利用互作组学数据揭示疾病间的关系[1]

研究背景

疾病的发生很少是由单一基因缺陷引起的，更多的是细胞内多个分子间作用相互影响的过程。而分子间的相互作用亦即构成了互作组学数据(interactome)，包括诸如：蛋白-蛋白、蛋白-脱氧核苷酸、代谢分子间互相作用等。这些分子、及分子间的互相作用可分别构成调控网络的点和边。一个疾病若是相关基因间相互作用的过程，那么在网络里，这些基因会聚集在一起，形成一个表示该疾病的疾病模块(disease module)，而疾病模块间在网络中的距离远近亦可用于刻画疾病间关系。本文的研究思路即是基于此模型来研究疾病，以及疾病间的关系。

在用疾病模块来表示疾病后，那么疾病模块里点缺失(Node deletion)、连边缺失(Link deletion)、连边权重减弱或加强(Modification of the interaction strength)等都可能引起疾病的发生。在实际研究的过程中遇到的挑战是：尽管高通量互作组学数据的不断产出以及对疾病相关基因了解的不断深入和扩展，但目前这两方面的知识还是不全面的、有些甚至是错误的。这可能导致一个疾病模块里，某些相关基因未被加入该模块、某些基因连边丢失，某些基因被错误地加入到该模块等。(图 1)

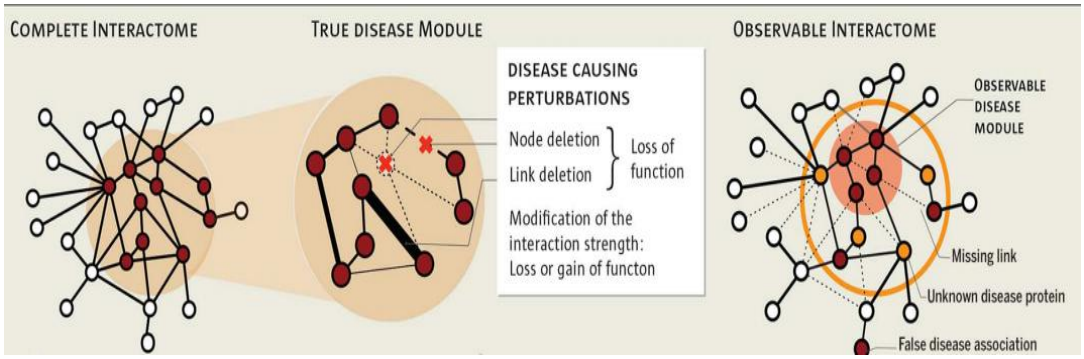


图 1：互作组学与疾病模块

疾病模块模型

1、疾病模块的刻画(图 2):

(1) 可观测到的疾病模块大小:S(The observable disease module size: S)，即疾病相关基因在网络中最大连通片(largest connected component)的顶点数。

(2) 疾病模块半径 $\hat{d}_S$：对于疾病模块中的每个基因，计算该基因到其他基因的距离，取其中的最短距离作为该基因在该模块内到其他基因的距离 d_S，得到每个基因的 d_S 后，取平均，作为该疾病模块的半径。

2、疾病模块间的分离程度 s_{AB} (图 3)

首先，计算疾病模块 A、B 各自的半径 d_{AA} 和 d_{BB}；

其次，计算疾病模块 A、B 间的距离 d_{AB}；计算方法同计算模块半径类似，对于疾病模块 A 中的每个基因，计算该基因到疾病模块 B 中各基因的距离，取其中的最短距离作为该基因到疾病模块 B 的距离 d_S；得到疾病模块 A 中每个基因的 d_S 后，再以相同的方法计算疾病模块 B 中每个基因到疾病模块 A 的距离 d_S。 得到疾病模块 A 和 B 中所有基因的 d_S 后，取平均，作为疾病模块 A、B 的距离。

最后，得到疾病模块间的分离程度：$s_{AB} = d_{AB} - (d_{AA} + d_{BB})/2$。

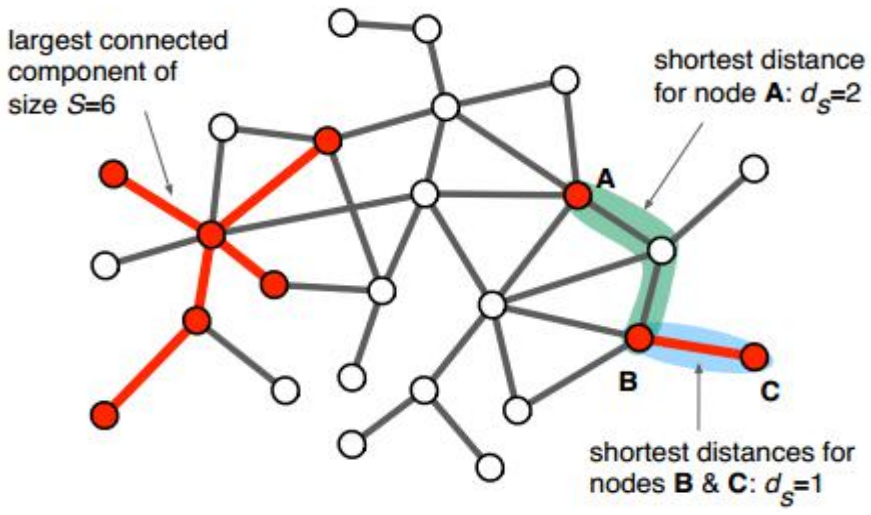


图 2：疾病模块的刻画：模块大小与模块半径。

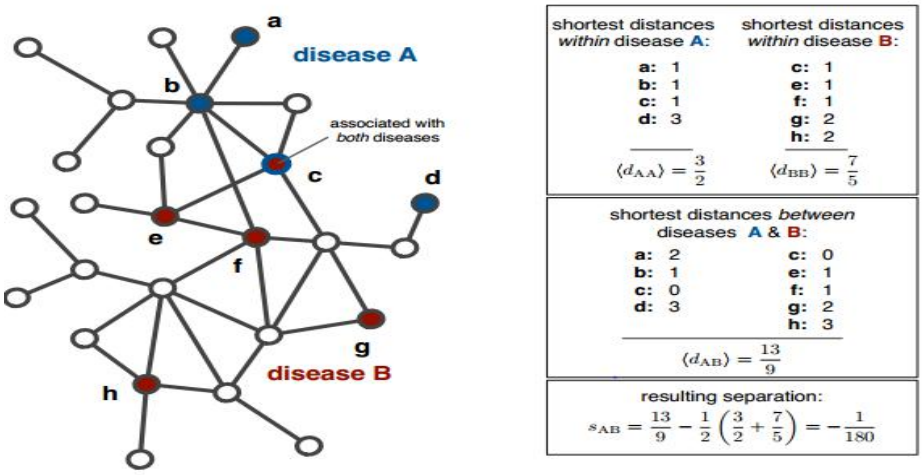


图 3: 疾病模块间的分离程度计算。

模型可靠性论证

分析思路:

1、证明确实存在疾病模块且疾病模块确实有生物学意义。

(1) 通过比较当前数据同随机选择基因情况下所形成的可观测到的模块大小 S 和模块内各基因的 d_S 分布情况; 发现当前数据所形成的可观测模块大小显著大于随机选择基因情况下所形成的可观测模块大小; 模块内各基因的 d_S 分布显著小于随机选择基因情况下的 d_S ，证明确实存在疾病模块;

(2) 发现疾病模块越聚集(即:可观测模块大小越大，疾病模块半径越小)，其中的基因功能越相似，证明得到的疾病模块确实有生物学意义。

2、证明疾病模块间的分离程度 s_{AB} 确实能反映疾病间的关系。

(1) 生物学相似性: 发现疾病模块间 s_{AB} 越小，其中的基因功能越相似;

(2) 共表达情况: 发现疾病模块间 s_{AB} 越小，相关基因表达量的相关系数越高;

(3) 疾病症状相似性: 发现疾病模块间 s_{AB} 越小，疾病症状的相似性越高;

(4) 共病性: 发现疾病模块间 s_{AB} 越小，同患两病的风险系数越高;

以上发现，说明疾病模块间的分离程度 s_{AB} 确实能反映疾病间的关系;

3、**为进一步说明该模型的用处，作者举了一例**: 哮喘(asthma)和脂泻病(celiac disease)，这两种疾病在病理和临床上没有明显的相关性(我注: 所以不易被想到一块)，但二者的分离

程度较小($S_{AB} < 0$)，说明二者在分子层面上可能有共同的根源。经进一步地检查，发现确实有相关的证据支持这一推测(如：二者有 3 个共同的基因; 1 个共同的 SNP; 共病风险系数高等)。

4、**最后，比较该模型和基于共同基因的分析方法。** 发现：有很多共同基因的疾病对，其疾病间的分离程度 S_{AB} 通常也会比较小，说明该模型还比较靠谱；同时，实际中有很多疾病间是没有共同基因的，那么基于共同基因来分析疾病间关系的分析方法此时就用不了，但该模型就不会受限于此，体现了该方法的优越性。接着作者就举了个这种例子(淋巴瘤和心肌梗死)来说明咯。

总结与讨论

虽然目前的数据还不全，但该文定义的两个度量刻画方法还是能定量地说明疾病模块确实存在，定义的疾病模块间的分离程度方法亦能反映疾病间生物学上的相似关系；接着，从实验所用数据可能有偏；所用的模型是基于无向图，而实际中信号传导等某些分子间互作过程是有向的；基于网络的社区发现方法是否能发现疾病模块等进行了补充讨论。最后，对该方法的用途作了进一步的阐述：如有助解读 GWAS 数据，帮助揭示药物新用途，助益发现那些未诊断出病因疾病的致病相关分子机制。可以预见，随着精确互作组学数据的增多，基于网络的分析方法在解读基因组变异相关数据上会更有用...

(终)

本文仅对该文的大体思路进行了描述，部分分析及结果未述及，具体实验细节可参考该文的补充材料，理解描述不当之处请指正。谢谢。

文献:

[1] J. Menche, A. Sharma, M. Kitsak, D. Ghiassian, M. Vidal, J. Loscazlo, A.-L. Barabasi Uncovering disease-disease relationships through the incomplete interactome Science 347:6224, 1257601-1 (2015).

在 http://www.barabasilab.com/pubs.php 可下载到该文全文 pdf，补充材料，实验数据及相关代码。Barabasi 实验室出产的关于复杂网络的文章都可在上述网址找到。

linwei@20160626