# 附录及批注

## 背景相关

RNA很重要，很多，microR，mRNA，ncRNA，lncRNA

### An overview is given for the biogenesis, decay and function of the most abundant RNAs:

tRNAs, ribosomal RNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), mRNAs, microRNAs (miRNAs), PIWI‑interacting RNAs (piRNAs) and long non‑coding RNAs (lncRNAs).

a | tRNAs are transcribed by RNA polymerase III (Pol III); the 5ʹ leader and 3ʹ trailer sequences are removed, introns are spliced, and the ends are joined. CCA nucleotides are added to 3ʹ ends, and nucleotide modifications — such as methylation (M), pseudouridylation (ψ) and deamination of adenosines to inosines (I) — are introduced before tRNA aminoacylation.

b | The 5S rRNA is transcribed by Pol III, whereas 28S, 18S and 5.8S rRNAs are transcribed as one transcript by Pol I. The precursor is processed by RNA exonucleases, endonucleases and the ribonucleoprotein (RNP) RNase MRP, guided by U3 small nucleolar RNP (snoRNP). Nucleotide modifications are introduced by snoRNPs. rRNAs are assembled together with ribosomal proteins into ribosomal precursor complexes in the nucleus and transported to the cytoplasm, where they mature to functional ribosomes.

c | Most snRNAs are transcribed by Pol II, capped and processed in the nucleus. When exported to the cytoplasm, they undergo methylation and assemble with LSM proteins into small nuclear ribonucleic particles (snRNPs) in a process aided by the survival motor neuron 1 (SMN1). These snRNPs are re‑imported into the Cajal body (CB) within the nucleus, where they undergo final maturation and snRNP assembly. U6 and U6atac snRNAs are transcribed by Pol III and are alternatively processed in the nucleus and the nucleolus. Mature snRNPs form the core of the spliceosome.

d | snoRNAs and small Cajal body-specific RNAs (scaRNAs) are processed from mRNA introns, capped and modified before they assemble into snoRNPs or scaRNPs in the CB. snoRNPs and scaRNPs carry out methylation and pseudouridylation of rRNAs, snoRNAs and snRNAs, or function in rRNA processing (for example, processing of U3 snoRNA).

e | mRNAs are transcribed by Pol II, capped, spliced, edited and polyadenylated in the nucleus. Correctly matured mRNAs are exported into the cytoplasm. Regulatory RNA‑binding proteins (RBPs) control correct translation, monitor stability, decay and localization, and shuttle mRNAs between actively translating ribosomes, stress granules and P bodies.

f | miRNAs are either transcribed from separate genes by Pol II as long primary miRNA (pri-miRNA) transcripts or expressed from mRNA introns (mirtrons) and processed into hairpin pre‑miRNAs in the nucleus. After transport into the cytoplasm, they are processed into 21‑nucleotide‑long double‑stranded RNAs. One strand is incorporated into Argonaute (AGO) proteins (forming miRNA‑containing RNPs (miRNPs)) and guides them to partially complementary target mRNAs to recruit deadenylases and repress translation.

g | piRNAs are ~28‑nucleotides‑long, germline‑specific small RNAs. Primary piRNAs are directly processed and assembled from long, Pol II-transcribed precursor transcripts, whereas secondary piRNAs are generated in the ‘ping pong’ cycle by the cleavage of complementary transcripts by PIWI proteins. Mature piRNAs are 2ʹ‑O‑methylated and incorporated into PIWI proteins. The piRNA–PIWI complexes (piRNPs) silence transposable elements (TEs) either by endonucleolytic cleavage in the cytoplasm or through transcriptional silencing at their genomic loci in the nucleus.

h | Most lncRNAs are transcribed and processed in a similar way to mRNAs. Nuclear lncRNAs play an active part in gene regulation by directing proteins to specific gene loci, where they recruit chromatin modification complexes and induce transcriptional silencing or activation. Other non‑coding RNAs (for example, 7SK RNA) regulate transcription elongation rates or induce the formation of paraspeckles (PS). Cytoplasmic non‑coding RNAs can modulate mRNA translation.

i | Incorrectly processed RNAs are recognized by several complexes in the nucleus and cytoplasm that initiate and execute their degradation. CPSF, cleavage and polyadenylation specificity factor; EJC, exon junction complex; hnRNP, heterogeneous nuclear RNP; NGD, no-go decay; NMD, nonsense-mediated RNA decay; NSD, non-stop decay; PABP, poly(A)-binding protein.

基于碱基的寡核苷酸存在一些入膜等问题

已经发现的一些小分子靶向结合RNA

* 力场的局限性和特异性，使得RNA的MD和docking并没有蛋白那么通用和出色。
* RNA在进行特异性结合，或者执行催化功能的时候，会发生构象变化，暴露出相关的活性位点，这种flexible的结构特性使得RNA没有蛋白质那么好研究
* RNA的flexible的特性使得在进行MD和docking的过程中，计算的开销变大，最终的到的结果中可供选择的RNA-小分子对也比较多，而常规的吉布斯自由能这种打分策略无法满足我们对真实可靠的靶向结合能力的筛选。
* RNA分子结合位点的静电势以带负电的RNA骨架和带正电的反离子为特征，这种口袋在绝大部分的RNA分子中很常见，因此，高亲电性配体可能以较低的结合选择性为代价，表现出强大的结合亲和力。
* RNA口袋通常比较浅，而且溶剂可及

## 综合总结

### Challenging:

1. RNA结构的多样性以及不确定性，已有的RNA结构解析手段并不如蛋白质结构解析那么简单清晰。（差别在哪？）为什么目前蛋白会比较好做？
2. 相关的样例比较少，数据规模小
3. 基于表型筛选的方法通量太低
4. 靶向亲和性的一个重要特征是结合亲和力以及结合特异性，如何保证预测器同时考虑到二者，个人认为docking更多进行的是一个结合亲和力的问题，我们需要一个complex的高得分以及其他分子对的低得分共同来验证这是一个特异性问题。

### RNA结构预测方法：

1. 基于物理原理；
2. 基于碎片结构拼接；
3. 基于二级结构数据补充；
4. 基于全序列的同源序列建模；
5. 基于进化近源相似；
6. 基于计算的纯denove或者半经验式的深度学习方法

### Docking methods review

1. Binding sites & pocket

* Geometry-based methods
* Energy-based methods
* Network and machine-learning approaches

1. Ligand角度

* Multi-conformer docking

生成多个ligand conformer，然后执行rigid docking

基于web-server以及database of bioactive ensembles

* Incremental construction

Anchoring rigid fragments through geometric matching and then incrementally building the ligand structure.

基于活性位点优先docking的策略，然后进行延伸以及修改

* Stochastic optimization

随机优化策略

主要包括The Monte Carlo (MC) and Genetic algorithms (GA)

问题——局部极小值；早期step会对后期step产生级联放大效应

1. Receptor角度

* Rigid docking
* Soft docking

降低空间位阻的能量惩罚性，允许空间位阻的存在

缺陷是，仅考虑到了局部小结构的变构性，对于整体的大变构没有做出策略

* Ensemble docking

生成多个可能的构像，基于这些构象进行rigid docking

* Fully flexible docking

其中比较出色的就是分子动力学模拟

耗时，计算资源消耗大

1. Score function

* Physics-based methods: physical principles of RNA–ligand binding lead to accurate scoring functions

Force-field approach，基于物理学原理构建的力场，环境，以及键能等

Empirical energy approach，能量经验公式，基于之前的研究总结的权重经验公式

* Knowledge-based scoring functions: statistical potential provides efficient scoring of binding modes

Statistical potential approach based on reference states

基于统计学的概率学公式构建的打分原则，其中参考很多重要的物理指标，如原子距离，键角，面角，键长，键能等

Iterative statistical potential approach

采用实验数据迭代优化概率学公式，解决了前一个策略受限于参考态的问题，能够修改部分由于参考态选择不正确带来的误差

缺陷依旧是数据问题，其一数据量不足，其二数据的丰度不足够用来表现所有复合体的模型

* Machine-learning based scoring method for RNA–ligand docking: an emerging scoring approach with high promise

Machine-learning model

SVM，RF，NN，DNN，CNN

feature engineering

* 溶剂介导的interaction

The sugar-phosphate backbone is negatively charged and polar, resulting in an accumulation of water molecules, cations, and water/ion-mediated RNA–ligand interactions.

1. 非docking的RNA-ligand binding方法

量子机制的方法

3D-RSIM(three-dimensional reference interaction site model) with Kovalenko-Hirata (KH) closure relation

运动学效应

### 方法学

1. icSHAPE

利用单链结合的蛋白阻滞合成，使得sequence终止于单链区段

icSHAPE值越高，代表该位点更偏向于单链

实验室已经有了一套完整的icSHAPE流程，参考网站:

http://olddriver.website/Applications/icSHAPE-pipe-tutorial.html

1. Transformer(Enformer)

能提取出更大量的信息，区别于普通卷积层的差异是，可以提取关系比较模糊，远距离的信息。

Self-attention层

1. Transfer learning (参考书目《迁移学习手册》)

a. Instances-Based Deep Transfer Learning，公共特征选取法，选取他们特征空间重叠部分的样本作为新的输入样本

b. Mapping-Based Deep Transfer Learning, 将source和target数据映射到新的特征空间中进行深度学习，loss function中加入DMM(Xs,Xt)来作为两个数据的映射后差异

c. Network-Based Deep Transfer Learning，

考虑模型的层进行迁移，对于两个相似性较高领域数据的深度学习，在模型的前部分层进行相似的特征提取，后部分层才进行区别学习，

在对target学习时，直接迁移source学习好的前部分层的weight和架构，然后将前部分层的LR降低或设为0，

d. Adversarial-Based Deep Transfer Learning，基于对抗的深度迁移学习，

特征提取器与目标判别器的对抗，最终目标是生成的特征尽可能帮助区分两个领域的特征，同时使得判别器无法对两个领域的差异进行判别。

## 单词表

### A

Antifungal: adj. 抗真菌的

Analogous: adj. 相似的 \*

Analogical: adj. 相似的，类推的

Alleviate: v. 减轻，缓和

Attribute: v. 归于，归因于，认为……具有某种特质

n. 属性，定语，特质

academia: n.学术界，学术生涯

Augment: v. 增大，增强，加强，补充

Auxiliary: adj. 辅助的，备用的

n. 助手，辅助人员，雇佣兵，志愿队，附属机构

albeit: adv. 虽然，尽管

arbitrary: adj. 任意的，随心所欲的，专横的，武断的

authentic: adj. 还原的，真迹的，传统的，正宗的，仿真的，可信的

artefactual: adj.

Atlas: n. 地图册，图谱

Asymmetrically: adv. 不对称地

Ambiguous: adj. 含糊的，不清楚的，有歧义的，模棱两可的

Applicability: n. 适用性，适应性

Avenue: n. 大街，林荫道，途径，渠道

Arithmetic: n. 算术，演算，计算，数据统计

adj. 算术的

annulus: n. 环，环形物，圆环，环面

adenocarcinoma: n. 腺癌，恶性腺瘤

allosterically: adv. 别构的

aberrant: adj. 异常的，畸变的，迷乱的

### B

Bonus: 津贴，额外的奖励

Burgeoning: adj. 迅速发展的，v. 急速生长

Burgeon: v. 萌芽，发芽

Backpropagation: n. 反向传递，反向传播，BP算法

Bewildering: adj. 使人困惑的，令人产生混乱的

Blunt: adj. 钝的，不锋利的，直言不讳的，迟钝的

n. 减弱，使变钝，使不锋利

### C

Conduit: n. 管道，管子

Configurational: adj. 结构的，构型的，外形的

Clash: n. 冲突，撞击

v. 交战，抵触，争论

corporate: adj. 公司的，法人的，共同的，全体的

census: n. 统计，人口普查，整体调查

chasten: v. 惩罚，抑制，磨炼

contaminant: n. 污染物，污秽物，杂质

crude: adj. 粗略的，大概的，粗糙的，天然的，自然的

covariance: n. 协方差，

conceptual: adj. 概念的，观念的

conjecture: n. v. 猜测，推测

conventional: adj. 依照惯例的，常规的，遵循习俗的，老一套的

circumvent: v. <正式>规避，回避；（为避免某物而）改道；克服

corroborate: v. 证实，确证 \*

coaxial: adj. 共轴的，同轴的

coronavirus: n. 冠状病毒

capitalize: v. 为……提供资金，变现，使资本化，利用，核定资本

### D

Disrupt: v.使中断，扰乱，使瓦解

de-emphasize: v. 不再重视，不再强调

discretize: v. 离散化

drastic: adj. 极端的，激烈的

drastically: adv. 大幅度地，极其，非常

discrepancy: n. 差异，不符 \*

decoy: n. 诱饵，诱骗，陷阱

v. 诱骗，被诱

diverge: v. 相异，出现分歧，岔开，分开，偏离

delineate: v. 描述(详细的)，解释，标明

decipher: v. 破译，解密，辨认(难认的东西)，理解(难懂的东西)

deplete: v. 大量减少，耗尽

denomination: n. 名称，命名

desiderata: n. 迫切需要之物，急需品

discernible: adj. 看的清的，觉察得出的

degeneracy: n. 退化，堕落，退步

dictate: v. 命令，规定，影响，支配

derivative: n. 派生物，衍生物，派生词，导数

### E

Equilibrium: n.均衡，平静，

Exhaustive: n. 详尽的，彻底的，耗尽的

Emanate: v. 产生，散发

Elastic: adj. 有弹性的，灵活的

Electrostatic: adj.静电的，静电学的

Euclidean: adj. 欧几里得的，欧式几何的

Encompass: v. 包含，包围，围绕，促成

Embark: v. 上船，上飞机，着手，开始

Elucidation: n. 说明，阐明

Enthalpic: adj. 焓的

Extrapolate: v. 外推，推演，推断

Elucidate: v. 阐释，解释，说明

### F

Fundamental: adj. 基本的，根本的，必需的

Forte: n. 长处，特长

adj. 强音的，响的

fixation: n. 加固，固定，专注，迷恋，定像

fidelity: n. 忠贞，忠诚，准确性，保真度

### G

Generalization: n. 概括，普遍化，泛化，一般化

### H

Hindrance: n. 障碍，妨碍

Hallmark: n. 特征，印记，

Harmonized: adj. 调和的，统一的，协调的

Heuristics: n. 启发法，启发式教学法

### I

Intercalation: n. 插入

Intuitively: adv. 直观地，直觉地

Intuition: n. 直觉，直觉感知

Implicit: adj. 含蓄的，隐含的，固有的

Intrinsic: adj. 内在的，固有的

Interdisciplinary: adj. 跨学科的，多学科交叉的

Interpretation: n. 解释，理解；表演，演奏；口译；说明

Industry: n. 工业界，产业，制造

Incompatibility: n. 不相容，不协调，不一致

Inevitably: adv. 不可避免的

Immortalized: adj.永生的，不灭的

Indiscriminately: n. 不加选择的，任意的

Incorporate: v. 包含，合并，混合，使具现化，体现

adj. 合并的，具体化的

inherent: adj. 内在的，固有的

intimately: adv. 熟悉地，亲切地，私下地

interplay: v. 互相作用，互相影响

impractical: adj. 不切实际的，不明智的，动手能力不强的

invaluable: adj. 极有用的，极宝贵的，无价的，不可估量的

ingredient: n. 成分，原料，要素，因素

adj. 构成组成部分的

intricacy: n. 错综复杂的事物，盘根错节，复杂精细

### J

### K

### L

Leverage: n. 影响力，杠杆 v. 充分利用，合理运用 \*\*

Lipophilicity: n. 亲油性，亲脂性

Lattice: n. 晶格，格子

Latent: adj. 潜在的，隐藏的，休眠的，潜伏的

Luciferase: n. 荧光素酶

### M

Malfunctions: 功能失调，失灵，故障

Magnitude: n. 大小，数量，规模，量级，重要性

Modulate: n. 调整，调节

Multitude: n.众多，大量，群众，民众

Mitigate: v. 减轻，减缓，缓和

Manipulate: v. 操纵，摆布，使用，篡改，移动

manipulation

motion: n. 运动，移动，手势，动作，提议，议案

mutual: adj. 相互的，彼此的，共同的，共有的，互助的

monolithic: adj. 整体的，巨石的，庞大的，完全统一的

n. 单块集成电路

mimicry: n. 模仿，模拟，拟态

metastasis: n. 新陈代谢，转移

metalloproteinase: n. 金属蛋白酶

### N

Noncanonical: adj. 非经典的

Nomenclature: n. 命名法，术语

### O

Orientation: n. 目标，定位，方向

### P

Physicochemical: 物理化学的

Physiology: n. 生理学，生理机能

Preconceived: adj. 预想的，先见的

Pathogenic: adj. 致病性的，发病的

Periphery: n. 表面，外围，边缘

Prerequisite: n. 先决条件，前提 \*

Penalty: n. 惩罚，处罚 \*

Pharmacophoric: adj. 药效的，药物性的

Pharmacophore: n. 药效团，药效基团

Pharmacokinetic: adj. 药代动力学的，PK

Pervasive: adj. 普遍的，流行的 \*

Permutation: n. 排列，置换，变换

Pivotal: adj. 关键的，中枢的

Plausible: 似乎是真的，貌似有理，似是而非的，花言巧语的

Plasticity: 塑性，可塑性，适应性，柔软性

Particle: n. 微粒，粒子，颗粒

Prudent: adj. 谨慎的，精明的，节俭的

Perturbation: n. 扰乱，不安

Proximity: n. 靠近，邻近，(时间，空间，关系的)

Pioneer: n.拓荒队，先锋少先队，工兵

adj. 先驱的，拓荒的

v. 倡导，开辟，

permeabilization: n. 透化，透化作用

productive: adj. 多产的，丰饶的，富有成效的，生产的

presumably: adj. 或许，大概，可能

putative: adj. 假定的，推定的

### Q

### R

Riboflavin: 核黄素，维生素B

Rigidify: v. 固化，使僵化

Reiterate: v. 重申，重述

Reversal: n.反转，逆转，颠倒

Rigorous: adj. 严密的，严格的，苛刻的，一丝不苟的

Repulsive: adj. 排斥的，令人厌恶的，冷淡的

Recapitulate: v. 概括，摘要，综述

Repertoire: n. 全部剧目，全部技能，全部节目

Relocalize: n. 重定位？

### S

Salient: adj. 突出的，显著的，凸出的

Sense: n. 感觉官能，感觉 v. 感觉到，检测到

Sophistication: n. 精明老练，精密

Sophisticated: adj. 见多识广的，先进的，复杂精巧的，时尚的，水平高的

Spatiotemporal: adj. 时空的

Synergistically: adv. 协同地，协同方式地 \*

Synergy: n. 协同，协同作用 \*

Scarce: adj. 缺乏的，不足的，稀有的。adv. 几乎不

Sparse: adj. 稀少的，稀疏的

Steric: adj. 空间的，立体的，位的

Stochastic: adj. 随机的，猜测的

Scalability: n. 可扩展性，可伸缩性，

Susceptible: adj. 易得病的，易受影响的，敏感的

Symmetry: n. 对称，相似，相等

Semantic: adj. 语义的，语义学的

Stringent: adj. 严格的，严厉的，紧缩的，短缺的

Subtlety: n. 细微差别，微妙之处，微妙，精细，机敏，狡猾，诡秘

Spectrum: n.范围，幅度，光谱，波谱，频谱

Semantic: adj. 语义的，语义学的

Stochasticity: n. 随机性

Substantially: adv. 大量地，大体上，基本上，客观地

Syntax: n. 句法，句法规则，语法构成

Superimpose: v. 添加，重叠，附加，安装

### T

triaged over: 筛选

therapeutics: n. 疗法，治疗学

tolerate: v. 容忍，忍受。宽容，允许

tremendous: adj. 巨大的，极大地，可怕的，精彩的

tweak: v. 扭，拧，稍微调整

therein: adv. 在其中，在那里，其中

telomeric: adj. 端粒的，末端着丝粒的

### U

Unravelled: 阐明，解决

Ubiquitously: adv. 无所不在的

Underlying: adj. 根本的，潜在的，表面下的，下层的

Unambiguously: adv. 不含糊地，明白地，单意义地

### V

Vital: adj.至关重要的，必不可少的，维系生命的

Vertice: n. 顶点（指图中的点，map G(V,E)）

Variance: n. 变异，变化，多样性，方差

Variability: n. 可变性，变化性

Variety: n. 多样性，变化，种类，品种

Antiviral: adj. 抗病毒的，

### W

### X

### Y

### Z

## 名词检索

### 生物

### 化学

### 物理

### 科技学说

## 文献检索

## 问题与思考

### 提问

1. 对于initial fragment hit的评判指标，目前没有很清楚

LE越高对应亲和力越大还是LE越低对应亲和力越大？

我们初始选取的fragment是应当低亲和力还是高亲和力，后续优化呢？

解答：IC50越低，表明该ligand亲和力越高，药效越好，而IC50越低LE会越高。故而LE与affinity是呈反比的。我们选择fragment时，尽量选择LE大亲和力低的，方便后续好进行优化

### 思考讨论

1. 如何表现高亲和与高特异性两种特性？
2. 小分子和RNA结合的关键点在于RNA结构，还是小分子结构呢（RNA）
3. 小分子靶向结合后一定会对RNA的后续产生调控作用吗？是一个问题
4. 老药新用是否违背小分子药物的特异性结合的性质？纯基于计算理论的结构匹配，不改变小分子结构和构象，无法跳脱出老药的枷锁。个人观点中，老药新用，应当加入一定量的生成模型，使得小分子药物能够发生变化，
5. 如何将transformer实施到目前的项目中，对小分子药物的smile格式进行学习，还是针对更长的RNA序列和结构信息进行整合

### 发现与观点

1. 考虑更多高通量的技术，比如如何设计一种小分子到RNA的seq手段，类似于
2. 我们目前的首要目标应该集中在结合亲和力上，也就是筛选结合RNA的小分子时，暂时忽略其类药性。因为在药物设计上，可以通过相似性的手段来获取类药性高的小分子药物。
3. 基于否定式的筛选模型，如果我们知道哪些小分子与RNA无法结合，可以用来做出一种筛选器，滤掉一些不能结合的或者低亲和力的样例。
4. 结构拆分，针对已有的样例，依据经验进行complex的结构拆分，把结合的关键集中到特定区域
5. RNA预筛选，针对RNA结构进行基于先验经验的筛选（我们预期包含低复杂性重复序列，多碱基repeats，以及pocket）
6. 在结构信息的构建过程中，是否考虑过多维度数据的输入，比如，输入的第一行表示序列信息，第二行表示二级结构信息（可以来自于icSHAPE-seq），第三行表示角度信息，诸如此类
7. 氢键等基于先验的经验知识可能对靶向结合具有比较重要的作用，一方面，在进行基于底层原子空间坐标架构的深度学习模型时，可能也需要考虑从原子相互作用力，原子相对位置，原子性质以及原子自身状态等角度来完善之前的模型。或者，基于底层原子空间坐标，我们是否能够识别出这些更加高纬度的经验知识（类似于文献中提到的，降维特征中表现出了氢键和范德华力的作用性）
8. 事实上，由于碱基和氨基酸的数量问题，简单排列组合就能知道，RNA的motif表现出的种类应当是远少于蛋白质的。这也就意味着，极大部分泛滥于RNA中的motif可能并不具备靶向亲和力（因为如果A有，B可能也有），我们可以考虑一些富集密度低的长motif，或者低复杂度的高repeats motif，结合这种策略来筛选RNA。
9. 依据repeats RNA小分子complex模型构建分子机器靶向特定的RNA序列，然后实施APEX-seq技术，达到单定位的RNA-seq。另外通过液液相分离体系的蛋白complex定位到特定空间位置，实施APEX-seq技术，也能够达到单定位的RNA-seq。