摘要：

动机：

蛋白质亚细胞定位是近二十年来计算生物学领域的一个重要研究课题。到目前为止，已经提出了各种计算方法来处理具有未知位置的蛋白质的大规模数据集。基于统计机器学习的方法是现有预测方法的一个主要分支，大量的研究表明，从生物领域知识中提取特征对提高预测精度非常有用。然而，领域知识，如基因本体论和功能领域，通常会导致冗余特征和高维特征空间，这可能会降低机器学习模型的性能。

成果：

在本文中，我们提出了一种新的特征表示协议，称为HCM（隐藏相关建模）。该方法考虑了领域知识库的结构层次和注释项之间的相关性，建立了更紧凑、更具判别性的特征向量。在四个基准数据集上的实验结果表明，与传统的GO法相比，HCM提高了约5%的预测精度。此外，该hcmdriven预测器，哼mploc 3，大大提高了预测的多区位蛋白在dbmloc数据集23%和16%在一个新建的基准设定精度。

1. 介绍

亚细胞定位知识对于理解蛋白质功能、调控机制和蛋白质相互作用是至关重要的。由于使用湿实验室实验来鉴定蛋白质的细胞间隔常常是费力而昂贵的，在硅片预测工具中，处理未知位置的蛋白质的大规模数据集是非常必要的。根据SWISS-PROT（博克曼et al.，2003）在2016年2月公布，其中超过550000的蛋白质，11.6%的实验定义和验证的亚细胞定位注释，而绝大多数蛋白质都不确定位置注释。自动工具，允许计算预测未知位置的蛋白质，在过去十年已经基本上发展起来。特别是，由于在公共数据库，蛋白质序列和注释数据的大量丰富的信息，例如氨基酸的统计，已知位置的同系物，演化序列信息、信号肽、理化性质和生物先验知识，已被纳入计算工具，导致强大的预测工具。例如，常见的在线访问的预测因子包括：BaCelLo（pierleoni et al.，2006），YLoc（briesemeister et al.，2010），MultiLoc（HöGlund et al.，2006），goasvm（万et al.，2013），WoLF PSORT（Horton et al.，2007）和Euk mPLoc（Chou沈，2007a）真核蛋白；和Hum mPLoc（Chou和沈，2006），hslpred（Garg et al.，2005）人类蛋白质，等。这些Web服务器提供了生物实验室的科学家极大的便利，面对日益增长的需求，从蛋白质组学及相关领域。我们此前发布的预测，哼mploc 2（申和Chou，2009），例如，电话量每年提高了近20000in 2010到80000 2015（补充图S1）。

一般来说，蛋白质亚细胞定位识别的计算方法可分为三类：i）基于同源搜索的排序，II）基于排序的信号，和iii）基于机器学习的方法。基于同源搜索的方法可以看作是一个最近邻预测器，其中两个蛋白质之间的距离通常是通过它们的序列同一性来度量的。通过搜索查询蛋白对大量注释的序列，这种方法发现前K最接近的蛋白质，并将其注释查询蛋白（2002奈尔和罗斯特，）。这是一个相当直接的协议，但它的性能在很大程度上取决于所检测到的同源目标（WAN等人，2013）。此外，暮光之城的现象也明显挑战该协议假设（加迪et al.，2003），即蛋白质序列同源性高可以有非常不同的结构和功能。

靶向信号的预测是一种生理透明决策模型。在这种方法中，蛋白质的分选信号（通常在末端）首先发现的，然后分为不同的位置信号，根据先验知识，如线粒体，叶绿体，和分泌途径的信号（PSORT，1997；霍顿et al.，2007）。然而，基于排序信号的方法不适用于我们的排序知识尚未到达的一些单元空间，也没有缺少前导序列的查询蛋白质。

基于机器学习的预测器是一类灵活的模型，它要求所谓的训练数据集通过统计学习算法学习分类规则。因此，训练数据的质量是这类方法的关键之一，它与学习的统计规则的质量密切相关。通过对蛋白质数据库中的亚细胞定位越来越可靠的注释，分类模型可以训练，通过大规模的训练数据的收集更充分。基于机器学习模型的另一个重要问题是如何表达蛋白质序列，因为大多数算法都需要数字特征向量作为输入。如何从原始蛋白质序列中提取鉴别特征以及相关的先验知识对最终性能有重要意义。在现有的用于预测亚细胞定位的机器学习工具中，已经使用了各种功能，包括：

i) 基于统计特征的残留，如k-mer频次（Cedano等。，1997；emanuelsson et al.，2000；公园和kanehisa，2003）、伪氨基酸组成（Chou和沈2006；沈和Chou，2007, 2008），特定位置的得分矩阵（PSSM）（谢等人，2005；Pierleoni et al.，2006；Chou和沈，2007b；南尼等人，2013）；

ii) 基于肽的特征，如功能域（Chou和蔡，2002；Marchler Bauer等，2005）和基序（史葛et al.，2004）；

III）语境词汇注释的基础功能，如基因本体论（GO）数据库（Ashburner et al.，2000；Chou和蔡，2003）。

由于GO术语包含领域知识的高级抽象，当有足够的注释可用时，它们往往比基于残基或基于肽的特征更精确。然而，该注释数据的同时也带来了新的算法的挑战。例如，利用伯努利事件模型为每个术语，即二进制存在/不存在一个GO术语编码，去为基础的方法往往在一个非常高维特征空间的结果，其中包括围棋术语数以万计（沈和Chou，2009；Blum et al.，2009）。随着GO数据库的不断扩展和更新，随着我们对蛋白质知识的不断拓展，维数将不断增加。高维特征向量增加了后续学习过程的复杂度，并考虑了标注数据库中的潜在噪声对预测性能的影响。虽然整个GO数据库庞大，但实际上每个蛋白质只包含几个术语。根据我们的统计，其中至少有一个GO术语在SWISS-PROT数据库的蛋白质注释的GO术语平均6。这将给我们提供一个稀疏的特征向量，它有成千上万个维度，但只有大约6个有用的组件。已经提出了不同的方法来处理这样的高维但非常稀疏的特征向量。例如，YLoc（briesemeister et al.，2010）只会选择去，这是典型的PROSITE模式特定的亚细胞位置。因此，它减少不必要的功能，并使结果更可解释，但它可能会遭受信息损失。的wegoloc（Chi和越南，2012）指定每个去权重可以突出有用的GO术语。

在这项研究中，我们使用GO相关信息对特征向量进行编码，而不是使用GO状态的存在状态或频率。众所周知，去的条件是三的丙烯酸图（DAG）的层次结构组织，即生物过程（BP）、分子功能（MF）和细胞组分（CC）。该条款是由不同类型的边缘相关的路径（即关系）在体大鼠。直到现在，很多定义的GO术语之间的语义相似度的方法已经被提出，如信息内容（米雷斯尼克et al.，1999；林，1998；江和Conrath，1997）和基于图的方法（吴等，2005；王等，2007；张等人，2006）。然而，据我们所知，很少有蛋白质亚细胞定位的预测考虑到术语相关信息。这促使我们研究GO术语之间的隐藏相关性，以便更好地测量两个高维稀疏GO特征向量之间的相似性。我们提出了一种新的协议，称为HCM（隐关联建模），挖掘蛋白质注释特征之间的隐含联系。为了应对缺乏去一些查询蛋白质注释，我们也纳入统计的残留特征，以及基于功能域的功能肽，是取自CDD（保守结构域数据库）。

有了这些新的优势特征的代表性，我们构建了一个新的预测，称为哼mploc 3，即我们先前开发的预测人类蛋白质的定位后，却赋予了完全新的设计特征表示。在这个新的系统，二元关联（BR）多标签分类的训练，处理多区位蛋白（Boutell et al.，2004）。HCM和嗡嗡的性能mploc 3已经与其他的特征提取方法和预测的一个新的测试集和三集相比其他基准。结果表明，HCM提高预测精度的5%左右，比传统的方法去，哼mploc 3显著提高精度的数据集上dbmloc 23%预测多区位和16%的蛋白质在新的测试集。

1. 材料和方法

2.1 数据集

与蛋白质亚细胞定位的大量计算工具不同，专门针对人类蛋白的预测因子很少。考虑到人类蛋白质亚细胞定位的知识在药物发现过程中对目标识别的需求迅速增加，本研究主要集中在人类蛋白质方面。我们构建了一个新的基准数据集的人类蛋白，命名为HumB，从2012年1月发布SWISS-PROT收集所有的人类蛋白质。为了提高数据质量，排除了没有亚细胞定位注释或具有类似“相似”、“潜在”和“可能”等关键词的不确定注释的蛋白质。为了提高数据质量，排除了亚细胞定位注释或具有类似“相似”、“潜在”和“可能”等关键词的不确定注释。该数据集包括在人类细胞中，12种常见的位置包括中心体、细胞质、细胞骨架、细胞内质网，溶酶体，高尔基体，线粒体，溶酶体，过氧化物酶体增殖和细胞膜，细胞核。位于其他舱室的蛋白质被过滤掉，而在这12个类别中有多个位置的蛋白质被包括在内。此外，冗余序列被删除使用双鱼座（2003王和dunbrack，）在序列同截止值为25%。最后，基准数据集包括3129个人类蛋白，其中2306个具有单一的亚细胞定位，其余的是多位置蛋白质。直观地说，每个位置可以被视为一个类标签，并且多个位置的蛋白质是一个多标记的样本。亨伯赛德郡共有4229个标签，每个标签的蛋白质有1.35平均。

除了基准设置驼峰，一个独立的测试集命名的HumT也准备绩效评估，其样本是从2015年5月到SWISS-PROT释放。已经在2012年1月释放的蛋白质被移除了。换句话说，humt没有重叠，HumB。此外，为了减少偏差，HumB和HumT之间的序列相似性小于25%。为准确评估的缘故，我们只考虑实验证据支持蛋白质的位置，即只有人类蛋白质的CC字段包含“生态：269“收集（证据规范本体，生态，是一种控制词汇的描述信息和生态：源项，269代表一个类型的实验证据）。最后，humt包括379人和541的蛋白质标签。（HumT的谦卑和数据分布在补充表一）所示。

虽然设计预测的人类蛋白质的位置，后面的嗡嗡声mploc 3思想可以应用到其他物种。展示新的预测模型的泛化能力，并与现有的最先进的预测工具的比较，我们不仅在我们的数据集的人类蛋白，而且一些行之有效的数据集发表的其他研究人员，包括动物蛋白在bacello数据集（pierleoni et al.，2006），动物蛋白HöGlund数据集（HöGlund et al.，2006）和dbmloc数据集（张et al.，2008）。这三套的详细资料在补充材料中。

2.2方法

本研究旨在开发一种基于机器学习的人蛋白亚细胞定位预测因子。图1显示了新预测器的总体结构，包括两个主要部分：特征提取和分类器构造。

新的特征表示协议HCM所产生的特征向量既包含剩余统计量又包含生物先验知识。对每一种特征的细节部分2.2.1、2.2.2和2.2.3，分别。

2.2.1 残留特征

残基的统计特性是亚细胞定位预测器特征向量的基本组成部分，尤其是当注释数据不可用时。在这里，残留的基础功能包括氨基酸组成（AAC）和进化信息的特定位置的得分矩阵（PSSM）表示。矩阵，PSSM（E。1）为每一个蛋白质序列，利用psiblast与0.001值搜索SWISS-PROT构造（Altschul等人。，1997），在那里Si；J代表，在序列的第i位残基变异jth最初氨基酸评分（1≤J≤20）在进化过程中，L代表的蛋白质序列的长度。

为了将矩阵压缩成固定长度的特征向量，需要将每个列平均化为一个值。注意，序列不同位置的残基通常具有不同的变异率，因此首先对每一行进行归一化处理以减少潜在偏倚。这里采用Z-score标准化（设备。2），其中，i，j表示归一化分数，n表示不同氨基酸的数目，即n等于20。然后对每一列，平均得分计算公式。三, 这两个操作后，该spssm转化为20在Equ朦胧的载体。4，然后，AAC和归一化的PSSM向量组合为40模糊向量，不仅引起氨基酸频率信息的蛋白质本身，而且从它的功能相关的蛋白残留量统计。

此外，考虑到定位的信息通常是所蕴含的N端和C端氨基酸序列（pierleoniet al.，2006），我们从两端，特别是多节段序列特征，第一10；20；：：：；N端60个残基，和最后10；20；：：：；100个残基的羧基末端。对于每个段，使用上述方法创建40个弱向量。通过连接所有这40个模糊向量（为全序列和16段），总维度变为680（40×17），其中可能包含一些冗余信息。因此，基于相关的特征选择（霍尔和史米斯，1999）（CFS）的方法，并最终导致一个43暗淡的特征向量。

2.2.2 GO 注解特征

特征提取过程由三部分组成，如图2所示。在挖掘相关信息之前，我们需要一个两两相似的矩阵。在GO数据库中构造这样一个超过40000个术语的矩阵是非常昂贵的。因此，我们只使用GO术语注释搜索SWISS-PROT蛋白质。GO注释包含实验支持和计算推断GO项。在这里，只有第一种类型被认为是保证注释的质量，包括10083 bp、3322 MF和1332 cc术语。通过改进的基于信息的度量（杨等人，2012），分别构造了三个相似性矩阵，分别用于bp、MF和cc。

在图2（b）中，可以观察到，我们不使用蛋白质本身的GO项，而是从同源蛋白质中检索每个蛋白质的代表GO项。这是基于许多蛋白质没有或很少去注释的考虑（申和Chou，2009）。梅（2012）和万等。（2013）在研究中采用同样的策略。具体来说，本同源，即有超过50%的查询蛋白的同源性和60%的阳性蛋白，通过爆破在SWISS-PROT搜索。围棋术语提取从SwissProt和InterPro数据库（Zdobnov和apweiler，2001）。

给定GO项的相关矩阵和代表每个蛋白质的GO项，将GO特征生成如下两个步骤。

1. 搜索最近邻查询蛋白质。

直观地说，根据GO注释，查询蛋白质将具有与最相似蛋白质相同的亚细胞位置的高概率。在此基础上，根据GO术语之间的语义相关性，确定训练集中查询蛋白质的近邻。具体来说，查询蛋白和KTH训练蛋白定义为平方相关查询的蛋白和组训练蛋白，K的GO术语每项之间的和的平方根之间的相似性，如Equ。5、在相关西（i代表去查询蛋白质的术语）和K的定义是在Equ。6、在时间都去，和K = FY1；Y2；··YMG·。

根据相似性度量，可以确定训练集中查询蛋白质的最近邻。由于BP，MF和CC在GO数据库各自的三天，他们可能在测量基因产物之间的相似性所起的作用不同。因此，我们将每个蛋白质的代表GO术语划分为7组，分别是bp、MF、cc、bp和MF、MF和cc、bp和cc和bp和cc和cc。在7组中分别计算相似性得分和10个最近邻。

2. 生成概率信息。

在这个步骤中，特征向量用概率信息表示。让ProA表示查询蛋白在位置A最初的概率，ProA定义为查询的蛋白质和其最近的邻居，找到在查询中的蛋白质和其所有的10个最近的邻居之间的相似性和相似性之和的比值，如Equ。7，查询蛋白质的所有最近邻居的索引集在何处（j j＝10），而INA是位于A处的最近邻居的索引集。

然而，由于缺乏GO注释，一些蛋白质可能在训练集中没有或很少有邻居。因此，我们应对平滑技术这一问题通过添加一个先验在Equ上映。8。先前等于蛋白质的比例位于，在NUMA和Num是蛋白质定位在和在训练集的蛋白质总数，分别。

对于7个GO组中的每一个，12个弱向量由所有12个位置的概率信息组成。最后，生成一个具有84个模糊度的特征向量。为了产生训练蛋白质的概率信息，进行了10次交叉验证。