Hans Journal of Computational Biology 计算生物学, 2015, 5, 11-16
Published Online March 2015 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2015.51002

Hans汉斯

The Prediction of Interaction Gene for Greensickness Based on Semi-Supervised Learning

Weijuan Zhang, Hongmei Zhang, Feng Chen

College of Information Science and Engineering, Henan University of Technology, Zhengzhou Henan
Email: zweijuan_0303@163.com

Received: Mar. 6th, 2015; accepted: Mar. 18th, 2015; published: Mar. 24th, 2015

Abstract

Greensickness is an incurable disease, which can cause huge economic losses every year. Using a series of biological experiments to develop resistant plants and get more interaction gene is unrealistic. In order to dig more reliable genes under the premise of only having very few interaction gene, the paper mainly uses bio-IT, such as statistical techniques of Canonical Correlation Analysis and data mining techniques of Semi-Supervised Learning to study the related gene and gets the prediction of the interaction gene at the end. The result of the study can effectively guide the direction of the greensickness research, narrow the scope of the research and improve the research speed.

Keywords

Bio-IT, Greensickness, Canonical Correlation Analysis, Semi-Supervised Learning

基于半监督学习的黄萎病互作基因对的预测

张伟娟，张红梅，陈　峰

河南工业大学信息科学与工程学院，河南　郑州
Email: zweijuan_0303@163.com

收稿日期：2015年3月6日；录用日期：2015年3月18日；发布日期：2015年3月24日

摘　要

黄萎病(Greensickness)属于不可治愈性病害，每年均造成巨大的经济损失。为了培育抗病植株、得到更多的互作基因对，逐次进行生物实验排除是不现实的。为了在已知少量关联基因的情况下挖掘更多的可靠基因对，本文主要使用统计技术典型相关分析(Canonical Correlation Analysis, CCA)和数据挖掘技术半监督学习(Semi-Supervised Learning, SSL)等生物信息技术对相关基因进行学习，最终实现对关联基因的预测。研究结果能够有效地指导黄萎病抗病研究的方向、精确研究范围、提高研究速度。

关键词

生物信息技术，黄萎病，典型相关分析，半监督

1. 引言

黄萎病[1]-[3]是危害较为严重的维管束病害，其病原菌主要为大丽轮枝菌和黑白轮枝菌。它会对棉花、西红柿、茶叶等多种植物造成不可治愈的伤害，严重影响产量。在植株本身的免疫系统中，当黄萎病病菌的致病基因作用时，特定的植株抗病基因被激活，发挥免疫功能。双方进行作用的对应基因为关联的互作基因对，下文简称为“互作基因”。

目前，较为成熟的基因功能检测技术有基因敲除[4]、基因转导技术[5]、基因芯片[6]等。但这些技术都存在着技术成本昂贵、复杂、重复性差、分析泛围较狭窄等问题。而新兴的生物信息学[7]能够有效地解决成本昂贵、分析范围狭窄等问题。因此文中使用生物信息技术求取黄萎病的互作基因。

为了根据已知的标记基因对得到更多标记对，本文使用半监督学习对数据集进行挖掘。由于已知的黄萎病抗病基因对数量较少，因此需要先使用典型相关分析增加数据的相关性，即使用典型相关分析和半监督学习在已知非常少量黄萎病互作基因的基础上求解准确率较高的可能互作基因[8]。

2. 研究方法

2.1. 方法原理简介

在这次研究中，采用典型相关分析法[9]对数据进行处理，使具有关联作用的互作基因的相关系数更高，进一步使用半监督学习[10]挖掘可能的互作基因。其中：

典型相关分析基本原理是：为了从总体上把握两组变量：$X=\{X_1,X_2,\cdots,X_p\}$ 和 $Y=\{Y_1,Y_2,\cdots,Y_q\}$ 之间的相关关系，分别在两组变量中提取有代表性的两个综合变量 U_k 和 V_K (分别为两个变量组中各变量的线性组合)，在 X，Y 两组变量中，分别构建若干有代表性的变量组成有代表性的综合变量，通过研究这两组综合变量之间的相关关系，来代替这两组变量间的相关关系，这些综合指标称为典型变量，利用这两个综合变量之间的相关关系来反映两组指标之间的整体相关性。

具体的求解过程不在这里赘述，有兴趣的可以参考文献 8，实验需要的是在保证变量 U，V 的相关系数最大，即

$$\rho_{U_1,V_1}=\frac{a_1^T C_{xy} b_1}{\sqrt{a_1^T C_{xx} a_1}\sqrt{b_1^T C_{yy} b_1}} \tag{1}$$

最大的前提下，得到典型系数 a、b、特征系数 λ。

半监督学习是机器学习的一中，在机器学习中，如果只使用有监督学习只关注少量的已标记数据，那么得到的学习模型不具有很好的泛化能力，同时会造成大量未标记数据样本的浪费。如果只使用无监督学习只关注大量的为标记数据，那么会忽略极具有价值已标记数据。因此，研究如何综合利用少量已标记数据和大量为标记数据来提高学习性能的半监督学习称为当前机器学习和模式识别的重要研究领域之一。这种学习方法符合现有大多数需要机器学习的情况。

2.2. 研究流程

文中根据主要的工作原理，制定了具体的研究方法，其具体流程如下所示：

1) 使用 BLAST (Basic Logical Alignment Search Tool) [11]，对采集自生物数据库的 DNA，蛋白质等序列进行对比，得到序列相似性参数。

2) 对数据进行离群处理、标准化后作为两组变量 X，Y 的初始值。

3) 为了扩大变量间的相关性，增大准确率，使用典型相关分析对变量 X，Y 进行初始标记，简述过程如下：

进行典型相关分析，得到参数 a、b、λ，假设有 m 个不同的特征值，那么存在 m 个映射，有 $a(1\cdots m)$，$b(1\cdots m)$，$\lambda(1\cdots m)$。

如图 1 中，(x_0, y_0) 可以看作是已标记的一组数据对其他数据如 (x_i, y_j) 进行标记。首先，使用公式 2 分别计算各基因的相似值，

$$\begin{cases} sx_i = a_2^T (x_i - x_0) \\ sy_j = b_2^T (y_j - y_0) \end{cases} \tag{2}$$

然后使用公式 3 计算基因对的相似度。

$$p(i, j) = \sum_{k=1}^{m} \lambda_k \sqrt{sx_i^2 + sy_j^2} \tag{3}$$

得到置信度 $p(i, j)$，若置信度极高，可进行标记。

得到置信度的结果如图 2 所示，其中左图表示每个基因对所对应的置信度，右图显示所有置信度值得分布区域，从中可以发现存在置信度非常高的少量基因对，它们可以作为初始标记的首选。大多数基因对的置信度在 0.35~0.5 之间，这些基因对具有一定的联系，但相关性不足以标记为关联基因对。

之后取值最大的 r 个基因对标记为 1，得到 $r + 1$ 个已标记数据，有效地扩大了已标记数据的个数。分别设置 r 为 100、500、1000、5000，比较不同的置信度对结果的影响。

4) 对扩展后的数据进行半监督学习，过程如下：

选择半监督算法，基于此次数据的特性，选择的是基于流型假设，够降维的半监督算法，这次实验使用的算法为 Laplacian [12]算法，对数据进行降维，然后生成拉普拉斯图，得到每个点的邻居。对每个点的邻居进行分析，使用 K 近邻算法[13]对未标记数据进行标记。得到所有正例，即标记出的互作基因。

3. 实验

3.1. 简介

实验采用的数据有两大集合，集合一为植物抗病基因集合包括 265 个数据，集合二是黄萎病致病基因集合包括 65 个数据。它采自主要源于生物数据库 NCBI、PDB 及其关联数据库。在已知的少量基因对中，选取互作基因对(gi|283764861 [14], gi|375968911 [15])为标记基因，使用文中提出的方法对其他基因

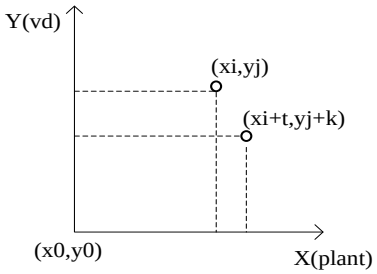

Figure 1. The similar relation of gene
图 1. 基因对相似关系

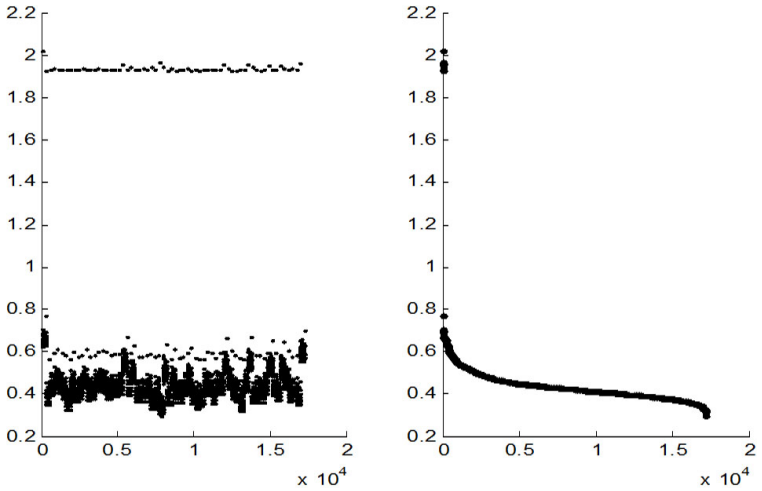

Figure 2. The confidence value of gene
图 2. 基因对的置信度

对进行标记，即对基因对$\left(x_i, y_j\right)$ (其中，$i = 1\cdots265$，$j = 1\cdots65$，即共有 $265 \times 65 = 17{,}225$ 个基因对)，通过实验标记其是否为互作基因。并通过比较预测的基因对已知的基因对的覆盖率检验方法的正确性，并得到最佳实验方案。

3.2. 实验结果

采取文中 2.2 节所描述的方案对实验数据进行互作基因的预测，其中，第三步初始标记后得到的标记结果如图 3 所示，图中 r 表示初始标记的个数。

第四步进行最终标记得到的互作基因结果图如图 4 所示，图中的四个子图分别对应图 3 中的进行 r 个初始标记后进行最终半监督学习得到的结果。

3.3. 结果分析

由上节的结果图可知：数据具有聚集性，即某些行或某些列数据密集，这样的基因具有普遍关联性，在抗病过程中起重要作用，如植物基因数据集中的第 27 个基因 gi|214011438 [16]、病菌中的第 1 个基因 gi|333352894 等基因具有普遍关联性。结果与现有研究符合，应该加大对它们的研究力度。

此外，数据具有独立性，即某些数据在其所在行列中只有自身，这种数据表示这对基因有一对一的关系，如基因对(gi|510708 complete genome111 [17], VDAG_05753T0)，具有其独特的功能。

对标记的结果进行统计分析，比较各种标记分案下最终结果对已知的少量基因对的覆盖率，结果如表 1 所示。

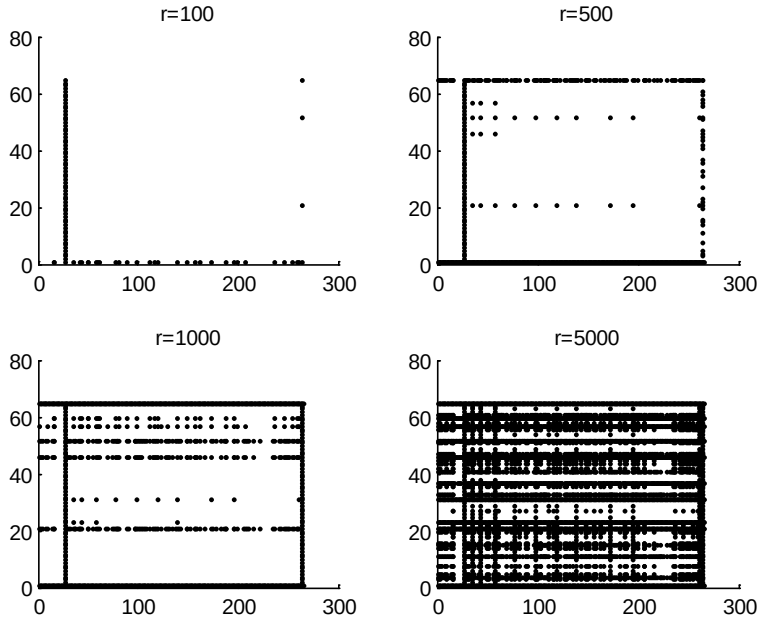

Figure 3. The result of initial label
图 3. 初始标记结果

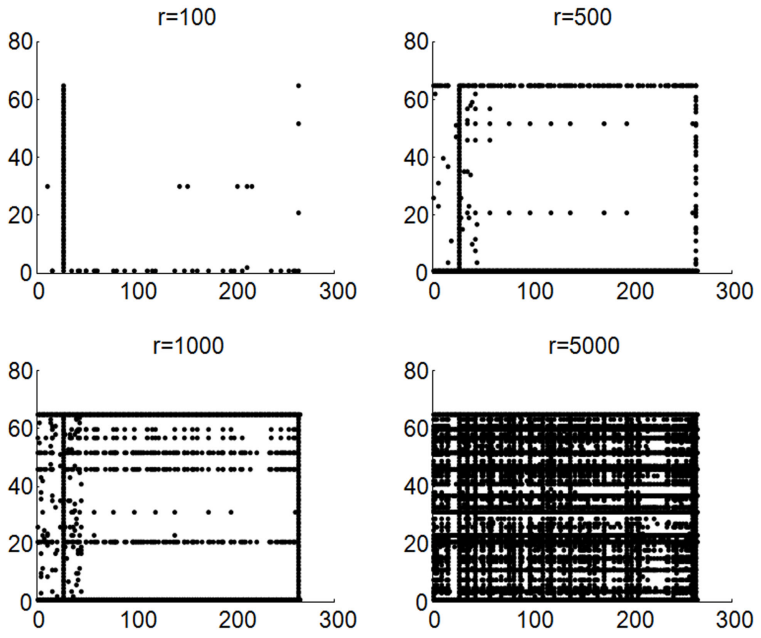

Figure 4. Final study result
图 4. 最终学习结果

Table 1. The coverage rate of result
表 1. 结果覆盖率

r	100	500	1000	5000
初始标记	12%	18%	20%	50%
半监督学习	27.5%	35%	56%	92%

由覆盖率结果可知，使用文中的方法预测黄萎病抗病基因对是可行的，预测结果覆盖已知的基因对的准确率较高，预测出的互作基因结果具有很高的研究参考价值。

综上可知，使用 CAA + SSL 的半监督学习方法可以在已知十分少量的互作基因对的情况下，以较高的准确率预测更多的基因对数据，挖掘结果具有可参考性。

4. 结束语

通过实验可以得到高准确率的基因对，为生物实验指明方向，减小研究的范围，有利于提高研究速度，早日攻克黄萎病。此外，此论文所使用的方法同样适用于其他在已知极少量标记数据的基础上求解其他基因对的情况。

基金项目

国家自然科学基金 61203265，河南省重点项目 122102110106。

参考文献 (References)

[1] 张保龙, 承泓良, 杨郁文 (2012) 棉花抗黄萎病研究进展. *中国农业科学技术出版社*, 北京.

[2] 宋学贞, 杨国正 (2013) 棉花抗黄萎病育种研究进展. *中国农业学报*, **21**, 16-22.

[3] 张志 (2010) 我国棉花抗枯、黄萎病育种存在的问题及对策. *河南农业*, **19**, 13-14.

[4] 陶果, 信吉阁 (2013) 肖晶等.基因敲除技术最新研究进展及其应用. *安徽农业科学*, **29**, 11605-11608.

[5] 袁莉, 付博, 陈香美等 (2004) 三种基因转导方法在不同代龄复制性衰老细胞中的比较研究. *中国生物化学与分子生物学报*, **2**, 257-263.

[6] 张骞, 盛军 (2008) 基因芯片技术的发展和应用. *中国医学科学院学报*, **3**, 344-347.

[7] 李霞, 李亦学, 廖飞 (2010) 生物信息学. 人民卫生出版社, 北京.

[8] Zhou, Z.H., Zhan, D.C. and Yang, Q. (2007) Semi-supervised learning with very few labeled training examples. *Proceedings of the National Conference on Artificial Intelligence*, Canada, 2007, 675-680.

[9] 孙权森, 曾生根, 王平安等 (2005) 典型相关分析的理论及其在特征融合中的应用. *计算机学报*, **9**, 1524-1533.

[10] 杨剑, 王钰, 钟宁 (2007) 流形上的 Laplacian 半监督回归. *计算机研究与发展*, **7**, 1121-1127.

[11] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

[12] Belkin, M. and Niyogi, P. (2003) Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15**, 1373-1396.

[13] 肖宇, 于剑 (2008) 基于近邻传播算法的半监督聚类. *软件学报*, **11**, 2803-2813.

[14] Kawchuk, L.M., Hachey, J., Lynch, D.R., Kulcsar, F., van Rooijen, G., Waterer, D.R., et al. (2001) Tomato *Ve* disease resistance genes encode cell surface-like receptors. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 6511-6515.

[15] de Jonge, R., van Esse, H.P., Maruthachalam, K., Bolton, M.D., Santhanam, P., Saber, M.K., et al. (2012) Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 5110-5115.

[16] Zhang, Y., Wang, X., Yang, S., Chi, J., Zhang, G.Y. and Ma, Z.Y. (2011) Cloning and characterization of a *Verticillium* wilt resistance gene from *Gossypium barbadense* and functional analysis in *Arabidopsis thaliana*. *Plant Cell Reports*, **30**, 2085-2096.

[17] Ayres, M.D., Howard, S.C., Kuzio, J., Lopez-Ferber, M. and Possee, R.D. (1994) The complete DNA sequence of *Autographa californica* nuclear polyhedrosis virus. *Virology*, **202**, 586-605.