Sequence motifs

汇报人：Lilian

2021/11/18

Briefings in Bioinformatics

Issues Advance articles Submit ▼ Purchase Alerts About ▼

BRIEFINGS IN BIOINFORMATICS ×

期刊影响因子™

2020	五年
11.622	10.288

JCR 学科类别	类别排序	类别分区
BIOCHEMICAL RESEARCH METHODS 其中SCIE 版本	3/78	Q1
MATHEMATICAL & COMPUTATIONAL BIOLOGY 其中SCIE 版本	2/58	Q1

Article Contents

Assessing deep learning methods in *cis*-regulatory motif finding based on genomic sequencing data

Shuangquan Zhang, Anjun Ma, Jing Zhao, Dong Xu, Qin Ma, Yan Wang ✉ Author Notes

Briefings in Bioinformatics, bbab374, https://doi.org/10.1093/bib/bbab374

Published: 05 October 2021 **Article history** ▼

Abstract

Introduction

- ✓ 转录因子 (TF) 通过在特定环境中调节基因活性而与疾病进展密切相关；
- ✓ TF 通过与特定的 DNA 或 RNA 序列（称为 TF 结合位点 (TFBS)）结合而具有独特的基因表达，对齐(aligned)的TFBSs被称为顺式调控基序（简称为基序）；
- ✓ 大量染色质免疫沉淀测序 (ChIP-seq) 数据并在公共领域免费提供，已被开发用于在基因组范围内发现 RNA 序列与其相应结合 TF 之间的相互作用；
- ✓ 现已有20余种深度学习模型，包括CNN,RNN,深度置信网络,图卷积神经网络等；主要分为两类：CNN和CNN,RNN/DBN混合网络，但是如何将这些现有工具适当地应用于癌症相关数据和单细胞数据仍然未知。

- ➢ 如何将这些工具适当地应用于癌症相关数据也未得到充分研究。

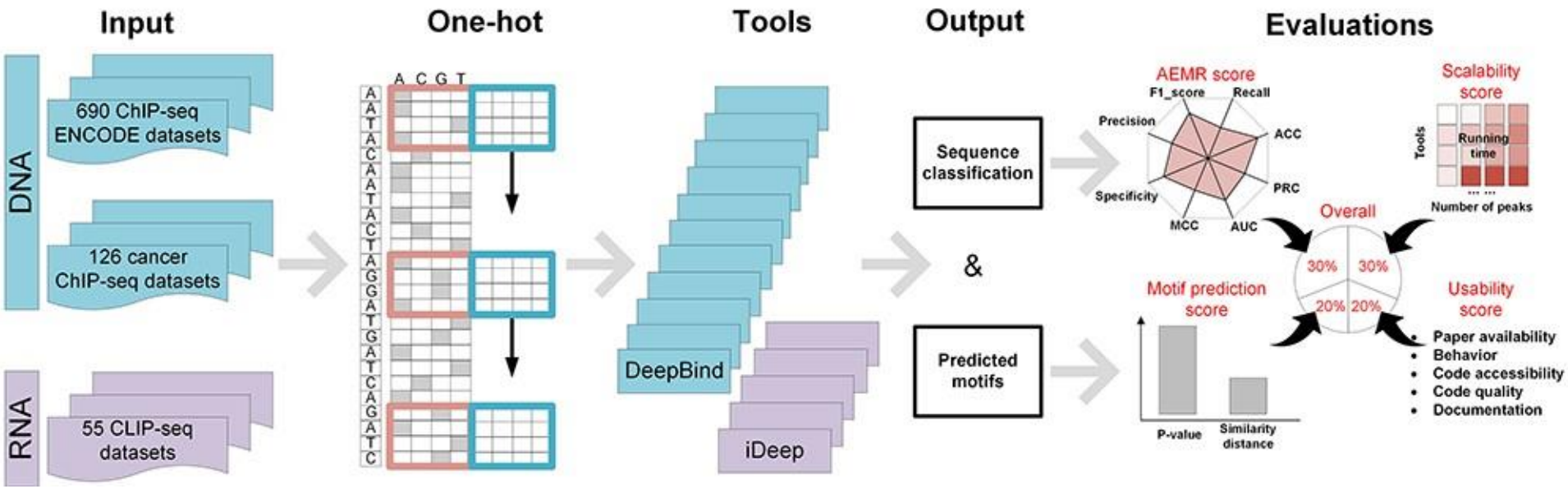
相关领域的研究人员如何为他们与基因调控相关的特定研究选择最合适的工具?

主要工作

- ✓ 使用 690 个 ENCODE DNA ChIP-seq 数据集（覆盖 161 个细胞系中的 91 个 TF）和 55 个 RNA CLIP-seq 数据集来评估 DNA/RNA 序列分类、motif预测、方法可扩展性和工具可用性的性能；

- ✓ 我们将这些工具应用于 126 个与癌症相关的 ChIP-seq 数据集，以评估这些 DL 方法在阐明九种癌症类型之间共享和特定基序的能力。

MATERIALS AND METHODS



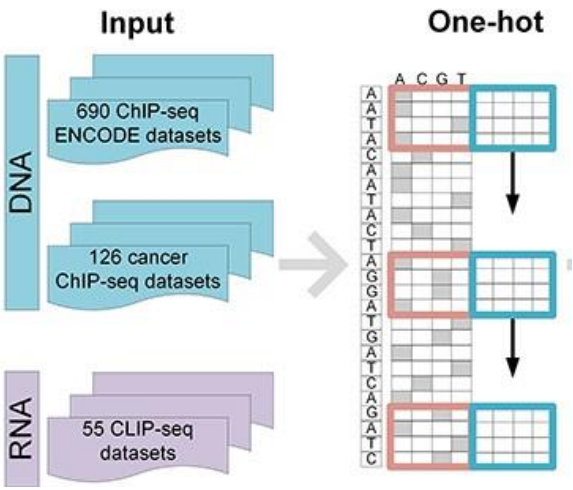
MATERIALS AND METHODS

Data collection

1. 对于DNA模型，从ENCODE获得690个ChIP-Seq数据集，该数据集涵盖了161个TF，涵盖了91个人类细胞类型，选择1,001bps长度的序列作为这些模型的输入。对于RNA模型，从文献中收集了55个CLIP-Seq数据集，并使用501 bps的固定长度作为输入。
2. 使用Cistrome数据浏览器来查询人类癌症ChIP-Seq数据集，并且从九种癌症类型中选择了126个ChIP-Seq。
3. 对于单细胞数据集，从NCBI中检索了172个单细胞CUT＆RUN数据集。除CLIP-Seq之外，其余数据集仅包含阳性样本(peak sequences), 按照DESSO中使用的策略产生阴性样品（例如，随机选择的序列）。正样本标记为1，负样本标记为0。

在原始研究中已经对ChIP-Seq和CLIP-Seq数据集进行了预处理，可直接获得固定长度的序列。深度学习模型需要二进制向量作为输入，因此每个输入序列都要转换成编码矩阵M=L×4，即A = [0，0，0，1]，G = [0，1，0，0]，C = [0，0，1，0]，T = [0，0，0，1]，其中L是输入序列的长度。对于126个癌症ChIP-Seq数据和172个单细胞CUT&RUN数据，每个数据集的峰长(the length of peaks)都不一样，因此使用下式对原始峰进行固定长度的修剪：

$$position = \left[\left\lceil \frac{\text{origin}\_start + \text{origin}\_end}{2} \right\rceil - 50, \left\lceil \frac{\text{origin}\_start + \text{origin}\_end}{2} \right\rceil + 49\right]$$



MATERIALS AND METHODS

DL model training

模型的第一层核被识别为基序检测器，通过扫描输入矩阵$M_{L\times4}$来识别激活的序列片段。结果f_i由公式给出：

$$f_i = \text{activation}\left(\text{conv}_i\left(M_{L\times4}\right) + \text{bias}_i\right), \qquad (2)$$

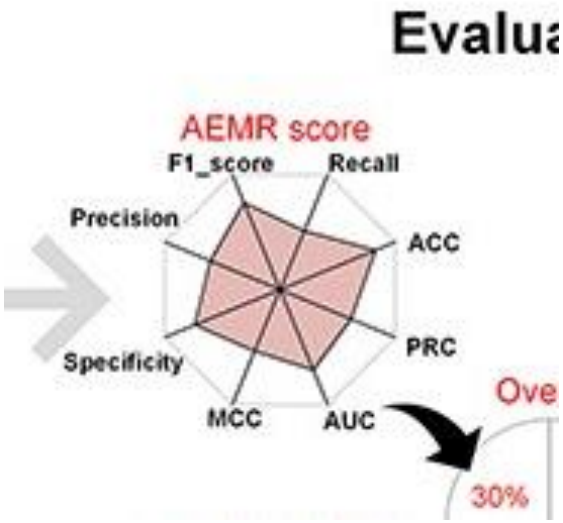
- activation表示激活函数， Convi是卷积核， biasi是阈值。
- 每个 DL 模型的卷积核数量因应用而异

MATERIALS AND METHODS

Evaluation criteria for sequence classification

AEMR 评分

AEMR 包括精度、召回率、F1_score、特异性、ACC、MCC、AUROC 和 AUPRC，它们评估了模型识别阳性和阴性样本以及对 DNA/RNA 序列进行分类的能力。



- ✓ 由八个等角辐条组成的雷达图，每个辐条代表上面定义的分数之一。
- ✓ 辐条的长度与数据点的分数相对于所有数据点的最大分数大小成正比。
- ✓ AEMR 分数是八角形雷达的总面积，按比例缩放为 0 到 1 的分数。
- ✓ AEMR 分数越高，该工具对序列分类的性能越好。

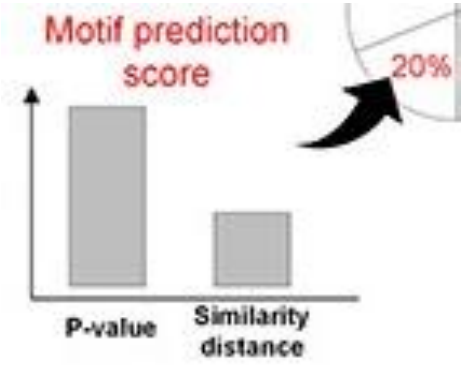
MATERIALS AND METHODS

Evaluation criteria for sequence classification

2. motif预测评分

识别motif的准确性，包含motif显著性的P值，E值和Q值。

- ✓ 使用 TFBSTools 根据每个碱基对的欧几里德距离的聚合来计算预测模体的 PWM 与记录的 TFBS 之间的相似性。
- ✓ 与预测模体的平均欧几里得距离最低的 TFBS 被视为匹配的 TFBS。



MATERIALS AND METHODS

Evaluation criteria for motif prediction

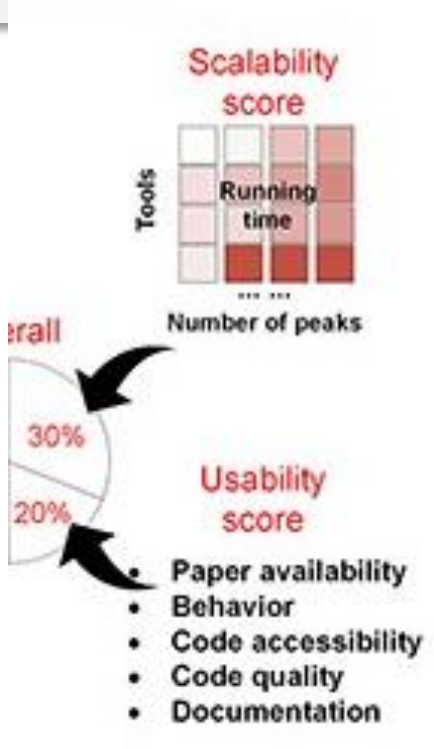
motif预测评分

对于每种 DL 方法，我们计算模体预测分数为。

$$motif\ prediction\ score = -\log(P_{value}) - \log(E_{value}) - \log(Q_{value})$$

- ✓ P 值表示与记录的基序宽度相同的随机基序具有与记录的基序一样好或更好的匹配分数的最佳对齐的概率。
- ✓ E 值表示到目前为止匹配中的预期误报数；
- ✓ Q 值是最小错误发现率.

MATERIALS AND METHODS



Evaluation of tool efficiency

在不同数据规模上执行的运行时间

从四种真实数据集开始，每种数据集包含十个具有固定数量峰值（10k、20k、30k、40k）的子数据集，在每种数据集上运行每个模型最多 12 小时。

可扩展性得分

为每种数据集中归一化训练时间的总和

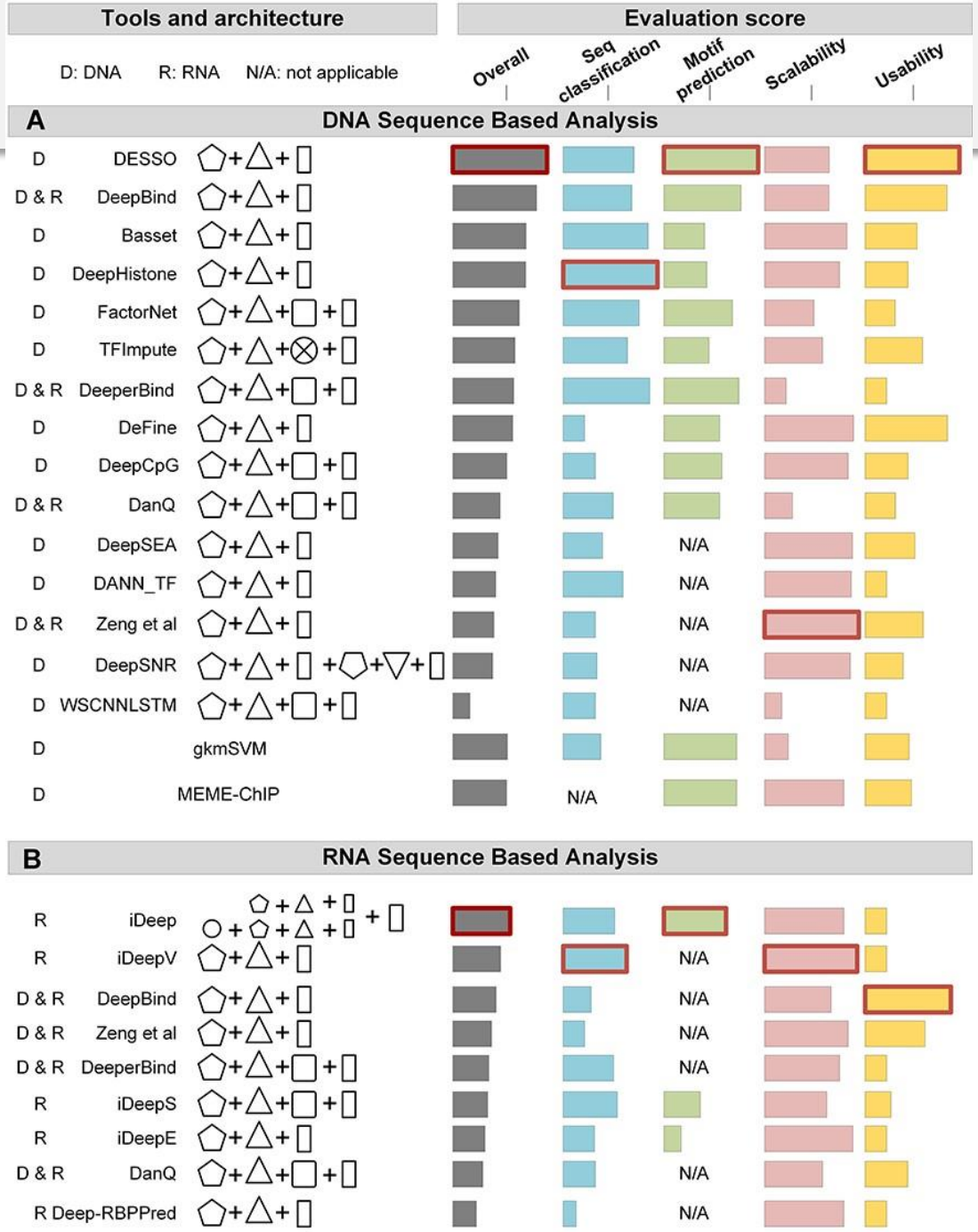
Evaluation of tool usability

涵盖可用性、行为、代码保证、代码质量和文档类别。具体来说，可用性检查是否可以轻松安装包和依赖项，以及该方法是否易于获得和使用。

RESULTS

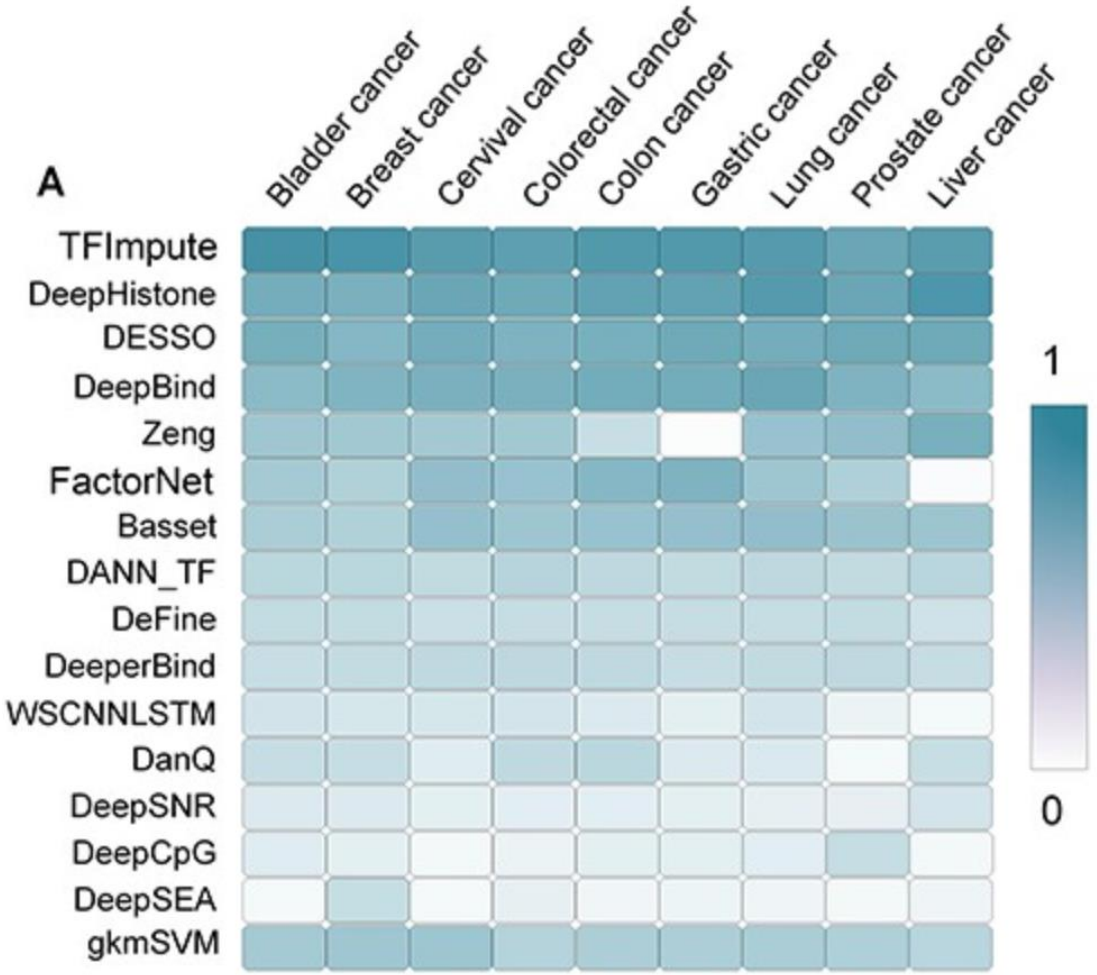
基于以上四个指标，计算了一个总分来评估 20 个工具的性能

DESSO在基于DNA序列的分析中总体得分最高，而**DeepBind**是基于RNA序列的最佳分析工具，也是第二好的DNA分析工具。



跨越九种癌症类型的 15 种深度学习方法的 AEMR 评分。

RESULTS

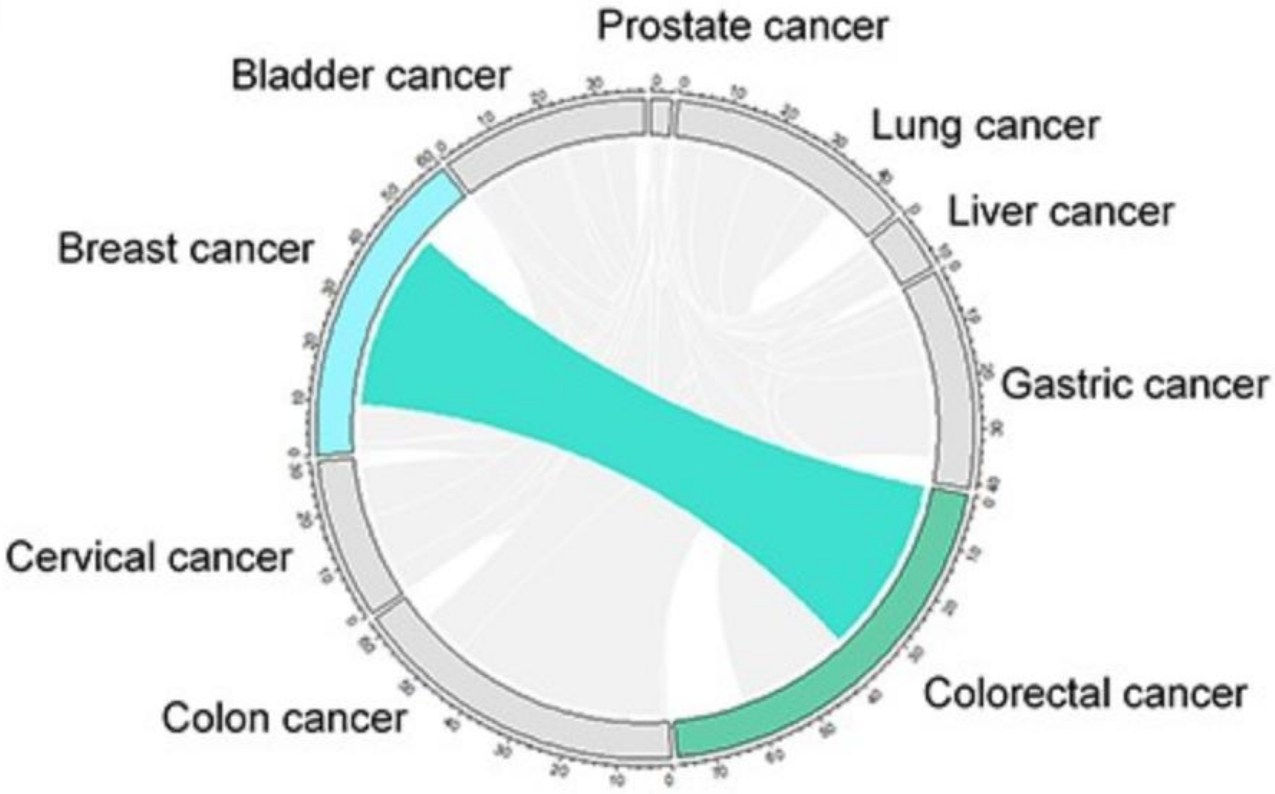


序列分类效果在各个工具中变化很大，TFimpute性能最好。DeepHiston和DeepBind分别在肝癌和肺癌数据上表现更好。

另一方面，每种工具在不同癌症类型之间的分类性能差异较小，这表明这些工具可以在不同数据集上稳定运行。

RESULTS

1、以结直肠癌为例来发现其与其他癌症类型的共享motif。在结直肠癌中鉴定的所有 77 个基序中，其中 38 个也在乳腺癌数据中被鉴定。



3、那些独特识别的motif可能是确定在特定癌症类型的发生和发展中起重要作用的基因特征的关键。例如在乳腺癌中独特鉴定的 ETV1 与正常组织相比具有更高的表达水平

2、在所有已识别的 132 个motifs中，只有 62 个是某些癌症类型所独有的，其他 70 个motifs在两种或更多癌症类型中被识别

nature computational science

Article | Published: 22 July 2021

Modeling gene regulatory networks using neural network architectures

Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng ✉ & Jianzhu Ma ✉

Nature Computational Science **1**, 491–501 (2021) | Cite this article

内容概述

本文作者提出了一个深度生成模型DeepSEM，它可以推断基因调控网络(GRNs)和单细胞RNA测序数据的生物学表示。

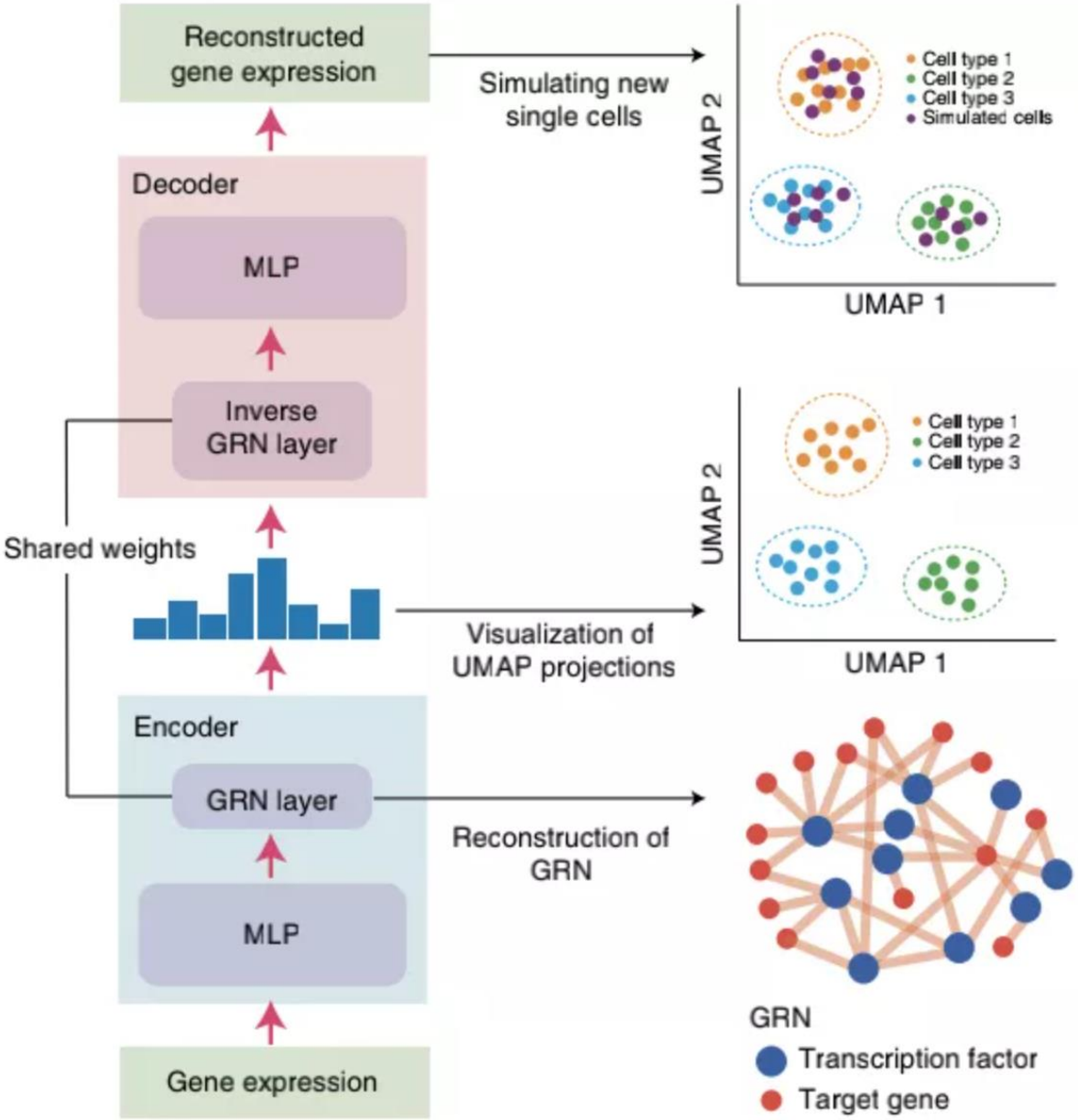
DeepSEM与最先进的方法相比，在各种单细胞计算任务上取得了优越的性能。此外，DeepSEM在小鼠皮层数据上进行验证，进一步证明了该模型的准确性和效率。因此，DeepSEM是分析细胞的scRNA-seq数据和推断GRNs的强大工具。

1、DeepSEM有两个神经网络层，命名为GRN层和逆GRN层，以明确地对GRN结构进行建模。

2、DeepSEM的VAE包含四个模块：编码器、GRN层、逆GRN层和解码器。

3、编码器和解码器都是以一个基因为输入的MLP，编码器和解码器的权重在不同基因之间共享。

3、GRN层和逆GRN层都是基因相互作用矩阵，它们显式地对GRN网络进行建模并引导神经网络的信息流。



✓ 左：DeepSEM两个主要模块，编码器(左下)和解码器(左上)。

✓ 右：DeepSEM通过利用不同的模块执行三个主要功能：

✓ (1)GRN 预测(右下)，(2)scRNA-seq 数据嵌入和可视化(右中)，以及(3)scRNA-seq模拟(右上)。

Conclusions

- ✓ DeepSEM与最先进的方法相比，在各种单细胞计算任务上取得了相当或更好的性能。
- ✓ DeepSEM在小鼠皮层数据上进行验证，进一步证明了该模型的准确性和效率。
- ✓ DeepSEM可以提供有用且强大的工具来分析细胞的scRNA-seq数据，同时可以推断细胞的GRN。

DeepSEM模型在单细胞生物学中的潜在的应用：

（1）DeepSEM可以通过利用GRN作为“桥梁”构建公共潜在空间来整合不同的单细胞模式。

（2）使用DeepSEM框架整合其他分子相互作用网络，例如蛋白质-蛋白质相互作用网络、开放染色质数据、DNA结合motifs和遗传相互作用网络，以进一步推断GRN并获得更高的准确性。

文献来源

nature biotechnology 影响因子：34.714

Published: 27 July 2015

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

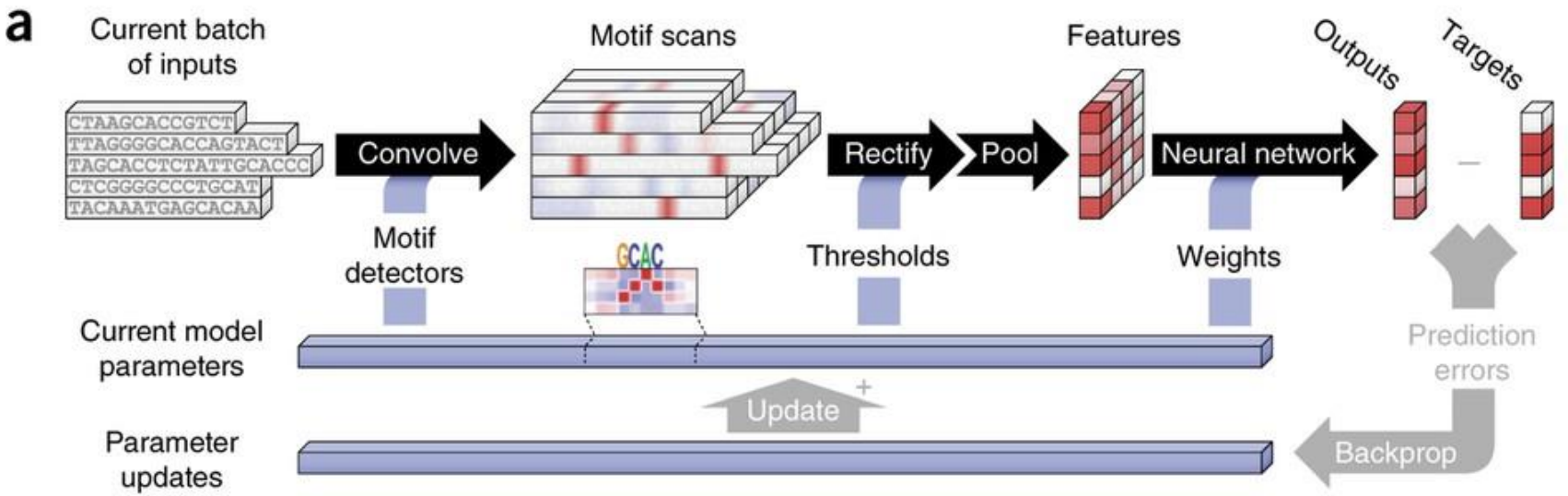
Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey

Nature Biotechnology **33**, 831–838 (2015) | Cite this article

内容概述

介绍了一项在基因组领域中使用卷积神经网络的开创性工作。文中表明可以使用深度学习从实验数据中确定序列特异性，无论是在体外数据训练还是体内测试中，深度学习都胜过其他最新方法。

作者根据此设计了一个软件——DeepBind，是全自动的，每个实验可处理数百万个序列。



- ✓ 把基因组序列窗口当作一个图，与由具有三个颜色通道(R,G,B)像素组成的图像不同，DeepBind把基因组序列看作是由(A, C, G, T)或(A, C, G, U)四个通道组成的定长序列窗口。 因此DNA蛋白结合位点预测问题就类似于图片二分类问题。
- ✓ 输入一条包含ATCG的序列，会返回一个标量值，值越高，说明序列是该种转录因子结合序列的可能性越高。
- ✓ 单个DeepBind模型并行处理五个独立序列，使用当前的模型参数来预测每个序列的单独分数。

- 尽管目前尚无用于评估序列特异性预测质量的单一度量标准，但作者认为DeepBind在各种数据集和评估度量标准上都超越了现有技术。重要的是，结果表明可以捕获核酸结合相互作用的真实特性
- DeepBind可以很好地扩展到大型数据集，对于ChIP-seq和HT-SELEX，作者发现从其他技术由于计算原因而丢弃的序列中可以学到有价值的信息。

文献来源

NATURE MACHINE INTELLIGENCE

期刊影响因子™

2020	五年
15.508	15.508

JCR 学科类别	类别排序	类别分区
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE *其中SCIE 版本*	2/139	Q1
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS *其中SCIE 版本*	1/111	Q1

来源: Journal Citation Reports™ 2020

nature machine intelligence

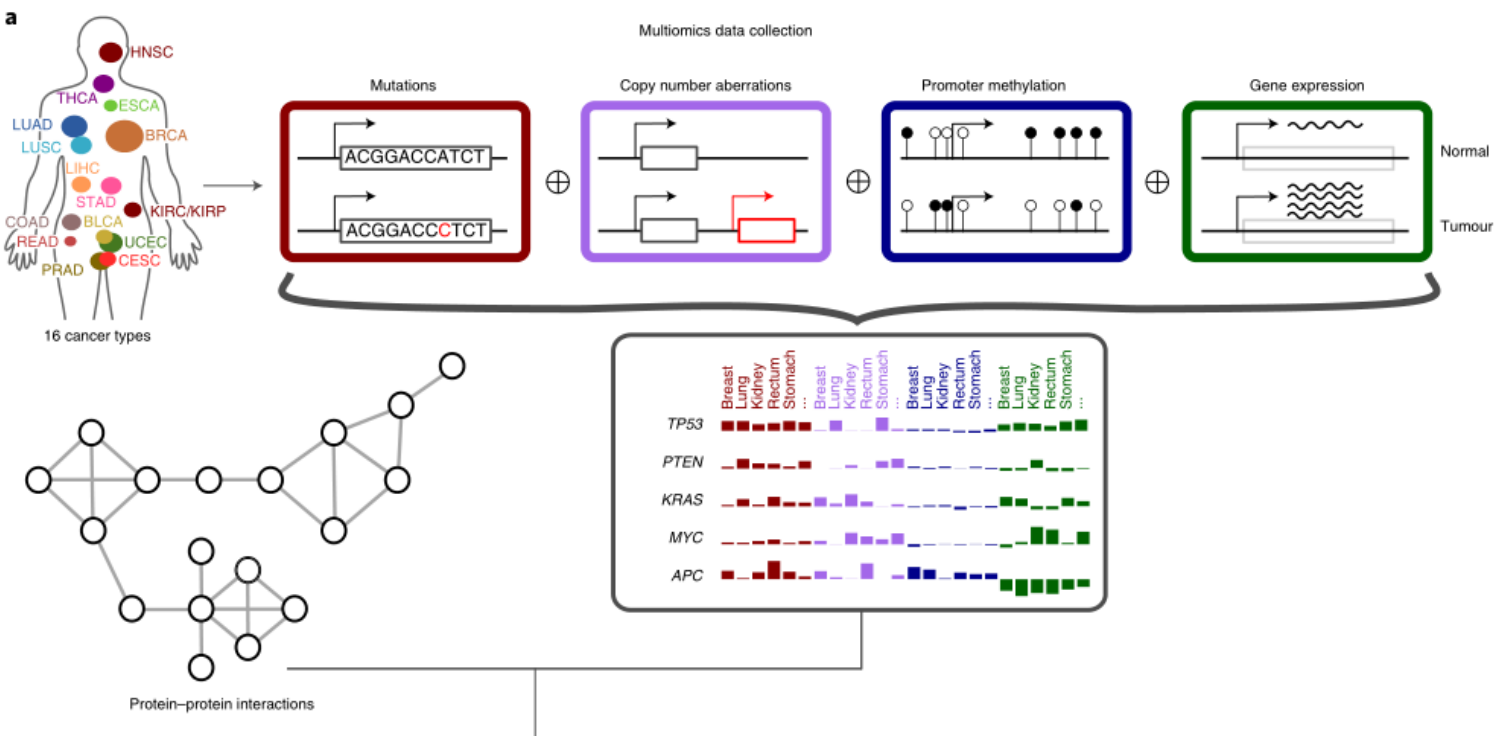
Explore content ⌄ About the journal ⌄ Publish with

nature › nature machine intelligence › articles › article

Article | Published: 12 April 2021

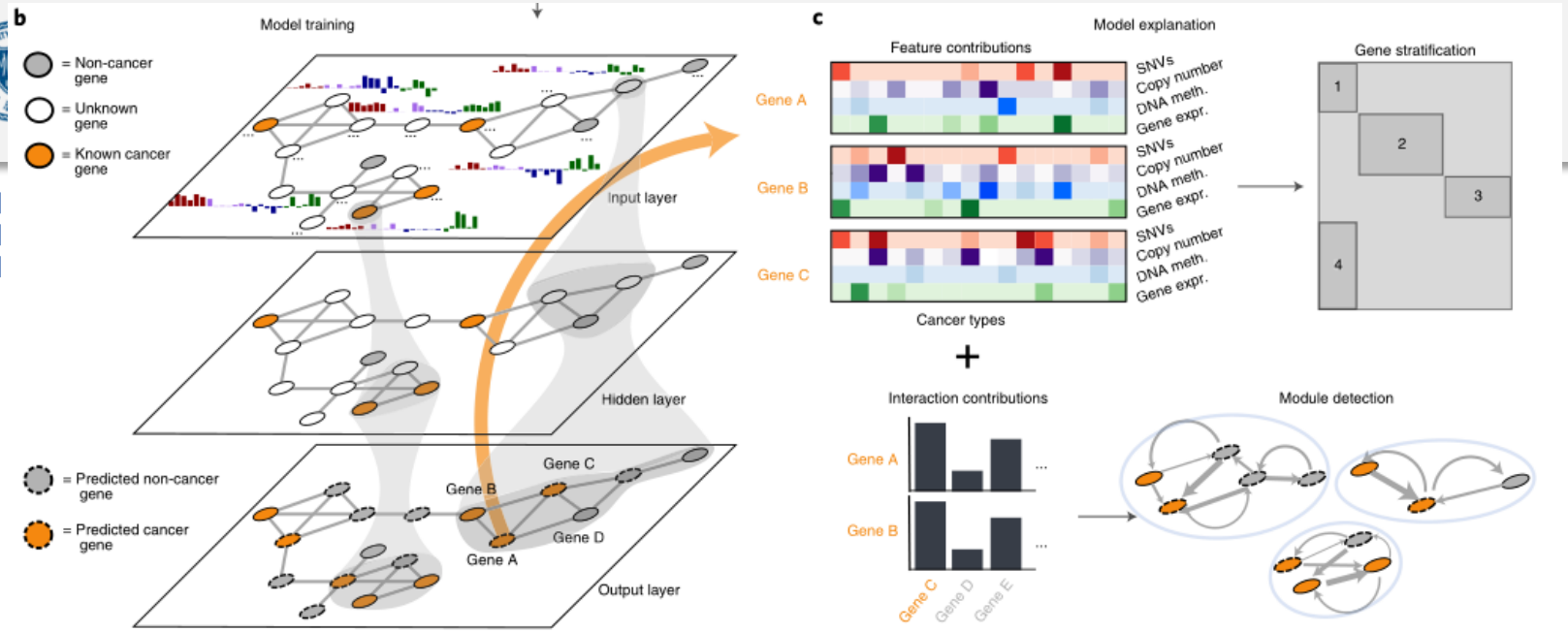
Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms

Roman Schulte-Sasse, Stefan Budach, Denes Hnisz & Annalisa Marsico ✉



- ✓ 研究团队开发了一款**基于图卷积网络（GCN）**的机器学习算法——EMOGI（Explainable Multiomics Graph Integration）。

- ✓ 该算法集成了从患者样本中生成的数以万计的数据集，这些数据集包括**突变的DNA序列数据、DNA甲基化、单个基因活性以及细胞通路中蛋白质相互作用信息。**在这些数据中，深度学习算法可检测导致癌症发展的模式和分子原理。



- ✓ 在EMOGI模型训练中，特征通过连续的图卷积层进行转换，输出层根据输出概率将基因分为预测癌症基因和非癌症基因。
- ✓ **结果：成功识别了165个先前未知的癌基因**，这些基因并不一定要发生突变才致癌，有些是通过表达失调致癌。所有这些新发现的癌基因都与已知的著名癌基因有紧密相互作用。而且细胞实验证实它们对肿瘤细胞的生存至关重要。
- ➢ 通过这一算法，找到了那些在癌症中并没有发生突变的基因，但是它们能够调控能量供应，因此与癌症发展密切相关。这些基因受到甲基化等方式的影响而表达失调，从而影响癌症发展。
- ➢ 这些基因是有潜力的癌症治疗靶标，但是由于它们隐藏很深，只有借助生物信息学和最新的人工智能算法，才能发现它们。
- ➢ 不光可用于癌症重要基因的筛选，可用于其他基因发挥重要作用的复杂疾病，例如糖尿病等代谢性疾病。

文献来源

nature biotechnology

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature biotechnology › primers › article

Published: April 2006

What are DNA sequence motifs?

Patrik D'haeseleer

nature biotechnology

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature biotechnology › primers › article

Published: 01 August 2006

How does DNA sequence motif discovery work?

Patrik D'haeseleer

内容概述

- ✓ 介绍sequence motif 的定义
- ✓ 解释共有序列（**consensus sequences**）、序列标识（**sequence logos**）
- ✓ 结合能（**Binding energy**）

- ✓ 如何从一组目标序列中以计算方式提取未知motifs（枚举、确定性优化、概率优化）
- ✓ 如何选择？几种方法的性能比对。