# Prediction of protein function using a deep convolutional neural network ensemble

用深度卷积神经网络集成预测蛋白质功能

# 摘要

**背景**：包含高分辨率三维（3D）蛋白质模型的大型数据库与功能注释的可用性允许利用先进的监督机器学习技术来自动预测蛋白质功能。

**方法**：在这项工作中，新颖的形状特征被提取出来，分别以角度和氨基酸距离的局部（每个氨基酸）分布形式表示蛋白质结构。 将每个多通道特征映射引入用于函数预测的深度卷积神经网络（CNN），并且通过支持向量机或基于相关的k-最近邻分类器来融合输出。 使用每个多通道特征集的一个CNN或每个图像通道的一个CNN来调查两种不同的体系结构。

**结果**：来自PDB数据库的单功能酶（n = 44,661）的交叉验证实验获得了90.1％的正确分类，证明了当不考虑序列相似性时相对于同一数据集的先前结果的改进

**讨论**：蛋白质功能的自动预测可以为广泛的数据集提供快速注释，为相关应用打开途径，如药理目标识别。 所提出的方法显示了对基于结构的蛋白质功能预测的希望，但是还没有足够的数据来正确评估方法在非同源蛋白质上的表现，并且因此减少了进化关系的混杂因素

**学科**：生物信息学，计算生物学，数据挖掘与机器学习

**关键词**：酶分类，函数预测，深度学习，卷积神经网络，结构表示

# 简介

宏基因组学已经导致蛋白质数据库的大量增加以及新蛋白质家族的发现（Godzik，2011）。 虽然新发现但可能是冗余的蛋白质序列的数量迅速增加，但实验证实，全基因组的功能注释仍然有限。 蛋白质结构，即氨基酸链的3D配置，是蛋白质功能的一个非常好的预测因子，事实上它比蛋白质序列更可靠的预测因子，因为它在自然界中更加保守（Illergård，Ardell＆Elofsson，2009年）。

到目前为止，具有功能注释和实验预测其天然状态结构的蛋白质的数量（例如，通过NMR光谱或X射线如何引用本文Zacharaki（2017），使用深度卷积神经网络集合预测蛋白质功能。 PeerJ Comput.Sci.3：e124; DOI 10.7717 / peerj-cs.124晶体学）足够大以允许机器学习模型被训练，从而能够对未注释蛋白进行自动功能注释（Amidi等，2017）。 而且，随着蛋白质序列数量的快速增长，绝大多数蛋白质只能以计算方式进行注释。 在这项工作中，我们考虑了来自蛋白质数据库（PDB）的酶结构，并将酶委员会（EC）编号用作相当完整的注释框架。 EC号是基于酶催化的化学反应的数字分类方案，由实验证据证明（Webb，1992）

自动酶注释的文献中有很多机器学习方法。 Sharma＆Garg（2014）对功能特性的各种计算方法的效用和推论进行了系统回顾，而机器学习方法的比较可以在Yadav＆Tiwari（2015）中找到。大多数方法使用来自氨基酸序列的特征并应用支持向量机（SVM）（Cai等，2003; Han等，2004; Dobson＆Doig，2005; Chen等，2006; Zhou等， ，2007; Lu等人，2007; Lee等人，2009; Qiu等人，2010; Wang等人，2010; Wang等人，2011; Amidi等人，2016），k-最近邻（kNN）分类器（Huang et al。，2007; Shen＆Chou，2007; Nasibov＆Kandemir-Cavas，2009），分类树/森林（Lee等，2009; Kumar＆Choudhary，2012; Nagao，Nagano＆Mizuguchi ，2014; Yadav＆Tiwari，2015）和神经网络（Volpato，Adelfio＆Pollastri，2013）。在Borgwardt等人（2005）的顺序，结构和化学信息被组合成蛋白质的一个图模型，其被SVM进一步分类。然而，仅基于结构信息的关于自动酶注释的文献中几乎没有工作。贝叶斯方法（Borro et al。，2006）使用结构派生性质进行酶分类，实现了45％的准确性。 Amidi等人（2016）在仅使用结构信息时从PDB数据库获得了39,251个蛋白质的73.5％的分类准确性。

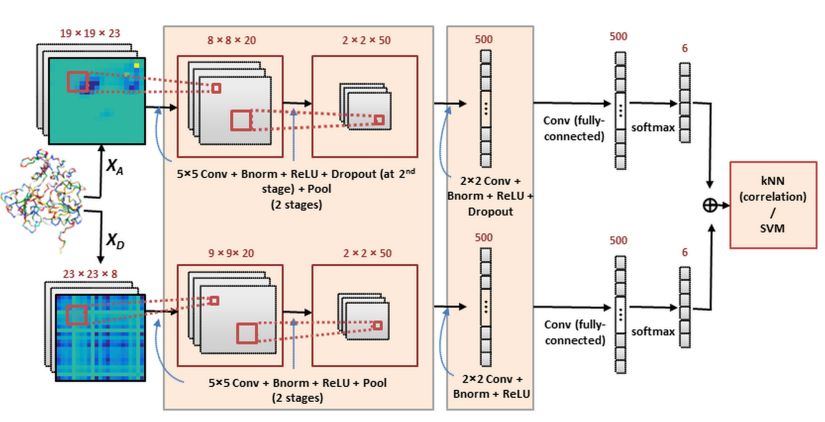
在过去的几年中，深度学习技术，特别是卷积神经网络，已经迅速成为解决许多具有挑战性的计算机视觉任务的首选工具，如图像分类（Krizhevsky，Sutskever＆Hinton，2012）。深度学习技术的主要优点是以无缝方式自动开发功能和调整性能，简化了传统的图像分析管线。 CNNs最近被用于蛋白质二级结构预测（Spencer，Eickholt＆Cheng，2015; Li＆Shibuya，2015）。在Spencer，Eickholt＆Cheng（2015）预测基于位置特异性评分矩阵（由PSI-BLAST生成），而在Li＆Shibuya（2015）中，1D卷积应用于与氨基酸序列相关的特征。此外，Lin，Lanchantin＆Qi（2016）提出了一种深度CNN架构来预测蛋白质性质。该架构使用多层移位和针迹技术来生成对蛋白质序列的完全密集的每位置预测。就作者所知，迄今为止，深CNN尚未用于预测蛋白质功能。

在这项工作中，作者利用实验获得的酶的结构信息并应用深度学习技术来生成基于结构预测酶功能的模型。引入新的几何描述符，并且通过将来自PDB数据库的44,661种酶的数据集分类成1 = 6种主要类别：氧化还原酶（EC1），转移酶（EC2），水解酶（EC3），裂合酶（EC4） ），异构酶（EC5），连接酶（EC6）。所提出的方法的新颖性首先在于将3D结构表示为由几何性质表征的“袋状原子（氨基酸）”，其次是通过深CNN对提取的特征图进行开发。尽管评估了酶促功能预测，但该方法不是基于酶特异性的，因此可以应用（再培训后）用于其他3D分子结构的自动大规模注释，从而为数据驱动分析提供有用的工具。在下面的章节中首先提供了关于实施的框架的更多细节，包括蛋白质结构的表示，CNN结构和网络输出的融合过程。然后介绍评估框架和获得的结果，然后进行一些讨论和结论

# 方法

数据驱动的CNN模型往往是域不可知的，并试图学习通过任何手工功能无法表现的附加功能。据推测，通过将“氨基酸特异性”描述符与深度学习的最新进展相结合，我们可以提高模型性能。该方法的主要优点是利用数据表示阶段和学习阶段的互补性。关于前者，该方法使用丰富的几何描述符，其将局部形状特征与表征该3D空间模型上的氨基酸相互作用的特征组合。形状表示由扭曲角度的局部（每个氨基酸类型）分布编码（Bermejo，Clore＆Schwieters，2012）。氨基酸相互作用由成对氨基酸距离的分布编码。虽然通常计算扭曲角度和距离图并绘制整个蛋白质的图（Bermejo，Clore＆Schwieters，2012），但在目前的方法中，它们分别针对每种氨基酸类型提取，因此表征了局部相互作用。因此，蛋白质结构被表示为一组多通道图像，其可被引入设计用于融合多个2D特征图的任何机器学习方案中。此外，应该指出，所利用的几何描述符对于蛋白质的整体翻译和旋转是不变的，因此不需要以前的蛋白质比对。

我们的方法构建了互为补充的深层CNN模型的集合。 将深度网络输出组合并引入到用于功能预测的基于相关的kNN分类器中。 为了比较，还实施了支持向量机用于最终分类。 如下所述，研究了两个系统体系结构，其中多个图像通道被联合或独立地考虑。 两种体系结构都使用相同的CNN结构（在突出显示的框内），如图1所示。



用于蛋白质分类的深度CNN集成方法。 在这个框架（体系结构1）中，每个多通道特征集被引入到CNN中，并且结果通过kNN或SVM分类进行组合。 网络包括执行卷积（Conv），批量归一化（Bnorm），整流线性单元（ReLU）激活，Dropout（可选）和最大池化（池）的层。 详细信息在“深度CNN分类”中提供。