题目

摘要

脓毒症是威胁患者生命的严重疾病，是重症监护病房患者死亡的主要原因。脓毒症的诊断和治疗不及时会给患者带来非常严重的后果，通过分子检测可以实现快速诊断，使早期干预成为可能，从而最小化死亡率。最近的研究表明，长非编码 RNA（lncRNA）可以调节促炎基因并且与脓毒症中的器官功能障碍有关。但是由于技术差异和系统实验偏差，确定具有绝对丰度的 lncRNA 特征是非常困难的。本论文提出了一种新型的诊断方案，利用在脓毒症患者和正常样本的对照之间反转的 lncRNA 对的相对表达（例如，在脓毒症患者中 lncRNAi > lncRNAj，而在正常对照中 lncRNAi < lncRNAj），以识别出 14 个 lncRNA 对作为脓毒症的诊断标志。之后，将该标志应用于独立的验证集中，评估其在不同年龄和标准化方法下的预测性能。与常见的机器学习模型和现有的诊断方法相比，本论文提出的方法在来自同一年龄组的验证数据集上都获得了更好的表现。

关键词： Sepsis, Diagnostics, Signature, Long non-coding RNA, Relative expression

1 引言

脓毒症是威胁患者生命的严重疾病，是重症监护病房（ICU）患者死亡的主要原因。根据美国疾病预防控制中心的报告，每年有超过 170 万人患脓毒症，每年有高达 27 万的美国人死于此病，每三名在医院死亡的患者中就有一名患有脓毒症。这是由免疫系统对感染的过度反应引起的。免疫系统释放的化学物质扩散到整个身体并导致炎症。脓毒性休克是脓毒症的一种亚型，当患者的血液循环和细胞代谢发生致命异常时。临床上脓毒症的诊断主要依据症状和一系列的医学检查，包括血液、尿液、伤口分泌物和粘液分泌试验。采取这些典型的检测可能会导致诊断和干预的延误。此外，基于现有的试验区分脓毒症和非感染性炎症具有挑战性。目前，机器学习方法被广泛应用于生物建模和生物标志物检测。目前的大部分与脓毒症相关的转录组研究都是基于检测到的转录本 (如 mRNA 或 lncRNA）的绝对表达丰度。但是由于系统性的实验偏差，使用绝对丰度进行诊断并不可靠，甚至存在偏差。此外，使用绝对表达值进行生物标志物的研究可能会受到批次效应和预处理方法的影响。本论文提出了一种利用样本内 lncRNA 对的相对表达量进行脓毒症诊断的方法。本论文首先进行样本内比较，得到所有可能的 lncRNA 对的相对表达量，并进行跨样本分析，筛选出在脓毒症和对照样本中显著改变方向的 lncRNA 对。然后，在脓毒症基因表达矩阵中评估识别到的 lncRNA 对，并与一系列机器学习模型进行比较。最后，分析对这些 lncRNA 对的生物学功能进行了研究。本论文第一次提出基于基因对（pair）的相对表达量的方法辅助传染病的诊断方法。对于很多使用传统诊断方法诊断传染病困难的传染病的诊断带来了一个新的思路。对于传染病的快速、准确诊断和降低传染病诊断成本有着重大的积极意义。

2 相关工作

以下是对脓毒症和基因表达量的一些基本介绍

2.1 脓毒症

脓毒症是威胁患者生命的严重疾病，是重症监护病房（ICU）患者死亡的主要原因。临床上脓毒症的诊断主要依据症状和一系列的医学检查，包括血液、尿液、伤口分泌物和粘液分泌试验。采取这些典型的检测可能会导致诊断和干预的延误。此外，基于现有的检测方法来区分脓毒症和非感染性炎症也具有一定的挑战性 [1]

2.2 基因表达量

基因表达量是指在细胞中某个基因的转录产物（RNA）或翻译产物（蛋白质）的数量。在研究生物学和遗传学时，了解基因在特定条件下的表达水平是非常重要的。

3 本文方法

3.1 本文方法概述

本论文提出了一种利用样本内 lncRNA 对的相对表达量进行脓毒症诊断的方法。本论文首先进行样本内比较，得到所有可能的 lncRNA 对的相对表达量，并进行跨样本分析，筛选出在脓毒症和对照样本中显著改变方向的 lncRNA 对。然后，在脓毒症基因表达矩阵中评估识别到的 lncRNA 对，并与一系列机器学习模型进行比较。最后，分析对这些 lncRNA 对的生物学功能进行了研究。基因的绝对表达丰度经常受到许多技术变异的影响，包括实验设计、样本处理、RNA 量和提取程序，以及标准化方法和批次效应。样本内基因间的相对表达是可靠且更有力地检测生物信号的方法。因此，本论文利用每对 lncRNAs 之间的表达水平来检索与疾病相关的 lncRNA 对。反向对被定义为在正常情况下具有相同相对表达顺序（lncRNAi > lncRNAj），而在脓毒症患者中具有相反的顺序（lncRNAi < lncRNAj）。之后，我们确定了反向 lncRNA 对，并将它们用作脓毒症的诊断预测标志。iPAGE 的工作流程如图 1所示。

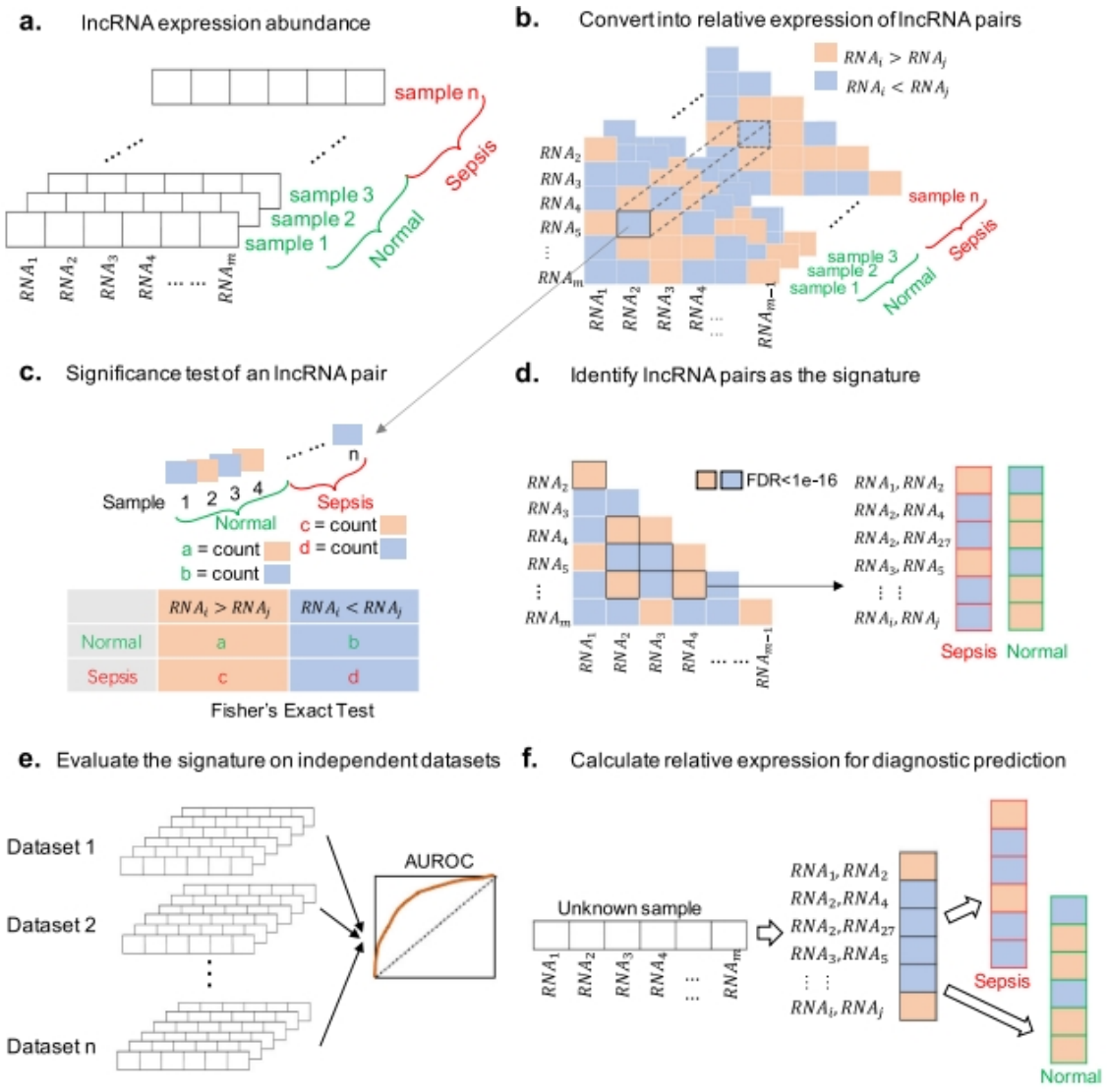

图 1. IPAGE 示意图

3.2 基因对（pair）的构建

在 iPAGE 算法中，基于基因表达量的绝对丰度进行了每个 lncRNAs 的配对比较。如图 1a 所示，由样本检测到的 lncRNAs 用向量表示：

$$G^{(k)} = \left(RNA_1^{(k)}, RNA_2^{(k)}, \ldots, RNA_m^{(k)}\right)$$

其中 RNAi 表示 lncRNAs 的绝对丰度，上标 (k) 代表给定数据集中的第 k 个样本。lncRNA 对（RNAi，RNAj）的相对表达定义为：

$$r^{(k)} = I\left(RNA_i^{(k)} - RNA_j^{(k)}\right)$$

其中 I(x) 用于指示 x 是否大于零。如果 RNAi 大于 RNAj，则 lncRNA 对（RNAi，RNAj）的相对表达将为 1。否则，相对表达将为-1。为了在每个样本中将绝对表达丰度转换为相对表达，对每两个 lncRNAs 执行减法运算（图 1b），即 RNAi-RNAj。R(k) 是由样本中所有 lncRNA 对的相对表达构成的向量。

$$
\begin{aligned}
R^{(k)} &= \left(r_{12}^{(k)}, r_{13}^{(k)}, \ldots, r_{1m}^{(k)}, r_{23}^{(k)}, r_{24}^{(k)}, \ldots, r_{2m}^{(k)}, \ldots, r_{(m-1)m}^{(k)}\right) \\
&= \left(I\left(RNA_1^{(k)} - RNA_2^{(k)}\right), I\left(RNA_1^{(k)} - RNA_3^{(k)}\right), \ldots, I\left(RNA_1^{(k)} - RNA_m^{(k)}\right),\right. \\
&\ldots, I\left(RNA_2^{(k)} - RNA_3^{(k)}\right), I\left(RNA_2^{(k)} - RNA_4^{(k)}\right), \ldots, I\left(RNA_2^{(k)} - RNA_m^{(k)}\right), \\
&\left.\ldots, I\left(RNA_{m-1}^{(k)} - RNA_m^{(k)}\right)\right)
\end{aligned}
$$

lncRNAs 的相对表达 R 与标签 Y 一起形成了一组训练数据 S

$$S = \{(R^{(1)}, Y^{(1)}), (R^{(2)}, Y^{(2)}), ..., (R^{(n)}, Y^{(n)})\},$$

其中对于正常样本，Y 等于 0，对于脓毒症样本，Y 等于 1。尽管绝对表达值在样本之间可能普遍存在偏差，但是样本内 lncRNA 对的相对表达是稳定的。

3.3 筛选反转基因对（pair）

为了筛选出脓毒症的样本内特征，进行了脓毒症和正常样本之间的跨样本分析（图 1c）。提取了在脓毒症和正常样本之间的相对表达中存在显著差异的 lncRNA 对。考虑了两种情况

$$r_{ij}^{(k)} = 1\,(RNA_i^{(k)} > RNA_j^{(k)}) \text{ and } r_{ij}^{(k)} = -1\,(RNA_i^{(k)} < RNA_j^{(k)})$$

对于正常组中的每个 lncRNA 对，计算了跨 n 个样本的 RNAi > RNAj 的数量，得到了如下的列联表

	$RNA_i^{(k)} > RNA_j^{(k)}$	$RNA_i^{(k)} < RNA_j^{(k)}$
Normal	$\mathrm{a} = \frac{1}{2}\sum_{k=1}^{n}\left(r_{ij}^{(k)} + 1\right) \times \left(1 - Y^{(k)}\right)$	$\mathrm{b} = \frac{1}{2}\sum_{k=1}^{n}\left(1 - r_{ij}^{(k)}\right) \times \left(1 - Y^{(k)}\right)$
Sepsis	$\mathrm{c} = \frac{1}{2}\sum_{k=1}^{n}\left(r_{ij}^{(k)} + 1\right) \times Y^{(n)}$	$\mathrm{d} = \frac{1}{2}\sum_{k=1}^{n}\left(1 - r_{ij}^{(k)}\right) \times Y^{(n)}$

接下来，利用 Fisher 精确检验来衡量每个 lncRNA 对区分脓毒症样本和对照样本的能力。p 值的计算公式为：

$$\mathrm{p} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!b!c!d!n!}$$

其中 n = a + b + c + d。然后，利用 Bonferroni 校正进行多重比较校正。基于调整后的 p 值，筛选出在正常样本和脓毒症样本之间显著改变的 lncRNA 对，作为特征（图 1d）。

3.4 分类和测试

经过筛选过后的 lncRNA 对可能可以用作诊断脓毒症的标志。本论文基于 lncRNA 对的相对表达量建立了一个分类器。当 RNAi > RNAj 的一对指示脓毒症时，r 被赋值为 1，否则被赋值为 −1。通过对所有差异 lncRNA 对的总和进行计算，得到表示脓毒症可能性的风险评分 。该分类器通过在八个独立测试集上进行 AUROC 测试（图 1e），之后可以用于诊断预测（图 1f）。

4 复现细节

4.1 与已有开源代码对比

原论文是对于脓毒症（Sepsis）进行的实验，我探究的是本论文中提出的 IPAGE 算法在莱姆病（lyme disease）上的应用，原论文并没有给出实验所用到的数据集，我在 GEO 数据集中筛选出了能够被用于本实验的莱姆病数据集，并对他们进行数据处理，处理完数据集之后再将最大的几个数据集合成为一个训练集，剩下的数据集作为测试集，之后参考原论文中给出的构建算法所需的 PAIR 代码修改应用到我的实验之中，并按照原论文中的思想筛选出对于莱姆病影响显著的基因对，并将它们记录下来，我还对与本论文中提出的算法进行了对照实验，通过基因的绝对表达量的 T test 实验筛选出在正常样本和患病样本中表达差异显著的基因对用于测试

4.2 收集数据、数据处理

在 GEO 数据集中筛选出莱姆病 (lyme disease) 相关的人类 Microarray 和 RNA-Sequence 数据集及其基因表达矩阵，从每个 Series 中选出莱姆病样本和正常样本并分别对其打上标签。找到每个 Series 对应的基因测序平台，将基因表达矩阵中的基因序列与基因名进行一一对应，再从中筛选出与人体免疫相关的基因。之后再按照 IPAGE 算法的步骤，导入数据，构建基因对（pair）并计算 p value，a，b，c，d，根据 p value 筛选 pair，再根据 pair 中基因对的相对大小作为输入数据进行训练和测试。

4.3 实验过程

按照 IPAGE 算法的步骤，导入数据，构建基因对（pair）并计算 p value，a，b，c，d，根据 p value 筛选 pair，再根据 pair 中基因对的相对大小作为输入数据进行训练和测试。对照实验：利用 T test 方法对莱姆病样本和正常样本中基因的绝对表达量筛选出在正常样本和脓毒症样本之间有着显著差异的基因，根据这些差异基因的基因表达量对传染病的诊断进行预测。最后将两种方法与不同机器学习方法的实验结果进行比较。

4.4 创新点

将原论文所提出的 IPAGE 算法应用到其他传染病的检测中。

5 实验结果分析

首先是 IPAGE 算法的基因相对表达量筛选出的 pair 方法作为输入特征加上一个分类器进行预测，如图 2 所示，pair 方法加上 Lasso 分类器的效果要明显强于 pair 加上其他传统机器学习分类器。

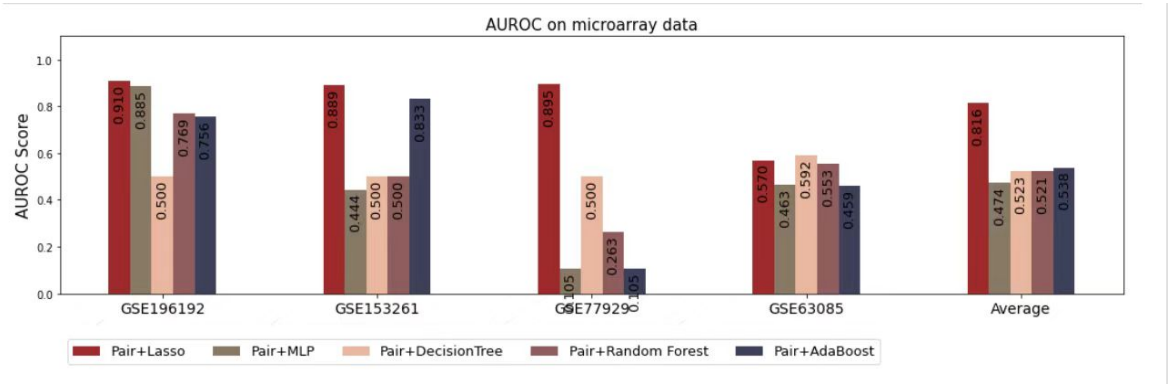

图 2. 实验结果示意

接下来进行 T test 加上机器学习的实验，在训练集上筛选出差异基因，再将差异基因的基因绝对表达量用于测试，并最终与 pair 方法加上机器学习方法的实验结果进行对比，发现 pair 方法加上 Lasso 分类器的效果要明显强于 pair 加上其他传统机器学习分类器，强于使用 T test 的基因绝对表达量诊断方法。

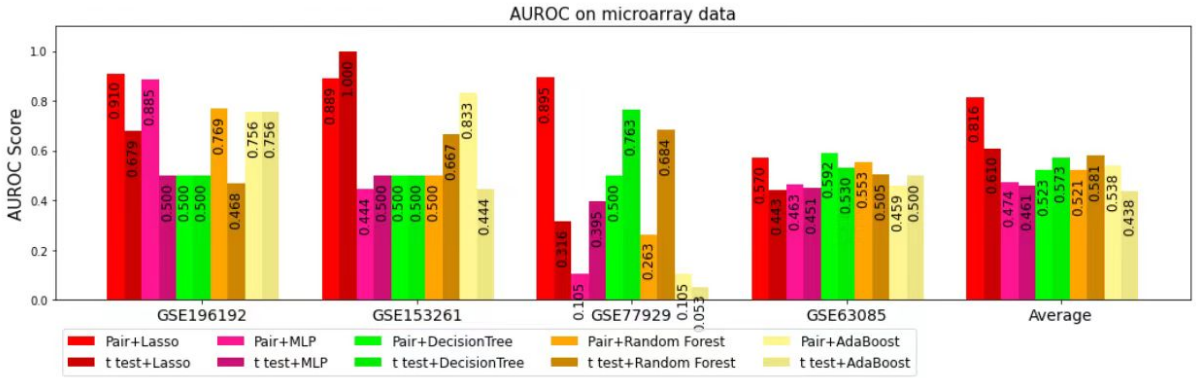

图 3. 实验结果示意

下图中展示了本实验中筛选出的用于诊断莱姆病的基因对，可以应用于后续的莱姆病诊断。

0	BST2	BCL2
1	EP300	CSF2RB
2	EP300	CSF3R
3	HLA-DQA1	CXCR4
4	MRPS5	EP300
5	PSMB9	ATM
6	PSMB9	IL24
7	PSMB9	JAK1
8	PSMB9	NFATC1
9	SERPINB2	CR1
10	SERPINB2	FCGR1A
11	SERPINB2	LTF
12	SLAMF7	EP300
13	SLAMF7	IGF2R
14	SLAMF7	RORA
15	SLAMF7	RUNX3

图 4. 莱姆病 Pair

6 总结与展望

本实验主要在 IPAGE 算法的基础上，将基因相对表达量的思想运用到传染病的检测中并获得了很好的效果。未来可以将基因相对表达量的传染病诊断应用于更多的传染病并且分别找到对各种传染病影响显著的基因对，并从他们之中找到能够直接诊断患者所患传染病的类型，从一个是和不是的二分类问题转化为一个直接诊断为哪一类传染病的多分类问题。由于基因数据集的获取相对困难，经过筛选后的数据集和数据量相对较小，未来希望可以在更大的数据集，更大的数据量上进行实验，找到对于各种病原体影响显著的基因对，对医学诊断做出贡献。

参考文献

[1] Cohen J, Vincent JL, Adhikari NK, Machado FR, Angus DC, and Calandra T. Sepsis: a roadmap for future research. *Lancet Infect Dis*, 15(5):581–614, 2015.