## 化合物与蛋白质结合强度的预测模型

祖松鹏

2015/6/11

**问题的重要性：**

1. 定量地描述化合物与蛋白质结合的强度，而不仅是预测是否有结合的可能。作为DrugCIPHER在强度量化上的一个拓展。
2. 更有效地预测活性化合物。中药化合物中部分化合物可能结合能力相对较弱。如果结合能力太弱，不利于后期的结构优化。

**该模型的创新性：**

1. 考虑到了同一家族的蛋白质间，与之结合的化合物存在结构上的相似性。整合蛋白质家族的信息进行定量预测。
2. 不依赖于蛋白质3D结构，可进行大规模的预测。传统方法会受限制在蛋白质需要较多的化合物结合信息。该模型因为考虑到同一家族的其他蛋白质的化合物结合信息，可以不限制于某一个蛋白质的化合物结合信息不足，从而进行大规模的预测。有利于从整体角度，对某一化合物进行全局的活性认识。

**目前的进展：**

1. 完成构建数学模型和相应的求解方法。
2. 在Peptide GPCR这一类蛋白质家族上完成测试。能比传统的单一模型最高减少50%的预测错误。
3. 利用RDKit化学信息学计算库，提取多角度的化合物结构信息(100多种宏观性质，5大类2D子结构特证和一类2D 药效团特征)。
4. 利用上述提取的多维特征，进行模型的训练。由于对计算机的计算能力和内存都要求较高，这部分目前进展比较缓慢。

**待做：**

1. 在其他蛋白质家族上完成模型的构建和测试。
2. 与SEA等模型进行方法比较。
3. 针对已知的中药化合物，进行大规模的定量模型预测，以抓取活性化合物，设计实验。
4. 尝试RIP3的化合物预测。

**讨论：**

1. 单纯不依赖于结构，对DrugCIPHER的改进能力有限。从研究方向上看，利用机器学习方法，或DrugCIPHER这类网络方法，不考虑化合物和蛋白质结合的结构信息，已经到了一个瓶颈。该类方法，只能得到一个**粗粒度**的预测，在银杏内酯这个案例上就如此，我们不能对精细的结构进行区分。但化学结构的些许差别可能活性就影响较大。
2. 利用PDB中已有的化合物与蛋白质结合信息，结合我们现在的方法，会是下一步靶标预测的有可能取得较大创新的地方。