

inMOTIFin: 一款用于调控序列的轻量级端到端模拟软件

Katalin Ferenc¹, Lorenzo Martini^{1,2}, Ieva Rauluseviciute¹, Geir Kjetil Sandve³ and
Anthony Mathelier^{1,4,5,*}

¹Norwegian Centre for Molecular Biosciences and Medicine (NCMBM), Nordic EMBL Partnership,
University of Oslo, 0318, Oslo, Norway

²Department of Control and Computer Engineering (DAUIN), Politecnico di Torino, 10129, Turin, Italy

³Department of Informatics, University of Oslo, Oslo, Norway

⁴Department of Medical Genetics, Institute of Clinical Medicine, Oslo University Hospital and University
of Oslo, Oslo, Norway

⁵Bioinformatics in Life Science (BiLS) initiative, Department of Pharmacy, University of Oslo, Oslo,
Norway

摘要

生物信息学框架的精确开发、评估、解释和基准测试依赖于控制模拟来验证其基础方法。然而，现有的模拟器通常缺乏端到端的灵活性或易于集成，这限制了它们的实际应用。我们介绍了 inMOTIFin，这是一个轻量级、模块化且用户友好的基于 Python 的软件，通过提供多样化的高效仿真和 DNA 调控序列的修改来解决这些差距。inMOTIFin 使用户能够高效地模拟或修改调控序列，以便自定义生成基序并精确控制其位置、共现性和间距以及直接修改真实序列，从而促进对基于基序的方法和解释工具进行综合评估。我们展示了 inMOTIFin 在评估从头开始基序发现预测、分析转录因子协同作用以及支持深度学习模型可解释性分析中的应用。inMOTIFin 确保了研究转录调控语法的稳健且可重复的分析。

inMOTIFin 可在 PyPI 的 <https://pypi.org/project/inMOTIFin/> 和 Docker Hub 的 <https://hub.docker.com/r/cbgr/inmotifin> 获取。详细文档可在 <https://inmotifin.readthedocs.io/en/latest/> 获取。用例分析的代码可在 https://bitbucket.org/CBGR/inmotifin_evaluation/src/main/ 获取。

关键词：基因调控 | 模拟 | 转录因子 | 调控语法规则

1

通讯作者：anthony.mathelier@ncmbm.uio.no

介绍

在开发算法或计算模型时，需要在一个严格控制的环境中评估它们的表现 [1]。虽然真实的生物序列可以作为性能和应用的最终测试标准，但它们的数量往往有限，并且缺乏所包含的基本事实真相。因此，它们可能不符合黄金标准，并且通常不足以全面评估模型在各种情况下的假设和行为，包括极端案例。

在调控基因组学的背景下，目标是破译顺式调控代码，该代码通过转录因子（TF）在结合位点 [2] 上的特异性结合来控制转录。具体而言，转录因子识别 DNA 序列模式，并在顺式调控区域（如启动子、增强子和沉默子）中协同作用。多年来已经开发了几种计算方法来识别嵌入顺式调控区域的 DNA 序列中的 TF 结合基序，以及它们的组成如何调节转录。例如，常见任务包括发现实验鉴定序列 [3] 内的 TF 结合模式从头开始，阐明两个 TFs [4] 之间的结合协同性，或最近解释预测 DNA 序列 [5, 2] 活性或功能的深度学习模型。由于底层真实情况通常未知，因此在开发和后续基准测试 [6, 7, 8, 9, 4] 期间，常用模拟基序和 DNA 序列来评估计算工具。实际生物序列经常被修改以支持模型评估，尤其是在用于深度学习模型解释性工具 [7, 10] 的情况下。

然而，据我们所知，没有单一的软件包提供一个自包含、易于使用的端到端模拟调控序列的方法，通过植入或修改用户提供的或模拟的背景序列中的基序实例（或位点）和语法。为了解决这一空白，我们开发了 inMOTIFin，这是一个灵活的 Python 软件包和命令行工具，旨在模

拟 DNA 基序和背景序列，并根据用户定义的语法插入基序实例。特别地，与现有的模拟器（补充表 1）不同，inMOTIFin 集成了灵活且可定制的基序语法、直接修改序列的能力以及模块化架构，从而促进对调控基因组学软件进行全面基准测试和系统评估。此外，inMOTIFin 允许创建包含分布在组中的基序实例的 DNA 序列以模拟通过转录因子协同结合的顺式调控语法规则。最后，该软件包支持修改真实的生物序列，例如修改单个核苷酸、屏蔽现有的基序位点或将基序位点在每个序列的基础上进行位置调整。在 inMOTIFin 内部，这些仿真任务可以根据各种用例轻松参数化。我们通过三个用例展示了如何使用 inMOTIFin：模拟不同长度和信息含量的基序，在用户提供的或随机序列中插入二聚体，并屏蔽序列中的现有基序实例以支持深度学习模型的可解释性。

特征与实现

主要特征：模拟带有 TF 结合基序实例语法的序列

inMOTIFin 的主要特征是模拟插入有共同发生基序的 DNA 序列，这些基序实例遵循共同发生的基序语法。用户可以定义基序语法以指定共同发生基序之间的间距（参见附加功能部分）。或者，为了更灵活的语法，在将基序插入序列之前会将其组织成组（参见补充第 3 节）。

对于一个模拟任务，用户准备一个包含一组参数的配置文件，这些参数控制背景序列

和基序的模拟和采样，以及最终序列中基序实例的数量、方向和位置分布。参数的全面文档可以在 https://inmotifin.readthedocs.io/en/latest/usage/command_line_options.html 中找到。inMOTIFin 首先准备数据集合和用于采样的分布。首先，创建一组背景序列池和一组基序池。然后将基序组织成组。接下来设置采样分布。具体来说，选择一个基序的概率被定义为一组多项式分布，在组级别（包括组的出现和共现）和组内的基序级别进行指定。插入的基序实例的方向遵循用户提供的成功概率的二项分布。实例的位置可以固定在中间或从均匀分布或高斯分布中采样。每条序列中的实例数量可能是特定的用户提供的值，或者从泊松分布中抽取。请注意，inMOTIFin 支持直接用户输入每个步骤，在适当的情况下，例如允许对具有已知转录因子结合基序实例的真实生物序列进行修改。经过这个初步准备阶段后，采样开始。

在采样过程中，模拟的每个组件都从其相应的分布中进行采样。为了实现这一点，在 MOTIFin 是基于 DagSim 构建的，这是一种用于因果模型的仿真框架 [11]。完整的模型表示为有向无环图，节点定义了随机过程或用户输入（补充图 1）。模拟的组件包括选择预期共现基序的组、选择基序、选择背景序列、采样基序实例、采样每个序列中的组和基序实例的数量以及采样基序实例的位置和方向。模拟出的值在每一轮传递给其下游对应部分，从而每次生成一个输出序列。总轮次数量对应于预期序列的数量。

通过 Python 接口，用户可以访问更多

功能，最值得注意的是对模拟的更好控制。这包括在每个序列中为每个基序实例设置确切位置，以及用任何核苷酸（包括 Ns）屏蔽 DNA 序列中存在的基序位点的选项。示例请参见 https://inmotifin.readthedocs.io/en/latest/usage/python_module.html。

该包输出一个包含所有插入了基序实例的序列的 FASTA 文件。此外，还提供了互补的输出文件以促进与下游工具的互操作性：

1. 提供序列中插入的基序实例位置的 BED 文件；
2. 一个包含所有模拟序列、基序、实例、方向和位置详细信息的 CSV 格式表格；
3. 一个包含所有采样实体计数的 JSON 文件：背景、组、基序、方向、基序实例序列、每个序列中的实例数量、位置和基序长度；
4. 提供所有模拟背景序列的 FASTA 文件（如适用）；
5. 一个包含模拟基序的 MEME 文件（如适用）；
6. 四个 TSV 文件详细说明了生成的基序-基序共现概率（如适用）。

附加功能

除了端到端的仿真功能，inMOTIFin 还支持随机序列和转录因子结合基序的特定仿真。为了生成随机背景序列，inMOTIFin 需要用户

费伦奇等。

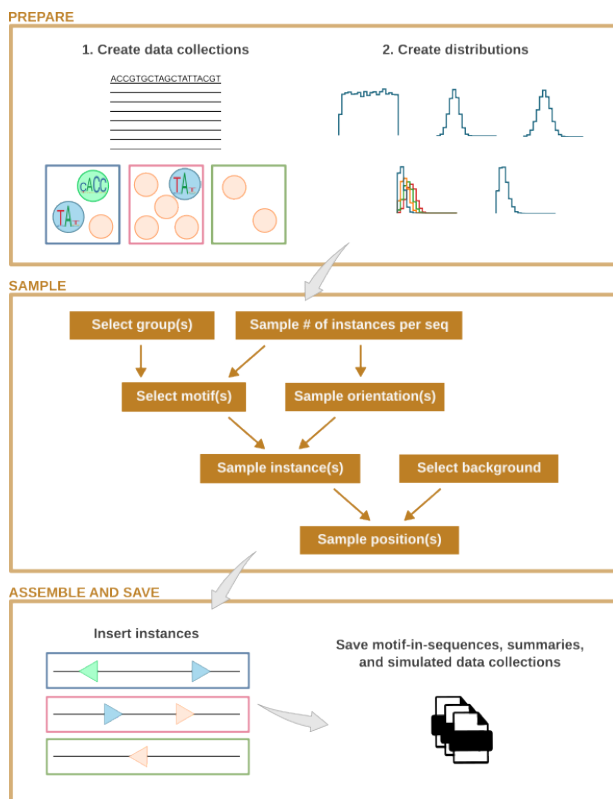


图 1: inMOTIFin 模拟框架: 在准备阶段, 创建背景和基序池, 并将基序分配到不同的组。设置各种实体的采样分布。在采样阶段, 首先对背景、基序组进行采样, 随后再对基序进行采样。如果两个基序被分配到了同一组, 则它们更有可能在同一序列中共现。当来自多个组的基序可以插入到同一个序列中时, 组与组之间的共现概率会根据第一个选定的组来条件化。从基序中抽取实例并以所选的方向和位置插入背景序列中。所有新创建的数据都会被保存到文件中, 并且还会生成额外的文件以方便后续分析。

提供序列长度、序列数量、字母表以及字母表中每个字母的频率。在生成转录因子结合基序时, 参数包括基序的最小和最大长度 (基序长度在最小和最大长度之间均匀采样)、基序的数量、字母表以及 Dirichlet 分布中每个字母的 α 值。实际上, inMOTIFin 依赖于 Dirichlet 分布来控制基序中每个位置的平均信息含量 (参见 https://inmotifin.readthedocs.io/en/latest/usage/motif_simulation.html 以获得详细解释和示例)。此外, inMOTIFin 支持模拟多聚体对应的基序 (例如, 在一对基序之间的固定间距)。这可以通过提供一组基序的文件以及一个多聚化规则文件来实现 (参见 https://inmotifin.readthedocs.io/en/latest/usage/command_line_options.html#multimerisation-of-motifs)。请注意, 间隔可以为负值, 允许压缩基序或与已知基序结合随机信号生成“噪声”基序。

实现

InMOTIFin 作为命令行工具和 Python 包两种形式实现。Python 包提供了访问构成仿真框架的各个类和函数的方法。仿真模块的时间性能为 $O(n)$ (参见补充图 7-9)。该实现是模块化的且文档齐全, 允许用户通过定义自己的类和函数来替换任何组件以利用多态性 (请参阅 https://inmotifin.readthedocs.io/en/latest/usage/python_module_usage.html 的完整文档)。

用例

为了展示 inMOTIFin 生成遵循特定共现基序语法的 DNA 序列的能力，我们产生了 50,000 个具有特定顺式调节语法属性的随机序列。我们的目标是生成包含三类中共现八个不同 TF 结合基序实例的序列：组 0 中的基序 0 – 2，组 1 中的基序 2 – 5，以及组 2 中的基序 1 和 6-7（详见补充部分 3 的详细概率）。对含有这八个基序实例的序列进行两两交集分析确认生成的序列正确地根据指定的软语法调节语法规则整合了基序实例（见补充图 2）。

为了进一步展示 inMOTIFin 的可能应用，我们展示了三个示例用例。这些例子强调了 inMOTIFin 的一些潜在用途，而不是提供一个全面的评估或针对模拟 DNA 序列使用的下游工具的基准测试。

用例 1：从头基序发现

此用例涉及将模拟的 TF 结合基序实例插入 DNA 序列中，以应用于从头开始基序发现工具。在补充部分 4.1 中，我们展示了如何评估一个的从头计算基序发现工具是否能够揭示由 inMOTIFin 生成并插入随机背景序列中的不同长度和信息内容 (IC) 的丰富基序。通过改变模拟基序的长度和 IC，可以确定给定工具的最佳检测边界。在这个实验中，我们特别使用了 RSAT 的从头开始发现算法 [12]，并使用 Tomtom[7] 将提取出的基序与插入的真实基序进行比较。如预期的那样，长且高 IC 的基序被一致识别，而较短或低 IC 的基序则更常被该工具忽略（补充图

3）。我们证明了发现的最低阈值是长度为 6（对应于 RSAT 用户定义设置）和每个位置平均 1.4 的 IC。如果 IC 较高，则可以发现较短的基序，反之，较长的低 IC 基序也可以被发现。这展示了 inMOTIFin 如何可靠地支持评估从头开始基序发现工具，并识别工具敏感性的潜在限制。

用例 2：生成反映成对合作转录因子的共现基序的序列

为了使复杂的转录调控机制得以实现，一些 TFs 以复合物的形式结合，协同控制转录 [13]。inMOTIFin 软件包支持多聚体基序的模拟以及将一组基序插入 DNA 序列中。在这个用例中，我们展示了 inMOTIFin 可以模拟一组二聚体基序，其中两个单独 TF 结合基序的组合由用户选择，对应于一种硬语法调控规则（参见补充部分 4.2）。我们模拟了一组包含不同比例定义好的完整序列集合中的各种二聚体插入的序列数据集。将我们的序列锚定在定义好的插入基序实例上，我们应用了 SpaMo[14] 来识别组成二聚体的二级基序。通过序列模拟，我们可以测试这些不同的基序组合能够被发现的程度（补充图 4）。总体而言，我们观察到对基序成对中的基序同一性没有影响，因为所有基序都被识别出。相比之下，当单个数据集中存在更多的基序组合时，插入特定定位点的恢复精度增加（补充图 4）。因此，inMOTIFin 稳健地促进了在各种调控复杂度下合作基序和基序实例检测的评估。

用例 3：深度学习模型的可解释性

深度学习模型的开发和使用在基因组研究中正在迅速增长，越来越多的框架被开发出来以执行各种分析 [15, 2]。其中，基于序列的模型尤其相关，其输入是基因组序列 [16]，模型经过训练可以推断特定功能，特别有助于解开转录语法 [2]。在这种情况下，inMOTIFin 有可能通过与基于序列的深度学习架构轻松集成来支持以基序为中心的研究，从而增强模型的可解释性 [17]。

一个典型的例子是输入扰动方法 [18]，它涉及选择性地修改输入的 DNA 序列，并测量由此产生的输出函数的变化。对于基于序列的模型，inMOTIFin 有助于自动创建背景序列（遵守特定条件，如%GC 含量）及其扰动。具体来说，inMOTIFin 允许在背景序列中控制插入 motif 实例以评估模型输出的变化。另外，屏蔽功能可以将真实序列中的已知 motif 实例替换为通用背景，甚至可以特异性地突变单个核苷酸。补充部分 4.3 展示了用于检测 GATA 和 TAL motif 的简单基于序列模型的应用。结果表明，在背景序列中插入实例如何改变了模型输出，强调了模型对不同 motif 的选择性敏感度（补充图 5-6）。因此，inMOTIFin 支持高级解释工作以阐明模型解释和敏感性。

结论

模拟是计算软件开发中重要但耗时的一部分。我们在此介绍的轻量级、健壮、易于集成和自动化的软件工具将有助于工具评估和基准测

试。InMOTIFin 是一款以模块化方式构建的开源软件，旨在支持扩展性，并重点关注未来用户的创新。除了这里概述的应用场景外，还有很多其他可能的应用，包括基于明确假设模拟新型数据来选择模型；用 CWM 或 seqlet 基础的 [19] 模拟替换 PWM 基础的模拟；或使用扩展字母表 [20] 的序列模拟。通过 inMOTIFin，我们旨在为生物背景下的模型评估铺平道路，在这种背景下，真实情况往往缺失且混淆因素可能会掩盖信号。

利益冲突

无竞争利益声明。

作者贡献

我们在这里遵循贡献者角色分类法 (CRediT)[21]。KF：概念化，方法论，软件开发，可视化，项目管理，撰写原稿；LM：方法论，软件开发，可视化，撰写原稿；IR：方法论，可视化，撰写原稿；GKS：概念化，监督，写作 - 审阅及编辑；AM：概念化，写作 - 审阅及编辑，监督，资金获取。

致谢

我们感谢软件质量研讨会的成员对该项目提供的反馈，Stefano Di Carlo 对手稿进行评审并提供了宝贵的建议，以及 Kuijjer 和 Mathelier 团队的成员进行了富有洞见的讨论。

本工作得到挪威研究理事会 [187615]、Helse SørØst 以及奥斯陆大学通过挪威分子生物科学与医学中心 (NCMBM, 前身为 NCMM) [Mathelier 小组] 和挪威癌症协会 [215027, 272930 至 Mathelier 小组] 的资助。

参考文献

- [1] Geir Kjetil Sandve and Victor Greiff. Access to ground truth at unconstrained size makes simulated data as indispensable as experimental data for bioinformatics methods development and benchmarking. *Bioinformatics*, 38(21):4994–4996, October 2022.
- [2] Julia Zeitlinger, Sushmita Roy, Ferhat Ay, Anthony Mathelier, Alejandra Medina-Rivera, Shaun Mahony, Saurabh Sinha, and Jason Ernst. Perspective on recent developments and challenges in regulatory and systems genomics. *Bioinformatics Advances*, page vbaf106, May 2025.
- [3] J. Van Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies 1 Edited by G. von Heijne. *Journal of Molecular Biology*, 281(5):827–842, September 1998.
- [4] Ieva Rauluseviciute, Timothée Launay, Guido Barzaghi, Sarvesh Nikumbh, Boris Lenhard, Arnaud Regis Krebs, Jaime A Castro-Mondragon, and Anthony Mathelier. Identification of transcription factor co-binding patterns with non-negative matrix factorization. *Nucleic Acids Research*, 52(18):e85–e85, October 2024.
- [5] Alexander Sasse, Maria Chikina, and Sara Mostafavi. Unlocking gene regulation with sequence-to-function models. *Nature Methods*, 21(8):1374–1377, August 2024.
- [6] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Régnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques Van Helden, Mathias Vandenberg, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, January 2005.
- [7] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford

- Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, February 2007.
- [8] David Simcha, Nathan D. Price, and Donald Geman. The Limits of De Novo DNA Motif Discovery. *PLoS ONE*, 7(11):e47836, November 2012.
- [9] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, October 2019. [arXiv:1704.02685](https://arxiv.org/abs/1704.02685) [cs].
- [10] Eva I. Prakash, Avanti Shrikumar, and Anshul Kundaje. Towards more realistic simulated datasets for benchmarking deep learning models in regulatory genomics. In David A. Knowles, Sara Mostafavi, and Su-In Lee, editors, *Proceedings of the 16th Machine Learning in Computational Biology meeting*, volume 165 of *Proceedings of Machine Learning Research*, pages 58–77. PMLR, 22–23 Nov 2022.
- [11] Ghadi S. Al Hajj, Johan Pensar, and Geir K. Sandve. Dagsim: Combining dag-based model structure with unconstrained data types and relations for flexible, transparent, and modularized data simulation. *PLOS ONE*, 18(4):1–9, 04 2023.
- [12] Walter Santana-Garcia, Jaime A Castro-Mondragon, Mónica Padilla-Gálvez, Nga Thi Thuy Nguyen, Ana Elizondo-Salas, Najla Ksouri, François Gerbes, Denis Thieffry, Pierre Vincens, Bruno Contreras-Moreira, Jacques van Helden, Morgane Thomas-Chollier, and Alejandra Medina-Rivera. RSAT 2022: regulatory sequence analysis tools. *Nucleic Acids Research*, 50(W1):W670–W676, July 2022.
- [13] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, 39(9):381–399, 2014.
- [14] Tom Whittington, Martin C. Frith, James Johnson, and Timothy L. Bailey. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Research*, 39(15):e98–e98, August 2011. Publisher: Oxford University Press (OUP).
- [15] Wardah S Alharbi and Mamoon Rashid. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum. Genomics*, 16(1):26, 2022.
- [16] Lucía Barbadilla-Martínez, Noud Klaassen, Bas van Steensel, and Jeroen de Ridder.

Predicting gene expression from DNA sequence using deep learning models. *Nat. Rev. Genet.*, 2025.

thorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155, April 2015.

- [17] J Bhagya, Jeena Thomas, and Ebin Deni Raj. Exploring explainability and transparency in deep neural networks: A comparative approach. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 664–669, 2023.
- [18] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.
- [19] Jacob Schreiber. tangermeme: Tutorial a4: Seqlets.
- [20] Coby Viner, Charles A. Ishak, James Johnson, Nicolas J. Walker, Hui Shi, Marcela K. Sjöberg-Herrera, Shu Yi Shen, Santana M. Lardo, David J. Adams, Anne C. Ferguson-Smith, Daniel D. De Carvalho, Sarah J. Hainer, Timothy L. Bailey, and Michael M. Hoffman. Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. *Genome Biology*, 25(1):11, January 2024.
- [21] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. Beyond au-