©2022 Chin J Biotech, All rights reserved

·生物技术与方法·

基于基因互作网络熵量化细胞分化状态

关天昊，高洁

江南大学 理学院，江苏 无锡 214122

关天昊, 高洁. 基于基因互作网络熵量化细胞分化状态. 生物工程学报, 2022, 38(2): 820-830.
GUAN TH, GAO J. Quantifying the state of cell differentiation based on the gene networks entropy. Chin J Biotech, 2022, 38(2): 820-830.

摘 要：细胞动态过程的研究表明，细胞在动态过程中会发生状态变化，主要由细胞内部的基因表达情况控制。随着高通量测序技术的发展，大量的基因表达数据能够在单细胞水平上获得细胞真实的基因表达信息。然而，现有大多数研究方法需要使用除基因表达以外其他的信息，带来了额外的复杂度和不确定性。此外，普遍存在的“缺失值”事件更是影响了对细胞动态发展的研究。为此，文中提出了基因互作网络熵方法，来量化细胞的分化状态，以此来研究细胞的发展动态。具体而言，通过借助网络的稳定性，依据基因间的关联来构造细胞特异性网络，并定义基因互作网络熵，从而将不稳定的基因表达数据转换成相对稳定的基因互作网络熵。该方法没有额外的复杂度和不确定性，同时在一定程度上规避了“缺失值”事件的影响，能更可靠地表征诸如细胞命运等生物学过程。将该方法应用于头颈部鳞状细胞癌和慢性髓细胞白血病两个单细胞 RNA-seq 数据集，不仅能够有效区分恶性细胞与良性细胞、有效区分不同分化时期，还可以有效反映疾病疗效过程，证明了利用基因互作网络熵方法研究细胞动态的潜力。该方法旨在探索单细胞混乱程度水平的动态信息，从而研究生物系统进程中的动态变化情况。研究结果能够为细胞分化、追踪癌症发展以及疾病对药物反应过程等研究提供科学建议。

关键词：单细胞 RNA-seq 数据；基因互作网络熵；细胞动态过程；分化状态

Received: February 11, 2021; **Accepted:** July 25, 2021; **Published online:** August 11, 2021
Supported by: National Natural Science Foundation of China (11831015, 91730301)
Corresponding author: GAO Jie. Tel: +86-510-85912033; Fax: +86-510-85910532; E-mail: gaojie@jiangnan.edu.cn
基金项目：国家自然科学基金 (11831015, 91730301)

Quantifying the state of cell differentiation based on the gene networks entropy

GUAN Tianhao, GAO Jie

School of Sciences, Jiangnan University, Wuxi 214122, Jiangsu, China

Abstract: Studies of cellular dynamic processes have shown that cells undergo state changes during dynamic processes, controlled mainly by the expression of genes within the cell. With the development of high-throughput sequencing technologies, the availability of large amounts of gene expression data enables the acquisition of true gene expression information of cells at the single-cell level. However, most existing research methods require the use of information beyond gene expression, thus introducing additional complexity and uncertainty. In addition, the prevalence of dropout events hampers the study of cellular dynamics. To this end, we propose an approach named gene interaction network entropy (GINE) to quantify the state of cell differentiation as a means of studying cellular dynamics. Specifically, by constructing a cell-specific network based on the association between genes through the stability of the network, and defining the GINE, the unstable gene expression data is converted into a relatively stable GINE. This method has no additional complexity or uncertainty, and at the same time circumvents the effects of dropout events to a certain extent, allowing for a more reliable characterization of biological processes such as cell fate. This method was applied to study two single-cell RNA-seq datasets, head and neck squamous cell carcinoma and chronic myeloid leukaemia. The GINE method not only effectively distinguishes malignant cells from benign cells and differentiates between different periods of differentiation, but also effectively reflects the disease efficacy process, demonstrating the potential of using GINE to study cellular dynamics. The method aims to explore the dynamic information at the level of single cell disorganization and thus to study the dynamics of biological system processes. The results of this study may provide scientific recommendations for research on cell differentiation, tracking cancer development, and the process of disease response to drugs.

Keywords: single cell RNA-seq data; gene interactions network entropy; cellular dynamic processes; differentiation state

当前，对细胞动态过程的研究表明，细胞在动态过程中会发生状态变化，产生表型稳定的细胞类型以及在表型之间转变的过渡类型，这一系列转变主要由细胞内部的基因表达情况控制。在生物意义上，识别生物组织谱系不仅提供关于正常组织发育和体内平衡的信息，还提供了关于癌症等病症的相关信息。而了解组织形成的谱系，本质上是研究细胞的动态发展过程。传统的研究方式是通过细胞的可遗传标记进行追踪，然而这些标记由于数量有限，可能会掩盖标记基因在细胞亚群中的异质性，这往往会导致细胞分化研究中的错误结论。由于单细胞实验可以产生横截面群体快照，能够反映细胞在其分化或过渡状态中的变化，为解决精确描绘细胞分化问题提供了新的方式。2009 年，Tang 等[1]首次发表单细胞 RNA-seq 技术，打破

了传统研究使用细胞群体均值来表示基因表达水平的情况,能够在单细胞水平上获得细胞真实的基因表达信息。此后,相关技术迅猛发展,2012 年 Hashimshony 等[2]开创了 cel-seq 技术,采用线性扩增的测序方法,数据的错误率比较低;2013 年 Picelli 等[3]发表的 smart-seq2 技术能够获得全长的转录测序数,测得的数据具有更好的敏感性,能够在每个细胞中检测到更多基因。随着这些高通量测序技术的发展,研究者能够获得在单细胞水平下大量的基因表达数据,使得从单细胞层面了解细胞状态和分化发育提供了可能。

但是,“缺失值”问题对人们分析与解读这些数据造成了困扰。事实上,早在群体组学数据时期,已经有学者关注“缺失值”问题。2006 年,Nie 等[4]提出了一个基于数据驱动的零膨胀泊松回归模型,通过转录组数据应对蛋白组数据的“缺失值”问题。而对于单细胞转录组数据,scImpute[5]与 scTSSR[6]方法都采取了借助其他基因表达信息的方式来应对“缺失值”问题。

此外,对单细胞数据的分析与解读已经使得人们在研究包括癌症在内的内在异质性疾病上取得了许多突破。2012 年,Torres-García 等[7]在单细胞水平上对细胞耗氧动态数据进行多参数分析、建模和特征提取,证实了不同类型细胞的差异。2014 年,Bendall 等[8]开发了 Wanderlust 算法,基于 k-近邻图与最短路径算法构建细胞发展轨迹,成功构建了从造血干细胞到人类 B 淋巴细胞的轨迹,准确预测了其发育轨迹。2016 年,Grün 等[9]提出了一种从单细胞转录组数据中推导谱系树的 StemID 算法,通过使用 k-medoid 来创建细胞簇聚类中心在高维空间中进行连接,将单细胞投影到网络上的边来生成谱系轨迹网络,能够在种群内所有可检测的细胞类型中识别出干细胞。2017 年,Guo 等[10]开发了 Slice 算法,用转录组熵作为分化的指标值,通过将细胞聚类进行简化,并以簇为单位构建最小生成树来形成分化轨迹的大框架,最终形成一个囊括所有细胞的分化轨迹。2018 年,Lummertz 等[11]开发了 CellRouter 算法,通过降维将基于 k-近邻图构建的网络图转换成流动网络图,将多状态谱系构建转化为最优化求解问题,探索复杂的细胞状态转变轨迹。2020 年,Wei 等[12]开发了一种基于扩散传播的轨迹推断方法 Dtflow,通过构建细胞的 k-近邻图,利用高斯核函数将其转换为马尔科夫转移矩阵,进而转换为巴氏核矩阵,基于该矩阵进行降维和伪时间排序,由于减少了对降维过程的依赖性,大大增强了分化轨迹的稳定性。

然而,这些算法需要使用除基因表达以外其他的信息,例如细胞的 k-近邻图、细胞聚类和多状态谱系等,带来了额外的复杂度和不确定性。当然,也有一些学者为了避免额外的复杂度,只使用基因表达数据进行研究。2020 年,Liu 等[13]基于恶性细胞较参考细胞间的差异分布提出了单细胞熵,结合高斯混合模型实现了恶性细胞与良性细胞的区分。2020 年,Zhong 等[14]提出了单细胞图熵 (single-cell graph entropy, SGE),在 5 个胚胎分化数据集中识别出了暗基因,即不同时期在基因水平表达无差异但对 SGE 值敏感的基因。在此基础上,本文构建细胞特异性网络,并使其与熵相结合,提出了基因互作网络熵 (gene interaction network entropy, GINE) 的方法来估算细胞的分化能力,以此来研究细胞的发展动态。通过 GINE,笔者完成了从借助单细胞“不稳定”的基因表达到借助相对“稳定”的 GINE 值研究细胞动态的转变,通过网络的稳定性在一定程度上规避了“缺失值”问题的影响。文章将 GINE 方法应用于头颈部鳞状细胞癌和慢性髓细胞白血病等 2 个单细胞 RNA-seq 数据集,不仅可以将恶性细胞与良性细胞分类、显著区分不同分化时期,而且通过

GINE 值有效反映了疾病疗效过程。相关结果证明了 GINE 方法的有效性以及其研究细胞动态的潜力，可以用于细胞分化、追踪癌症发展以及疾病对药物反应过程等研究。

1 材料与方法

1.1 数据

本文选取的两组真实数据集都来自美国国立生物技术信息中心基因表达数据库 (gene expression omnibus, GEO)。第一组数据为头颈部鳞状细胞癌 (head and neck squamous cell carcinoma, HNSCC) 患者的单细胞 RNA-seq 数据集 (编号为 GSE103322)，包含了来自 18 例患者的 5 902 个细胞。第二组数据是慢性髓细胞白血病 (chronic myeloid leukaemia, CML) 患者的单细胞 RNA-seq 数据集 (编号为 GSE76312)，反映了经酪氨酸激酶抑制剂 (tyrosine kinase inhibitor, TKI) 治疗 CML 过程中不同阶段的基因表达情况，该数据集包含 2 055 个癌症干细胞，囊括了诊断期以及经 TKI 治疗后不同阶段的基因表达情况。对于这两组数据，本文在使用前作了一定的质量控制 (quality control, QC)，针对数据中存在多个同名基因的情况进行整合，取其均值作为其表达水平。此外，数据中存在许多基因的计数为零的情况，我们选择保留在 10 或更多细胞中表达的基因。

1.2 基因互作网络熵

1.2.1 基因互作网络

在先前的研究中，如 SLICE 算法，通过借助蛋白互作网络 (protein-protein interaction networks, PPI) 中蛋白质 i 与蛋白质 j 间的关联强度，定义了对应编码基因之间边的权重。然而，由于不同类型细胞中的蛋白互作网络的差异或是缺失，会带来对应基因间权重的不确定性。为此，本文构建一个基因互作网络 (gene-gene interaction network, GGI)，对于给定的基因 i 与基因 j，假定两者之间的关联程度表现为它们在整体样本中的皮尔逊相关系数，并简记为 r_{ij}。相关系数 r_{ij} 的取值范围为–1 到 1，非零值表示 2 个基因间存在相互作用关系，其绝对值则定义为相互关联强度。为了降低单细胞测序数据中的噪声干扰，首先，进行相关性检验，保留具有统计学意义 (取 P 值小于 0.05) 的相关系数作为 GGI 网络中的边；其次，本文认为不是关联强度非零就能表征基因相互关联，并假设小于阈值的关联强度值是噪声带来的影响。因此，在相关系数检验后进一步进行筛选，通过对已知时序数据进行分类的准确度来进行调整，选择使其分类准确度最高的值作为阈值。对于绝对值大于阈值的，两基因间存在相关性，反映为 GGI 网络中存在关联边。反之，则不存在关联边。对于给定的 M 行 N 列基因表达矩阵，本文先进行对数化处理，然后通过 R 语言 Hmisc 包中的 rcorr 函数进行计算，最终得到的 M 行 M 列的基因关联矩阵 $(r_{ij})_{M\times M}$ 即可构建本文需要的 GGI 网络。

1.2.2 基因互作网络熵

对于给定的 M 行 N 列的基因水平表达矩阵 $E_{M\times N}$，其中 M 代表基因维数，N 代表样本维数。首先将其进行标准化处理，得到新的矩阵 $EN=\log_2(E+1)=(e_{ij})_{M\times N}$。基于先前构建的 GGI 网络，笔者构建了样本特异的基因互作网络熵值的矩阵 $H=(GINE_{ij})$，算法流程可参见图 1。具体而言，首先将 e_{ij} 定义为给定样本 j 中基因 i 的表达水平，将 $N(i)$ 定义为 GGI 网络中以基因 i 为中心的邻近点集合 (含基因 i 本身)。对于给定的邻近点集合 $k\in N(i)$，本文假定基因 k 与基因 i 的关联强度近似于乘积 $r_{ik}\times e_{ij}\times e_{kj}$。进一步地，进行归一化处理以确保 $\sum_k p_{ik}=1$，得到

$$p_{ik}=\frac{r_{ik}e_{kj}}{\sum\limits_{n\in N(i)} r_{in}e_{nj}} \quad \forall k\in N(i) \tag{1}$$

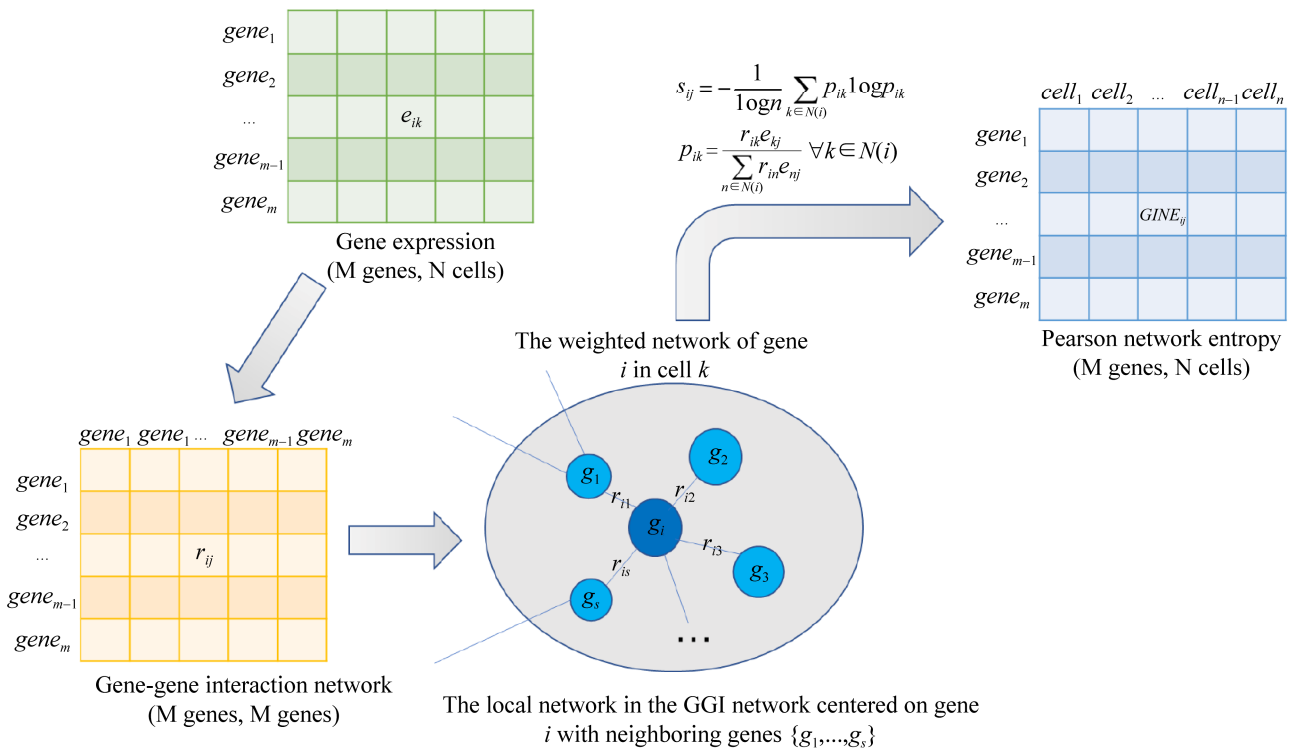

图 1　基因互作网络熵计算流程示意图

Figure 1　Schematic diagram of the GINE calculation process.

如果 $k \notin N(i)$ 或者 $k \in N(i)$ 且 $e_{kj}=0$，那么 $p_{ik}=0$。接着，可以借此定义 GGI 网络中基因 i 的样本特异的基因互作网络熵，计算公式如下：

$$s_{ij} = -\sum_{k \in N(i)} p_{ik} \log p_{ik} \tag{2}$$

其中 s_{ij} 反映了 GGI 网络中以基因 i 为中心的局部网络的不确定性水平。同时，考虑到不同局部网络之间中心结点度数的差异，定义了归一化的局部网络熵：

$$GINE_{ij} = -\frac{1}{\log n} \sum_{k \in N(i)} p_{ik} \log p_{ik} \tag{3}$$

其中 n 为集合 $N(i)$的模。在此步骤后，本文得到基因 i 在样本 j 中对应的基因互作网络熵 $GINE_{ij}$，该值不仅取决于局部网络中心基因的表达值，而且受邻近基因影响。最终，得到样本特异的基因互作网络熵矩阵 $H=(GINE_{ij})_{M\times N}$。

2　结果与分析

2.1　有效区分癌细胞与正常细胞

为了证明基因互作网络熵的适用性，本文测试了 HNSCC 数据集。该数据集包含了来自不同患者共 5 902 个细胞，其中的癌细胞通过一种基于拷贝数差异的方法进行识别。为了减少计算量，本文随机选取 HNSCC 数据集中 1 000 个样本进行区分癌与非癌细胞的测试，为了避免取样带来的偏差，本文采用简单不重复随机取样从全部数据的矩阵中选取了维度为 1 000×1 000 的基因-样本的子矩阵，进行了多次测试。下文的图 2A、C、E 反映了基于原始基因表达矩阵进行区分的结果，图 2B、D、F 则对应以基因互作网络熵矩阵进行区分癌细胞与正常细胞的结果，这里组别 1 指代 HNSCC，组别 0 指代正

常细胞。事实上，图 2A、C、E 达到了大致区分癌与非癌细胞的效果，但在降维形成的聚类交界处无法使两者彻底分隔，在图 2B、D、F 中，癌与非癌细胞之间则存在更为明显的分界。此外，本文还比较了每一个细胞的 GINE 期望、方差与其 T-分布邻域嵌入 (T-distributed stochasticneighbor embedding, Tsne) 降维分析的关联性，发现 GINE 期望值与 Tsne1 分量呈正相关，如图 3A 所示，类似地，图 3B 也证实了细胞的 GINE 方差与 Tsne2 分量呈正相关。

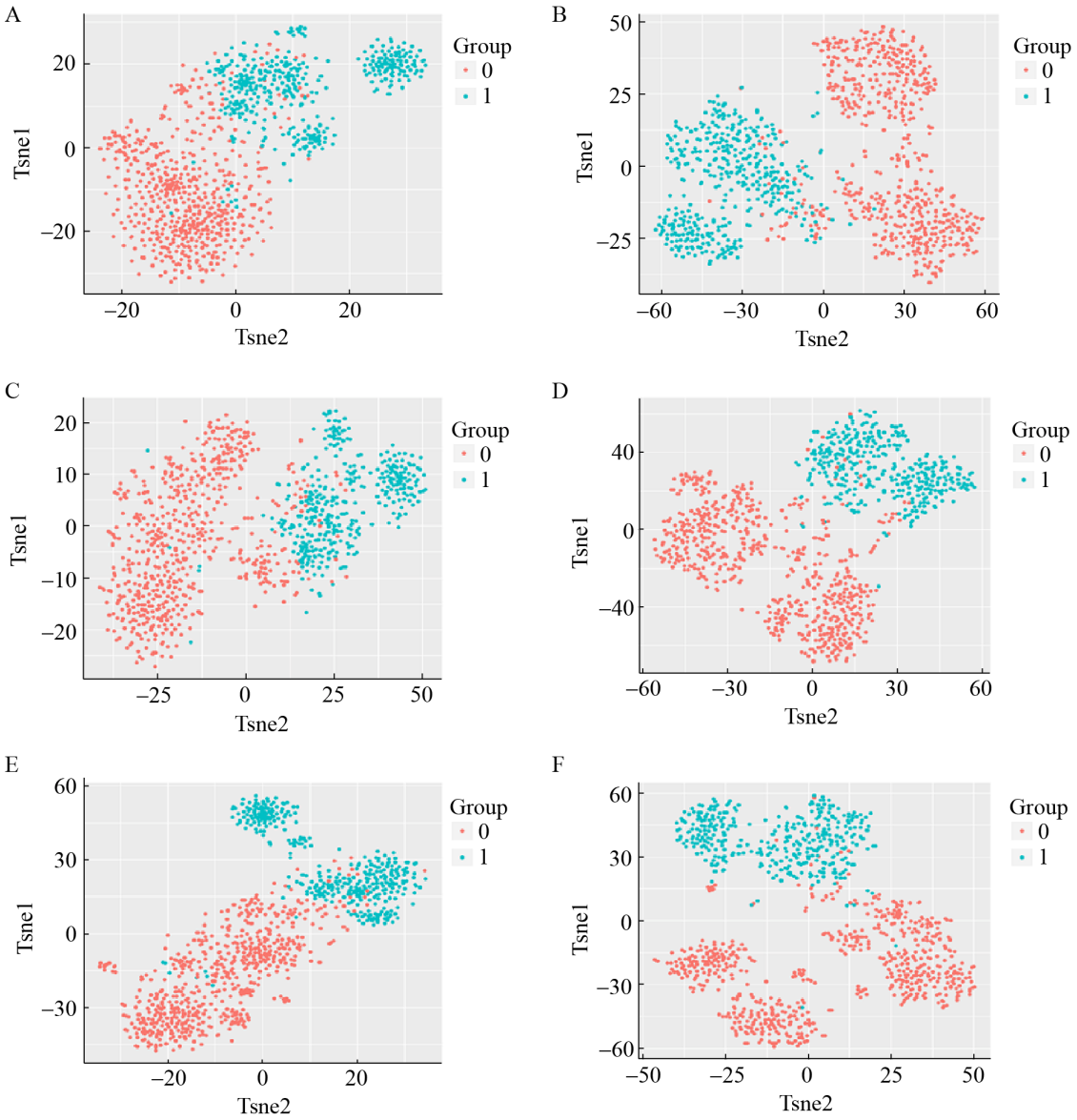

图 2　HNSCC 数据基于基因表达矩阵和基因互作网络熵矩阵的癌与非癌细胞区分结果图

Figure 2　Results of differentiation between cancer and non-cancer cells based on gene expression matrix and GINE matrix for HNSCC data. (A, C, E) Visualization results based on gene expression matrix. (B, D, F) Visualization results based on GINE matrix.

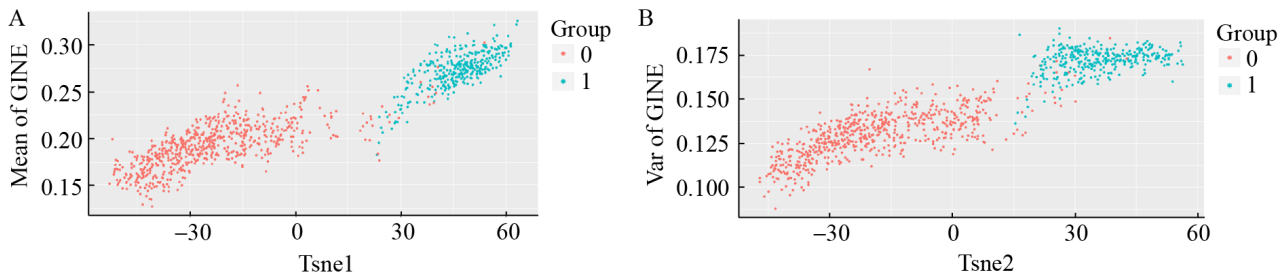

图 3 HNSCC 细胞的 GINE 统计数据与 Tsne 的散点图 A：HNSCC 细胞的 GINE 均值与 Tsne1 的散点图；B：HNSCC 细胞的 GINE 方差与 Tsne2 的散点图

Figure 3 Scatter plots of GINE statistics versus Tsne for HNSCC cells. (A) Scatter plot of GINE mean versus Tsne1 for HNSCC cells. (B) Scatter plot of GINE variance versus Tsne2 for HNSCC cells.

因此，本文认为细胞的 GINE 期望与方差可以视为 GINE 熵矩阵降维后的一维随机邻居嵌入，表明 GINE 方法是对于单细胞 RNA-seq 数据进行维度降低的有效方法。

2.2 有效聚类时间序列数据集

Tsne 是目前来说效果最好的数据降维与可视化方法，已广泛用于对单细胞数据的分析；为了阐明 GINE 值与基因表达值间的表现差异，本文将对真实数据集 GSE76312 的基因表达矩阵和基因互作网络熵矩阵进行 Tsne 降维可视化效果对比。图 4 是降维可视化后的结果，图中紫、绿、蓝、红、黄色的分组分别代表着诊断时期以及 CML 患者经 TKI 治疗 3、6、12、18 个月时的细胞，图 4A 是基于基因表达值矩阵进行降维可视化的结果，图 4B 则是基于基因互作网络熵矩阵的降维可视化的结果。显而易见的是，基于基因互作网络熵值的降维方法成功区分出了不同时期的 CML 细胞，然而基于基因表达水平的降维无法做到，不同时期的细胞完全混合在一起，无法进行有效区分。通过基因互作网络熵方法对数据集的聚类分析结果表明 GINE 具有比原始基因表达更好的表现，证实了基于 GINE 值在时序聚类方面的优越性，验证了基因互作网络熵在反映生物学过程差异方面的可靠性，从而为探索细胞群体的动态信息提供了新的方法。

2.3 有效反映疾病疗效过程

本文又研究了基于 GINE 值不同阶段的 CML 细胞异质性情况，如图 5 所示。值得注意的是，GINE 均值可以反映细胞的多能性程度，即细胞分化成所有主要细胞系的能力。细胞多能性越高，越不会表现出对特定细胞系信号路径的偏好，因此所有信号路径会具有相似的活性，对应的 GINE 均值越高。对于一个分化的细胞，或者对于一个致力于特定谱系的细胞，其不确定性是降低的，因为这需要激活反映该谱系选择的特定的信号途径，对应的 GINE 均值越低；由于终端分化细胞的多能性是低于肿瘤干细胞的，因此，GINE 均值下降，表明细胞群体越靠近正常细胞状态。按照不同时期计算了转化后细胞的 GINE 均值，如图 5A 所示，CML 患者在 TKI 治疗后的 3–12 个月，细胞的 GINE 均值整体呈下降趋势，反应了 TKI 治疗出现好转的情况。而 TKI 治疗 18 个月后，GINE 均值的提高表明整体的状态又在偏离正常细胞状态，因此代表细胞产生了耐药性。图 5A 显示，细胞在诊断时期的 GINE 均值最低，这表明细胞不进行基因筛选就计算 GINE 均值

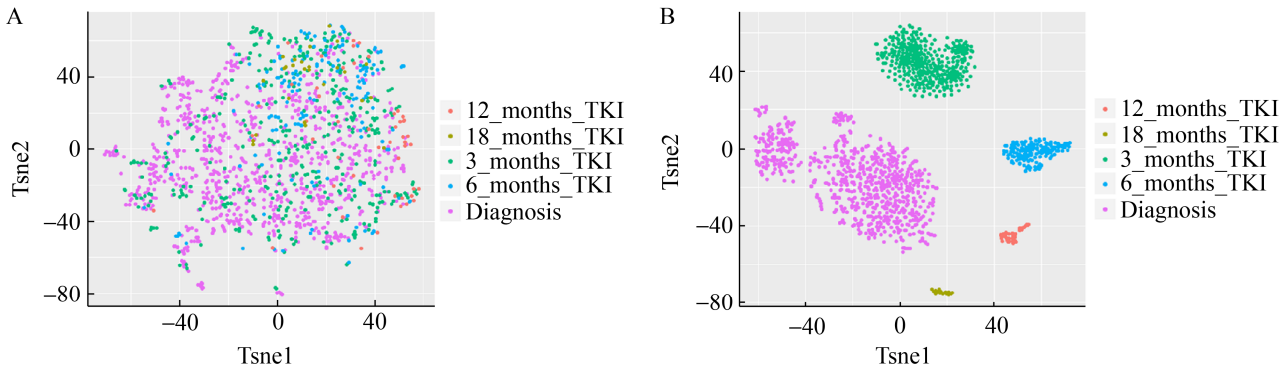

图 4　CML 数据的 Tsne 降维效果图　A：基于基因表达矩阵的降维可视化图；B：基于基因互作网络熵值矩阵的降维可视化图

Figure 4　Plot of the effect of Tsne dimensionality reduction for CML data. (A) Dimensionality reduction visualization plot based on the gene expression matrix. (B) Dimensionality reduction visualization plot based on the GINE entropy matrix.

是不合适的。事实上，由于基因的选择性表达，基因会处于开启或者关闭的状态，而细胞的状态主要是由处于开启状态的高表达基因所决定，这需要选取反映细胞真正状态的高表达基因。为确保正确反映细胞状态，更好地呈现不同时期细胞对 TKI 的反应状态，本文又对不同时期细胞分别取其前 5%和 1%的 GINE 均值并考察其变化情况，如图 5B 和 5C 所示，变化曲线与整体情况大致保持一致，且曲线变化更加明显。由于前 1%的 GINE 均值更能凸显 TKI 治疗后不同时期的状态变化，在取前 1%的 GINE 均值作为细胞状态的基础上，计算同一时期所有细胞的 GINE 均值 (mean of stage, MOS)，以此作为该状态的指标值。图 5D 是以 MOS 曲线来衡量 CML 患者经 TKI 治疗后不同时期的状态变化，横轴上的数字 1–5 分别代表诊断时期，以及 CML 患者经 TKI 治疗 3、6、12、18 个月等 5 个时期。如图 5D 所示，整体曲线呈下降趋势，尤其是在第 2 个和第 3 个时期，MOS 值大幅度下降，而在第 5 个时期 MOS 值开始有回升迹象。因此，有理由相信 CML 患者经 TKI 治疗后 3–6 个月是药物反应最有效的时期，而到 18 个月将可能产生耐药性。这个发现为研究疾病对药物的反应，进而研究细胞的动态提供了方法上的支持。

2.4　基因功能分析

在上一小节中，本文识别了 CML 患者经 TKI 治疗 3–6 个月为药物反应最有效的时期，故将该时期前 5%差异表达基因作为 GINE 关键基因集，当然这些基因是本文选取 *P* 值为 0.05 得到的。为了进一步研究验证这些基因的生物功能，使用在线生物信息学数据库 KOBAS，对这前 5%的差异表达基因进行京都基因与基因组百科全书 (kyoto encyclopedia of genes and genomes, KEGG) 通路分析，表 1 为通路分析的结果中 *P* 值小于 0.01 的部分，图 6 则是 GINE 基因集富集分析。根据 KEGG 通路分析结果，所选基因大多与代谢通路 (hsa01100)、PI3K-Akt 信号通路 (hsa04151)、病毒蛋白与细胞因子和细胞因子受体的相互作用通路 (hsa04061)、MAPK 信号通路 (hsa04010)、癌症途径通路 (hsa05200) 等紧密相关。值得注意的是，PI3K-Akt 信号通路的 *P* 值小于 0.01，MAPK 通路的 *P* 值也不超过 0.02。事实上，PI3K-Akt

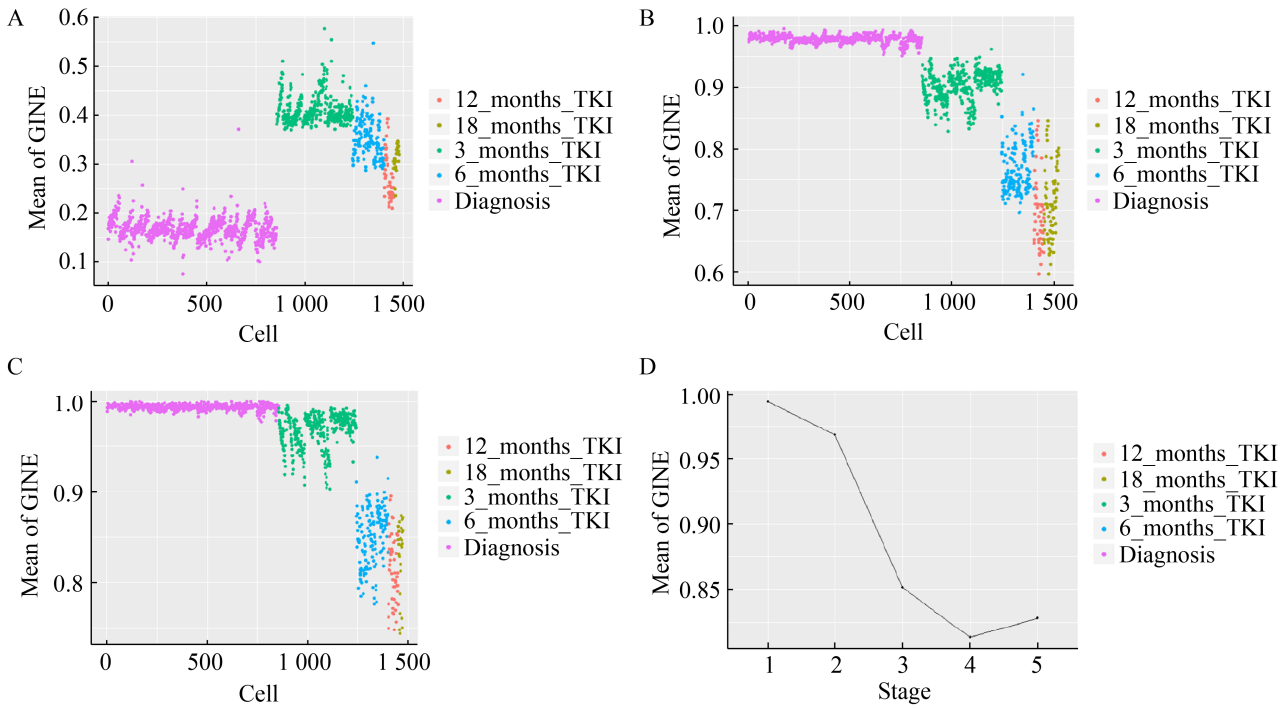

图 5　基于基因互作网络值不同阶段的 CML 细胞异质性情况

Figure 5　Heterogeneity of CML cells at different stages based on GINE values. (A) Change in mean GINE value of CML cells. (B) GINE mean change of the top 5% genes in CML cells. (C) GINE mean change of the top 1% genes in CML cells. (D) Change in mean GINE values across time. Numbers 1–5 on the horizontal axis represent the 5 periods of CML, representing the period of diagnosis, and the periods of 3, 6, 12 and 18 months after treatment with TKI, respectively.

表 1　GINE 基因集的部分 KEGG 通路分析

Table 1　Partial KEGG pathway analysis of the GINE gene set

Term	Database	ID	Input number	Background number	*P*-value	Corrected *P*-value
Olfactory transduction	KEGG PATHWAY	hsa04740	33	448	0.000 000 002 96	0.000 000 749 00
Neuroactive ligand-receptor interaction	KEGG PATHWAY	hsa04080	21	338	0.000 024 600 00	0.003 115 415 00
Protein digestion and absorption	KEGG PATHWAY	hsa04974	10	90	0.000 045 400 00	0.003 825 140 00
cAMP signaling pathway	KEGG PATHWAY	hsa04024	15	214	0.000 100 190 00	0.006 337 165 00
ECM-receptor interaction	KEGG PATHWAY	hsa04512	8	86	0.000 772 500 00	0.032 174 924 00
Cytokine-cytokine receptor interaction	KEGG PATHWAY	hsa04060	16	294	0.000 860 890 00	0.032 174 924 00
PI3K-Akt signaling pathway	KEGG PATHWAY	hsa04151	18	354	0.000 890 220 00	0.032 174 924 00
Focal adhesion	KEGG PATHWAY	hsa04510	12	199	0.001 643 860 00	0.051 987 024 00
Axon guidance	KEGG PATHWAY	hsa04360	11	181	0.002 384 920 00	0.067 042 711 00
TGF-beta signaling pathway	KEGG PATHWAY	hsa04350	7	94	0.005 172 660 00	0.127 505 302 00

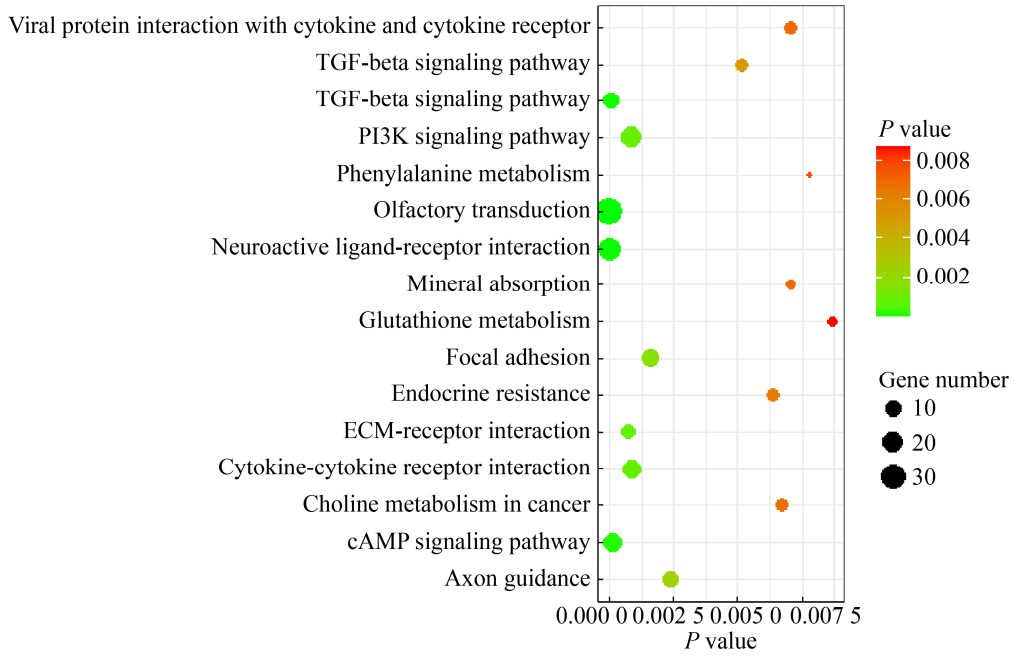

图 6　GINE 基因集富集分析

Figure 6　GINE gene set enrichment analysis.

信号通路会被多种类型的细胞刺激或毒性刺激所激活，并调节细胞的基本功能，如转录、翻译、增殖、生长和存活[15]。进一步的研究表明，PI3K-Akt 信号通路还参与了 CML 的发病机制[16]。而 MAPK 通路是细胞增殖、应激、炎症、分化、功能同步化、转化、凋亡等信号转导通路的共同交汇通路之一，把胞外信号经受体、G 蛋白/小 G、蛋白激酶、转录因子等组成的信号网络，传递到胞内，参与细胞增殖、分化、癌变、转移、凋亡等。

此外，经过文献挖掘，本文筛选出的诸多基因已经显示和 CML 或者其他肿瘤相关。例如，趋化因子受体 CCR5 在急性髓细胞白血病 (acute myeloid leukaemia, AML) 的恶性特征确定和髓外浸润的定位方面发挥作用[17]。血红素加氧酶 1 (heme oxygenase-1, HO-1) 在多种实体瘤和慢性白血病具有细胞保护作用，它的过表达可抑制伊马替尼诱导的细胞凋亡[18]。在加入 HO-1 诱导剂 (cobalt protoporphyrin, CoPP) 后，*NTRK1* 基因表达水平显著增加。基因 *PDGFB* 更是从 22 号染色体转位到 9 号染色体的证据，是判定慢性粒细胞白血病的重要依据[19]。

3　讨论

在研究细胞分化与疾病发展方面，研究生命过程中细胞动态是生物学和临床重要的任务，了解这种细胞命运的变化有助于构建个体特异的疾病模型，有助于针对与细胞分化相关的复杂疾病设计具有高度特异性的疗法。在基于单细胞数据研究分化、致癌转化等的方法中，大多都需要大量的信息，例如，在蛋白作用强度与相应编码基因相关性的假设下引入 PPI 网络，带来了不确定性和复杂度。当前用于分析单细胞数据的大多数现有方法都是基于基因的表达，然而由于基因表达的不稳定性，对其表征生物过程带来了困难。本文通过构建基因互作网络，依据基因在细胞中的整体表达确定相互关联的基因，借助相关基因信息来应对基因不表达或者低表达的情况，即“缺失值”问题。事

实上，在不同的时间、不同的条件下，网络具有更好的稳定性，与单个基因相比，通过对基因进行集聚，网络水平的基因表达更加稳定，一定程度上规避了“缺失值”事件的影响。本文开发的 GINE 方法，通过将不稳定的基因表达矩阵转换为相对稳定的 GINE 熵矩阵，可以可靠地表征动态生物学过程的状态，探索来自单细胞数据的基因-基因关联的动态信息，从而研究细胞动态。GINE 为单细胞分析提供了新的计算见解，并有助于发现细胞命运的潜在信号。本文的分析结果表明，GINE 在表征时间序列数据的细胞聚类方面的效果比基因表达数据要好。其次，GINE 值的变化还可以识别疾病对药物反应过程中的差异，展示了时间序列数据中样本的动态变化，依据细胞 GINE 值的分布特征，可以识别疾病演变过程中的关键时期，进一步可以识别其关键分化途径，这有助于追踪细胞异质性并阐明细胞分化的分子机制。

REFERENCES

[1] Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods, 2009, 6(5): 377-382.

[2] Hashimshony T, Wagner F, Sher N, et al. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep, 2012, 2(3): 666-673.

[3] Picelli S, Björklund ÅK, Faridani OR, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods, 2013, 10(11): 1096-1098.

[4] Nie L, Wu G, Brockman FJ, et al. Integrated analysis of transcriptomic and proteomic data of *Desulfovibrio vulgaris*: zero-inflated poisson regression models to predict abundance of undetected proteins. Bioinformatics, 2006, 22(13): 1641-1647.

[5] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun, 2018, 9(1): 997.

[6] Jin K, Ou-Yang L, Zhao XM, et al. scTSSR: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation. Bioinformatics, 2020, 36(10): 3131-3138.

[7] Torres-García W, Ashili S, Kelbauskas L, et al. A statistical framework for multiparameter analysis at the single-cell level. Mol Biosyst, 2012, 8(3): 804-817.

[8] Bendall SC, Davis KL, Amir el-AD, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell, 2014, 157(3): 714-725.

[9] Grün D, Muraro MJ, Boisset JC, et al. *De novo* prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell, 2016, 19(2): 266-277.

[10] Guo M, Bao EL, Wagner M, et al. SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res, 2017, 45(7): e54.

[11] Lummertz da Rocha E, Rowe RG, Lundin V, et al. Reconstruction of complex single-cell trajectories using CellRouter. Nat Commun, 2018, 9(1): 892.

[12] Wei JY, Zhou TS, Zhang XN, et al. DTFLOW: inference and visualization of single-cell pseudotime trajectory using diffusion propagation. Genom Proteom Bioinform, 2021.

[13] Liu JX, Song Y, Lei JZ. Single-cell entropy to quantify the cellular order parameter from single-cell RNA-seq data. Biophys Rev Lett, 2020, 15(1): 35-49.

[14] Zhong JY, Han CY, Zhang XH, et al. Predicting cell fate commitment of embryonic differentiation by single-cell graph entropy. bioRxiv, 2020. DOI: 10.1101/2020.04.22.055244.

[15] West KA, Sianna Castillo S, Dennis PA. Activation of the PI3K/Akt pathway and chemotherapeutic resistance. Drug Resist Updat, 2002, 5(6): 234-248.

[16] Li Q, Wu Y, Fang S, et al. BCR/ABL oncogene-induced PI3K signaling pathway leads to chronic myeloid leukemia pathogenesis by impairing immuno-modulatory function of hemangioblasts. Cancer Gene Ther, 2015, 22(5): 227-237.

[17] Mirandola L, Chiriva-Internati M, Montagna D, et al. Notch1 regulates chemotaxis and proliferation by controlling the CC-chemokine receptors 5 and 9 in T cell acute lymphoblastic leukaemia. J Pathol, 2012, 226(5): 713-722.

[18] Tibullo D, Barbagallo I, Branca A, et al. Mechanisms of heme oxygenase 1-induced resistance to imatinib in CML cells. Blood, 2010, 116(21): 3385.

[19] Gradl G, Tesch H, Schwieder G, et al. Translocation of *c-abl* oncogene and *PDGFB* (c-sis) gene in a case of CML with 46, XY, t(22;22). Blut, 1989, 58(6): 279-285.

(本文责编 郝丽芳)