文献阅读报告

毛鑫 2012010285

作为第三组的成员，我研读了生物信息网络的相关文献，并从drugCIPHER一文中获得较大收获。

drugCIPHER一文着眼于发掘药物和蛋白质之间的靶向关系，并将重点放在整合pharmacological及genomic两空间，利用多数据集的互补关系，谋求更精确的配对结果。这里提到的pharmacological空间由Therapeutic相似性及Drug Chemical相似性两部分组成，分别根据临床效果或化学成分定义药物之间的相似性。而genomic空间则定义了蛋白质间相互作用的网络，并在此基础上计算药物与蛋白之间的closeness，进而度量药物之间的相似关系。此前的研究往往在前两者中只取其一。而本研究则发现，虽然有一定的关联，但临床效果或化学结构的相似性与靶向相似性并不完全相关，而化学结构与临床效果的相似性也并非同一件事。故两者只取其一必然精度较低，两者结合则可互有长短，相互补充。在此基础上，结合已经发现的genomic空间相似性，使用更加整合的数据，采用更加整合的方法，便有望在结果的精度和可解释性方面取得进展。

具体的操作可分为三步。第一步为计算药物与任意蛋白的closeness，第二步为关联pharmacological及genomic两空间，第三步为构建整合空间上的回归模型并据此发掘药物的靶向蛋白。第一步采用以下公式进行。

$$\varphi_{pd} = \sum_{p_k \in \mathrm{T}(d)} e^{-L_{ppk}^2},$$

在此式中，ϕ_{pd} 表示药物d与蛋白p之间的closeness。pk是d已知的靶向蛋白质。L_{ppk} 表示p和pk在蛋白相互作用网络中的最短距离。通过观察此式的构造，我们可以发现closeness计算的准确度与对应药物已发现的靶向蛋白质占所有靶向蛋白质的比例有关。由于这个原因，相关药物的标靶蛋白必须相当明确。

第二步需先选定若干靶向明确，临床成熟，成分确定的药物。将之作为“坐标系”，用以关联两空间，统一药物及蛋白的表示形式。采用Therapeutic相似性及Drug Chemical相似性的药物表示形式如下所示。

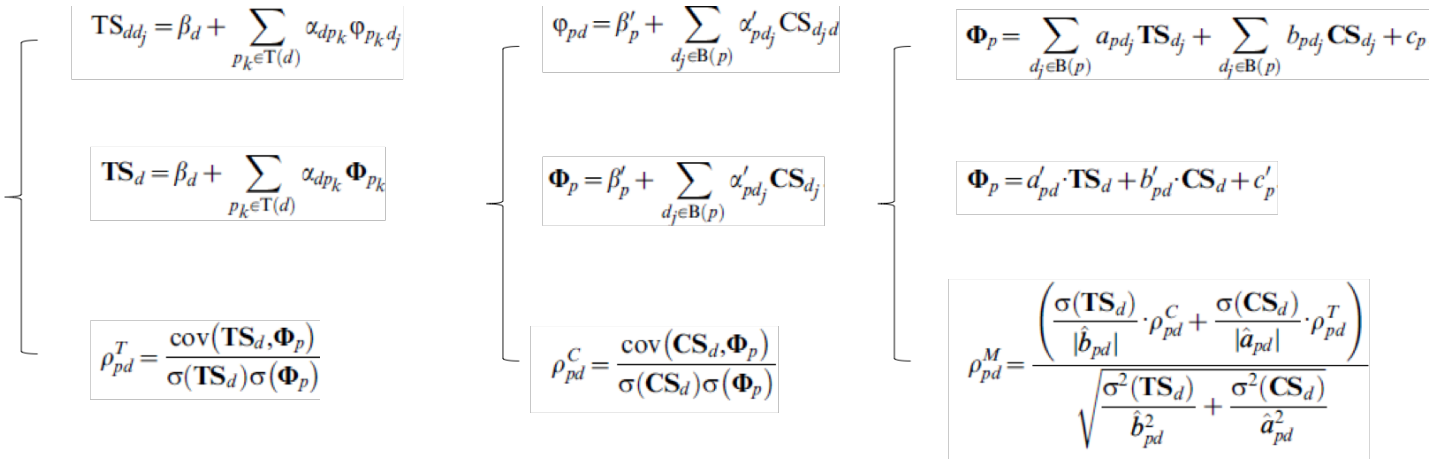
$$TS_d = \{TS_{dd1}, TS_{dd2}, \ldots TS_{ddn}\}$$

$$CS_d = \{CS_{dd1}, CS_{dd2}, \ldots CS_{ddn}\}$$

采用closeness的蛋白质表示形式如下所示。

$$\Phi_p = \{\phi_{pd1}, \phi_{pd2}, \ldots \phi_{pdn}\}$$

第三步建立的回归模型可分为三种。分别为DrugCIPHER-TS，DrugCIPHER-CS与DrugCIPHER-MS。前两者利用的pharmacological空间数据只有Therapeutic相似性或Drug Chemical相似性。最后一种则采用了上述两种数据，故在理论上和实际上都取得了更高的精确度。三者的回归模型分别如下所示。



其中，各式的 ρ_{pd} 皆表示对应情境下某一蛋白p与某一药物d之间的concordance值。值越高者对应的自变量在模型中越显著，相应的蛋白成为药物标靶的可能性也就越大。故在最后的处理中，需人为设置一阈值，使大于该阈值的concordance值对应的配对被采纳。

通过阅读此文，我认识到了对具有差异的数据集进行整合可以提供更有价值的研究数据。同时，我也有自己的思考。在承认文章思路，即整合pharmacological及genomic两空间的同时，我认为为了提高计算的整合度，可以采用主成分分析或LASSO辅助回归等方法对变量进行整体分析，直接去掉关联性较弱的自变量，而不用逐一计算concordance值再作筛选。