挖掘疾病相关基因：一种基于异构网络边预测的数据整合方法[1]

作者: Daniel S. Himmelstein, Sergio E. Baranzini

该文 20150709 发表于 PLOS Computational Biology

一、文章思路:

1. 整合多种类型数据，构建异构网络;

(1) 网络中的点:

Metanode	Count	Source	References
Disease	99	Disease Ontology	[43]
Gene	19,116	HGNC (coding)	[44]
Tissue	77	BRENDA (BTO)	[45]
Pathophysiology	8	manual	–
Positional	326	MSigDB (C1)	[44]
Perturbation	3,402	MSigDB (C2)	[36,37]
BioCarta	217	MSigDB (C2)	[46]
KEGG	186	MSigDB (C2)	[38]
Reactome	674	MSigDB (C2)	[39]
miRNA Target	221	MSigDB (C3)	[41]
TF Target	615	MSigDB (C3)	[40,41]
Cancer Hood	427	MSigDB (C4)	[47]
Cancer Module	431	MSigDB (C4)	[48]
GO Process	825	MSigDB (C5)	[42]
GO Component	233	MSigDB (C5)	[42]
GO Function	396	MSigDB (C5)	[42]
Oncogenic	189	MSigDB (C6)	[49]
Immunologic	1,910	MSigDB (C7)	[49]

Metanodes. The kind, number of corresponding nodes, and data source for each type of node.

doi:10.1371/journal.pcbi.1004259.t001

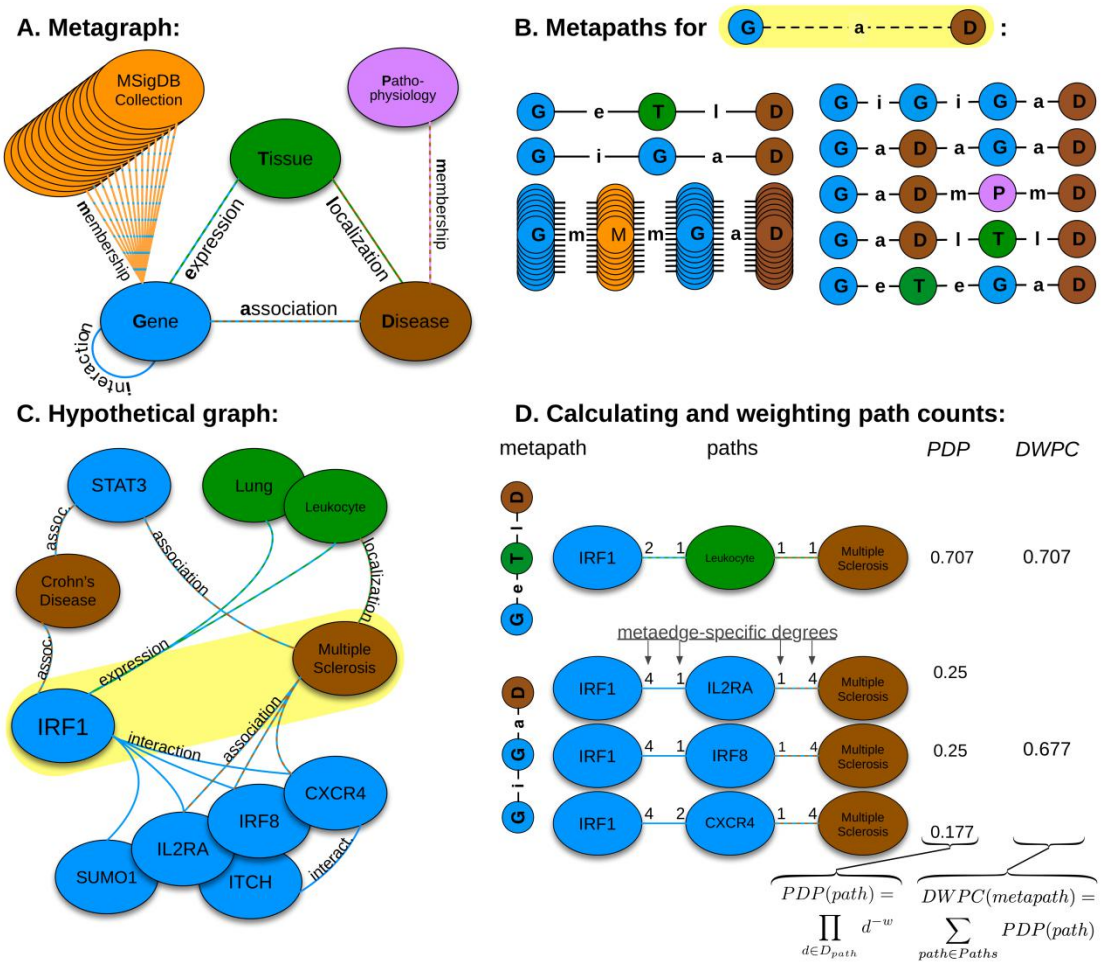
(2) 网络中的边:

Metaedge	Count	Source	References
Disease—association—Gene	938	GWAS Catalog	[5]
Disease—membership—Pathophysiology	90	manual	–
Disease—localization—Tissue	1,086	CoPub 5.0	[50]
Gene—expression—Tissue	251,366	GNF BodyMap	[51]
Gene—interaction—Gene	97,938	iRefIndex	[52]
Gene—membership—Positional	18,343	MSigDB (C1)	[44]
Gene—membership—Perturbation	366,211	MSigDB (C2)	[36,37]
Gene—membership—BioCarta	4,456	MSigDB (C2)	[46]
Gene—membership—KEGG	12,656	MSigDB (C2)	[38]
Gene—membership—Reactome	35,597	MSigDB (C2)	[39]
Gene—membership—miRNA Target	33,455	MSigDB (C3)	[41]
Gene—membership—TF Target	161,258	MSigDB (C3)	[40,41]
Gene—membership—Cancer Hood	41,913	MSigDB (C4)	[47]
Gene—membership—Cancer Module	48,220	MSigDB (C4)	[48]
Gene—membership—GO Process	75,155	MSigDB (C5)	[42]
Gene—membership—GO Component	34,880	MSigDB (C5)	[42]
Gene—membership—GO Function	23,578	MSigDB (C5)	[42]
Gene—membership—Oncogenic	30,166	MSigDB (C6)	[49]
Gene—membership—Immunologic	370,862	MSigDB (C7)	[49]

Metaedges. The kind, number of corresponding edges, and data source for each type of edge.

doi:10.1371/journal.pcbi.1004259.t002

2. 利用 MetaPath 方法，从异构网络中提取疾病到基因的特征;



说明:

(1) 选取长度小于等于 3 的基因到疾病的 MetaPath(上图 B);

(2) 最终利用 MetaPath 构建了 24 种特征，除了 2 种直接采用 path count 外，其余的采用了一种改进的 path count 方法:Degree-Weighted Path Count(DWPC)，该方法考虑了 MetaPath 经过节点的度数，计算方法如上图 C: 首先，列出每种 MetaPath 特征下基因到疾病所有的路径，并得到该路径各条边涉及的节点的度数; 其次，计算每一条路径的 path-degree product(PDP); 最后将每个基因到疾病各条路径的 PDP 相加，得到 DWPC 作为该基因在此 MetaPath 特征下的值;

3. 采用 Elastic Net(Regularized logistic regression model)方法，找出关键特征及基因;

分析时，选取的数据为至少含有 10 个相关基因的 29 种疾病，其中正例关系共 698 对，反例关系 551823 对. 将这些数据按 3:1 的比例划分为训练集和测试集，用到的评价指标主要为 ROC 曲线下的面积(AUROC).

Disease	Pathophysiology	HC-P	HC-S	LC-P	LC-S
Crohn's disease	immunologic	67	179	4	2
multiple sclerosis	immunologic	50	43	38	29
type 2 diabetes mellitus	immunologic	49	49	20	15
breast carcinoma	neoplastic	43	65	2	6
ulcerative colitis	immunologic	40	96	2	3
prostate carcinoma	neoplastic	34	202	3	4
type 1 diabetes mellitus	immunologic	33	56	9	6
rheumatoid arthritis	immunologic	30	27	20	11
coronary artery disease	metabolic	29	43	15	9
obesity	metabolic	28	22	34	18
celiac disease	immunologic	24	32	9	8
systemic lupus erythematosus	immunologic	22	35	14	8
refractive error	degenerative	21	11	2	1
primary biliary cirrhosis	immunologic	20	16	2	0
vitiligo	immunologic	20	27	4	0
age related macular degeneration	degenerative	18	30	11	18
metabolic syndrome X	metabolic	17	11	1	0
asthma	immunologic	17	23	13	4
psoriasis	immunologic	16	14	5	5
schizophrenia	psychiatric	15	27	20	13
chronic lymphocytic leukemia	neoplastic	14	16	3	4
migraine	unspecific	13	15	38	58
Alzheimer's disease	degenerative	12	11	27	18
Graves' disease	immunologic	12	15	1	1
Parkinson's disease	degenerative	12	21	8	13
atopic dermatitis	immunologic	11	15	5	1
bipolar disorder	psychiatric	11	34	26	74
lung carcinoma	neoplastic	10	14	6	6
ankylosing spondylitis	immunologic	10	5	6	6

Diseases. Associations were predicted for 29 diseases with at least 10 positives. For these diseases, the number of high-confidence primary (HC-P), high-confidence secondary (HC-S), low-confidence primary (LC-P), and low-confidence secondary associations (LC-S) that were extracted from the GWAS Catalog is indicated.

doi:10.1371/journal.pcbi.1004259.t003

注: Elastic Net

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

The elastic net penalty

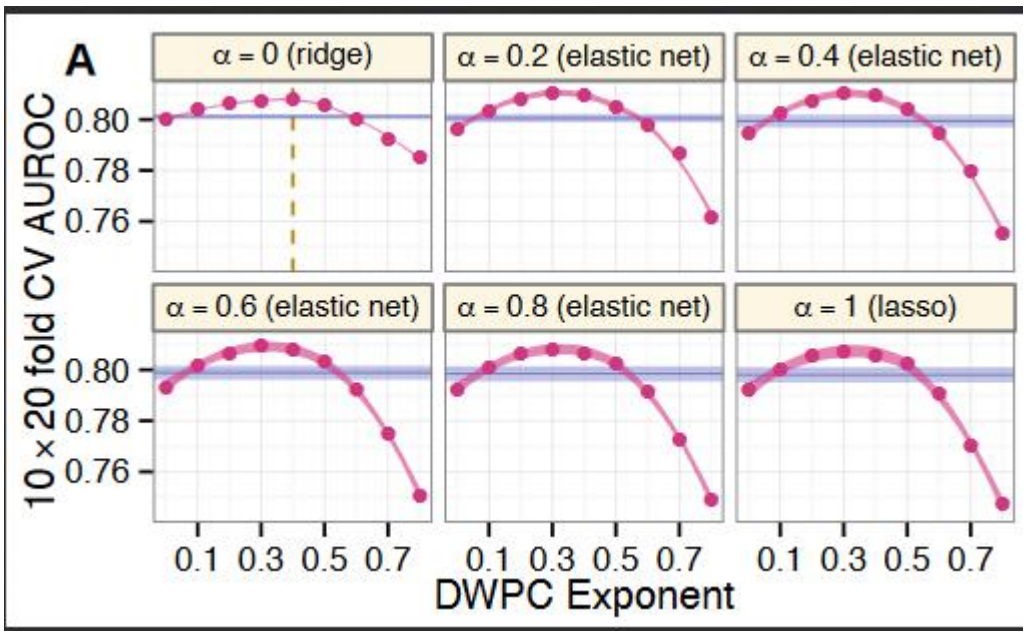
$$J(\beta) = \alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$$

$$(\text{with } \alpha = \frac{\lambda_2}{\lambda_2+\lambda_1})$$

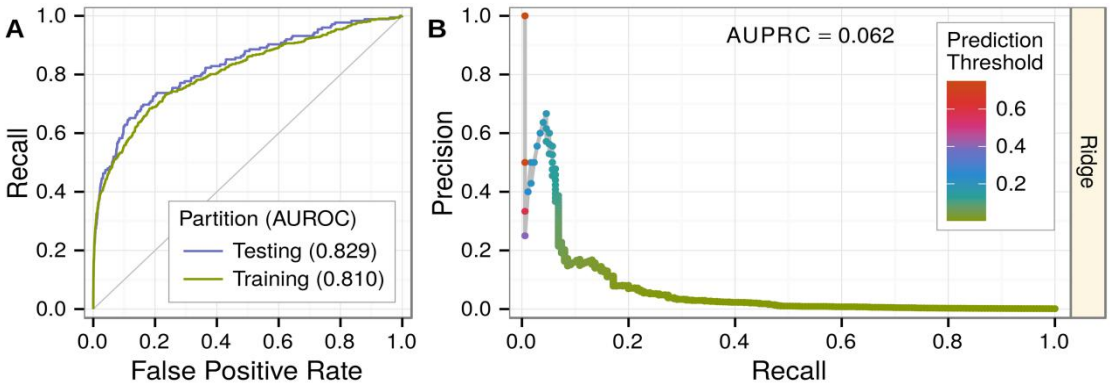
$$\min_{\beta} \|\mathbf{y}-\mathbf{X}\beta\|^2 \text{ s.t. } J(\beta) \leq t.$$

二、分析结果

(1) 在利用训练集调参时，发现 ridge regression model 得到的结果较为一致，因此后面主要利用 ridge regression 进行分析;



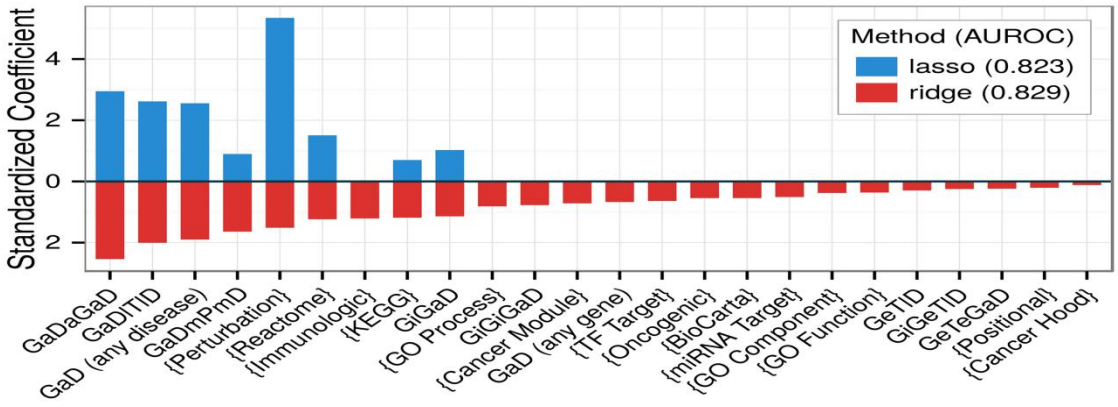
(2) Ridge regression 在测试集的表现:



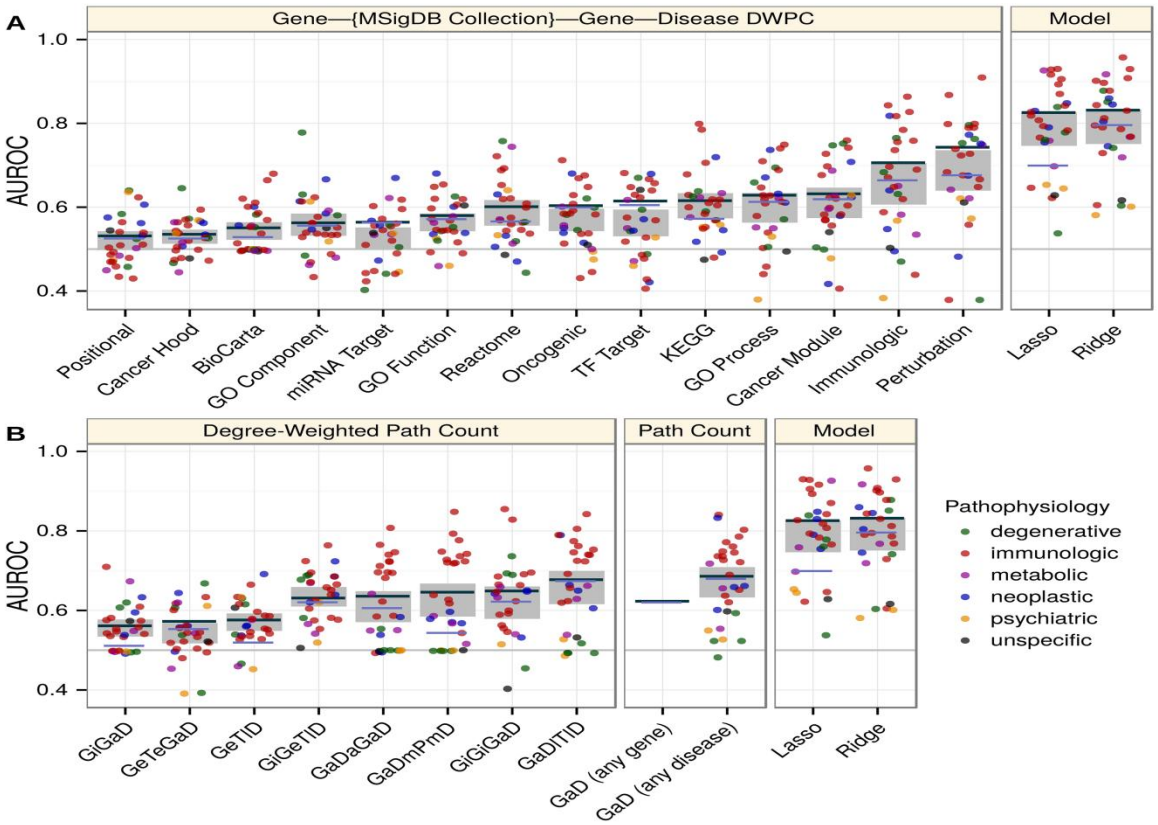
说明: A 图中模型在测试集和训练集的表现较为一致，说明没有过拟合, AUC 也较大，模型表现不错; B 图为 precision-recall 曲线图，是对 ROC 曲线的补充，因为该实验数据为类别不平衡数据，从 PRC 看，相应的 AUC 是比较低的，当预测阈值较为严格时，该模型的准确率还是较高的，作者说: 当 recall 为 10%时，准确率为 16.7%，而正例在样本中的比例为 0.13%，那么在均匀分布假设下此准确率是其 132 倍(两数相除为 128.5)...

(3) 利用全部数据(训练集+测试集)，探索各个特征的重要性:

I. Lasso 和 ridge 回归模型得到的参数大小比较(Lasso 在性能接近时选了 8 个特征);



II. 各个单独特征下得到的模型性能比较;

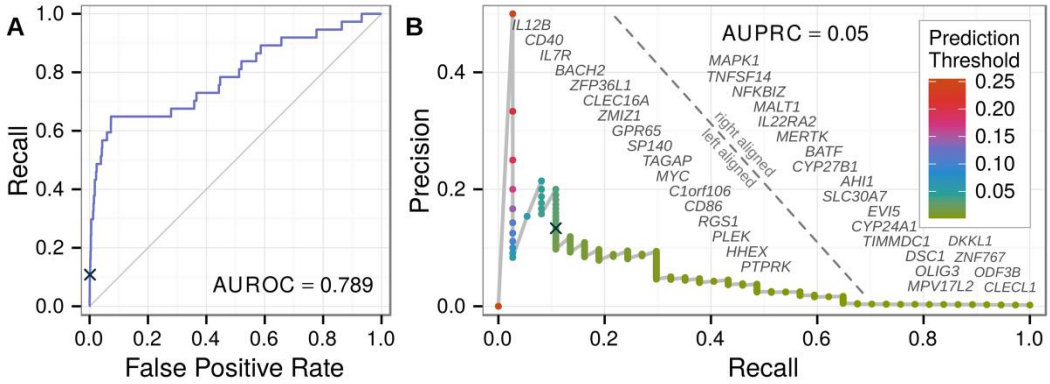


说明: 图中每点表示一种疾病, 点颜色表示疾病的类型, 黑色线段表示模型在所有疾病的整体表现(原文为 global performance, 不知为何物, 是否为 median?), 紫色线段表示模型在所有疾病的平均表现(原文为 average performance), 灰框为均值的 95%置信区间(不知如何计算, 置信区间不是应该会包含均值的吗?). A 图为来自 MSigDB 的特征, B 图为非 MSigDB 的特征.

(4) 案例分析: 多发性硬化症

思路: 1.在 The WTCCC2 multiple sclerosis (MS) GWAS 研究中, 鉴定出了 37 个相关基因, 将这些基因从已知数据集中去掉, 然后利用其余数据构建异构网络来提取特征, 以及训练 ridge 回归模型; 2. 将

WTCCC2 研究作为测试数据集，以其中的 37 个相关基因作为 true positives 来评估训练得到的模型性能；结果：在 WTCCC2 上得到的 AUROC(0.789)和之前测试集(0.83)的表现较为一致，而这 37 个相关基因的部分基因在 ridge 回归模型得到的结果中比较排序结果靠前(大部分基因落在小于下图中所选阈值(打叉)...)；



三、总结:

1. 论文整合了多种数据;
2. 利用 MetaPath 提取特征，改进后的方法考虑了节点的度数;
3. 由 Lasso 回归模型的表现可知，整合的部分数据之间是相关的，多种数据提供了冗余信息，数据并非越多越好;
4. ROC 曲线用于评估类别不平衡数据有缺陷，存在 AUROC 较大，但 AUPRC 很小的情况；目前该方法的表现有待进一步提高;
5. 缺乏同其他方法的比较.

文献:

[1] Himmelstein DS, Baranzini SE (2015) Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. PLoS Comput Biol 11(7): e1004259. doi:10.1371/journal.pcbi.1004259

文章链接:
journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004259#pcbi.1004259.s003

linwei@20161122