们称某一基因组的这一距离为临界距离 d_{crt}，它是从概率上区分操纵子基因对与边界基因对的阈值．

b．调控信号．原核生物操纵子以转录单元的形式出现，涉及转录、翻译过程相关的调控信号．这些信号主要有操纵子上游的启动转录的启动子(promotor)、下游终止转录的终止子(terminator)以及每一个基因 5′端上游起始蛋白质合成的翻译起始信号．其中，启动子区大多由转录起始位点上游 –10 bp 的 pribnow 框(pribnow box)TATAAT 和 –35 bp 的共同序列(consensus sequence)TTGACA 组成．终止子包括不依赖 ρ(ρ-independent)的终止子和依赖 ρ(ρ-dependent)的终止子两种：前者由高 GC 的发卡结构和随后的 T 富含区域组成，后者则无明显的序列特征[23]．翻译起始信号中最常用的是 shine-dalgarno(SD)信号，而当基因组内存在多种翻译起始机制时，操纵子基因对与边界基因对的翻译起始信号会有显著差异[24-26]．本文通过权重矩阵(weight matrix)来刻画这三类信号．对每对相邻基因首先提取翻译起始信号序列(下游基因的上游 20 bp 内的区域)、启动子信号序列(下游基因的上游 20 bp 到上游基因终止位点之间的区域，若多于 60 bp，只取[–80, –20] bp 区域，若少于 10 bp，则只取[－30, －20] bp 区域)以及终止子信号序列(上

游基因的终止位点到下游基因的起始位点之间的区域，若多于 100 bp，则取 100 bp，小于 20 bp，则只取 20 bp)，然后对这三类信号分别建立操纵子权重矩阵(operon weight matrix)和边界权重矩阵(boundary weight matrix)以及它们在操纵子基因对和边界基因对出现的频率．需要说明的是，作为数学模型，这里的翻译起始信号序列、启动子信号序列、终止子信号序列等只是从概率上描述了相应的调控信号及其区域．

再进一步，定义 *OSP*、*OPP*、*OTP* 分别为操纵子基因对出现翻译起始信号、启动子信号和终止子信号的概率，定义 *BSP*、*BPP*、*BTP* 分别为边界基因对出现翻译起始信号、启动子信号和终止子信号的概率，W_{sd}、W_{pr}、W_{tr} 分别为翻译起始信号、启动子信号和终止子信号的长度．另外，定义矩阵 $OSW_{i,j}$(i=1, 2, …, W_{sd}; j=1, 2, 3, 4)、$OPW_{i,j}$(i=1, 2, …, W_{pr}; j=1, 2, 3, 4)和 $OTW_{i,j}$(i=1, 2, …, W_{tr}; j=1, 2, 3, 4)分别为操纵子基因对翻译起始信号、启动子信号和终止子信号的权重矩阵．这里矩阵的行数表示信号的长度，行号为信号的相应位置，矩阵的列分别为 A、C、G、T 出现的概率．因此 $OSW_{i,j}$ 表示操纵子翻译信号第 i 个位置出现 j 列对应碱基的概率，其他矩阵依此类推．定义矩阵 $BSW_{i,j}$(i=1, 2, …, 10; j=1, 2, 3, 4)、$BPW_{i,j}$(i=1, 2, …, 10; j=1, 2, 3, 4)和 $BTW_{i,j}$(i=1, 2, …, 10; j=1, 2, 3, 4)分别为边界基因对翻译起始信号、启动子信号和终止子信号的权重矩阵．向量 BG_i(i=1, 2, 3, 4)为全基因组非编码区的 A、C、G、T 含量．模型中设定操纵子基因对可能有启动子信号和终止子信号是基于算法稳定的考虑．由于算法是根据相应信号区域的序列进行迭代自学习，上述设定并不影响操纵子基因对之间一般没有启动子和终止子的大多数情形．

设一对相邻基因对的翻译起始信号序列 S_{sd}、启动子信号序列 S_{pr}、终止子信号序列 S_{tr} 长度分别为 N_{sd}、N_{pr}、N_{tr}，进而假定三个信号在相应序列上的任意位置有相同的概率分布．由此可得到给定间距为 d 的操纵子基因对与非操纵子基因对出现 S_{sd}、S_{pr}、S_{tr} 的概率，如给定间距为 d 的非操纵子基因对出现终止子信号 S_{tr} 的概率为：

$$P(S_{tr}|nop, d)=1-BTP+BTP\times\frac{1}{N_{tr}-W_{tr}+1}\sum_{i=0}^{N_{tr}-W_{tr}}\prod_{j=1}^{W_{tr}}\frac{BTW_{j,k}}{BG_k} \quad (3)$$

其中，k 为对应信号序列(S_{sd}, S_{pr}, S_{tr})第 $i+j$ 个位置对应碱基的序号(A、C、G、T 分别对应 1、2、3、4)．类似地，可计算其他所有情况的概率．公式及推导详见网络版附录(http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20100686&flag=1)．

c．COG 功能注释信息．处于同一操纵子的基因通常被认为是功能相关或处于同一代谢通路．已有的研究表明，功能信息有助于提高操纵子预测的精度[4, 7]．鉴于已测序物种的 COG 功能信息较全面且易获取，对于有 COG 注释的基因组，本文算法自动添加其 COG 信息．因此，我们定义基因对(p_1, p_2)的 COG 打分函数为

$$cog=\begin{cases}1, (p_1, p_2)\text{的 COG 分类属于同一类}\\ 0, (p_1, p_2)\text{的 COG 分类不属于同一类}\end{cases} \quad (4)$$

进一步地，定义 *OCP*、*BCP* 分别为操纵子基因对、边界基因对 COG 分类属于同一类的概率．于是，可以计算下列概率：

$$P(cog=1|opr)=OCP,\ P(cog=0|opr)=1-OCP \quad (5)$$

$$P(cog=1|nop)=BCP,\ P(cog=0|nop)=1-BCP \quad (6)$$

1.3 迭代自学习算法

基于上述综合基因间距离、调控信号及 COG 信息等操纵子结构特征建立的数学模型，本文进一步设计了迭代自学习的优化算法．迭代自学习的算法在基因组结构信息的建模分析中具有很好的效果，尤其适合于处理对实验确认数据比较缺乏的信号分析[25, 27-28]．本文迭代自学习算法的目标是寻找最大化似然函数(1)的参数．对于一个任意的已知基因注释信息的基因组，整个算法采用如下步骤的最大期望值(expectation maximization，EM)迭代算法：

a．获得该基因组的全基因组序列和基因注释信息，提取所有同链上的相邻基因，得到全部的基因对的数据．

b．提取基因对的基因间距离 d、信号序列(S_{sd}, S_{pr}, S_{tr})及 COG 分类信息．

c．初始化参数：基因间距的参数通过 *E. coli* 实验确认的操纵子基因对的间距作 Logistic 回归得到 α=2.44, β=–0.0415；调控信号的 *OSP*、*OPP*、*OTP*、*BSP*、*BPP*、*BTP*，参数均为 0.5，而权重矩阵 *OSW*、*OPW*、*OTW*、*BSW*、*BPW*、*BTW* 则随机初始化．COG 的 *OCP*、*BCP* 初始化值为 0.5；这里初值的选取可以是任意的，EM 算法可以保证迭代收敛．虽然因为参数空间结构复杂，不同的初值有可能得到不同的收敛结果，但 EM 算法本质上是一种优化算法，对于不同初值得到的迭代结果，可以通过比较最终似然率的方法来进行选择．这里选取初始参数的原则主要是为了加快收敛速度．

d．E 步迭代：根据上述定义的概率模型，用已有参数值计算各个基因对属于(或不属于)同一操纵子的概率；属于(或不属于)同一操纵子并有(或没有)翻译信号的概率；属于(或不属于)同一操纵子并有(或没有)启动子信号的概率；属于(或不属于)同一操纵子并有(或没有)终止子信号的概率．注意这里计算的是条件概率．例如，一对基因对属于同一操纵子的概率是：

$$P(opr|F, \Omega)=\frac{P(opr, F|\Omega)}{P(F|\Omega)} \quad (7)$$

这里 F 代表所有特征的全体，$P(F|\Omega)$可以从(1)式得到，$P(opr, F|\Omega)$对应于(1)式的前半部分．而一对基因不属于同一操纵子并有终止子信号的概率为：

$$P(nop, TR|F, \Omega)=P(nop|F, \Omega)\times P(TR|nop, F, \Omega) \quad (8)$$

其中，

$$P(TR|nop, F, \Omega)=\frac{BTP\times\frac{1}{N_r-W_r+1}\sum_{i=0}^{N_r-W_r}\prod_{j=1}^{W_r}\frac{BTW_{j,k}}{BG_k}}{1-BTP+BTP\times\frac{1}{N_r-W_r+1}\sum_{i=0}^{N_r-W_r}\prod_{j=1}^{W_r}\frac{BTW_{j,k}}{BG_k}}$$

对应于(3)式．其余概率可类似得到．

e．M 步迭代：通过上一步得到的各个概率值重新估计参数．对于 d 采用 Logistic 回归，对于其他参数直接采用计数结果相除．例如，计算边界基因对含有终止子信号的概率 BTP 时，先将所有基因对不属于同一操纵子的概率相加得到边界基因对的总数，有：

$$Count(nop)=\sum P_n(nop|F, \Omega) \quad (9)$$

再将所有基因对不属于同一操纵子并有终止子信号的概率相加，得到所有含终止信号的边界基因对的个数．有：

$$Count(nop, TR)=\sum P_n(nop, TR|F, \Omega) \quad (10)$$

二者相除即得到边界基因对含有终止子信号的概率：

$$BTP=\frac{Count(nop, TR)}{Count(nop)} \quad (11)$$

其他概率可类似得到．

f．返回到步骤 c，重新初始化参数，直至达到收敛的迭代结果．比较所有迭代结果中对数似然率，对数似然率最高的迭代为最优结果．再由(12)计算出每一个基因对是属于同一个操纵子的概率．高于一定阈值的被判定为操纵子基因对．

2 结果与讨论

2.1 操纵子预测结果及评价

随着人们对原核基因组序列的复杂结构尤其是操纵子结构研究的不断深入，用于预测操纵子结构的算法也在不断发展．Brouwer 等[29]对当前已有的 30 种操纵子结构预测方法进行了系统地测试和比较，结果表明 Dam 等[7]在 2007 年发展的预测算法具有最好的预测水平．Dam 等发展的方法综合采用基因对间距、基因对在不同物种中的保守性、操纵子上下游序列信号、基因对的长度比等重要特征，由此设计了高效的模式识别算法．鉴于此，我们选择 Dam 等方法与本文方法进行预测水平的比较．

为评价算法的预测能力，人们通常使用敏感性(sensitivity，SN)和特异性(specificity，SP)两个评价指标．对于某一算法的判别结果，SN 和 SP 是此增彼减、互为补充的关系，简单地比较其中某一个指标难以客观全面地评价算法的水平．为更全面地评价操纵子预测的能力，本文引入当前广泛使用指标(accuracy，AC)来衡量预测的精度，它综合反映了算法在 SN 和 SP 两个指标上的总体水平[7]：

$$AC=\frac{Tp+Tn}{Wo+Tub} \quad (12)$$

其中，Tp 是预测正确的操纵子基因对个数，Tn 是预测正确的操纵子边界基因对个数，Wo 是测试集中全部操纵子基因对的个数，Tub 则为测试集中全部操纵子边界基因对的个数．

我们首先比较两种方法对本文构建的 7 个物种测试集的预测精度．对于测试集所涉及的 7 个基因组及其基因注释，本文方法的预测过程如前所示．概括地说，算法不依赖于任何已知的操纵子结构信息作为训练集，通过基于最大期望值的迭代自学习，最终获得符合迭代条件的最大化似然函数的各参数，同时对所有基因对是否属于操纵子基因对进行判别．对 Dam 等[30]方法的结果由作者根据计算预测而建立的 DOOR 数据库获得，该方法依赖于已有的操纵子数据作为训练集，通过机器学习得到模型的参数[7]．两种方法对基因对的预测结果与 7 个物种的测试集中的基因对注释分别进行比较，即可得到各自的预测精度．比较的结果表明，本文算法的敏感性 SN 高于 Dam 等方法，而特异性 SP 低于 Dam 等方法，综合的预测精度 AC 要高于 Dam 等方法(本文算法对 7 个物种的平均预测精度为

83.3%，Dam 等方法为 82.2%)，如表 1 所示(敏感性 *SN* 和特异性 *SP* 的结果数据详见网络版附录(http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20100686&flag=1)的表 S1)．显然，本文方法的预测精度在总体上高于 Dam 等方法．

应该指出的是，与 Dam 等方法相比，本文方法的最大优越性在于没有依赖任何物种的已知操纵子结构信息，而 Dam 等方法是以已知的操纵子数据作为训练集来得到其模型的参数，进而用于其他物种基因对的判别[7]．该方法的训练集来自 *E. coli* 和 *B. subtilis* 两个物种，其中 *E. coli* 的操纵子数据来源于 RegulonDB[16]，共 707 对操纵子基因对和 497 对边界基因对，*B. subtilis* 的操纵子数据来源于 DBTBS[17]，共 850 对操纵子基因对和 775 对边界基因对．由于来源数据库的完全一致及选取方法的类似，Dam 等的训练集数据与表 1 的测试集有大量的重合，其预测精度自然会被高估．因此，表 1 所示的本文方法相对于 Dam 等方法的总体优势是被低估的．

Table 1　Comparison of prediction accuracies (*AC*) between Dam′s method and our algorithm

Species	*E. coli*	*B. subtilis*	*P. aeruginosa*	*S. coelicolor*	*M. pneumonia*	*S. solfataricus*	*G. sulfurreducens*	Average
Dam′s method	84.3%	82.7%	84.0%	85.0%	84.1%	79.7%	75.3%	82.2%
Our algorithm	85.3%	85.7%	84.9%	85.0%	84.2%	79.0%	79.1%	83.3%

实际上，不依赖于学习集的自学习方法不仅在算法上具有优越性，也能更合理地反映和刻画不同物种操纵子结构的特异性．为说明这一点，我们进一步讨论本文算法在自学习和依赖特定物种学习等两种情况下的操纵子预测水平．Dam 等将 *B. subtilis* 或 *E. coli* 物种的操纵子数据训练得到的参数模型运用于其他基因组的操纵子结构预测，进而构建 DOOR 数据库[30]．为此，本文也用 *B. subtilis* 自学习得到的操纵子模型(简称“*B. subtilis* model”)来预测其他 6 个物种，并根据测试集计算预测精度．比较结果表明(表 2)，根据各物种自身进行自学习的预测水平要显著高于 *B. subtilis* model 的预测水平，这一优势对于 *S. solfataricus*、*S. coelicolor* 以及 *M. pneumonia* 等物种尤其明显．以 *S. coelicolor* 为例，*B. subtilis* 在系统分类上属于变形菌门(*Proteobacteria*)，*S. coelicolor* 属于放线菌门(*Actinobacteria*)，两者在进化距离上相隔很远．比较 *B. subtilis* 和 *S. coelicolor* 通过自学习迭代得到的操纵子上游调控区的信号，二者表现出显著的差别(图 2)．本文还考察了二者的操纵子下游调控区信号(图 3)，显然，*B. subtilis* 具有典型的 ρ-independent 终止子信号，而 *S. coelicolor* 的转录终止机制尚不清楚，但与 ρ-independent 的终止子信号明显不同．信号的比较分析表明，*B. subtilis* 和 *S. coelicolor* 的操纵子结构具有较大差异．同样 *S. solfataricus* 属于古细菌，它与 *B. subtilis* 在进化距离上更加遥远，而转录的起始、终止调控信号及基因对间距分布都有很大差异．在建立数学模型和设计预测算法过程中，应该充分考虑到这些物种特异性．

Table 2　Comparison of prediction accuracies (*AC*) of our algorithm between the model by self-learning method and the *B. subtilis* model

Species	*E. coli*	*P. aeruginosa*	*S. coelicolor*	*M. pneumonia*	*S. solfataricus*	*G. sulfurreducens*	Average
Self-learning	85.3%	84.9%	85.0%	84.2%	79.0%	79.1%	82.9%
B. subtilis model	83.1%	82.1%	80.7%	80.4%	65.1%	78.7%	78.4%

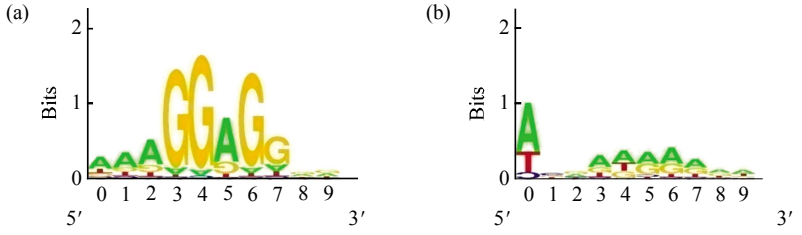

Fig. 2　The logos of operon-upstream regulation signals for (a) *B. subtilis* and (b) *S. coelicolor*

The operon annotations for both genomes are calculated by our algorithm with self-learning strategy.

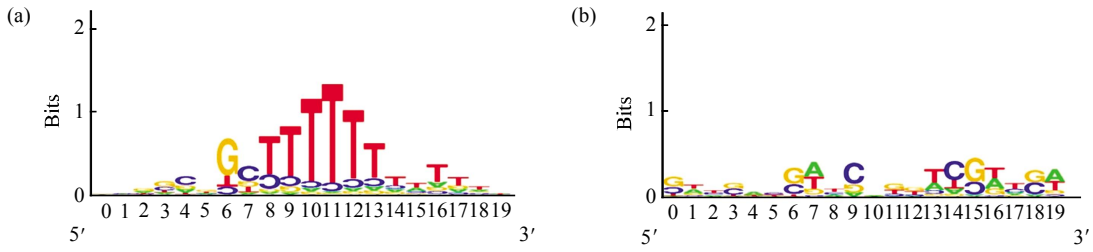

Fig. 3　The logos of operon-downstream regulation signals for (a) *B. subtilis* and (b) *S. coelicolor*

The operon annotations for both genomes are calculated by our algorithm with self-learning strategy.

应该说明的是，在本文图 2 及其他所有图中的信号 logo 图并非将序列以基因的 5′端或 3′端对齐得到的一致保守序列，而是通过本文的迭代寻优算法，在调控区得到的某一宽度窗口的相对位置权重矩阵，横坐标是该窗口内的核苷酸位置.

2.2　操纵子模型参数分析

通过对测试结果的分析，可以看出本文发展的不依赖于学习集的迭代自学习方法具有很好的操纵子结构预测水平，而模型对物种特异性的刻画是本文方法能够高效预测操纵子结构的关键. 更重要的是，由此可以得到某一基因组的一套具有物种特异性的模型参数，对这些模型参数的进一步分析，将有助于对原核生物操纵子结构的普遍特征和多样性的理解.

我们首先考察反映基因间距离的相关参数. 如前所述，基因间距离被认为是刻画操纵子结构的最有效特征[4, 7]. 在我们分析的 762 个基因组中，操纵子基因对的基因间距离都表现出普遍一致的紧凑分布趋势. 根据迭代结束时的 α、β 值进行分析，99.6%的物种都是 $\beta<0$，这说明绝大多数原核基因组的操纵子内部的相邻基因保持了一个普遍的规律，即间距越大，属于同一操纵子的可能性就越小. 在本文模型中的临界距离 $d_{crt}=-\alpha/\beta$ 是某一基因组从概率上判别操纵子基因对与边界基因对的重要参数，我们也统计了 762 个物种 d_{crt} 值的分布，如图 4 所示(平均值 82 bp，标准差 19 bp). 由此可见，大部分物种构成操纵子的相邻基因之间的距离临界值集中在 80 bp，但是不同基因组仍然有一定的差异.

操纵子结构的特征还反映在操纵子上下游以及基因上下游调控区的各类信号. 本文基于对 762 个

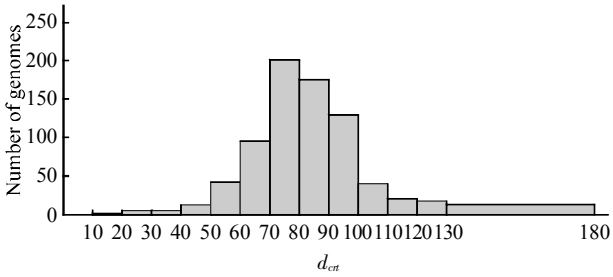

Fig. 4　Distribution of critical distance d_{crt} of 762 genomes

原核基因组的操纵子预测结果，进一步对这些信号进行分析，得到了一些新的认识. 比如，在很多物种特别是原壁菌(*Firmicutes*)、变形杆菌(*Proteobacteria*)中的操纵子下游转录终止调控区，得到的信号具有典型的 ρ-independent 终止子的高 GC +T 富含特征(图 5a)；而古细菌和细菌 *Actinobacteria* 所对应的终止信号既不含此特征也非 ρ-dependent 的终止子，但却有显著的特征，如图 5b、c 所示. 前面对 *S. coelicolor* 的操纵子下游转录终止子的分析也属于这一情况(图 3). 尽管目前还缺乏对这一类转录终止信号的实验研究证据，但由此可见转录终止还有可能存在未被人们认识的机制. 在操纵子上游的调控区，转录启动子是十分重要的调控信号. 但是，由于启动子本身的多样性和序列非高度保守性，在转录启始位点难以明确的情况下，只通过对齐翻译起始位点的上游区进行分析是很难发现转录启动子的. 有趣的是，我们的算法在部分物种(既有细菌，也有古细菌)的操纵子上游调控区可以发现十分显著的启动子 TATA 盒信号.

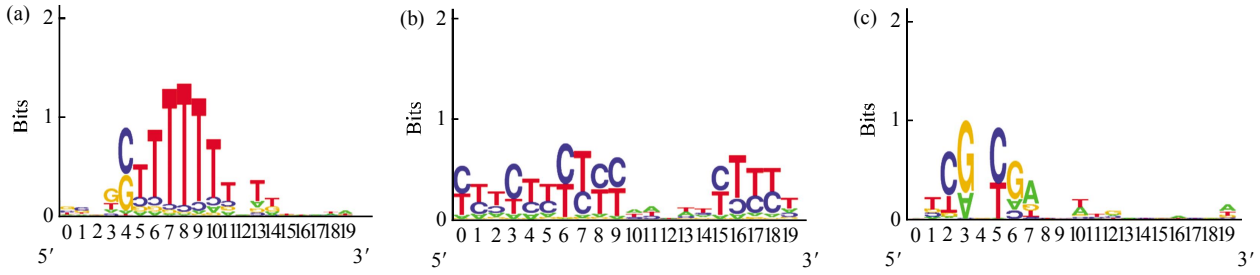

Fig. 5　Transcription terminators logo in the operon downstream

(a) *Firmicutes* and *Proteobacteria*. (b) Archaea. (c) *Actinobacteria*. Where, *Firmicutes* and *Proteobacteria* have typical ρ-independent; while signals of Archaea and *Actinobacteria* are quite different.

本文在对基因组的全部基因对建立数学模型中引入了蛋白质 COG 功能信息．诸多的研究表明，操纵子结构确实在很大程度上与蛋白质功能相联系．通过大规模的预测结果分析，我们进一步给出了这一联系的定量描述．在本文预测的 762个原核基因组中，各物种的操纵子基因对属于同一 COG 类的比例平均达到 43.5%(标准差 7.2%)，而边界基因对属于同一 COG 类的比例平均为 14.5%(标准差为 3.9%)．由此可见，操纵子基因对在功能上的关联要显著高于边界基因对．同时，我们也对操纵子结构在基因功能上的含义作出了一个具体的估计．

综上所述，通过对本文算法所预测的各基因组操纵子模型参数进行大规模比较分析，进一步提高了对原核生物基因组中操纵子结构的普遍特征和物种特异性的认识．概括地说：a．构成操纵子的基因间距离相对紧凑的这一特点普遍适用于各基因组，但仍表现出一定范围的分布；b．操纵子和构成操纵子基因的上下游调控信号主要与转录和翻译过程相关，这些调控信号表现了很复杂的物种特异性；c．构成操纵子的基因在功能上的相关是原核生物的一个重要特征，对这种功能注释信息的关注有助于设计高效的操纵子预测算法．

3　结　　论

操纵子结构的预测研究是当前原核生物基因组学的一个挑战性难题．我们基于对操纵子结构的认识，从转录相关的调控信号、基因间距离以及 COG 注释等特征出发，建立描述操纵子复杂结构的概率模型，并提出不依赖于训练集的迭代自学习算法．通过对实验验证的操纵子数据集的测试比较，结果表明本文的算法对于预测操纵子结构非常有效．在不依赖于任何已知操纵子信息的情况下，我们的算法在总体预测水平上超过了目前最好的操纵子预测方法．我们还证明了这种自学习的预测算法要优于依赖特定物种进行训练的预测方法．这些特点使得本文算法能够适用于任意新测序的物种(算法可根据 COG 注释与否采用或者不采用 COG)，有助于发展高效的基因组操纵子自动注释工具，对正在快速推进的原核生物基因组测序计划及其基因组学研究具有十分重要的意义．我们将在后续的论文中报告进一步的工作结果.

在本文的高效预测结果基础上，我们得以对细菌和古细菌的基因组进行大规模比较分析．这一系统性的研究进一步提高了对原核生物基因组操纵子结构的普遍特征和物种特异性的认识，为原核生物的比较基因组学和生物进化研究提供了新的思路．尤其值得关注的是操纵子和构成操纵子基因的上下游调控信号，这些调控信号主要与转录和翻译过程相关，表现了很复杂的物种特异性，可能有目前还未被人们认识到的新机制，对它们的深入研究将有利于从转录机理上认识操纵子的结构，可能是今后比较基因组学的重要研究内容．构成操纵子的基因在功能上的相关是原核生物的一个重要特征，对这种功能注释信息的关注不仅有助于设计高效的操纵子预测算法，也可以为功能基因组学的基因功能分析和基因调控网络研究进一步提供线索.

致谢　感谢胡钢清博士、曲姣、李冠辰、谭验等的有益讨论.

参　考　文　献

[1]　Overbeek R, Fonstein M, D′souza M, *et al*. The use of gene clusters

to infer functional coupling. Proc Natl Acad Sci USA, 1999, **96**(6): 2896-2901

[2] Zheng Y, Szustakowski J D, Fortnow L, *et al*. Computational identification of operons in microbial genomes. Genome Res, 2002, **12**(8): 1221-1230

[3] Sorek R, Cossart P. Prokaryotic transcriptomics: A new view on regulation, physiology and pathogenicity. Nat Rev Genet, **11**(1): 9-16

[4] Salgado H, Moreno-hagelsieb G, Smith T F, *et al*. Operons in *Escherichia coli*: Genomic analyses and predictions. Proc Natl Acad Sci USA, 2000, **97**(12): 6652-6657

[5] Charaniya S, Mehra S, Lian W, *et al*. Transcriptome dynamics-based operon prediction and verification in *Streptomyces* coelicolor. Nucl Acid Res, 2007, **35**(21): 7222-7236

[6] Bockhorst J, Craven M, Page D, *et al*. A Bayesian network approach to operon prediction. Bioinformatics, 2003, **19** (10): 1227-1235

[7] Dam P, Olman V, Harris K, *et al*. Operon prediction using both genome-specific and general genomic information. Nucl Acid Res, 2007, **35**(1): 288-298

[8] Price M N, Huang K H, Alm E J, *et al*. A novel method for accurate operon predictions in all sequenced prokaryotes. Nucl Acid Res, 2005, **33**(3): 880-892

[9] Ermolaeva M D, White O, Salzberg S L. Prediction of operons in microbial genomes. Nucleic Acids Res, 2001, 29(5): 1216-1221.

[10] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, **25**(1): 25-29

[11] Tran T T, Dam P, Su Z, *et al*. Operon prediction in Pyrococcus furiosus. Nucl Acid Res, 2007, **35**(1): 11-20

[12] Yada T, Nakao M, Totoki Y, *et al*. Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. Bioinformatics, 1999, **15**(12): 987-993

[13] Janga S C, Lamboy W F, Huerta A M, *et al*. The distinctive signatures of promoter regions and operon junctions across prokaryotes. Nucl Acid Res, 2006, **34**(14): 3980-3987

[14] Westover B P, Buhler J D, Sonnenburg J L, *et al*. Operon prediction without a training set. Bioinformatics, 2005, **21**(7): 880-888

[15] Rogozin I B, Makarova K S, Natale D A, *et al*. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucl Acid Res, 2002, **30**(19): 4264-4271

[16] Salgado H, Gama-castro S, Peralta-gil M, *et al*. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucl Acid Res, 2006, **34**(Database issue): D394-D397

[17] Sierro N, Makita Y, De Hoon M, *et al*. DBTBS: A database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. Nucl Acid Res, 2008, **36**(Database issue): D93-D96

[18] Okuda S, Katayama T, Kawashima S, *et al*. ODB: A database of operons accumulating known operons across multiple genomes. Nucl Acid Res, 2006, **34**(Database issue): D358-D362

[19] Guell M, Van Noort V, Yus E, *et al*. Transcriptome complexity in a genome-reduced bacterium. Science, 2009, **326**(5957): 1268-1271

[20] Wurtzel O, Sapra R, Chen F, *et al*. A single-base resolution map of an archaeal transcriptome. Genome Res, **20**(1): 133-141

[21] Qiu Y, Cho B K, Park Y S, *et al*. Structural and operational complexity of the Geobacter sulfurreducens genome. Genome Res, **20**(9): 1304-1311

[22] Okuda S, Kawashima S, Kobayashi K, *et al*. Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. BMC Genomics, 2007, **8**(1): 48-59

[23] Lesnik E A, Sampath R, Levene H B, *et al*. Prediction of rho-independent transcriptional terminators in *Escherichia coli*. Nucl Acid Res, 2001, **29**(17): 3583-3594

[24] Zhu H Q, Hu G Q, Yang Y F, *et al*. MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. BMC Bioinformatics, 2007, **8**: 97-110

[25] 胡钢清，刘永初，郑晓斌，等. 原核基因翻译起始点位预测的新方法，生物化学与生物物理进展，2008, **35**(11): 1254-1262
Hu G Q, Liu Y C, Meng X B, *et al*. Prog Biochem Biophys, 2008, **35**(11): 1254-1262

[26] Hu G Q, Zheng X B, Yang Y F, *et al*. ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes. Nucl Acid Res, 2008, **36**(Database issue): D114-D119

[27] Zhu H Q, Hu G Q, Ouyang Z Q, *et al*. Accuracy improvement for identifying translation initiation sites in microbial genomes. Bioinformatics, 2004, **20**(18): 3308-3317

[28] Sun Z X, Sang L J, Ju L N, *et al*. A new method for splice site prediction based on the sequence patterns of splicing signals and regulatory elements. Chin Sci Bull, 2008, **53**(21): 3331-3340

[29] Brouwer R W, Kuipers O P, Van Hijum S A. The relative value of operon predictions. Brief Bioinform, 2008, **9**(5): 367-375

[30] Mao F, Dam P, Chou J, *et al*. DOOR: A database for prokaryotic operons. Nucl Acid Res, 2009, **37**(Database issue): D459-D463

Operon Prediction Based On an Iterative Self-learning Algorithm*

WU Wen-Qi**, ZHENG Xiao-Bin**, LIU Yong-Chu, TANG Kai, ZHU Huai-Qiu***

(*State Key Laboratory for Turbulence and Complex Systems and Department of Biomedical Engineering, College of Engineering, Center for Theoretical Biology, Center for Protein Science, Peking University, Beijing* 100871, China)

Abstract As a specific functional organization of genes in prokaryotic genomes, operon contains a set of adjacent genes under the control of the corresponding regulatory signals, and is expressed as the transcript unit. It has been found that genes in an operon usually tend to have related functions, or belong to the same pathway in cell. Therefore the study of operon structure is significant to understand the gene functions and regulatory networks for prokaryotes. However with the current limitation of data acquisition of operons verified by experiments such as prokaryotic transcriptomics, computation methods to annotate the operons in a newly sequenced genome have so far been the major source of operon data, and will continue to be an important mission. Over the past decade, a set of computational approaches to operon prediction have been proposed, however mainly based on experimental operons as their training sets. Nevertheless the lack of experimental operon dataset has been the bottleneck of operon prediction. The authors employ an iterative self-learning algorithm which is independent of training set with known operon dataset. The algorithm develops based on a probabilistic model using features including gene distance, regulation signals of gene expression and functional annotation such as COG. The test result compared against the experimental operon data indicates that the algorithm can reach the best accuracy without any training set. Besides, this self-learning algorithm is superior to the algorithm trained on any species with known operons. Accordingly, the algorithm can be applied to any newly sequenced genome. Moreover, comparative analysis of bacteria and archaea enhances the knowledge of universal and genome specific features of operons.

Key words genome analysis, operon, iterative self-learning, evaluation of prediction

DOI: 10.3724/SP.J.1206.2010.00686

---

*This work was supported by grants from The National Natural Science Foundation of China (30970667, 30770499, 10721403), The MOST Project of China (2009ZX09501-002), The Excellent Doctoral Dissertation Supervisor Project of Beijing (YB20101000102), and National Basic Research Program of China (2011CB707500).

**These authors contributed equally to this work.

***Corresponding author.

Tel: 86-10-62767261, E-mail: hqzhu@pku.edu.cn

Received: December 28, 2010 Accepted: April 22, 2011

附　录

1　操纵子调控信号的概率模型及其推导

给定间距为 d 的操纵子基因对与非操纵子基因对，分别出现翻译起始信号序列 S_{sd}、启动子信号序列 S_{pr} 和终止子信号序列 S_{tr} 的概率如下：

$$P(S_{sd}|opr,\ d)=1-OSP+OSP\times\frac{1}{N_{sd}-W_{sd}+1}\sum_{i=0}^{N_{sd}-W_{sd}}\prod_{j=1}^{W_{sd}}\frac{OSW_{j,k}}{BG_k},\quad (S1)$$

$$P(S_{pr}|opr,\ d)=1-OPP+OPP\times\frac{1}{N_{pr}-W_{pr}+1}\sum_{i=0}^{N_{pr}-W_{pr}}\prod_{j=1}^{W_{pr}}\frac{OPW_{j,k}}{BG_k},\quad (S2)$$

$$P(S_{tr}|opr,\ d)=1-OTP+OTP\times\frac{1}{N_{tr}-W_{tr}+1}\sum_{i=0}^{N_{tr}-W_{tr}}\prod_{j=1}^{W_{tr}}\frac{OTW_{j,k}}{BG_k},\quad (S3)$$

$$P(S_{sd}|nop,\ d)=1-BSP+BSP\times\frac{1}{N_{sd}-W_{sd}+1}\sum_{i=0}^{N_{sd}-W_{sd}}\prod_{j=1}^{W_{sd}}\frac{BSW_{j,k}}{BG_k},\quad (S4)$$

$$P(S_{pr}|nop,\ d)=1-BPP+BPP\times\frac{1}{N_{pr}-W_{pr}+1}\sum_{i=0}^{N_{pr}-W_{pr}}\prod_{j=1}^{W_{pr}}\frac{BPW_{j,k}}{BG_k},\quad (S5)$$

其中，k 为对应信号序列(S_{sd}, S_{pr}, S_{tr})第 $i+j$ 个位置对应碱基的序号(A、C、G、T 分别对应 1、2、3、4)，N_{sd}、N_{pr}、N_{tr} 为 S_{sd}、S_{pr}、S_{tr} 的长度．变量 OSP 及矩阵 $OSW_{i,j}$ 等的定义参见正文．

以计算 $P(S_{sd}|opr,\ d)$为例来说明上述公式详细推导过程．这里，$P(S_{sd}|opr,\ d)$为给定间距为 d 的操纵子基因对出现翻译起始信号序列 S_{sd} 的概率，可分为两部分：序列 S_{sd} 不出现翻译起始信号的概率 $P_{non\text{-}sd}$ 及出现翻译起始信号的概率 P_{sd}．其中，

$$P_{non\text{-}sd}=1-OSP,\quad (S6)$$

$$P_{sd}=OSP\times\frac{1}{N_{sd}-W_{sd}+1}\sum_{i=0}^{N_{sd}-W_{sd}}\prod_{j=1}^{W_{sd}}\frac{OSW_{j,k}}{BG_k},\quad (S7)$$

当序列 S_{sd} 出现翻译起始信号时，将 S_{sd} 按翻译起始信号的长度 W_{sd} 依次划分成 $N_{sd}-W_{sd}+1$ 个片段．由于假定信号在相应序列上的任意位置有相同的概率分布，故每个片段出现翻译起始信号的概率为 $OSP\times P_i/(N_{sd}-W_{sd}+1)$．其中，$P_i$ 指第 i 个片段翻译起始信号的强度，即相应的权重矩阵与背景概率之比$\prod_{j=1}^{W_{sd}}\frac{OSW_{j,k}}{BG_k}$．最终，$P_{sd}$ 即为 $N_{sd}-W_{sd}+1$ 个片段出现翻译起始信号的概率之和，即 S7 式．$P(S_{sd}|opr,\ d)$即为(S6)、(S7)式的求和．

类似地，可以计算得到(S2)、(S3)、(S4)、(S5)式．

2　本文算法与 Dam 等方法的预测精度的详细比较

两种方法对基因对的预测结果与 7 个物种测试集中的基因对注释分别进行比较，即可得到各自的预测精度，如表 S1 所示．参数敏感性 (sensitivity, SN)、特异性(specificity, SP)的定义如下：

$$SN=\frac{Tn}{Tub},\quad SP=\frac{Tp}{Wo}.$$

其中，Tp 是预测正确的操纵子基因对个数，Tn 是预测正确的操纵子边界基因对个数，Wo 是测试集中全部操纵子基因对的个数，Tub 则为测试集中全部操纵子边界基因对的个数．参数 AC(accuracy)的定义参见正文．

Table S1　The prediction comparison between Dam′s method and our algorithm

	SN		*SP*		*AC*	
	Dam′s method	Our algorithm	Dam′s method	Our algorithm	Dam′s method	Our algorithm
E. coli	90.1%	89.8%	75.8%	78.8%	84.3%	85.3%
B. subtilis	73.4%	82.8%	95.6%	89.6%	82.7%	85.7%
P. aeruginosa	92.5%	94.0%	69.2%	69.2%	84.0%	84.9%
S. coelicolor	85.6%	87.0%	83.6%	80.3%	85.0%	85.0%
M. pneumonia	87.5%	86.3%	77.9%	80.4%	84.1%	84.2%
S. solfataricus	66.4%	88.8%	95.7%	70.8%	79.7%	79.0%
G. sulfurreducens	72.7%	85.9%	82.1%	61.1%	75.3%	79.1%
Average	81.2%	87.8%	82.8%	75.7%	82.2%	83.3%