(2014) 158–167

主页：www.intl.elsevierhealth.com/journals/cmpb

基于序列的DNA中蛋白质结合位点的预测：两种SVM模型的比较研究

Byungkyu公园[a], *Jinyong Im*[b], *Narankhuu Tuvshinjargal*[b], *Wook Lee*[b], *Kyungsook韩*[b,*]

[a]*韩国仁川仁荷大学信息与电子研究所*
[b]*韩国仁川仁荷大学计算机科学与工程系*

文章信息

文章历史：
收到2013年12月7日收到修改后的表格2014年7月17日
2014年7月18日接受

关键词：
DNA-蛋白质相互作用结合位点
蛋白质结合核苷酸预测模型

摘要

由于蛋白质-DNA复合物的许多结构在过去几年中已知，已经开发了几种计算方法来预测蛋白质中的DNA结合位点。 然而，其反向问题（即预测DNA中的蛋白质结合位点）受到的关注较少。原因之一是核苷酸的相互作用倾向之间的差异远小于氨基酸之间的相互作用倾向。 另一个原因是DNA表现出比蛋白质更少的多样性序列模式。 因此，预测蛋白质结合DNA核苷酸比预测DNA结合氨基酸要困难得多。 我们使用蛋白质-DNA复合物的广泛数据集计算了核苷酸三联体与氨基酸的相互作用倾向（IP），并开发了两种支持向量机（SVM）模型，其仅通过序列数据预测蛋白质结合核苷酸。 一个SVM模型仅使用DNA序列数据预测蛋白质结合核苷酸，另一个SVM模型使用DNA和蛋白质序列预测蛋白质结合核苷酸。 在使用1519个DNA序列的10倍交叉验证中，使用DNA序列数据的SVM模型仅以67.0%的准确度，67.1%的F-度量和Matthews相关系数（MCC）预测蛋白质结合核苷酸，为0.340。 对于未用于培训的181种DNA的独立数据集，其准确率为66.2%，F-measure为66.3%，MCC为0.324。 使用DNA和蛋白质序列的另一种SVM模型在1519个DNA序列和859个蛋白质序列的10倍交叉验证中达到69.6%的准确度，69.6%的F-测量值和0.383的MCC。 使用181个DNA和143种蛋白质的独立数据集，其显示出67.3%的准确度，66.5%的F-测量值和0.329的MCC。 在交叉验证和独立测试中，使用DNA和蛋白质序列数据的第二个SVM模型显示比使用DNA序列数据的第一个模型更好的性能。 据我们所知，这是首次尝试从单独的序列数据中预测给定DNA序列中的蛋白质结合核苷酸。

*通讯作者。 电话：+82 32 860 7388；传真：+82 32 863 4386.电邮地址：khan@inha.ac.kr（K. Han）。

http://dx.doi.org/10.1016/j.cmpb.2014.07.009
0169-2607 /

1. 介绍

DNA和蛋白质之间的相互作用在许多细胞过程中起着基础作用[1]. 例如，结合DNA特定区域的蛋白质通过激活或抑制DNA的基因表达而充当转录因子。 因此，识别DNA中的蛋白质识别部分或蛋白质中的DNA识别部分将有助于理解各种细胞过程[2,3]. 蛋白质-DNA相互作用已经在一些理论或实验研究中进行了研究，但蛋白质-DNA相互作用的机制尚未完全了解。 许多关于蛋白质-DNA相互作用的研究集中在研究转录因子在其DNA靶标中的结合偏好[4–7]. 在一些研究中已经使用机器学习方法预测蛋白质序列中的结合残基。 BindN [3] 使用SVM根据氨基酸的化学性质预测蛋白质序列中的RNA或DNA结合残基。 DNABindR[8] 使用朴素贝叶斯分类器来预测蛋白质中的DNA结合残基和DP-Bind[9] 使用具有位置特异性评分矩阵（PSSM）和氨基酸特性的SVM来预测蛋白质中的DNA结合残基。 最近发布的Web服务器称为PiDNA[10] 计算给定结构的蛋白质-DNA复合物的位置权重矩阵（PWM），并使用具有小能量变化的结构模型所建议的PWM来预测蛋白质-DNA相互作用。 然而，基于结构的方法如PiDNA不能用于预测蛋白质和DNA与未知结构的相互作用。

虽然有几项关于预测蛋白质中DNA结合残基的研究，但很少有尝试仅使用序列数据来预测DNA序列中的蛋白质结合核苷酸。其原因之一是核苷酸在蛋白质-DNA相互作用中显示出比氨基酸少得多的相互作用倾向。 另一个原因是DNA中只有四种类型的核苷酸，而蛋白质中有20种类型的氨基酸。 对于长度为n的序列，DNA具有4个"可能的序列模式，但蛋白质具有20个"可能的序列模式，这是5倍"倍。 因此，预测蛋白质结合核苷酸比预测DNA结合氨基酸更困难。

在这项研究中，我们分析了蛋白质-DNA复合物的最近结构，并使用支持向量机（SVM）计算了核苷酸三联体与氨基酸之间的相互作用倾向。 我们建立了两个使用相互作用倾向预测蛋白质结合位点的支持向量机模型

DNA序列。 一个SVM模型预测单独来自DNA序列数据的蛋白质结合核苷酸，另一个预测

SVM模型预测来自DNA的蛋白质结合核苷酸和蛋白质序列数据。 两种模型都采用单一的DNA

序列作为输入，所以应给予双链DNA

该模型分别作为单链DNA。 如本文后面所示，具有广泛数据集的实验结果表明，使用DNA和蛋白质序列的SVM模型比单独使用DNA序列的模型在交叉验证和独立测试中表现出更好的性能。 本文的其余部分介绍了我们的方法和实验结果的细节。

2. 材料和方法

2.1. 对绑定网站的定义

从核酸 – 蛋白质相互作用数据库（NPIDB）中提取蛋白质-DNA相互作用数据，[11]. 在由NPIDB提供的三种类型的核酸 – 蛋白质相互作用中，我们使用氢键（H键）和水桥。 NPIDB的疏水相互作用数据不用于两个SVM模型的比较分析，因为一个SVM模型不能匹配疏水性簇内的疏水性核酸 – 蛋白质相互作用[12].

对于蛋白质-DNA结合位点，将两种类型的DNA-蛋白质相互作用结合到每个蛋白质-DNA复合物序列中。 例如，在10MH A4（腺嘌呤在第4位）的DNA链中，G5，G7，C8和A9参与与蛋白质的相互作用。 核苷酸A4，G5，G7和C8通过H键相互作用和水桥与氨基酸结合，而A9仅通过H键相互作用与氨基酸结合。 在下面的例子中，结合核苷酸用符号'+'标记，非结合核苷酸用符号' – '标记。

	01	02	03	04	05	06	07	08	09	10	11	12
DNA sequence:	G	T	C	A	G	C	G	C	A	T	G	G
H bonds:	-	-	-	+	+	-	+	+	+	-	-	-
Water bridges:	-	-	-	+	+	-	+	+	-	-	-	-
Binding sites:	-	-	-	+	+	-	+	+	+	-	-	-

2.2. 核苷酸三联体的相互作用倾向

对于核苷酸三联体的相互作用倾向（IP），我们计算了三联体中的中间核苷酸与氨基酸相互作用的倾向。 在下面的例子中，蛋白质序列中第10个氨基酸（E10）的DNA序列中的第5个核苷酸（C5）的IP代表

C在三联体ACA中与E结合的优选性

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17
DNA sequence:	C	T	C	A	C	A	C	G	T	G	G	G	A	C	T	A	G
Protein sequence:	M	K	R	E	S	H	K	H	A	E	Q	A	R	R	N	R	L

Binding pair: (C-5, E-10)

核苷酸三联体t与氨基酸a之间的相互作用倾向IP_{TA}由方程(1).

$$IP_{TA} = \frac{1}{\angle} \cdot \frac{N_{I=三重峰}\; N_{i} \bullet_{j=AA} N_j}{H\ Acos\ (DAH)\quad Nt \cdot Na \cdot_{I=三重峰,\ J=AA} N_{IJ}} \quad 绑定对 \tag{1}$$

在等式(1), ∠DAH是供体 - 受体 - 氢 (D-A-H) 角。 HαAcos (∠DAH) 是受体 (H-A) 在供体 - 受体 (D-A) 上的投影长度，$N_{i,j}$是结合的核苷酸三联体的总数到任何氨基酸，N_t是核苷酸的总数

到任何氨基酸，N_t是核苷酸的总数

N_a是氨基酸总数，N_t是核苷酸三联体t的数目，N_a是数据集中氨基酸a的数目。 使用H-A投影长度的原因是考虑H-A距离及其与D-A的关系。 有4个^3= 64个核苷酸三联体和20个氨基酸，因此我们计算了核苷酸三联体和氨基酸之间的1280个IP。

2.3. 特征矢量

使用三种类型的特征来预测蛋白质结合核苷酸： (1) DNA序列的全局特征， (2) 核苷酸的局部特征，以及 (3) 相互作用配偶体的特征 (即蛋白质特征)。 全局特征包含

整个DNA序列信息：序列长度 (L) 和核苷酸组成 (C)。 核苷酸组成包括腺嘌呤，胞嘧啶，鸟嘌呤和鸟嘌呤的数目

目标DNA序列中的胸腺嘧啶。 因此，特征向量对于DNA序列长度总是具有一个元素

四个元素为核苷酸组成作为全局特征。

局部特征包含DNA序列中核苷酸的性质：分子量 (M)，核苷酸pK_a (P) 和具有20个氨基酸的核苷酸三联体IP。 滑动窗口的第一个和最后一个核苷酸不形成核苷酸三联体，因此它们的IP值在特征向量中被设置为0。

合作伙伴特征 (A) 包含相互作用蛋白质序列的信息。 这个特征是由方程式计算的。(2) 和(3) 并表示为特征矢量中的20种氨基酸的20个元素。

$$P_{b\in\{20\text{氨基酸}\}} = \frac{\text{序列长度}}{I,\ \text{双向}=B}\ \text{归一化位置 (bi)} \tag{2}$$

$$\text{标准化头寸 (i)} = \frac{\text{职位 (i)}}{\text{序列长度}} \tag{3}$$

如前所述，我们建立了两个支持向量机 (SVM) 模型来仅根据序列数据预测蛋白质结合核苷酸。 对于仅使用DNA序列数据预测蛋白质结合核苷酸的SVM模型，蛋白质特征 (特征A) 不包括在其特征向量中。 对于预测蛋白质结合的另一种SVM模型

使用DNA和蛋白质序列的核苷酸，
特征特征 (特征A) 被编码在其特征向量中。

当我们使用9个核苷酸的窗口在特征向量中包含蛋白质特征时，特征向量包含从T (i-4) 到T (i + 4) 的9个核苷酸三联体 (参见图。1对于特征向量的结构)。 对于给定的DNA和蛋白质序列对，编码5个全局特征元素 (L，CA，CC，CG，CT) 和20个蛋白质伴侣元素 (A_R，...，A_Y)。 对于9个核苷酸编码22个局部特征元素 (M，P和20个IP) (对于T (i-4) 和T (i + 4)，0代替三联体IP)。 因此，表示9个重叠核苷酸三联体的窗口的特征向量具有总数

223 (= 5 + 20 + 198) 个特征元素。 如果我们不考虑蛋白质特征，则排除20个元素。 因此，一个特征向量包含203个 (= 5 + 198) 个元素。

2.4. 绩效评估

2.4. 绩效评估

(Sn)，回忆 (Recall)，特异性 (Sp)，准确性 (Ac)，阳性预测值 (PPV，精确度)，阴性预测值 (NPV)，F-测量F) 和马修斯相关系数 (MCC)。 真阳性 (TP) 是被正确预测为结合核苷酸的结合核苷酸，真阴性 (TN) 是预测为非结合核苷酸的非结合核苷酸，假阳性 (FP) 是非结合核苷酸，错误地预测为结合核苷酸，并且假阴性 (FN) 是被错误地预测为非结合核苷酸的结合核苷酸。

$$\text{灵敏度} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specicity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{准确性} = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$PPV = \frac{TP}{TP + FP} \tag{7}$$

$$NPV = \frac{TN}{TN + FN} \tag{8}$$

$$F = \frac{2 \times Sn \times Sp}{Sn + Sp} \tag{9}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{(TP + FP)\ (TP + FN)\ (TN + FP)\ (TN + FN)} \tag{10}$$

灵敏度是正确预测的结合核苷酸与实际结合核苷酸的比率。 特别是
正确预测的非结合核苷酸与实际非结合核苷酸的比率。 准确度是正确的比例

预测核苷酸到所有核苷酸。 阳性预测值 (PPV) 测量正确预测的结合核苷酸与预测为结合的所有核苷酸的比率。 阴性预测值 (NPV) 测量正确预测的非结合核苷酸与预测为非结合的所有核苷酸的比例。

3. 结果与讨论

3.1. 数据集

我们获得了蛋白质-DNA复合物的结构数据，这些数据是通过X射线晶体学以及来自Protein Data Bank (PDB) 的分辨率为3.0A°或更好的分辨率确定的，[13]. 截至2013年5月，共有1691个蛋白质-DNA复合物，这些复合物用于计算核苷酸三联体与氨基酸的相互作用倾向 (IP)，由方程(1).

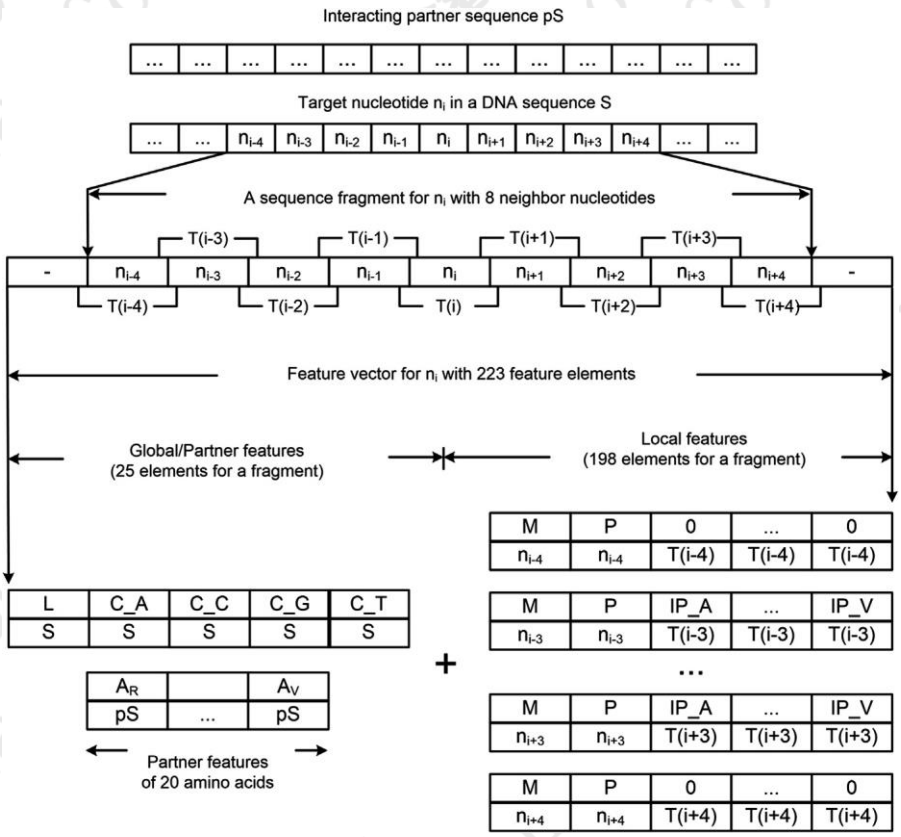

图1 - 具有9个核苷酸窗口的特征向量的结构。 具有9个核苷酸窗口的特征向量的结构，其覆盖9个重叠的核苷酸三联体：T（i-4），T（i-3），.... . ...，T（i + 4）。 5个全球特征元素（L，CA，
CC，CG，CT）和20种蛋白质特征元件（A_R，...，A_V）针对一对DNA和蛋白质序列编码一次。 22
对9个核苷酸中的每一个编码局部特征元素（M，P和20个IP）。 对于结尾核苷酸三联体的IP，窗口T（i-4）和T（i + 4），0被编码，因为它们的IP值没有定义。 对于9个核苷酸的窗口，特征向量中元素的数量是223（= 5 + 20 + 198）。 当没有使用20蛋白质特征时，a的数量
特征向量变成203（= 5 + 198）。

从NPIDB获得1691个蛋白质-DNA复合物的结合信息[11]. 从1691个蛋白质-DNA复合物中，我们提取了5346个蛋白质-DNA序列对，其中包含1700个DNA序列和1002个蛋白质序列。 通过运行CD-HIT[14] 在1700个DNA序列中，我们构建了181个DNA序列（以下称为D250）与其他序列相似度低于80％的测试数据集。剩余的1519个DNA序列被用于构建训练数据集（称为D5096）。D250数据集包含181个DNA序列和143个蛋白质序列。 D5096数据集包含1519个DNA序列和859个蛋白质序列。

我们的训练数据集包含202个转录因子（23.5％）和其他蛋白质。 此外，测试数据集包含44个转录因子（30.8％）和其他蛋白质。 因此，我们的模型不仅可以预测转录因子结合位点，还可以预测其他蛋白质-DNA结合位点。 在这里解释构建数据集的过程图2。

我们应用了基于特征的冗余去除方法（F方法）[15] 到D5096数据集来构建两个不同的训练数据集。 对于仅使用DNA序列数据的SVM模型，我们构建了一个训练数据集TD1。 同样，对于使用DNA和蛋白质序列的SVM模型，我们构建了一个训练数据集TD2。 使用TD1和TD2，我们尝试了从11到39的几种不同窗口大小。

3.2. *支持向量机*

SVM已被用于解决生物信息学中的许多预测问题。 Shi等人[16]，例如，使用SVM预测蛋白质磷酸化位点。 我们使用SVM库（LIBSVM）[17] 构造以径向基函数（RBF）为核心的支持向量机模型。 在训练SVM模型时，必须调整两个参数值：C和y。 参数C根据决策表面的简单性折衷了训练数据的错误分类。 参数y表示RBF的宽度。 本文所示的所有结果都是针对TD1和TD2的C = 10和y = 1 /#特征向量获得的。 作为正特征向量和负向特征向量的权重，我们使用w_1（正特征向量的权重）= 1.1和w_{-1}（负特征向量的权重）= 1和w_1向量）= 1.45，对于TD2，w_{-1}（负特征向量的权重）= 1。

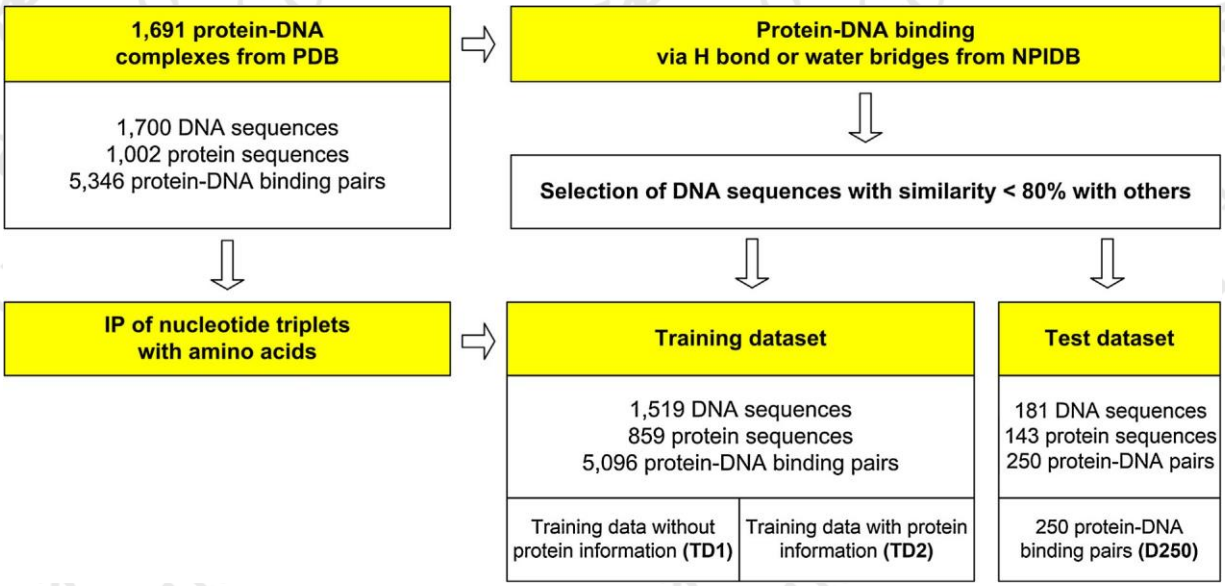

图2 – SVM模型数据集的构建。 蛋白质-DNA复合物的结构数据通过HPLC测定从PDB获得分辨率为3.0 A°或更好的X射线晶体学。 截至2013年5月，共有1691个蛋白质-DNA复合物，并且这些复合物被用于计算核苷酸三联体与氨基酸的相互作用倾向（IP）。 1691个蛋白质-DNA复合物含有5346个蛋白质-DNA结合对，5346个蛋白质-DNA结合对用于构建本研究中使用的训练和测试数据集。 5346蛋白质-DNA结合对含有1700个DNA序列和1002个蛋白质序列。 从1700个DNA序列中，首先通过选择与其他序列相似性低于80%的181个DNA序列构建称为D250的测试数据集。 剩余的1519个DNA序列被用于构建训练数据集D5096。 D250数据集包含181个DNA序列和143个蛋白质序列。 D5096数据集包含1519个DNA序列和859个蛋白质序列。

3.3. *两种预测模型的性能*

对于仅使用DNA序列数据的SVM模型，我们尝试了各种窗口大小，并且表1e对从11到39个核苷酸的不同窗口大小的模型的预测性能进行了抑制。 在10倍交叉验证中，35个核苷酸的窗口表现出最佳性能（67.0%的准确度和0.340的MCC）。 当在D250的181个DNA序列上测试SVM模型时，在31个核苷酸的窗口（准确度66.2%和MCC 0.324）下观察到最佳性能。

另一方面，表2 显示了使用DNA和蛋白质序列的SVM模型的结果。 如图所示表2，在TD2交叉验证中，35个核苷酸的窗口显示出最佳性能（准确性69.6%，MCC 0.383）。 当SVM模型在D250的181个DNA序列和143个蛋白质序列上进行测试时，在大多数测量中（以67.3%的准确度和0.329的MCC为准），其显示出最好的表现，其中31个核苷酸窗口。

随着窗口尺寸增加（在D250上进行测试），MCC趋于增加，但不会随着尺寸为31或更大的窗口而增加。 这是因为更大的窗口在窗口的两端包含更多的空核苷酸。 支持D250的支持向量机模型的测试细节可在附录1中找到。正向预测值（PPV）与负向预测值（NPV）之间的差异表1和2 可以解释如下。 使用不同窗口大小的TD1的10倍交叉验证导致PPV和NPV之间几乎没有差异（小于10%）（表格1）。 但是，测试使用31个核苷酸窗口的D250模型显示PPV和NPV之间的差异为12.4%。 使用具有31个核苷酸的窗口的TD2的10倍交叉验证显示PPV和NPV之间的20.2%的差异（表2），并且使用31个核苷酸的窗口在D250上测试模型在PPV和NPV之间显示出16.2%的差异。 使用DNA和蛋白质序列的SVM模型比仅使用DNA序列更准确地预测非结合核苷酸。 因此，使用DNA和蛋白质序列的预测模型导致PPV和NPV之间的差异大于只使用DNA序列的差异。

我们的预测方法不限于转录因子的DNA结合位点，因此我们研究中使用的训练和测试数据集都包括转录因子和其他蛋白质以及DNA的结合配偶体。 实际上，训练数据集D5096中的859个蛋白质序列中的202个（23.5%）是转录因子，并且测试数据集D250中的143个蛋白质序列中的44个（30.8%）是转录因子。预测模型的实验结果表明，它们不仅可以预测转录因子的DNA结合位点，还可以预测其他类型蛋白质的DNA结合位点。

为了测试我们对未知结构DNA序列的预测模型，我们在三个数据集上进行了测试：（1）来自JASPAR的数据[18]，（2）随机DNA序列，和（3）除去转录因子结合位点的Eschrichia coli基因组。 首先，我们从JASPAR数据库中提取了163个转录因子及其DNA结合位点[18]. PDB不含蛋白质-DNA复合物，其包括163种转录因子中的任何一种和163种转录因子

表1 - 仅使用DNA序列数据的SVM模型的性能。 共使用了5个特征（2个全局特征和3个局部特征）。

WS	Ac	锡	Sp	*F*	PPV	NPV	AUC	MCC
用TD1进行10倍交叉验证								
11	62.9%	65.1%	60.9%	62.9%	58.9%	67.0%	0.6803	0.2589
15	63.9%	65.4%	62.7%	64.0%	60.0%	67.9%	0.6965	0.2798
19	65.4%	64.8%	66.0%	65.4%	61.7%	68.9%	0.7082	0.3069
23	65.9%	66.4%	65.5%	66.0%	61.7%	70.0%	0.7188	0.3179
27	66.4%	67.3%	65.7%	66.5%	61.2%	70.9%	0.7256	0.3283
31	66.7%	67.7%	65.9%	66.8%	61.8%	71.4%	0.7292	0.3341
35	67.0%	67.9%	66.4%	67.1%	62.1%	71.8%	0.7304	0.3401
39	67.0%	67.9%	66.2%	67.1%	61.9%	71.9%	0.7305	0.3396
在D250上进行测试								
11	63.8%	65.0%	62.8%	63.9%	57.3%	70.0%	0.6859	0.2761
15	64.4%	65.2%	63.7%	64.4%	58.0%	70.5%	0.6931	0.2869
19	64.7%	64.3%	64.9%	64.6%	58.4%	70.3%	0.7049	0.2902
23	65.3%	67.2%	63.9%	65.5%	58.8%	71.7%	0.7090	0.3078
27	65.5%	66.9%	64.5%	65.7%	59.1%	71.7%	0.7162	0.3108
31	66.2%	67.3%	65.3%	66.3%	59.9%	72.3%	0.7174	0.3239
35	65.1%	66.3%	64.2%	65.2%	58.7%	71.3%	0.7126	0.3025
39	65.2%	66.4%	64.3%	65.3%	58.8%	71.4%	0.7108	0.3044

WS：窗口大小，Ac：准确度，Sn：敏感度，Sp：特异性，F：F-测量值，PPV：阳性预测值，NPV：阴性预测值，AUC：ROC曲线下面积，MCC：马修斯相关系数。

而他们的绑定数据根本不用于训练预测模型。 对于163个转录因子，该模型正确预测了其98.5%的DNA结合位点。 附加文件2中提供了163个JASPAR ID和预测结果。

对于第二个测试数据集，我们使用FaBox生成了1000个具有相同核苷酸组成的40-100个核苷酸的随机DNA序列[19]. 我们测试了我们的模型并用假设的方式评估了它的特性

随机DNA序列不含蛋白质结合位点。 随机序列模型的特异性为63.8%，略低于实际数据（表格1）。 附加的1e 3显示了1000个随机DNA序列和预测结果。

对于第三个测试数据集，我们删除了所有已知的转录因子结合位点[20] 来自大肠杆菌基因组。 假定剩余的大肠杆菌基因组不含蛋白质结合位点，因为目前没有蛋白质结合的证据。

表2 - 使用DNA和蛋白质序列的SVM模型的性能。 共使用了6个特征（2个全局特征，3个局部特征和蛋白质特征）。

WS	Ac	锡	Sp	*F*	PPV	NPV	AUC	MCC
用TD2进行10倍交叉验证								
11	67.6%	69.4%	66.4%	67.9%	55.9%	77.9%	0.7453	0.3485
15	68.3%	67.8%	68.6%	68.2%	57.1%	77.6%	0.7497	0.3556
19	69.1%	66.9%	70.4%	68.6%	58.1%	77.6%	0.7521	0.3648
23	69.4%	68.1%	70.2%	69.2%	58.4%	78.2%	0.7576	0.3748
27	69.7%	69.1%	70.2%	69.6%	58.7%	78.7%	0.7601	0.3829
31	69.7%	69.3%	70.0%	69.6%	58.6%	78.8%	0.7605	0.3826
35	69.6%	69.7%	70.0%	69.6%	58.4%	79.0%	0.7603	0.3832
39	69.4%	69.9%	69.1%	69.5%	58.1%	79.0%	0.7591	0.3802
在D250上进行测试								
11	64.0%	64.1%	64.0%	64.0%	54.0%	73.0%	0.6901	0.2754
15	64.6%	61.7%	66.5%	64.0%	54.8%	72.4%	0.6950	0.2768
19	66.1%	62.9%	68.1%	65.4%	56.6%	73.6%	0.7043	0.3061
23	65.7%	63.3%	67.4%	65.3%	56.2%	73.5%	0.7106	0.3017
27	67.1%	63.9%	69.3%	66.5%	57.9%	74.4%	0.7120	0.3275
31	67.3%	63.5%	69.8%	66.5%	58.1%	74.3%	0.7160	0.3288
35	67.1%	64.1%	69.0%	66.4%	57.8%	74.4%	0.7144	0.3264
39	67.0%	63.4%	69.3%	66.2%	57.7%	74.1%	0.7107	0.3229

WS：窗口大小，Ac：准确度，Sn：敏感度，Sp：特异性，F：F-测量值，PPV：阳性预测值，NPV：阴性预测值，AUC：ROC曲线下面积，MCC：马修斯相关系数。

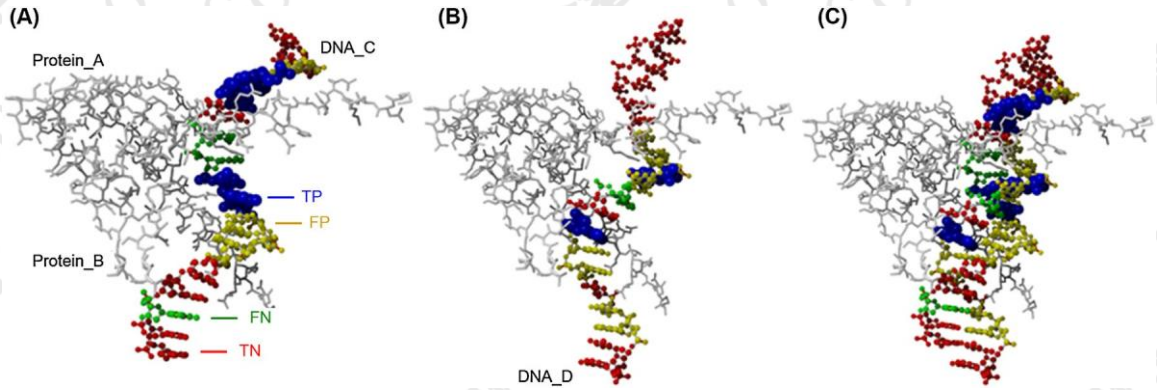

图3 - 仅使用DNA序列数据预测DNA链中的蛋白质结合位点的实例。 灰色棒代表蛋白质-DNA复合体1AN4的蛋白质序列。 (A) 在DNA链C中，该模型正确预测了4个结合核苷酸 (TP) 和8个非结合核苷酸 (TN)。 6个核苷酸实际上是非结合的，但被预测为结合 (FP)。 3个核苷酸实际上是结合的，但被预测为非结合性 (FN)。 (B) 在DNA链D中，模型预测了2个结合核苷酸 (TP) 和10个非结合核苷酸 (TN)。 8个核苷酸实际上是非结合的，但被预测为结合 (FP)。 1个核苷酸实际上是结合的，但被预测为非结合性 (FN)。 (C) 叠加结构显示DNA链C和D中的蛋白质结合位点。TP：真阳性 (蓝色)，TN：真阴性 (红色)，FP：假阳性 (黄色)，FN：假阴性 (绿色)。 (为了解释这个图例中的颜色引用，读者可以参考本文的网页版。)

从剩余的大肠杆菌基因组中，我们随机提取了1000个40-100个核苷酸的DNA序列，并在其上测试了我们的模型。 该模型显示了66.3%的特异性，接近65.3%的最佳特性 (在D250上测试，WS = 31表格1)。其他文献4显示了1000个大肠杆菌序列和预测结果。

3.4. 两个模型预测的蛋白质结合核苷酸的例子

图3 显示了仅使用DNA序列数据预测DNA链中蛋白质结合位点的实例。图4 显示了另一种使用DNA和蛋白质序列的SVM模型预测的DNA链中的蛋白质结合位点。 显然，由两个模型预测的蛋白质结合位点是不同的，因为其中一个使用关于DNA的结合配偶体的额外信息来预测结合位点。 虽然使用DNA序列的模型只能正确预测6 (= 4 + 2) 个结合核苷酸和18 (= 8 + 10) 个非结合核苷酸，另一个使用DNA和蛋白质序列的模型预测5 (= 3 + 2) 正确结合核苷酸和20 (= 9 + 11) 个非结合核苷酸。 同时使用DNA和蛋白质序列的模型的假阳性 (FP) 和假阴性的数量 (= 5 - 4) 减少了2 (= 14 - 12)

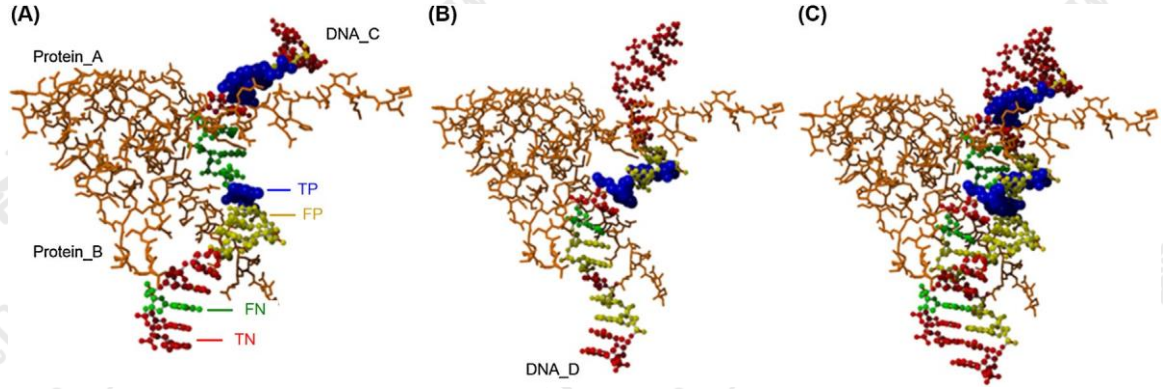

图4 - 使用DNA和蛋白质序列预测DNA链中蛋白质结合位点的实例。 蛋白质-DNA复合物1AN4的两条蛋白质链A和B (由橙色棒代表) 具有相同的氨基酸序列。
(A) 在DNA链C中，模型预测了3个结合核苷酸 (TP) 和9个非结合核苷酸 (TN)。 5个核苷酸实际上是非结合的，但被预测为结合 (FP)。 4个核苷酸实际上是结合的，但被预测为非结合性 (FN)。 (B) 在DNA链D中，该模型正确预测了2个结合核苷酸 (TP) 和11个非结合核苷酸 (TN)。 7个核苷酸实际上是非结合的，但被预测为结合 (FP)。 1个核苷酸实际上是结合的，但被预测为非结合性 (FN)。 (C) 叠加结构显示DNA链C和D中的蛋白质结合位点。TP：真阳性 (蓝色)，TN：真阴性 (红色)，FP：假阳性 (黄色)，FN：假阴性 (绿色)。 (为了解释这个图例中的颜色引用，读者可以参考本文的网页版。)

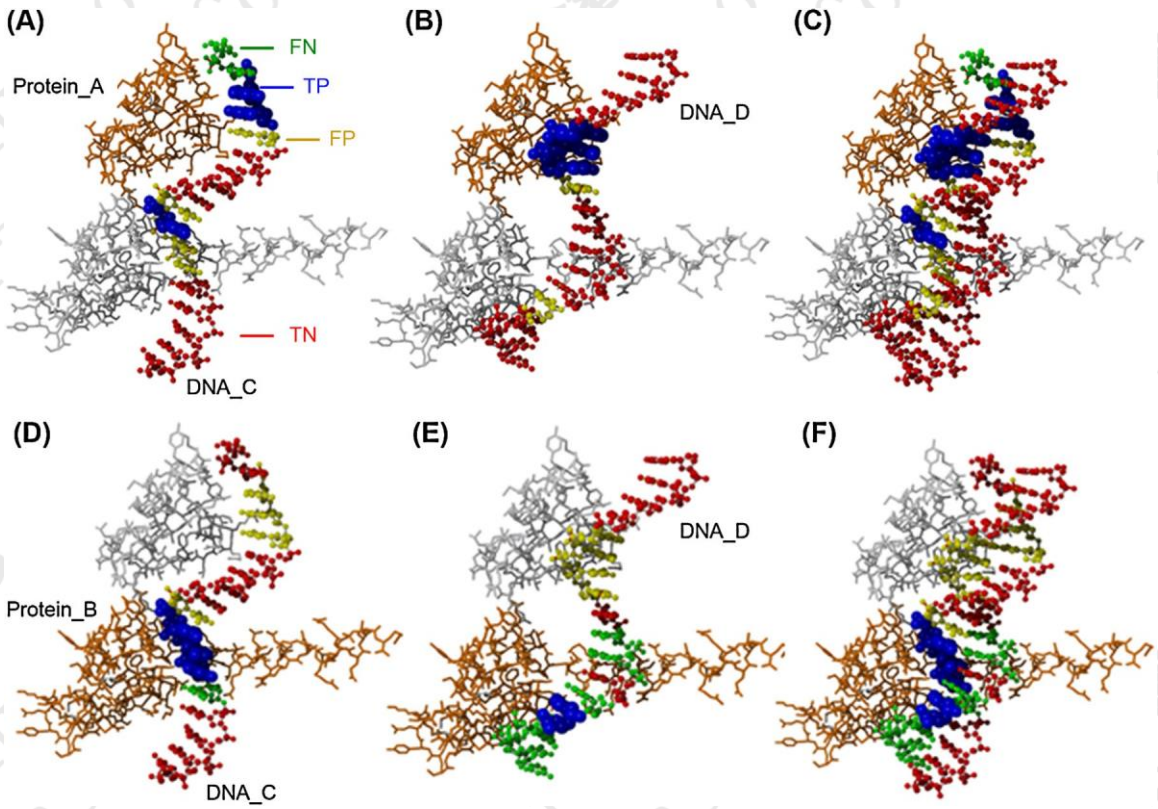

图5 - 用2NLL的不同蛋白质A和B预测DNA链C和D中结合位点的实例。 橙色棒代表被指定为DNA的结合配偶体的2NLL的蛋白质，并且灰色棒代表未被选择作为结合配偶体的蛋白质。 (A) 当蛋白质链A作为DNA链C的结合配偶体给予时，该模型正确预测3个结合核苷酸（TP）和9个非结合核苷酸（TN）。 4个核苷酸不具有约束力，但被预测为结合（FP）。 2个核苷酸实际上是结合的，但被预测为非结合性（FN）。 (B) 当蛋白质链A作为DNA链D的结合配偶体给予时，其预测3个结合核苷酸和13个非结合核苷酸正确。 2个核苷酸是结合的，但被预测为非结合的。 (C) 显示DNA链C和D与蛋白质链A的结合位点的叠加结构。 (D) 当蛋白质链B作为DNA链C的结合配体给予时，其预测3个结合核苷酸和10个非结合核苷酸正确。 4个核苷酸实际上是非结合的，但被预测为结合。 1个核苷酸实际上是结合的，但被预测为非结合的。 (E) 当蛋白质链B作为DNA链D的结合伴侣被给予时，其预测了1个结合核苷酸和6个非结合核苷酸正确。 4个核苷酸不具有约束力，但被预测为结合。 7个核苷酸结合但被预测为非结合。 (F) 显示DNA链C和D中具有蛋白质链B的结合位点的叠加结构。TP：真阳性（蓝色），TN：真阴性（红色），FP：假阳性（黄色），FN：假阴性绿色）。 (为了解释这个图例中的颜色引用，读者可以参考本文的网页版。)

底片（FN）。 因此，同时使用DNA和蛋白质序列的模型具有更多的TP + TN，并且在FP + FN中含有更少的数目。 使用DNA和蛋白质序列的模型比仅使用DNA序列预测非结合核苷酸的模型略好。

图5显示了用不同蛋白质伙伴预测DNA链中结合位点的另一个例子。 即使是一样的

DNA序列中，DNA序列中的蛋白质结合位点可以
当其绑定合作伙伴发生变化时更改。 具有2NLL蛋白A的DNA链C与蛋白A和DNA链D的平均预测结果分别为PPV 51.5%和NPV 90.9%，以及DNA链C与蛋白质B和DNA链D与蛋白质B是PPV 31.5%和NPV 68.6%。 另一方面，DNA链C和DNA链D的平均预测结果仅为PPV 66.7%和NPV 66.7%。 同样，使用DNA和蛋白质序列的模型比仅使用DNA序列预测非结合核苷酸的模型更好。 此外，这些预测结果可在附加文件1中找到。

4. 结论

我们使用广泛的数据集计算了核苷酸三联体与氨基酸的相互作用倾向（IP）

的蛋白质-DNA复合物，并开发了两种支持

预测蛋白质结合的矢量机器（SVM）模型仅来自序列数据的核苷酸。 一个SVM模型pre-使用DNA序列数据单独记录蛋白质结合核苷酸，另一种SVM模型使用DNA和蛋白质序列预测蛋白质结合核苷酸。 在使用1519个DNA序列的10倍交叉验证中，使用DNA序列数据的SVM模型仅以67.0%的准确度预测蛋白质结合核苷酸，67.1%的F-度量和0.340的Matthews相关系数（MCC）。 对于未用于培训的181种DNA的独立数据集，其准确率为66.2%，F-measure为66.3%，MCC为0.324。 使用DNA和蛋白质序列的另一种SVM模型在1519个DNA序列和859个蛋白质序列的10倍交叉验证中达到69.6%的准确度，69.6%的F-测量值和0.383的MCC。 使用181个DNA和143种蛋白质的独立数据集，其显示出67.3%的准确度，66.5%的F-测量值和0.329的MCC。

当预测DNA序列中的蛋白质结合位点时，使用DNA和蛋白质序列的预测模型表现出比单独使用DNA序列更好和更可靠的性能。 与基于结构的方法不同，我们的方法不假设DNA的结构是已知的。 如果相互作用蛋白的序列数据可用，那么使用蛋白质序列数据也可以更好地预测蛋白结合核苷酸。 我们的方法不限于转录因子的DNA结合位点。 在我们的预测模型的训练和测试数据集中，转录因子和其他蛋白质都被包括作为DNA的结合伴侣。 当它的结构不知道时，找到DNA中的蛋白质结合位点将是有用的，并且序列是唯一可用的信息。 据我们所知，这是第一次尝试单独用序列数据预测蛋白质结合DNA核苷酸。

利益冲突

没有声明。

致谢

本研究由韩国国家研究基金会（NRF）通过科学，信息通信技术和未来规划部（NRF-2012R1A1A3011982）资助，部分由教育部（2010-0020163）资助基础科学研究计划，仁荷大学。

附录A. 补充数据

与本文相关的补充数据可以在网上找到 http://dx.doi.org/10.1016/ j.cmpb.2014.07.009.

[1]

引用

X.M. 丁，X.Y. P一个，C. Xu, HB. 沉，计算 prediction 的DNA-PRotein 间动作：a review, Curr. COMPUT。计算机辅助药物德。6 (3) (2010) 197–206.

[2] S.Y. 何, F.C. 宇, CY. 张，HL黄，设计的accur吃 predictors 对于DNA-结合网站in proteins 运用hybrid SVM-PSSM 方法，生物系统90 (1) (2007) 234–241.

[3] LJ. W昂，S.J. BroWN，BindN：a weB系工具对于efficient prediction 的DNA 和RNA 捆绑网站in 氨基酸 序列，核酸酸RES。34 (W1) (2006) W243-W248。

[4] Z. 钱，L. 鲁，X. 刘，Y.-D. 蔡，Y. 李，一个approach 至pr法令 transcription 因子DNA 捆绑现场specificity 基于上 g烯和transcription 因子实用美食gorization，生物信息学23 (18) (2007) 2449–2454.

[5] S. Y昂，HKYalamancHILI，X. 李，K.-M. 雅o, P.C. 假，MQ 张，J. W昂，科尔高昂evolution 的transcription 因素 和其捆绑网站，生物信息学27 (21) (2011) 2972–2978.

[6] G. 郑，Q. 刘，G. 丁，C. WEI，Y. 李，至wards biologiCal中c哈日acters 的间行动between transcription 因素和 其DNA柏油gets in 哺乳动物，BMC 基因组学13 (1) (2012) 388.

[7] P. 阿桑娜希雅，A. Malousi，S. Kouidou，N. 嘛glav呃如，GrEMET：一个integrative 工具对于该prediction 的mutation 效果上g烯regulation，COMPUT。方法镨ograms 生物医学。111 (1) (2013) 214–219.

[8] CH Yan，M. Terribilini，F. Wu，RL Jernigan，D. Dobbs，V. Honavar，Predicting DNA-binding sites of proteins from amino acid sequence，BMC Bioinform。 7 (2006)，http://dx. doi. org/10. 1186/1471-2105-7-262。

[9] S. Hw昂，ZK郭台铭，I.B. Kuznetsov, DP-绑定：a 我们b serv呃对于 基于序列prediction 的DNA-结合residues in DNA-结合proteins，生物信息学23 (5) (2007) 634–636.

[10] CK林，CY. 陈，PiDNA: predicting protein-DNA 间行动同structur人楷模，核酸酸RES。41 (W1) (2013) W523-W530。

[11] D.D. Kirsanov, O.N. 赞恩g在一个，EAAksianov, S.A. 施普瑞，如. 嘉yag在一个，A.V. 麦酒xeevski，NPIDB：nucleic 酸-PRotein 间行动数据基础, 核酸酸RES。41 (D1) (2013) D517-D523。

[12] A. 麦酒xeevski，S. 施普瑞，D. 麦酒xeevski，O. Klychnikov, A. Ershova, M. Titov, A. 嘉yag在一个，CluD, a progr上午对于 决心的hydrophobic 集群in 3D structures 的protein 和protein-Nucleic 酸compleXES，Biophysics 48 (S1) (2004) 146–150.

[13] P.W. 玫瑰, B. 小檗碱一个，C.X. 双，W.F. 布卢姆，D. 迪米特尔opoulos，D.S. 古德塞尔，A. 普尔利奇，M. 克萨达G.B. 奎因J.D. WestbrOOK，J.Y翁荣南，B. Yukich, C. 扎尔十二月KI，HM 伯曼P.E. 畛域, 该RCSB protein 数据银行：redesigned 我们b 现场和我们b 服务，核酸酸RES。39 (D1) (2011) D392-D401。

[14] Y. 黄，B.F. 牛，Y. 高, LM福W.Z. 李，CD-HIT 套房：a 我们b serv呃对于集群和比较biologiCal中 序列，生物信息学26 (5) (2010) 680–682.

[15] S. Choi，K. Han，Prediction of RNA-binding amino acids from protein and RNA sequences，BMC Bioinform。 12 (S13) (2011)，http://dx. doi. org/10. 1186/1471-2105-12-S13-S7。

[16] Y. 施，B. YUAN，G. 林，D. Schuurmans，镨otein 磷ylation 现场prediction via feature 迪斯科v呃y 支持v埃克特苹果电脑海因, 清华科学。Technol。17 (6) (2012) 638–644.

[17] CC张，CJ. 林，LIBSVM：a 溴化锂ary 对于支持v埃克特 苹果电脑海因斯，ACM Tr答。INTELL。SYST。Technol。2 (3) (2011) 27.

[18] A. Mathelier, X. 赵, AW. 张，F. 霸rcy, R. WorsleY型亨特，D.J. 氩enillas，S. 愎HMAN，C.-y. 陈，A. 周杰伦，H. Ienasescu，J. 林，C. 嘘年, G. T一个，M. 周，B. Lenhard, A. 桑德林，

W.W. Wasserman，Jaspar 2014: 一个extensiv埃尔y expanded 和 更新开放式访问数据基础的transcription 因子 捆绑 prOLES，核酸酸RES。42 (D1) (2014) D142-D147。

[19] P. Villesen，Fa博x: 一个线上toolbox 对于fasta 序列，摩尔。ECOL。笔记7 (6) (2007) 965–968.

[20] H. Salg忙乱, M. P呃阿尔塔吉尔，S. 伽马，科技成果转化o, A. 桑托斯-ZAvaleta，L. Muiz-Rascado, J.S. 碣CA-索特洛, V. WEISS，
H. 索拉诺 - 李尔a, I. 马丁内斯 - 弗洛尔ES，A. 梅迪纳里v呃a, G. SalgADO-奥索里奥, S. Alquicir一个埃尔南德斯，K.
Alquicir一个埃尔南德斯，A. 洛佩斯 - 富恩特斯，L. Porrn - 索特洛, 上午 韦尔塔C. 博纳志愿组织 - 马丁内斯，Y.I. 巴尔德AS-马丁内斯，L. Pannier, M. Olv呃a, A. 啦巴斯迪达，V. Jimnez-Jacinto, L.
Veg一，阿绿ar忙乱, V. 德尔莫尔al-ChvEZ，A. 埃尔南德斯，阿绿arEZ，E. 莫尔eTT的部份，J. 科拉多-V集成开发环境，回覆gulondb v8.0: 组学数据套， evolutionary 水土保持通报通货膨胀，regulatory phrASES，
cross-validated 金standards 和铁道部e, 核酸酸RES。41 (D1) (2013) D203-D213。