药物靶标识别方法简述

一、 方法介绍

1. 二分局部模型法（BLMs）

BLMs 是一种预测药物靶标的监督学习方法。它利用了了药物的化学结构信息和基因的序列信息，整合了之前预测方法的优势，将局部模型的理念转化为二分网络问题，并基于此学习和预测药物-基因相互作用网络。总体来说，BLMs 将边预测问题转化为熟知的带标签的点的二分问题，而这种方法在预测 PPI(protein-protein interaction)、代谢网络以及调控网络方面都取得了不错的效果。

BLMs 需要建立药物相似性矩阵和基因相似性矩阵。其中药物相似性是依据基于比对算法的 SIMCOMP（Hattori et al.,2003）方法来计算的，药物c和c'的相似性由$s_c(c,c') = |c \cap c'|/|c \cup c'|$给出；基因序列相似性的计算则依据 Smith-Waterman 算法（1981 年 Smith 和 Waterman 提出的一种用来寻找并比较具有局部相似性区域的动态规划算法），特定基因g和g'的 Smith-waterman 分数由公式 $s_g(g,g') = SW(g,g')/\sqrt{SW(g,g)}\sqrt{SW(g',g')}$ 给出。通过上述方法，我们便可以得到药物相似性矩阵s_c和基因相似性矩阵s_g。

在当前药物集$V_d = \{d_1, d_2, \cdots d_m\}$ 和基因集$V_t = \{t_1, t_2, \cdots t_n\}$的基础上，我们利用边$e_{ij}$表示药物$d_i$ 和靶标基因t_j的相互联系。利用已知的 drug-target 信息，我们就可以构建出一幅二分图（图中边的端点分属于两个不相交的点集合）。不同的局部 drug-target 信息可以训练出不同的局部模型，在对应的局部模型中，采用以下算法流程来预测未知边e_{ij}的存在与否。

- 将靶标基因t_j分离出来，对剩余的基因集中的每个基因，若其是药物d_i的已知的靶标基因，则贴上标签+1，否则为-1。
- 标签值可视为分类结果，而基因的相似性矩阵可视为观测特征。这样便可寻求监督方法对基因进行靶标分类。
- 利用上述分类器，我们便可以预测t_j的标签值（也即分类结果），并决定药物 d_i 和基因t_j是否存在边e_{ij}。
- 同样的方法，我们先将靶标基因t_j重新并入二分图，然后将药物d_i分离出来，对剩余的药物集中的每个药物，若其靶标基因中包含t_j，则贴上标签+1，否则为-1。
- 标签值可视为分类结果，而药物的相似性矩阵可视为观测特征，这样便可寻

求监督方法进行药物分类。

- 利用上述分类器，我们便可以预测d_i的标签值（也即分类结果），并决定药物d_i 和基因t_j是否存在边e_{ij}。

对于待预测的药物d_i 和基因t_j，如果d_i没有已知的靶标基因而t_j至少有一个已知靶向药物，或者t_j没有已知的靶向药物而d_i至少有一个已知的靶标基因，那么上述算法只能做单向预测；如果d_i至少有一个已知的靶标基因并且t_j至少有一个已知靶向药物，上述算法便能得到两个相互独立的边预测值，利用函数方法（比如 max{x,y}）来整合多个预测分数，得到全局分数。

在局部分类方法中我们采用支持向量机（SVM），对分离基因t_j或者分离药物d_i得出一个-1 到+1 之间的连续值表征边e_{ij}存在的可能性。对每个待预测边所得到的分数做排序，得分越高，则药物与基因之间靶向作用的可能性也就越高。在 SVM 的方法下，我们可以通过定义二分图任意两点核函数k(u,v)的方式来处理非向量形式的点的信息，提高了方法的普适性。

2. DrugCIPHER-MS

基于药物表型效应以及基于化学结构的药物靶标识别方法虽然都有成果，但都具有一定的局限性：单纯基于药物表型效应无法区分同一 pathway 的不同 target；单纯基于药物化学结构并不具备理论的完备性，药物的结构相似性并不一定能延伸到靶标。而 DrugCIPHER-MS 则整合了 TS(药物表型相似性)和 CS(药物化学结构相似性)，理论上假设并证实了 TS 和 CS 通过 PPI 相互联系，并依此建立线性回归模型，预测药物靶标。

根据 ATC codes 相似性，我们根据下式建立药物d_i和d_j之间的表型相似性$\mathrm{TS}(d_i,d_j)$，其中S(i,j)定义了 ATC codes 的相似性。

$$\mathrm{S}(i,j)=\frac{2*\log(\Pr(prefix(i,j)))}{\log(\Pr(i))+\log(\Pr(j))} \qquad \mathrm{TS}(d_1,d_2)=\max_{i\in \mathrm{ATC}(d_1),j\in \mathrm{ATC}(d_2)}(\mathrm{S}(i,j))$$

药物d_i和d_j之间的化学结构相似性则依据谷本系数确定。此外，我们根据 PPI 网络定义药物和基因之间的相似性。

$$\varphi_{pd}=\sum_{p_k\in \mathrm{T}(d)} e^{-L^2_{pp_k}} \qquad \mathrm{R}_{d_1d_2}=\frac{\sum_{p_i\in \mathrm{T}(d_1)}\varphi_{p_id_2}}{\mathrm{No.T}(d_1)+\mathrm{No.T}(d_2)}=\frac{\sum_{p_j\in \mathrm{T}(d_2)}\varphi_{p_jd_1}}{\mathrm{No.T}(d_1)+\mathrm{No.T}(d_2)}$$

其中φ_{pd}是依据L_{ppk}(PPI 网络中基因p_k 和p之间的最短距离)得到的药物 d 和基因 p 的相似性。R_{d1d2}则通过药物靶标和药物的平均距离来度量药物之间的相似性。

DrugCIPHER-MS 假设给定基因 p 在 PPI 网络中的距离向量可以被特定的药物 d 的 TS 和 CS 线性表征：

$$\mathbf{\Phi}_p = a'_{pd}\cdot\mathbf{TS}_d + b'_{pd}\cdot\mathbf{CS}_d + c'_p.$$

利用最小均方误差的方差估计出a'_{pd} 和b'_{pd}后便可以估计药物 d 和基因 p 的匹配分数（分数值越高，则基因 p 在药物 d 的生物过程中越重要，同时也越可能是药物 d 的靶标）：

$$\rho^M_{pd} = \frac{\left(\frac{\sigma(\mathbf{TS}_d)}{|\hat{b}_{pd}|}\cdot\rho^C_{pd} + \frac{\sigma(\mathbf{CS}_d)}{|\hat{a}_{pd}|}\cdot\rho^T_{pd}\right)}{\sqrt{\frac{\sigma^2(\mathbf{TS}_d)}{\hat{b}^2_{pd}} + \frac{\sigma^2(\mathbf{CS}_d)}{\hat{a}^2_{pd}}}}.$$

其中ρ^C_{pd} 和ρ^T_{pd}分别是单独对 CS 和 TS 作线性回归得到的匹配分数：

$$\rho^T_{pd} = \frac{\mathrm{cov}(\mathbf{TS}_d, \mathbf{\Phi}_p)}{\sigma(\mathbf{TS}_d)\sigma(\mathbf{\Phi}_p)} \qquad \rho^C_{pd} = \frac{\mathrm{cov}(\mathbf{CS}_d, \mathbf{\Phi}_p)}{\sigma(\mathbf{CS}_d)\sigma(\mathbf{\Phi}_p)}$$

二、 方法比较

在 BLMs 文献测试的 4 个药物集的基础上，根据药物靶标数量设置权重，计算得到平均 AUC 为 0.9676。而基于整体 PPI 和药物网络的 DrugCIPHER-MS 方法则达到了 0.988 的 AUC。一定程度上反映了 DrugCIPHER-MS 相对于 BLMs 的优势。

三、 方法探讨

BLMs 和 DrugCIPHER-MS 本质上有很大的差异：BLMs 是建立局部模型，单独利用 CS 和 PPI 信息，将预测靶标问题转化为二分问题；DrugCIPHER-MS 则是建立整体模型，整合 TS、CS 和 PPI 信息，利用已有药物靶标信息建立线性回归模型。DrugCIPHER-MS 理论上的缺陷在于药理空间和基因空间的关系并不能完全用线性表达；BLMs 方法的缺陷在于建立的是局部模型，没有考虑局部空间之外的影响。

基于这两种方法的思想，我提出一种在 BLMs 方法框架上的改进方案。

- 将二分图由局二分图扩展为全局二分图，也即整个 PPI 网络对应基因集合和整个药物集合。（这样做会使得二分图变得更加稀疏）
- 二分图中的药物节点信息采用[TS CS]表征（仿照 DrugCIPHER-MS 整合多维度信息）。
- 采用神经网络(或者非线性 SVM)的分类方法来进行单对药物-基因匹配训练，

从而描述边预测或者说二分问题中的非线性。

当然，也可以在 DrugCIPHER-MS 的基础上，采用非线性模型拟合。将待预测的药物的 TS 和 CS 和基因在 PPI 空间的相似性特征向量 Φ_p 作为神经网络的输入，训练集输出为 0 或者 1 表征匹配分数。

总的来说，在现有知识的基础上，我们可以将药物靶标识别问题从不同的角度转化为分类问题或者回归问题。显然，在深度学习的领域中，我们可以建立更为复杂的模型来给特定的药物或者基因作二分（可以延伸到多分）问题求解。在 BLMs 的文献中，我们看到不同的方法本身也是可以整合的，只要最终可以提高预测精度。但抛开模型本身来讲，现有数据信息的获取，也即特征的提取和选择也是保证预测精度的重要前提。生物过程本身就是个极其复杂且易受干扰的系统，随着对其认识的深入，相信我们也能够逐渐获得更为深层的特征和发现更加准确的预测方法。

三、 参考文献

【1】 K Bleakley，Y Yamanishi. Supervised prediction of drug-target interactions using bipartite local models[J].Bioinformatics,2009,25(18):2397-2403

【2】 Zhao S, Li S. Network-Based Relating Pharmacological and Genomic Spaces for Drug Target Identification[J]. Plos One, 2010, 5(7):: e11764.