基于快速、准确的双正则化单类协同过滤的大尺度离靶点识别及其在药物再利用中的应用

摘要

靶点的筛选是药物发现的主要方法之一，但会有预期靶点和脱靶的情况存在，即除了我们预期的靶点外，其实药物化学成分还有很多其他的靶点，简称脱靶，这些没有认识到的靶点，会产生副作用和治疗，那么如果找到了全部的脱靶处，然后针对其中进行筛选，选出能进行治疗的靶点，就可以完成对药物的重新利用。这个方法相比于其他方法更加简单准确，与其他方法相比，其他方法比的是同基因族，而跨基因族的无法进行验证，本方法的 ReMap，能够连续的探索化学成分空间、蛋白质空间和相互作用组，是一个有价值的补充。

介绍

单药单基因药物的开发耗时耗财，往往得不偿失，且药物不可能只与预期的靶点相结合，作者使用了一个例子，证明了其他非预期靶点的存在，同时有时候的副作用，在一定情况下是一种治疗。找到治疗靶点，识别非靶标相互作用，是药物发现和开发的重要的步骤，可以降低药物的损耗率，加快药物发现和开发的过程，最终生产出更安全，更便宜的药物。

化学相似度的应用及其广泛，但有些参数（TL 等），无法用此相似度建立图谱表达，因为化学的微小的结构改变可能会导致功能的改变，所以，化学-蛋白质相互作用应该由：蛋白质、化学结构共同决定。所以，应该将蛋白质纳入考虑的范围。同时作者解惑，为什么不用深度学习，因为此方法是否适用于基因还有待考察。

作者对比了技术，机器学习的方法和其他的技术有很多，但很多有以下的缺点：1、计算需要耗费大量的时间和内存。2、有些方法依赖于需要负例，但化学和蛋白质的关联有时候是稀疏的，对于负例较少的，如果为了计算随机生成，会对性能产生影响。3、很多方法是针对与同一个基因组的，在跨基因的运算效果，还无法确定。

作者方法的优点：1、无需负数据。2、有很好的可扩展性，准确性较高。3、创新了新的基准集，可以用于评估。4、此方法可以用于药物重定位。

方法

1、问题的公式化

有三个数据来源

化学-蛋白质关联网络、化学-化学相似性和化学-蛋白质相似性，最终希望找到化学-目标蛋白质相互作用的可能性。

会生成以下三个矩阵 n*m 的矩阵 R，代表化学和蛋白质是否关联，Ri,j=1 表示相关，否则表示不相关。n*n 矩阵 C，代表化学相似度，Ci,j 代表相似度得分。m*m 矩阵 T，Ti,j 代表蛋白质相似度得分。

最后得到矩阵 P，Pi,j 代表第 i 个化学成分和第 j 个蛋白质相似性作用的预测得分。

2、remap 综述

还没读完

讨论

重新预测提高了脱靶预测和药物再利用的预测能力

对于脱靶预测，本方法的能力优于其他方法，但对于矩阵分解算法相比，它有几点的改进：1、因为定义为一类协同过滤问题，所以训练不需要负数据。2、将含有已知负数据的先验知识引入到矩阵分解中，进行 imputation 和加权处理。最后，利用全局插补和加权算法，在不显著牺牲算法性能的前提下，提高了算法的计算效率。

剩余问题及未来方向

有很高的潜力,但存在数据冷启动问题,即需要需要一定的数据量才能运行起来，2、次优蛋白的相似度度量问题。针对第一个问题，REMAP 没有显示出比 PRW 更好的性能。 此外，如果新化学物质的目标有 5 个或更少的已知配体，则 REMAP 的回收率低于 0.5。 当新的化学物质与数据库中的化学物质相似时，REMAP 的回收率达到 90%以上。 这些结果表明，在实践中，现有的基于矩阵因式分解的方法，包括 REMAP，如果感兴趣的化学品没有任何已知的目标，则不是最佳选择。为了解决这个问题,可以设计一种将 PRW 或其他算法的优点与 REMAP 相结合的算法。此外，REMAP 的时间和记忆效率使得应用主动学习来克服冷启动问题成为可能。针对第二个问题，REMAP 的次优性能可能是由于缺乏分子水平的生化细节来推导蛋白质-蛋白质相似性度量。作者也做出了一些操作来克服。