



# Report

 汇报人: Lilian



## 文献来源

Genome analysis

# iSOM-GSN: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps

Nazia Fatima and Luis Rueda  \*

School of Computer Science, University of Windsor, Windsor, ON N9B 3P4, Canada

\*To whom correspondence should be addressed.

Associate Editor: Wren Jonathan

Received on December 18, 2019; revised on April 27, 2020; editorial decision on May 5, 2020; accepted on May 7, 2020

**iSOM-GSN: 一种通过自组织映射将多组数据转换为基因相似性网络的综合方法**

文献来源: bioinformatics/ 2020.8 /温莎大学计算机科学学院



## Abstract & Introduction

由于使用高分辨率微阵列和下一代测序，诸如“癌症基因组图谱” (TCGA) 等大型项目已经产生了大量的多维数据。这导致了多样化的多维数据，其中需要设计降维和表示学习方法来集成和分析这些数据。

### 已有的一些方法:

- 算法 iCluster 和 iCluster+, 利用潜变量模型和多组学数据的主成分分析，旨在将癌症数据聚类为亚型；
- 结合基因表达和 DNA 甲基化来识别共表达基因的模块
- 应用有监督的深度机器学习来解决一个非常相关的问题，例如 deepDrive，它根据基于突变的特征和基因相似性预测候选驱动基因网络 (GSN)
- 自编码器
- .....



## 已有方法的缺点

尽管这些工作被设计为使用嵌入和传统的机器学习方法，但在多组学数据集成中使用深度神经网络仍处于起步阶段。此外，这些方法不足以概括它们的多组学数据以预测疾病状态。此外，这些工作中的大多数缺乏揭示每种类型或癌症或特定临床变量或疾病状态的基因相关性的目的。

在基因交互数据上应用图卷积神经网络 (CNN) 的主要挑战之一是**缺乏对它们所属的向量空间的理解**，以及在显着较低的维度上表示这些交互所涉及的固有困难，即欧几里得空间。在处理各种类型的异构数据时，挑战变得更加普遍。



## 本文工作:

在本文中，提出了一种基于深度学习的方法，并用于通过整合多组学数据来预测疾病状态，称之为 **iSOM-GSN** 的方法。

利用 **SOM** 的强大降维功能，通过使用基因表达数据将多组学数据转换为 **GSN**。然后将这些数据与其他基因组特征相结合，以提高预测准确性并帮助可视化。据我们所知，这是**第一个使用 SOM 将多组学数据转换为 GSN 进行表征学习**，并使用 **CNN** 对疾病状态或其他临床特征进行分类的深度学习模型。

**iSOM-GSN**将具有更高维度的“多组学”数据转换到二维网格上。之后，再应用 **CNN** 来预测各种类型的疾病状态。

这项工作的主要贡献可以总结如下：

1. 一种使用 **iSOM-GSN** 预测肿瘤侵袭性和进展的深度学习方法；
2. 通过 **SOM** 获得 **GSN** 的新策略；
3. 使用 **iSOM-GSN** 来识别相关的生物标志物
4. 解释和可视化多维、多组学数据的增强方案；
5. 图表示学习和降维的有效模型。

## Methods & Results

考虑了两个数据集：TCGA 前列腺腺癌 (PRCA)和 TCGA 乳腺癌 (BRCA),PRCA 和 BRCA 的样本总数分别为 499 和 570, 数据集包含大量基因表达特征, 约 60 000 个特征。

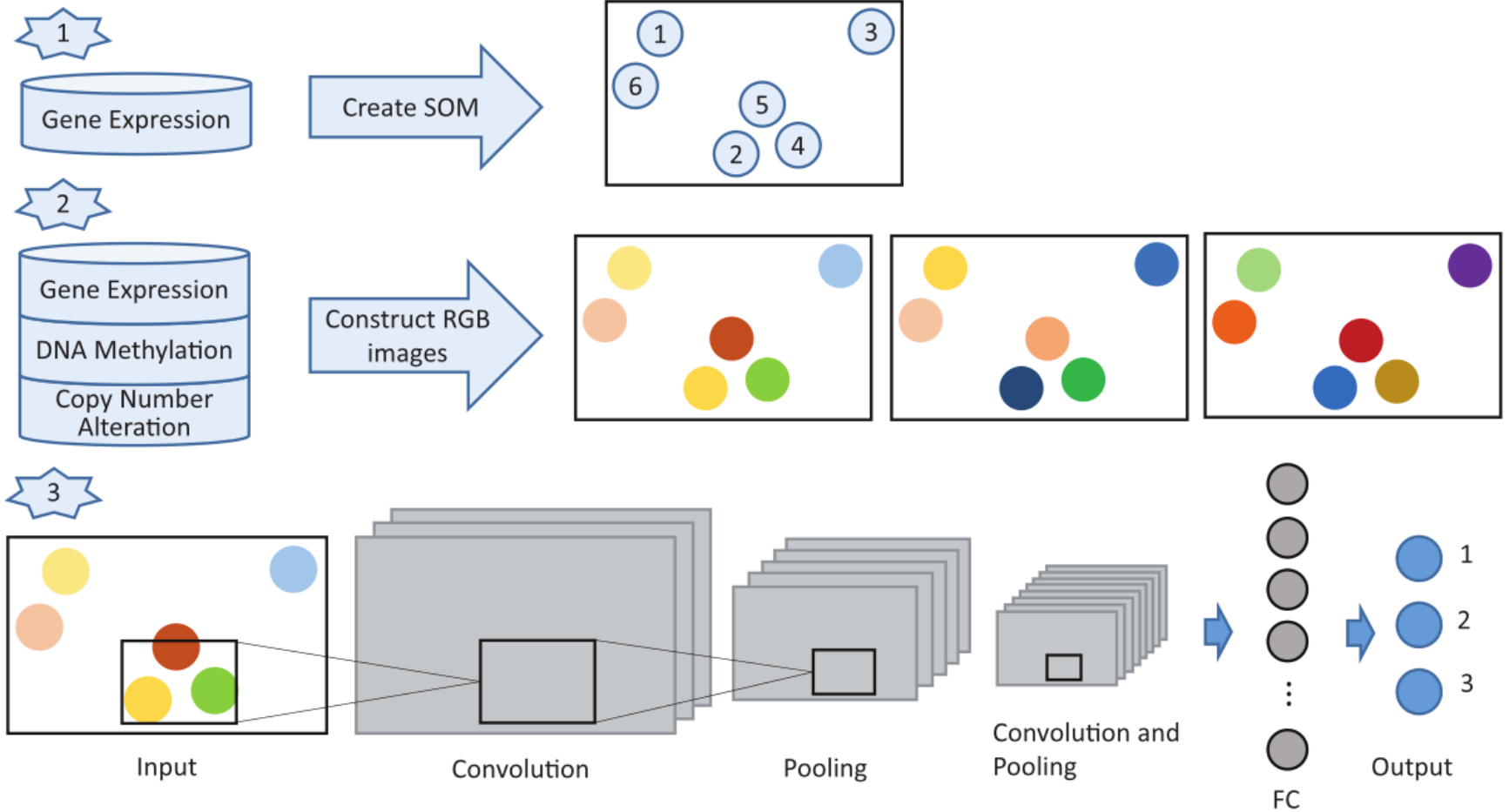
1.虽然有些特征的值全为零, 但有些特征非常稀疏。通过删除方差低于 0.2% 的那些特征来应用过滤步骤。结果, 具有至少 80% 零值的特征被删除, 特征数量减少到 16 000 个。

2. 然后将数据在所有组学的通用尺度上标准化, 包括 DNA 甲基化和拷贝数改变 (CNA) 数据。基因名称以 HUGO 格式保存, 并且删除了 HUGO 认为不相关的名称。

3. 然后根据患者 ID 组合所有三种类型的数据, **分别产生 387 名和 392 名患者的 PRCA 和 BRCA 数据, 包含所有三种所需的组学数据**



# iSOM-GSN框架





# Gene similarity network (GSN的构建)

Journals & Magazines > Proceedings of the IEEE > Volume: 78 Issue: 9

## The self-organizing map

**Publisher: IEEE**

[Cite This](#)

[PDF](#)

T. Kohonen **All Authors**

**5090**

Paper  
Citations

**72**

Patent  
Citations

**16867**

Full  
Text  
Views

芬兰赫尔辛基大学神经网络专家Kohonen, 1990





Kohonen网络是2001年芬兰科学家Kohonen提出的，是一种称为自组织特征映射的网络(Self-Organizing feature Map, SOM),属人工神经网络的范畴，是数据挖掘中的无指导学习算法。

聚类算法主要涉及如何测度数据点之间的“亲疏程度”以及以怎样的方式实施聚类。SOM解决这两个问题的基本策略如下：

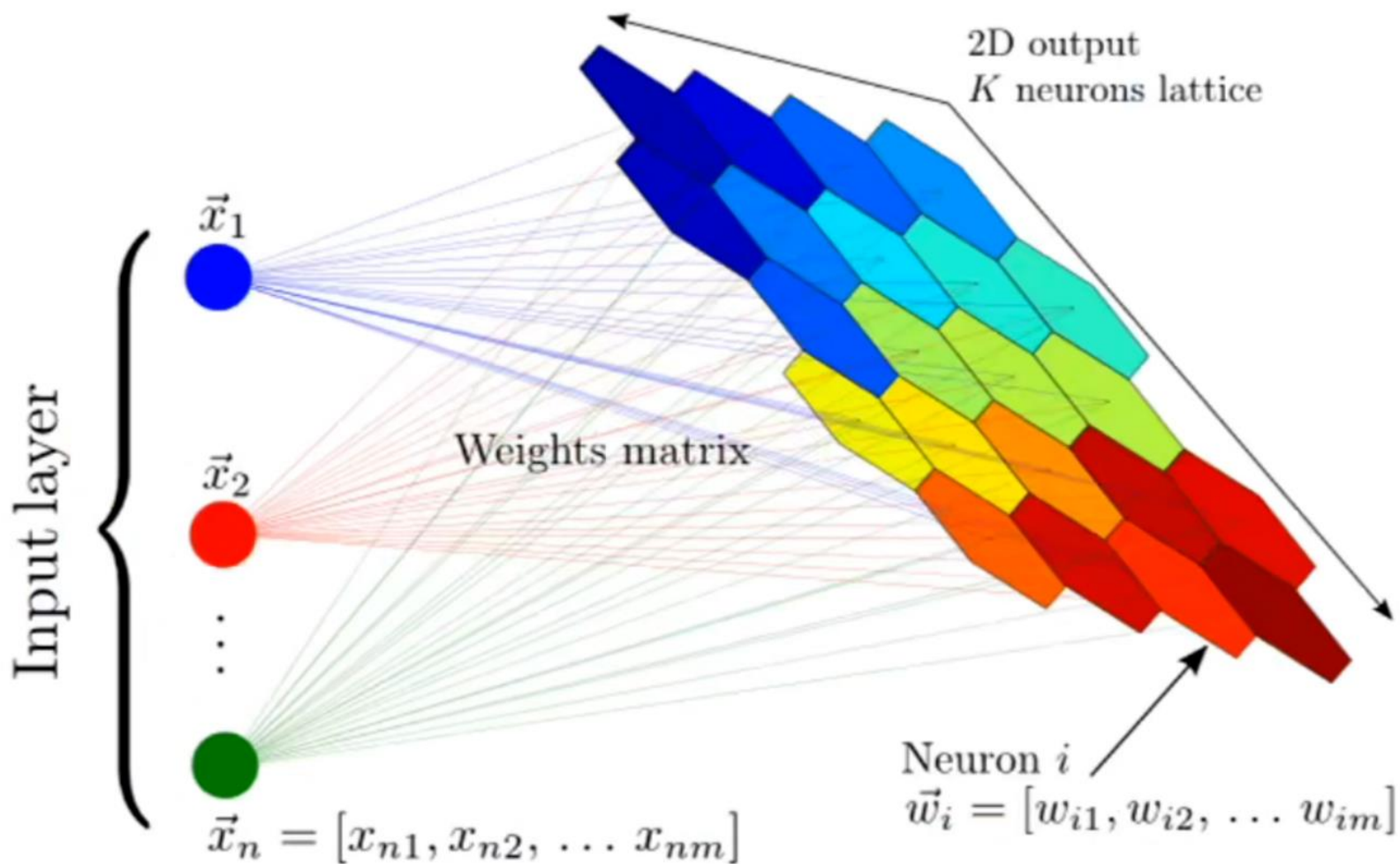
第一，采用欧氏距离作为数据点“亲疏程度”的测度,通常适合于数值型聚类变量,但也能够处理重新编码后的分类型聚类变量。

第二，模拟人脑神经细胞的机理，引入竞争机制，巧妙实现聚类过程。

SOM是一种无监督的人工神经网络。不同于一般神经网络基于损失函数的反向传递来训练，它运用竞争学习(competitive learning)策略,依靠神经元之间互相竞争逐步优化网络。且使用近邻关系函数(neighborhood function)来维持输入空间的拓扑结构。

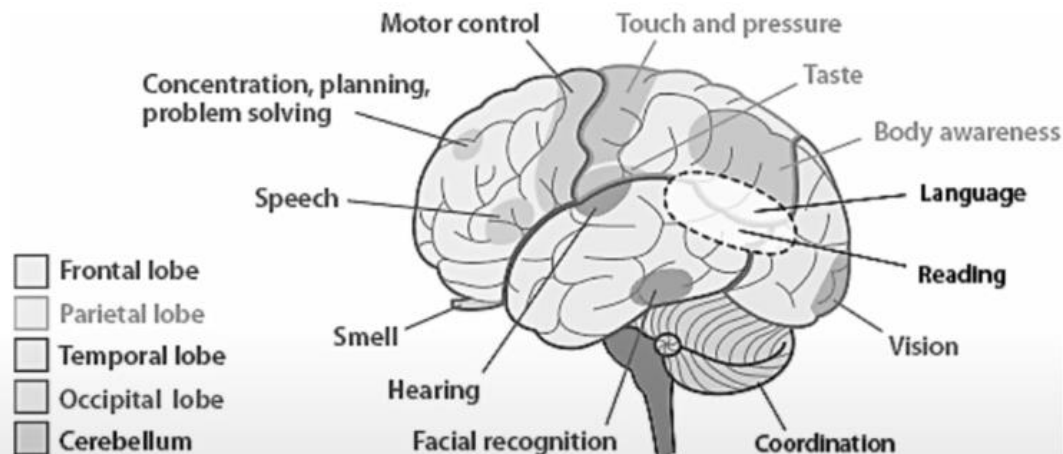


SOM网络采用两层、前馈式、全连接的拓扑结构





- Transform complex inputs into easy to understand two-dimensional outputs



<http://mlexplore.org/2017/01/13/self-organizing-maps-in-go/>

1. 神经细胞的组织是很有序的，通常呈二维空间排列
2. 空间中处于不同区域的神经细胞控制着人体不同部位的运动。
3. 空间中处于邻近区域的神经细胞之间存在侧向交互性。
4. 空间中处于不同区域的神经细胞对不同刺激信号表现出不同的敏感性

网络的输出神经元之间相互竞争以求被激活，结果在每一时刻只有一个输出神经元被激活。这个被激活的神经元称为竞争获胜神经元，而其它神经元的状态被抑制，故称为Winner-Take -All。

**竞争学习**

## SOM训练的过程:

1、权连接初始化，对所有从输入结点到输出结点的连接权值赋予随机的小数，置时间计数 $t = 0$

2、对网络输入模式 $x_k = (x_1, x_2, \dots, x_n)$

3、使用欧几里得距离选择获胜神经元：

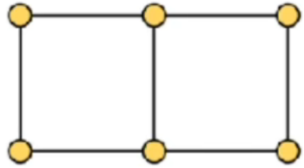
遍历竞争层中每一个节点：计算 $X_i$ 与节点之间的相似度(通常使用欧式距离)

选取距离最小的节点作为优胜节点(winner node)，有的时也叫BMU(best matching unit)

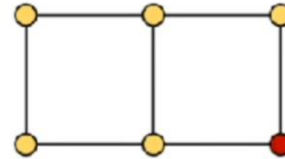
$$d_j = \| \hat{X} - \hat{W}_j \| = \sqrt{\sum_{j=1}^m [X - \hat{W}_j]^2}$$

4、根据邻域半径 $\sigma(\text{sigma})$ 确定优胜邻域将包含的节点；并通过neighborhood function计算它们各自更新的幅度；

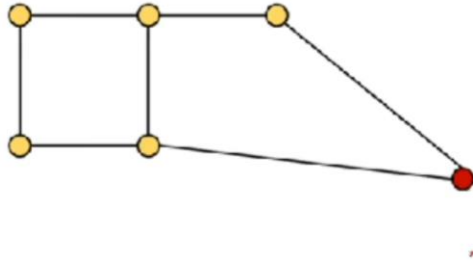
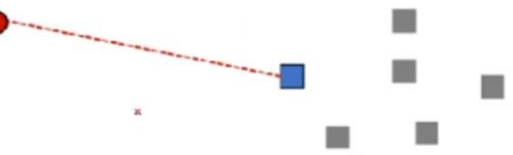
5、更新优胜邻域内节点的Weight



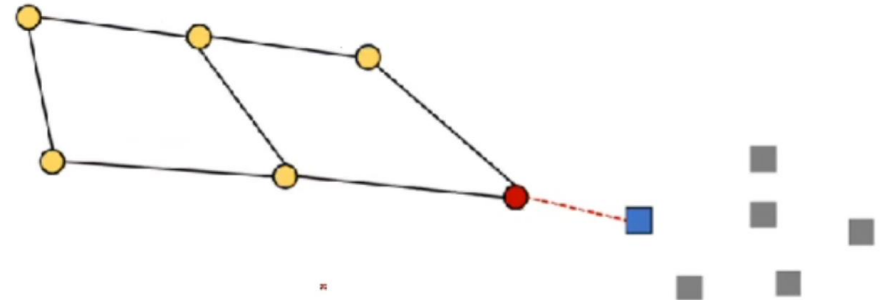
Step 1: Select one data point (blue).



Step 2: Identify Best Matching Unit (red).



Step 3: Move BMU closer to data point.



Step 4: Move BMU's neighbors closer to data point, with farther neighbors moving less.

## Gene similarity network (GSN的构建)

**1** 随机初始化：用分配给每个神经元  $c_k$  的随机权重初始化  $m$  个神经元，其中  $k = 1, 2, \dots, m$ ，其中  $m$  是考虑中的基因数量，在我们的例子中， $m = 14$ 。

**2** 计算每个基因  $g_j$  与其神经元  $c_k$  之间的欧几里得距离，并确定获胜神经元，即与其各自神经元距离最小的神经元，如下所示：

$$d_j = \left[ \sum_{i=0}^n (s_{1j}^{(i)} - c_k^{(i)})^2 \right]^{1/2}$$

**3** 假设  $c_k$  是获胜神经元，即它最接近基因  $g_j$ 。然后，更新  $c_k$  的权重。获胜的神经元也称为最佳匹配单元 (BMU)：

$$c_k(t+1) = c_k(t) + \theta_j(t)L(t)(s_{1j}(t) - c_k(t)),$$

$$L(t) = L_0 \exp\left(\frac{-t}{\lambda}\right) \quad t = 1, 2, \dots, e,$$



4 使用定义如下的邻域函数更新靠近BMU、 $\mathbf{c}_k$ 的神经元的权重:

$$\Theta(t) = \exp\left(\frac{d_j^2}{2\sigma^2(t)}\right) \quad t = 1, 2, \dots, e.$$

5 重复步骤2-4进行 $e$ 次迭代或直到所需的收敛（即权重保持不变或变化小于阈值）。

最后，获得 $m$ 个神经元，它们代表2D空间中的 $m$ 个基因:

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_m, \mathbf{y}_m).$$

运行训练算法的结果是，获得一个SOM，其中基因根据其相似性进行组织，表示GSN。



### PRCA





# MutSigCV

## nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [letters](#) > article

[Published: 16 June 2013](#)

### Mutational heterogeneity in cancer and the search for new cancer-associated genes

[Michael S. Lawrence](#), [Petar Stojanov](#), ... [Gad Getz](#)  [+ Show authors](#)

[Nature](#) **499**, 214–218 (2013) | [Cite this article](#)

**142k** Accesses | **3370** Citations | **212** Altmetric | [Metrics](#)

MutSigCV 算法通过基于基因表达数据构建患者特异性突变模型来识别显著突变的基因。该方法将整个基因组或外显子组序列作为输入，并识别突变更频繁的基因。通过观察基因中的突变是否显著超过基于背景模型的预期计数来发现显著性水平（p值）。然后计算错误发现率（q值），并且基因与 $(q \leq 0.1)$ 被分离为显著突变。从 MutSigCV 获得的前 14 个突变基因被考虑用于其余的实验。

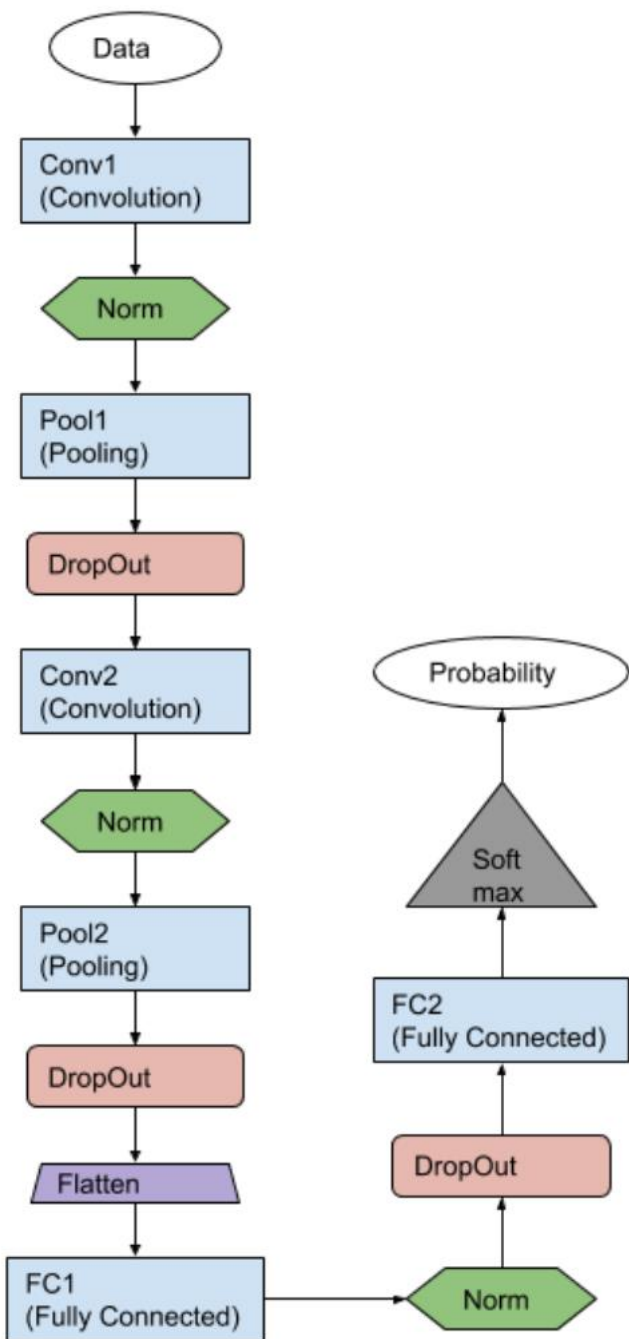


## Integrating multiple data types

iSOM-GSN 的第二步是整合多种数据类型。使用第一步生成的 GSN 作为模板图像。然后在具有预定义半径的点周围扩展一个圆形区域，并使用不同类型的组学为圆圈着色。

通过将每个组分数据视为RGB配色方案的组成部分，对每个圆圈进行着色，其中红色表示基因表达，绿色表示DNA甲基化，蓝色表示CNA。





1、应用dropout learning，即将网络神经元的输出值随机设置为零。该网络包括三个 dropout 层，dropout 比率为 0.5。

2、通过随机输入图像来使用数据增强，并在训练过程中对其进行缩放和镜像。

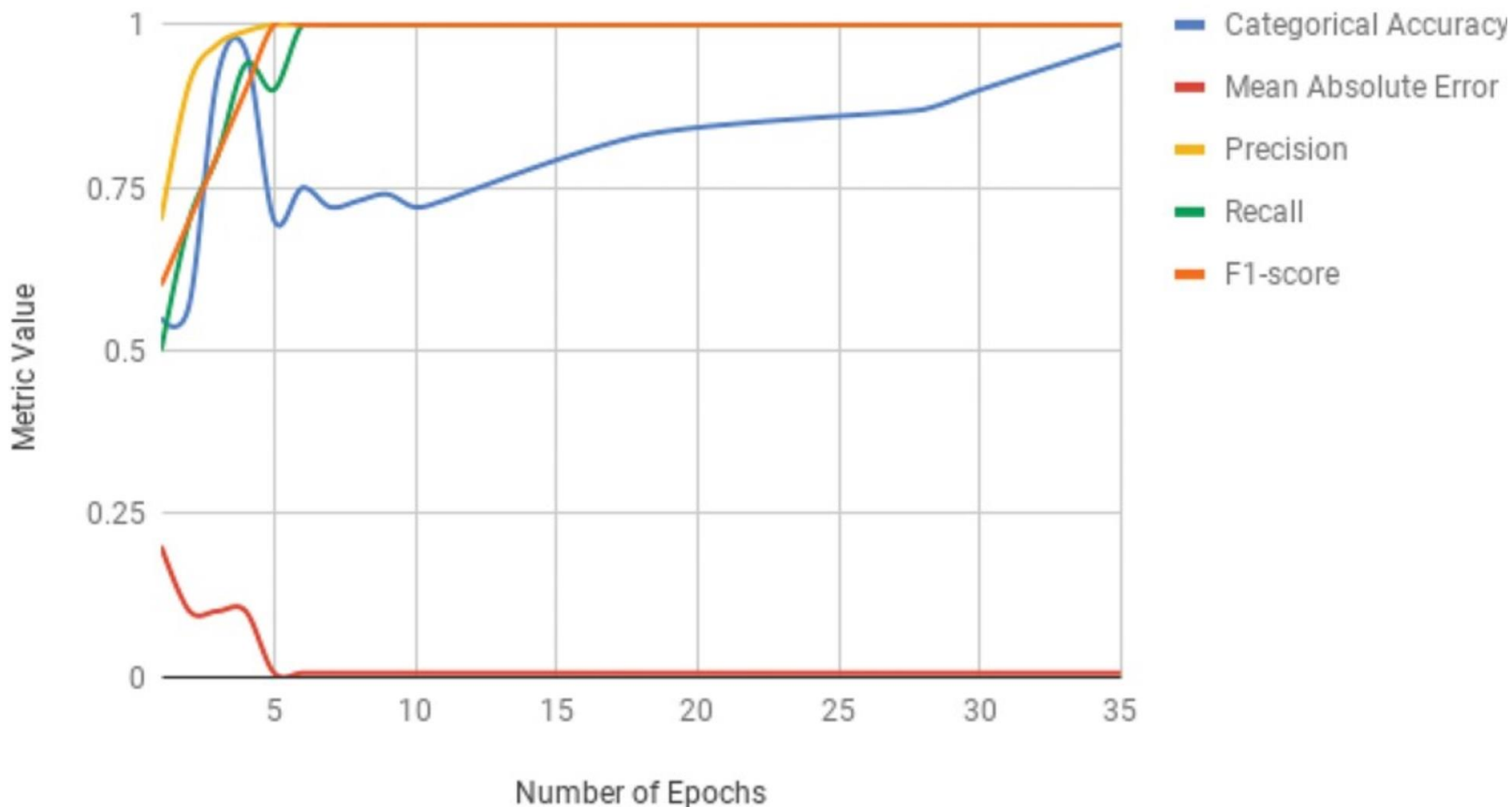


## Results

各种参数的预测性能在 94-98% 的范围内

iSOM-GSN 在两个多组学数据集（即 PRCA 和 BRCA）上运行。

### PRAD





**Table 3.** Most relevant genes found to predict Gleason groups of PRCA and stages of BRCA, along with a description of each relevant pathway

PRCA	Pathways	BRCA	Pathways
SPOP	Signaling by Hedgehog and Hedgehog pathway	RUNX1	Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds and embryonic and induced pluripotent stem cell differentiation pathways and lineage-specific markers
TP53	Apoptosis modulation and signaling and glioma	PIK3CA	Glioma and development dopamine D2 receptor trans-activation of EGFR
FOXA1	Embryonic and induced pluripotent stem cell differentiation pathways and lineage-specific markers	TP53	Apoptosis modulation and signaling and glioma
CTNNB1	Beta-adrenergic signaling and blood–brain barrier pathway: anatomy	SF3B1	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3 and mRNA Splicing—major pathway
MED12	Gene expression and RNA polymerase II transcription initiation and promoter clearance	PTEN	Glioma and metabolism of proteins
PITPNM2	Glycerophospholipid biosynthesis and metabolism	CBFB	Regulation of nuclear SMAD2/3 signaling and ATF-2 transcription factor network
PTEN	Glioma and metabolism of proteins	CDH1	Arf6 trafficking events and integrated breast cancer pathway
ATM	Apoptotic pathways in synovial fibroblasts and integrated cancer pathway	MAP2K4	Apoptosis modulation and signaling and tacrolimus/cyclosporine pathway, pharmacodynamics
NKX3-1	Endometrial cancer and pathways in cancer	MAP3K1	Apoptosis modulation and signaling and tacrolimus/cyclosporine pathway, pharmacodynamics
ZMYM3	Diseases associated with ZMYM3 include dystonia 3, torsion, x-linked and myasthenic syndrome, congenital, 6, presynaptic	NCOR1	Signaling by NOTCH1 and transcriptional activity of SMAD2/SMAD3–SMAD4 heterotrimer
SALL1	Transcriptional regulation of pluripotent stem cells and developmental biology	CDKN1B	CDK-mediated phosphorylation and removal of Cdc6 and PI3K-AKT-mTOR signaling pathway and therapeutic opportunities

*Note:* Genes associated with relevant pathways are shown here.



## 文献来源

# nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [letters](#) > article

[Published: 16 June 2013](#)

## Mutational heterogeneity in cancer and the search for new cancer-associated genes

[Michael S. Lawrence](#), [Petar Stojanov](#), ... [Gad Getz](#)  [+ Show authors](#)

[Nature](#) **499**, 214–218 (2013) | [Cite this article](#)

**142k** Accesses | **3370** Citations | **212** Altmetric | [Metrics](#)

癌症中的突变异质性和寻找新的癌症相关基因

文献来源: nature /2013.7/哈佛-MIT博德研究所



# 文献来源



## nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [articles](#) > article

Article | [Published: 30 May 2022](#)

## Brain motor and fear circuits regulate leukocytes during acute stress

[Wolfram C. Poller](#) , [Jeffrey Downey](#), [Agnes A. Mooslechner](#), [Nargis Khan](#), [Long Li](#), [Christopher T. Chan](#), [Cameron S. McAlpine](#), [Chunliang Xu](#), [Florian Kahles](#), [Shun He](#), [Henrike Janssen](#), [John E. Mindur](#), [Sumnima Singh](#), [Máté G. Kiss](#), [Laura Alonso-Herranz](#), [Yoshiko Iwamoto](#), [Rainer H. Kohler](#), [Lai Ping Wong](#), [Kashish Chetal](#), [Scott J. Russo](#), [Ruslan I. Sadreyev](#), [Ralph Weissleder](#), [Matthias Nahrendorf](#), [Paul S. Frenette](#), [Maziar Divangahi](#) & [Filip K. Swirski](#)  — [Show fewer authors](#)

[Nature](#) (2022) | [Cite this article](#)

文献来源: nature /2022.5/西奈山伊坎医学院  
脑运动和恐惧回路在急性应激期间调节白细胞



## 概述

先前已经推测中枢神经系统如何在急性压力期间控制白细胞，但很少关注将大脑网络与白细胞动力学联系起来的过程。

该研究首次展示了大脑中的特定区域如何在急性应激下并感染 COVID-19 或流感时控制身体的细胞免疫反应。更具体地说，该研究表明急性应激促使来自称为室旁下丘脑的区域的神经元触发白细胞（免疫细胞或白细胞）从淋巴结到血液和骨髓的大规模迁移。这会削弱对 COVID-19 和流感等病毒的免疫反应，使身体对抗感染的抵抗力降低，并使其面临更大的并发症和死亡风险。这一将大脑与免疫系统联系起来的基本发现，使人们更好地了解应激如何影响身体对病毒的反应，以及为什么有些人可能更容易患上严重的疾病和更糟糕的结果。

研究人员观察了一组放松和紧张的小鼠模型，并分析了它们的免疫系统。与放松的小鼠组相比，经历急性应激的小鼠在几分钟内表现出免疫系统的巨大变化。具体来说，应激会导致体内免疫细胞从一个位置迁移到另一个位置。使用光遗传学和化学遗传学等复杂工具，研究人员发现来自室旁下丘脑的神经元正在促使免疫细胞从淋巴结迁移到血液和骨髓中。





## 文献来源

# nature computational science

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature computational science](#) > [articles](#) > [article](#)

Article | [Open Access](#) | [Published: 23 May 2022](#)

## Large-scale microbiome data integration enables robust biomarker identification

[Liwen Xiao](#), [Fengyi Zhang](#) & [Fangqing Zhao](#)

[Nature Computational Science](#) **2**, 307–316 (2022) | [Cite this article](#)

**2553** Accesses | **1** Citations | **35** Altmetric | [Metrics](#)

大规模微生物组数据集成可实现可靠的生物标志物识别  
2022/5 北京生命科学研究院, 中国科学院

## 概述

肠道菌群失调与人类疾病之间的密切联系正日益得到认可。缺乏公正的数据整合方法阻碍了从不同人群中发现与疾病相关的微生物生物标志物。

这里提出了一种算法 NetMoss，用于评估微生物网络模块的变化，以识别与各种疾病相关的稳健生物标志物。与以前的方法相比，NetMoss 方法在消除批次效应方面表现出更好的性能。

通过对模拟数据集和真实数据集的综合评估，证明 NetMoss 在识别疾病相关生物标志物方面具有很大优势。

基于对泛病菌群研究的分析，在全球人群中，多疾病相关细菌的流行率很高。作者认为大规模的数据整合将有助于从更全面的角度理解微生物组的作用，准确的生物标志物识别将极大地促进基于微生物组的医学诊断。

