阅读日期：20190505

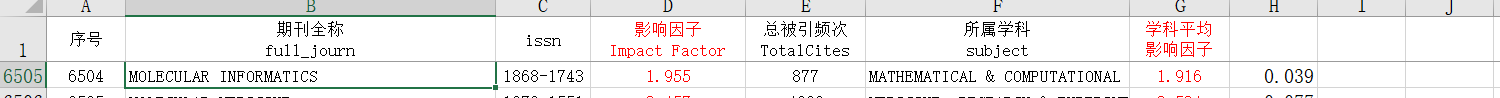
发表日期：2014

发表杂志：molecular informatics

影响因子：1.955(+0.039)

平均影响因子：1.916

标签：



|  |  |
| --- | --- |
| 文章名 A Machine Learning Approach to Explain Drug Selectivity to Soluble and Membrane Protein Targets | |
| 中文译名 | 一种解释药物对可溶性和膜蛋白靶点选择性的机器学习方法 |
| 网址 |  |
| 关键词 | Keywords: Drugs · Machine learning · Drug selectivity · Targetome · Protein-ligand recognition |
| 摘要 | Abstract:  Improved understanding of the forces that determine drug specificity to their targets is important for drug design and discovery, as well as for gaining knowledge  about molecular recognition. Here, we present a machine  learning approach that includes all approved drugs with  a known protein target. The drugs were characterized  using easily interpretable physico-chemical descriptors. Employing the Random Forest method, we were able to predict whether a drug binds to a soluble or membrane protein with an average accuracy of 84% and an average area  under curve of 0.91. The high average performance suggests that there exist some general physico-chemical differences between drugs that bind to membrane and soluble  protein targets. Variable importance measures in combination with permutation tests were used to find the most influential descriptors. This resulted in six outstanding descriptors, that all involve drug flexibility and lipophilicity, suggesting that drugs binding to membrane protein targets are in general more flexible and lipophilic, and conversely, drugs binding to soluble protein targets are more rigid and hydrophilic. With the notion that ligands in general are blueprints of their protein pockets, we may also draw  general conclusions about the protein-pocket properties  which may add to the understanding of molecular recognition.  总结：1机器学习中随机森林的方法。2药物弹性强，亲脂，易于结合膜蛋白。3药物僵硬，亲水，易于结合可溶性蛋白。4使用了排列组合的方法。 |
| 启发 | 1,3D结构取决于蛋白质所处的环境和肽链上氨基酸的组成。  2，蛋白质分为两类，第一类是可溶性蛋白比如球蛋白和纤维蛋白，比如酶和胶原。  第二类是由膜蛋白组成，比如G受体蛋白和孔（porins）.  3,评价前人工作：Such methods to describe proteins are useful for certain applications but provides very little information on the actual binding pocket of the protein.（这些工作用了蛋白质的性质，但是没有提供蛋白质口袋的信息）  4为了避开蛋白质3D信息的缺失问题，本文主要关注点转向了药物的3D结构和药物的理化属性。判断药物可以和可溶性蛋白质结合还是可以和膜蛋白结合。 |
| 摘录 | 1数据集收集：drugbank中得到，12年的数据，只要小分子，不要抗体和蛋白质。Sdf file(看起来是药物结构文件)，并且每个药物结构需要一个可以结合的单一的compound。手动检查每个文件，若一个文件有多个可结合的结构，则删去已知的结构（我理解为在之前出现过的结构。）特别常见的结构比如盐和buffer(缓冲物)也被去除了。还有一些别的规则用于处理数据。最后留下了1370个compound。这些compound中每个计算得到了3224个化学描述符。  2 为了留下只和蛋白质相互作用的药物小分子，有很多要求被设立了。可结合蛋白质的保健品被当作小分子药物得以保留。结合了非蛋白的比如核苷酸或者脂质的药物被除去了。药物被分为3类，第一类是可与水溶性蛋白结合，第二类是可与膜蛋白类结合，第三类是两者都可。  3 膜蛋白的识别是这样定义的：只要含有跨膜片段就算膜蛋白。G家族跨膜蛋白有7个折叠的部分是在膜上，（大部分在膜上）。而络氨酸激酶这种蛋白则包含很大一片区域是水溶蛋白。我们当前不能判断药物是作用在跨膜部分还是水溶部分，所以采用了这种简单的分类方式：只要有跨膜区域，就算膜蛋白。  4，药物这样被分为3类：可结合跨膜类，可结合水溶类，模棱两可类  5，241 ambiguous, 620 membrane, and 351 soluble.这些数据的3224个属性中选取了34个相关的属性选出来做实验，适合药物设计相关的属性。这34个属性中选取了17个关联性较弱的属性。    数据集筛选流程。左侧筛选的是混合物结构。右侧筛选的是药物的分类。之后使用了随机森林进行实验 |
| 相关论文 | 1【这篇文章用了氨基酸亲水性的性质】 [16] H. Strçmbergsson, G. J. Kleywegt, BMC Bioinformatics 2009, 10, S13.【】 |

总结：这是一篇提出新观点的论文。

它的创新点在于，从药物的角度出发，针对药物不同性质，预测并证实了不同性质的药物和水溶性或脂溶性蛋白的结核性不同。使用随机森林的方法做了分类。

给我的启发是：当使用别人没有使用过的数据时，我们要得出不同的结论。

本文中心结论：

drugs binding to membrane protein targets are in general more flexible and lipophilic

drugs binding to soluble protein targets are more rigid and hydrophilic.

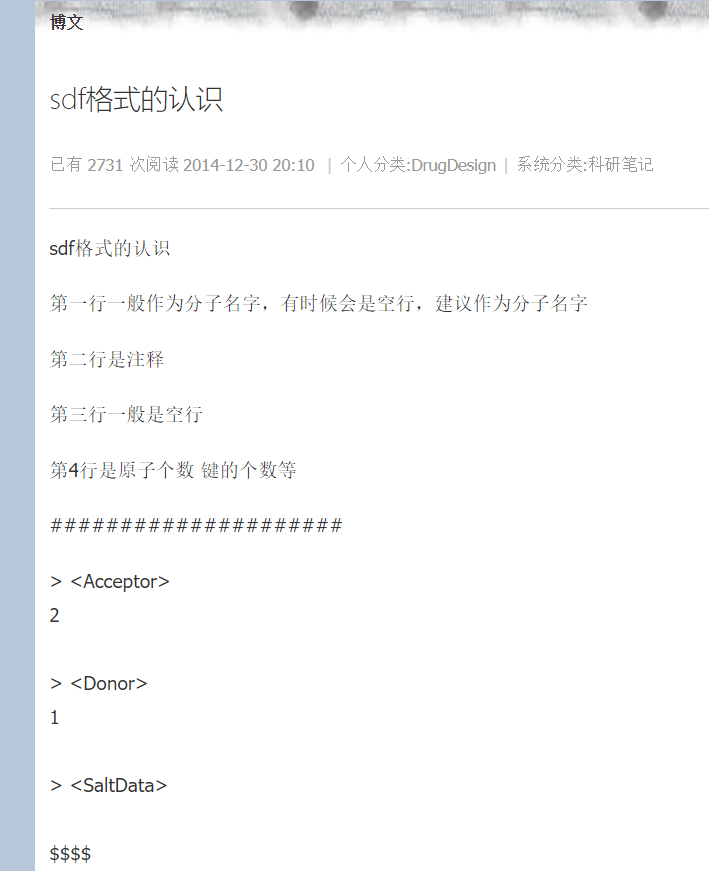
结合此论文，针对当前课题，可能得到的结论：

1结合位点所在域的种类不同，可结合的药物种类不同。

2结合位点所处的蛋白质种类不同，可结合的药物种类不同。

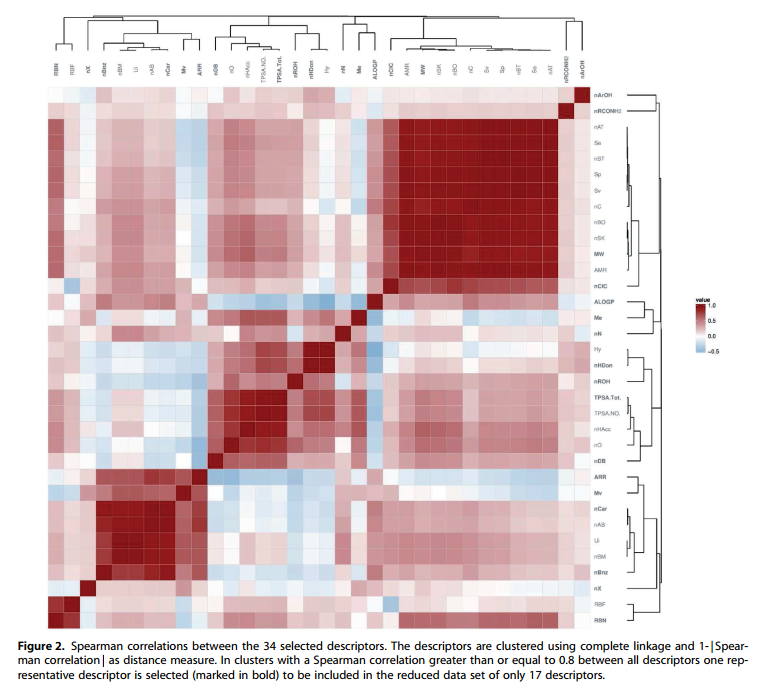
3可不可以 把 膜蛋白，分为2类：第一类是全部膜蛋白，第二类是部分跨膜。针对药物结合位点，探究其可结合药物的结合作用是否发生在对应的区域，来做研究。

【图一】



Sdf格式是药物分子的结构格式。

图二：



皮尔逊相关系数用于度量两个变量之间的相关联程度。

<https://baike.baidu.com/item/%E7%9A%AE%E5%B0%94%E9%80%8A%E7%9B%B8%E5%85%B3%E7%B3%BB%E6%95%B0>