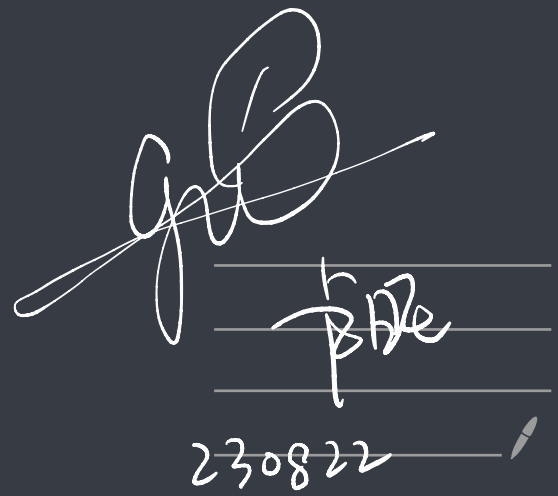
BioBit 暑期培训



0822 钱珑 DNA数字信息存储 北大CQB

淀粉保存

00 → C　3.4nm

01 → G　10.5个base

10 → A　21 bit

11 → T　"空间体积小好运"

dNTP ~330g/mol

$\frac{1g}{330g/mol} \times \frac{1}{2} \times N_A \doteq 0.91 \times 10^{21}$ base

↙×2

1.8 zeta bits

÷8 ↘

225 ExaBytes

1g是很大用量 DNA通常是ng级

✓ 脱氧核苷酸平均分子量是氨基酸的3倍

3 6

K MByte

9 12 15 18 21

GB TB PB EB ZB YB

2023 126.33 ZB whole world

30.02 ZB in China

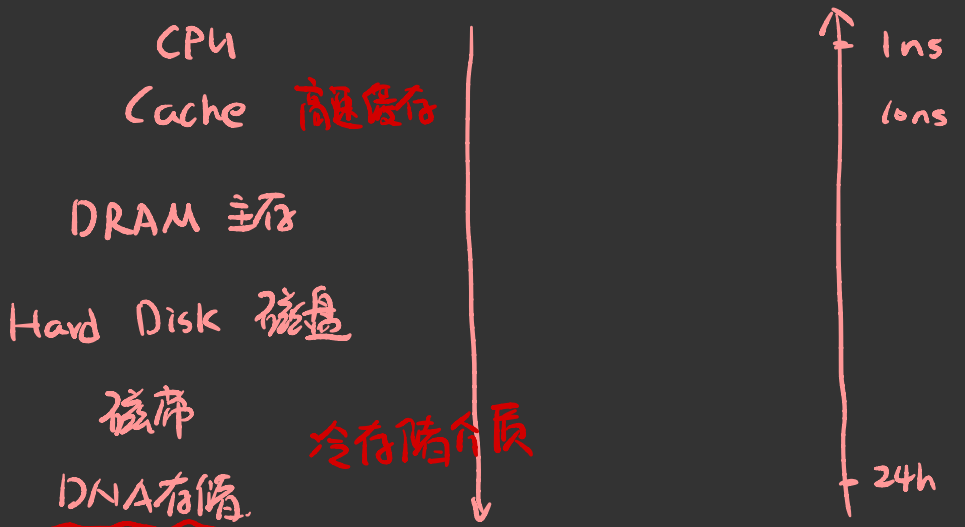
25 188.03

27 284.30

指数增长

储存量不足

耗电量 3%-13% (2023)



✓ 阿里云的机房在杭州附近的一个湖中，从而确保冷却

每5年倒换一次，防止介质损坏

EB级存储中心：~3.8万个机柜 1.68亿度电/年 5年寿命 电磁干扰（安全问题）

1988：首个DNA存储 // 2012：Church 1Gb DNA存书

2013：Huffman编码

Cell population 存一个5帧动图

霍夫曼编码

近几年：存储介质、实例应用　　2021年6月：《DNA存储白皮书》
微软等企业发布的

Q&A：DNA存储能否落地？

2019年3月　1w$　ATCG合成模块　8.4hr　哈佛 or MIT 出了一套完整DNA存储装置。

存5个字母 hello　存储

花10个小时　测序模块（纳米孔测序）36 min

目前 DNA storage：200 MB 存储量。10B/s – 3kB/s 写　28$/KB

1GB/s 读　$0.4–4/GB

合成、测序技术能否推动！→ 决定该技术能否落地

Q&A：技术难题是什么？

DNA合成技术：化学循环合成

合成 → 保护 → 脱保护 ↺

Microsoft：程式合成　　酶合成技术

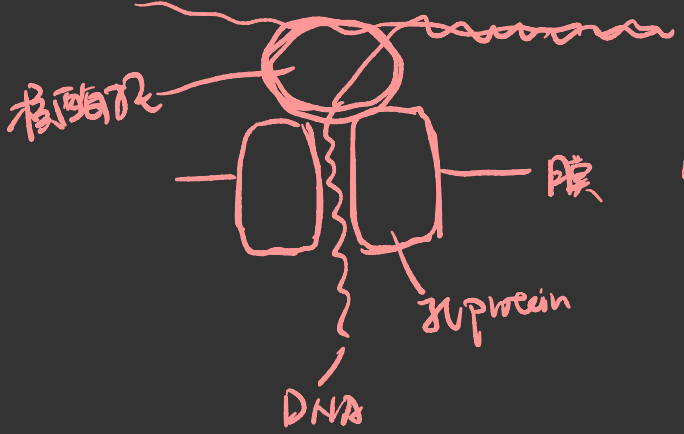
阵列合成技术
- 光化学法：用光做活合成活性　　光学分辨率 亚微米级
- 热敏法
- 电化学法：微电极控制 ⟶ 最密反应腔 2万M/cm²
- 喷墨打印：喷印极少量碱基合成 ⟶ 主流技术 ✓

阵列上反应发生的密度决定了该技术的成本

测序
- 一代 Sanger
- 二代 Illumina　非 de-novo 合成　边测边合　短片段
- 三代 单分子 直接测序
  - PacBio　查
  - 纳米孔　Roche

长片段测序　贵很多
↳ 不需组装　序列相关性保留

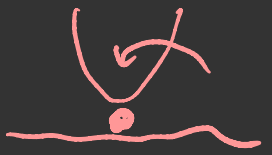
突破！边测边分析



错误率较二代高10-100倍

内外膜电压不同

查 滚环复制 + 荧光碱基



DNA拼接

活字印刷

Catalog 公司 16GB

Wikipedia 数据存储

武汉病毒所.

编码：信道编码器 → 调制写入单元 合成、组装

纠错码 + 纠错码

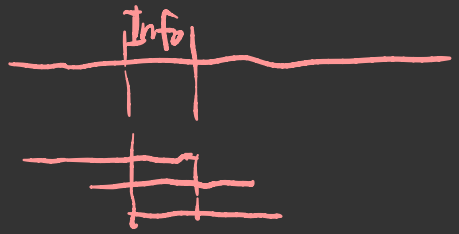
噪声 → 信道

封装 降解 稳定性

测序解码

* 物理冗余

↳ 在获取时 备份三份



* 逻辑冗余 Reed Solomon Code (2015. Grass)

索引位.

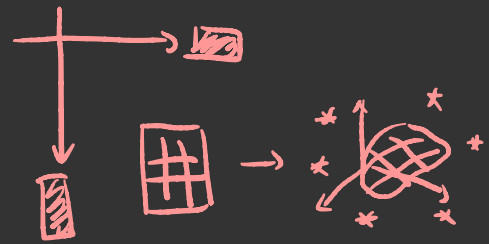
突变

mismatch / indel

DNA 断裂

查

分组、每组加3位 RS码



喷泉码 DNA fountain (2017) Dropout 输液!

多个 数据包 → XOR → 丢失后仍可恢复
(只需要一定量数据包)

冗余度最低

纠错率 0.000 n % 效率当年最高 Best one!

图像编码 (2022) 等

关于 Indel : 纳米孔测序易发 Indel

出孔速率不稳定 ✓ 电流变化产生 Indel

2023年 钱老师文章.

长片段纠错编码体系

GC含量. 避免 poly A/C …

避免生物活性、避免酶结合、避免 blacklist

Random access

数据随机访问

引物 → 索引序列 ⇝ PCR

- ✓ 磁盘上有物理位置（移动磁头）
- ✓ 溶液中需要随机分子碰撞

  100μl中至少 10个 copy 的分子

- ✓ 2021年 将DNA包在硅质小球中（SiO2壳面？）
  - ↳ 从而物理上精准提取

- ✓ Gene 纳米颗粒 ⇝ 文件夹系统 可以变组文件逻辑结构

（钱老师工作）

↓ 5个filter
↓ fountain 喷体
↓
有限个位的 "A"
（水印）

RS纠错
↑
mismatch 错误
↑
most likely hood
↑

将DNA吸附在表面时，人为提高 ΔE，测序精度↑

文件锁定、删除、重命名　也陆续发布.

1g DNA 很难达到 EB 级存储 ＜ 物理冗余　拷贝数、分子数 / 逻辑冗余

↳ 可能降低信息密度更重要!

✓ 隐私便携性（封装）

DNA SiO_2 球 ⟶ 3D打印 热塑材料 ⟶ 兔子 (↗ 眼镜片)

(2020, Nat Biotech)　塑料兔子.

溶液中不能长期存，会断!

✓ 细菌基因组存储

噬菌菌中的整合酶

(钱老师工作)

↳ 插入基因组.

2000代 三个月

10^{-10} 突变率 保真度高

(2013, Adv Sci)

天大 酶母人工染色体.

保留时间不久

枯草芽孢杆菌 ⟶ 芽孢状态封装DNA

Oligo Pool	体内
max 2-3 PB/mL	max 1TB/mL
	50 kb inf DNA

✓ 分级加密 or 销毁

↓

DNA代替传统

转化 → 多层信息存储

↳ 依赖于DNA索引

95℃ 加热破坏索引.

(2019 类 Nat comm)

折纸术 仅信息存储

✓ 痕迹追踪、（DNA稳定不易坏）

DNA barcode 标签！ （2020 Science ⟶ 芳龄喷在表面.

（2020 Nat comm ⟶ 将DNA喷在表面. 便用电信号差异大的短序列做编码. 虽然短序列但可以用 Nanopore 测序.

✓ 动态写入 & 擦除

2021，电信号 ⇒ 数字信号

(Nat Chem bio-l) ↳ 诱导 CRISPR 系统

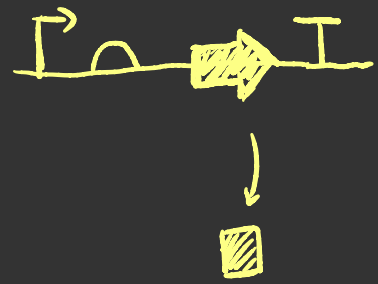
2021. 22联盟 信号可视化

0822 陈昊荣

北大 微电子

普渡 DNA Origami

MIT 博后 合成生物学



"Cello" MIT 导师的工作

硬件描述语言

↓ Parsing

↓ Logic

↓ Gate assignment

↓

DNA

DH10β 中可用

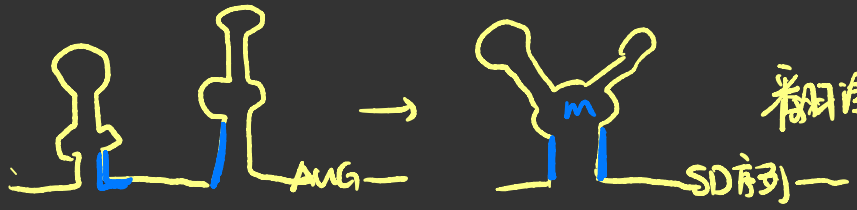
How to more general?

RNA world!

转录

Riboswitch

天然存在



翻译

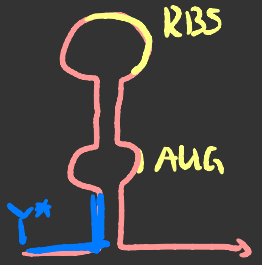
用配体m的结合力
翻转构象

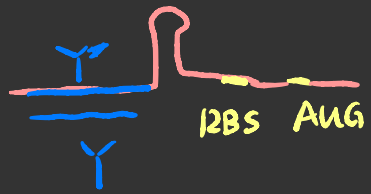
RNA 元件感知小分子！

~40nt的筛选库
Library pool

SELEX

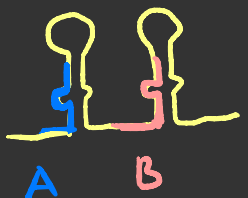
筛 → ✓





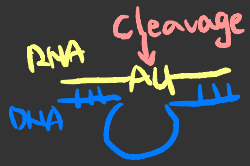
哈佛

成体系的核酸元件



OR Gate

AND Gate



DNAzyme

Mg++ 即可切

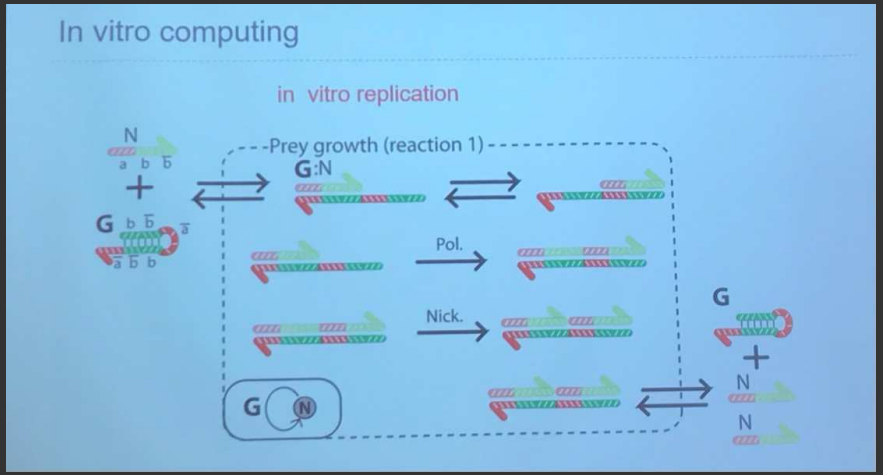
不仅能剪核酸分子.

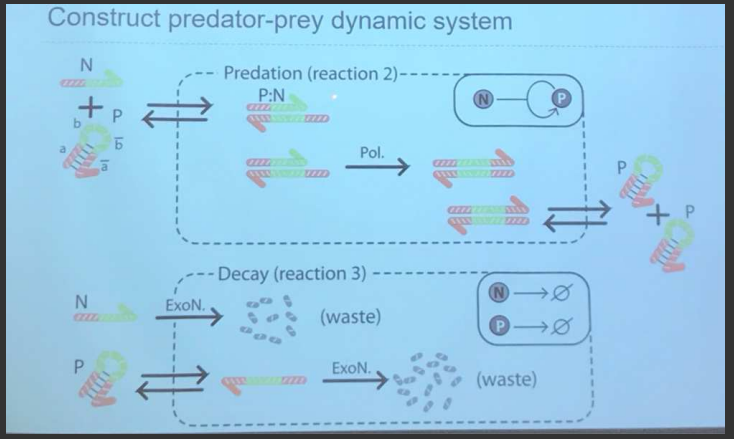
一种化学原是DNA的RNA酶，可以定向地切割RNA分子

NOT Gate

A AND (NOT B) ✓

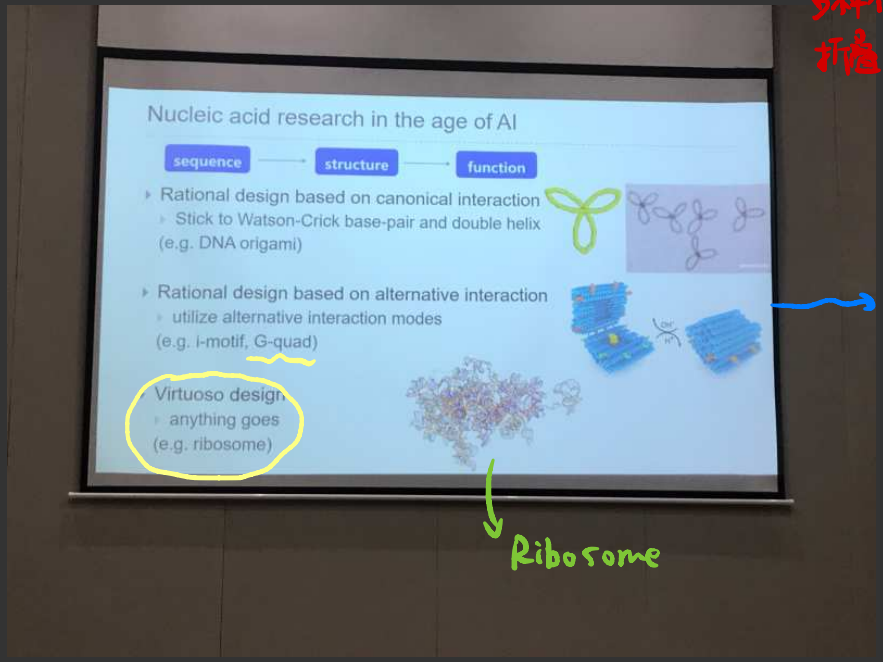
可振荡！





Sequence → structure → function

多种核酸折叠技术
- DNA brick
- DNA origami
- ss-origami 单链DNA



→ 有更多折叠

M-fold ✓

0.3 off-ratio

0822 马大稳 川大 → 清华 → MIT → 斯坦 → 之江

博后

CRISPR：脱靶、靶向范围窄 // AAVS：容量小

↓

PAM 限制

Cas9 同源 protein - 32个

PAM区不保守

↓

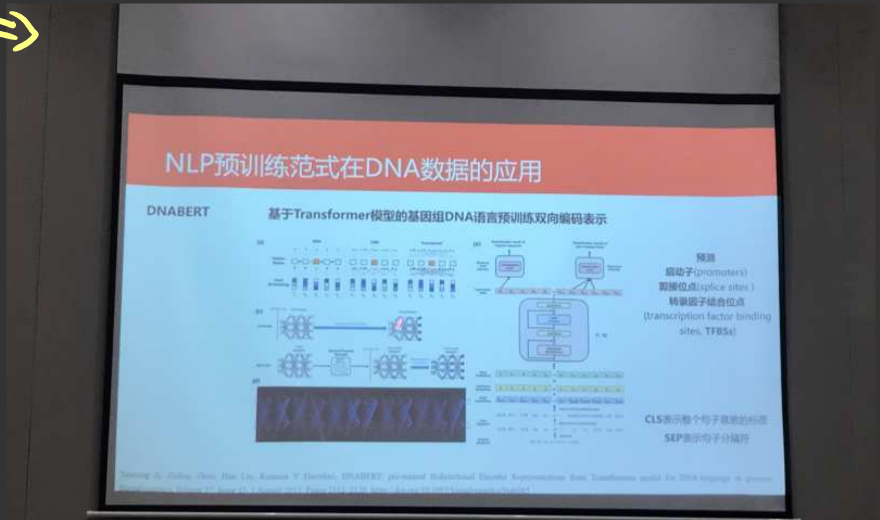
嵌合Cas9 ← 互作短肽移植

(SaCas9)

与DNA类似？ ⟹

NLP 预训练范式:

- 预训练阶段:
- 微调阶段



雷帕霉素: 免疫抑制剂

(不可临床)

PROTAC 改造!

看看合作的点在哪?

↑

当时漏掉信息

现在看似乎就是宝

设无师的工作.

microRNA 与人工启动子

内含肽 编入"与门"

(2016 Nat comm) microRNA toggle switch

二维分子结构 → Graph ？ 空间关系如何识别？

CGT 细胞基因治疗

Qian：如何从模型中提取知识？

钱鹏贝提问

↳ 没有明确回答

230822 傅帅

BT + IT 融合时代

ation ×

Biological Computer ✓

→ 碳基替代硅基

Computational Biology？

→ AlphaFold2

Bioinfomatics ？

→ 数据获取、加工、存储、分析

Biological computation 信息处理能力

明确的研究范式

“星际穿越”

“烟救援”

DBTL循环

Design → Build → Test → Learn

230823 传化科技园

泰州国家园区　22年来传化

生物技术产业新势力 <
- 合成生物学　工程化流程　技术分类　60%-70%
  产业驱动力
- 新材料&先进疗法 ATMP
- AI崛起　AI+医疗 商业模式
  - 医疗器械类
  - 非医疗器械类
  - AI制药企业"共85家"（截至23年7月）（北京 深圳 居多）
- 医药产业关键技术

↓

大力支持+强监督

"挤压销售泡沫"

反腐

纺化化学

高端改性

纺织、纸、塑料

……

竹林式发展

纺织油剂 打破垄断

降本20%

无氯防水 冲锋衣

智能工厂

物流

生鲜基地

公路港

"货运网"

信息平台

交易平台

服务平台

奥的斯电梯

柔性供应链

农业

pǔ yáng

花卉 兰花

玉米 大豆

种苗

科技城

500强

1500多人

一城一产 全球科学家

生活 工作

创新药 基因治疗 合成生物学

400w方 — 200w方产业

200w方

浅滩智谷

就只是一小片

园区名

05年入股新安化工

Si P Cl

隔离剂 双硫环

产业集群

60+亿营收

0823 Angel Cellular computing with contextual dependencies towards environmental applications

@ Angel GMoreno angel.goni @ upm.es

Turing machine — 解读位（操作符）
\ 输入内容 } → 输出.

↙ electronic implementation

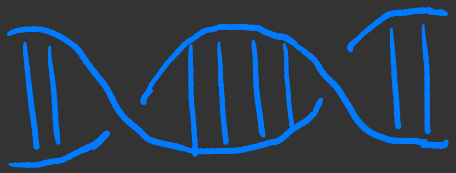
↘ cellular implementation (2019, Nat commu)

"Combinatorial logic"

"Boolean logic gate"

(Wang, 2011, nat commu)

10.1038/ncomms1516 ↵ doi



(Nielsen A 2016, Science) 3 input ⇒ 1 output

"Contextual dependencies"

My python environment is so messy.

Prediction are stopped by non-linearities

(Huseyin, 2021, Nat commu) Different Plasmid, Different Host

↓

Parts 的描述是因环境而异的. "different result"

Predicting compatibility

Pseudomonas putida 假单胞菌

Understanding

↓

Using

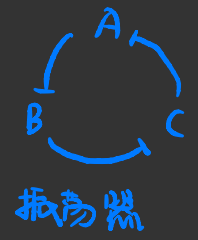
(Nikel, P, 2015, mBio)

(Garcia, 2017, eLife) ⟶ How to use nosie?

2023. pBLAM 基因组插入工具

(Gorfi-Moreno, 2017, ACS synthic biology) 6(7). 1359-1369

0824 王宏俊 大判作！



振荡器

石油基原料 → 合成化学工业 → 产品 （20世纪）

生物基原料 → 合成生物学 → 产品 （21世纪）

模块 系统

DNA → 生物元件 → 生物器件 → 基因线路 ⋯⋯ 细胞、组织

合成生物学 ⇒ 工程生物学

模块化、标准化、层次化 DBTL循环

第三次生物技术革命

BioTech Giants : Microsoft Cambridge / Redmond

做合成 做DNA存储.

Orthogonal split inteins 蛋白胶水

（下面有图）

基础研究 or 应用研究

digital-like circuits ⇒ genetic logic gate

- 标准化的逻辑门库
- 电子设计自动化辅助工具.

帝国理工博士

(2007 – 2011)

/ˈædəm/ Adam, Krishna /ˈkrɪʃnə/

Q: 对比一下不同层次的Gate? 如何选择?

DNA　RNA (核糖开关)　protein (内涵肽)

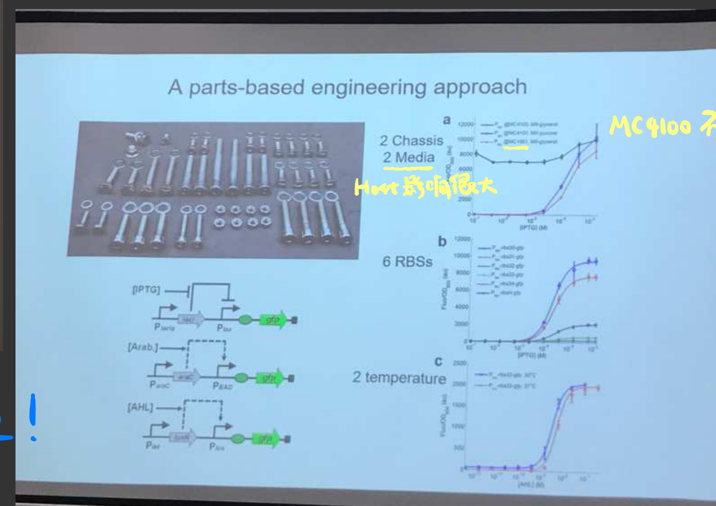
有没有什么principle?

两个TF都存在时才激活 ⟶ 复合体 RS ⟶ 招募 σ54

元件来自于植物(正交化)

✓ 安全性 和 速率问题

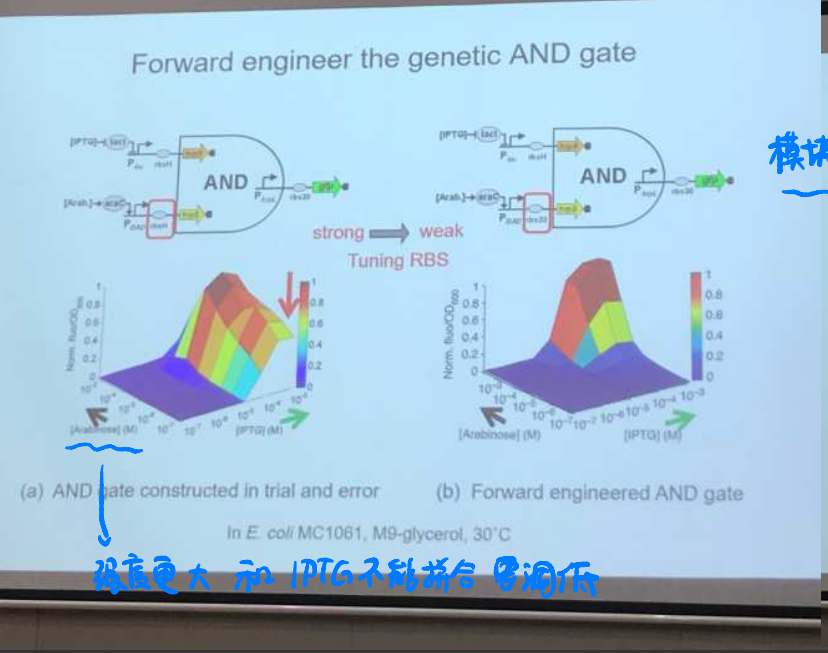
AND Gate



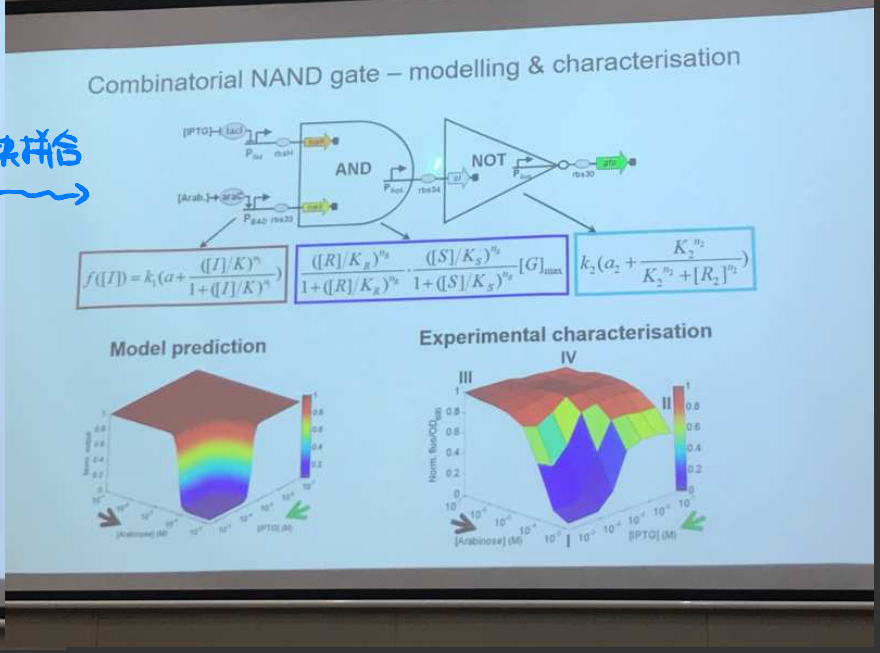
MC4100 不work

Host影响很大

Forward engineer the AND Gate!



模块拼合



强度更大 和 IPTG不能拼合 导通低

Q：拼合时需要关注什么问题

如何标准化？

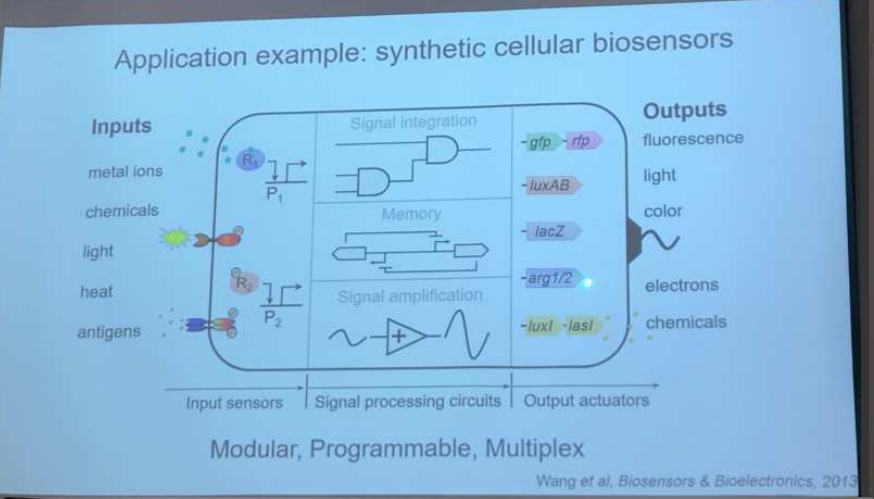
接口

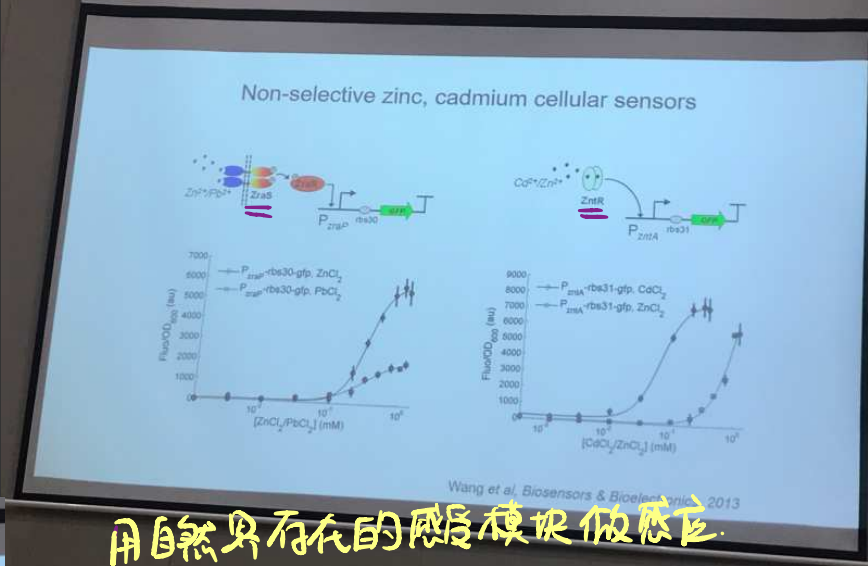
主要用实验测量表征

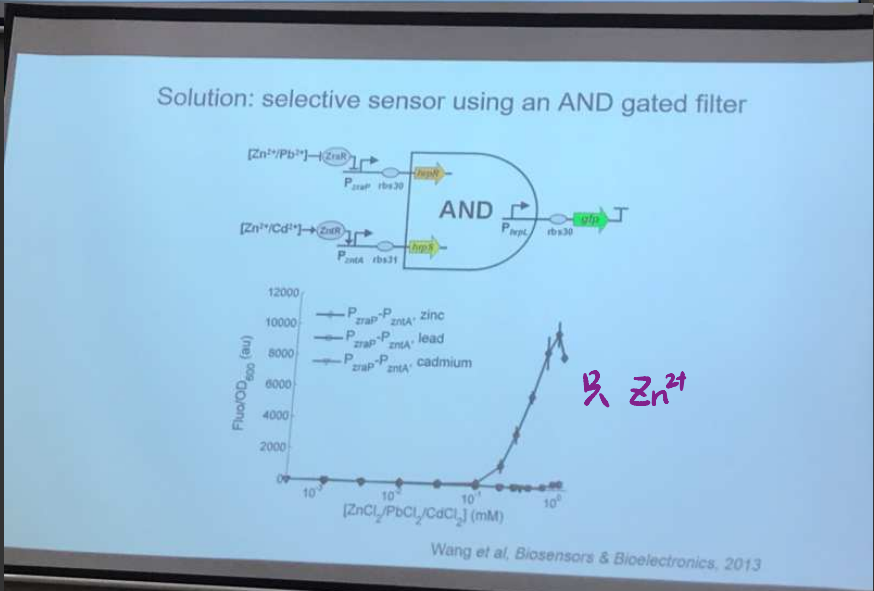
也许有在

隔离序列等方法

来标准化







用自然界存在的感受模块做感应.

↳ 不够特异.

↓

用AND Gate 提高特异性

Q: 不线性如何检测?

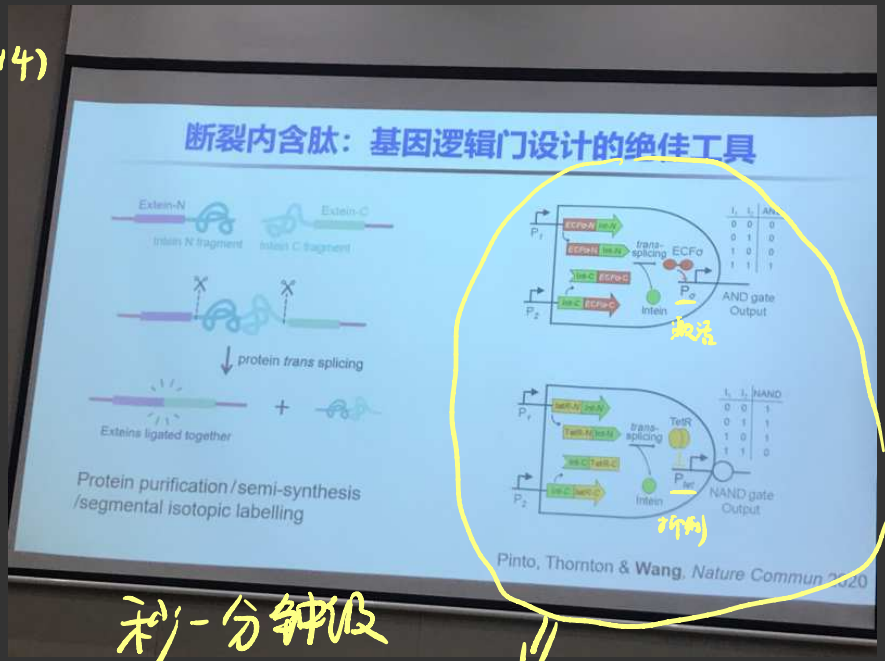
Liu Y et al, Nat Commun 5, 5393 (2014)

国内 癌cell检测并定向清杀

断裂内含肽

不依赖 cofactor 、无残留、自切割

纯化、半合成领域中广泛使用



秒-分钟级

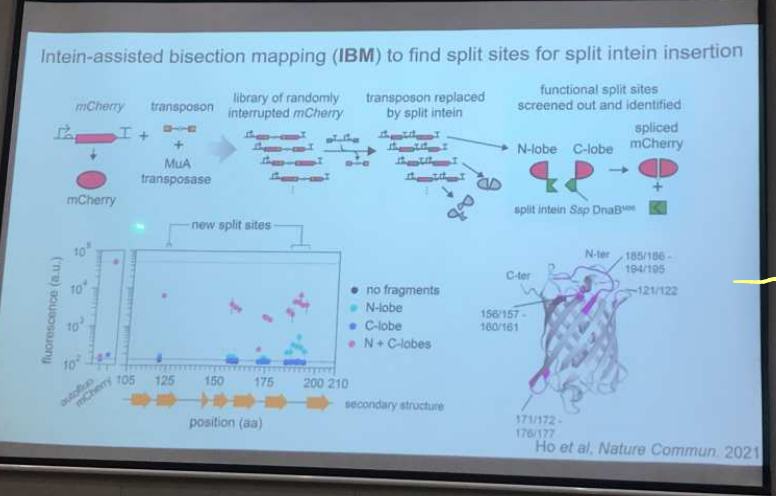
可用内源肽搭建

与门、与非门

15对改内含肽被发现

↓ NEXT

如何在哪里断裂？



查一下定点怎么系统筛

系统的方法

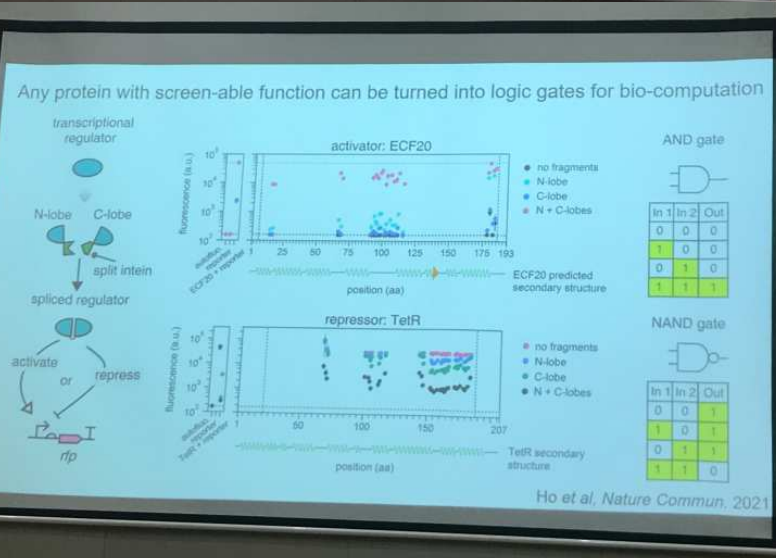
Q: 真核中work?

全细胞反应?

荧光强度比较WT? 查

没能直接问出来，但下来查查一下

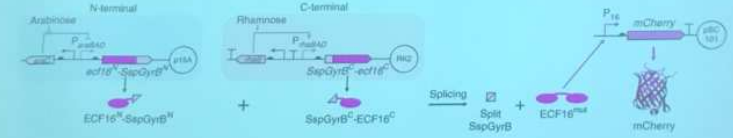
看看是否有细胞应用潜力

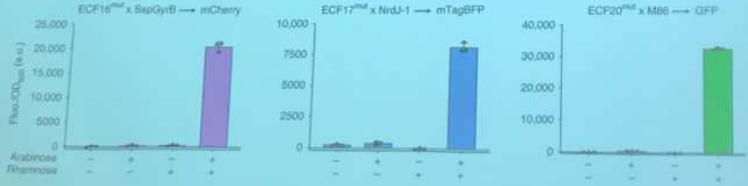


提供了系统化的

断裂位点筛选方法

基于断裂内含肽的正交逻辑与门设计



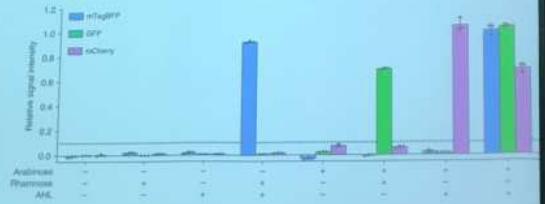


Split inteins can be coupled to **transcription factors** to engineer **logic gates**

基于断裂内含肽的复杂逻辑线路设计—多输入组合区分

3-input 3-output majority voting circuit

Experimental data match the predicted behavior



Split inteins enable gene circuits **scalability and complexity**

Pinto *et al, Nature Commun* 2020

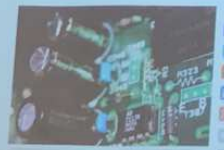
政性不错.

Analogue circuits —— "transcriptional amplifiers" 模拟电路

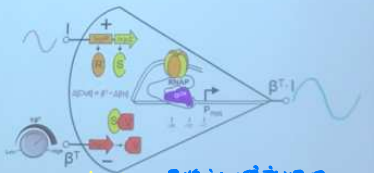
模拟线路 — 生物放大器的设计

Imperial College London

First ever biological amplifier created by Imperial scientists



Modular transcriptional amplifier for ultrasensitive chemical sensors



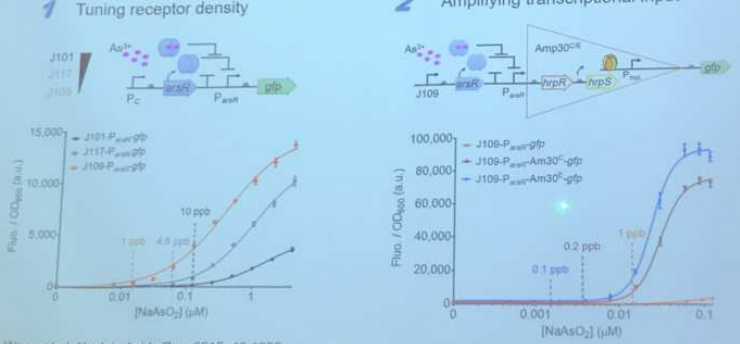
Imw → 而增放大倍率调低

Wang *et al*, *Nuclear Acids Res*, 2014

多层放大器级联的超敏感砷污染传感器

1 Tuning receptor density

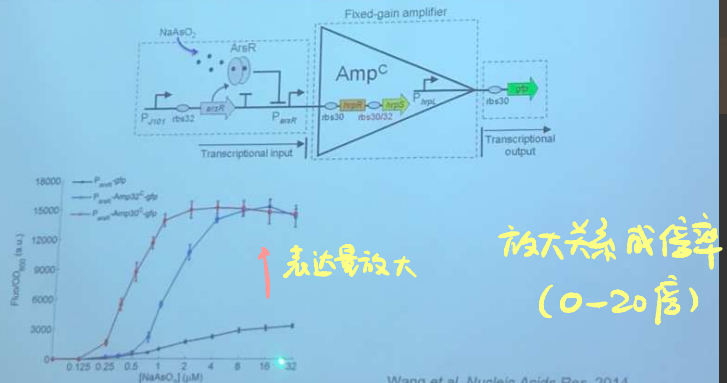
2 Amplifying transcriptional input



Wang *et al*, *Nucleic Acids Res*, 2015, 43:1955

雪摘清楚这几个为因的生物学限强

具有转录信号放大功能的砷污染传感器

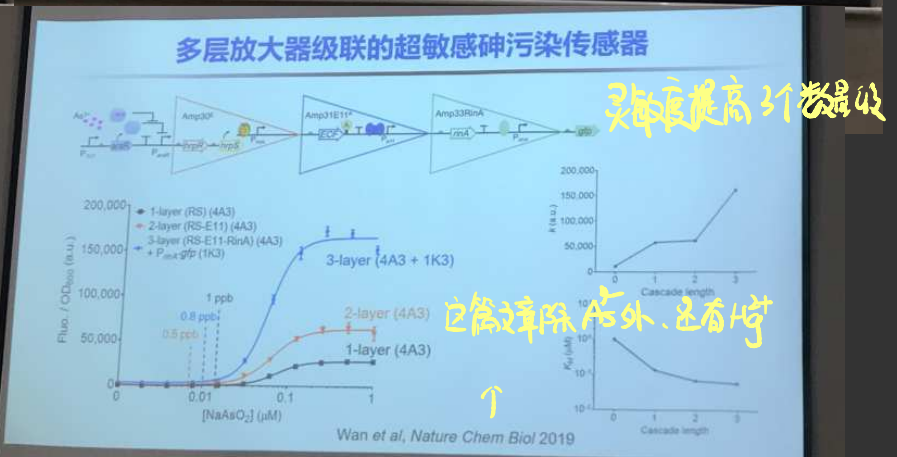


↑ 表达量放大

放大关系 成倍率 (0-20倍)

Wang *et al*, *Nucleic Acids Res*, 2014

多层放大器级联的超敏感砷污染传感器

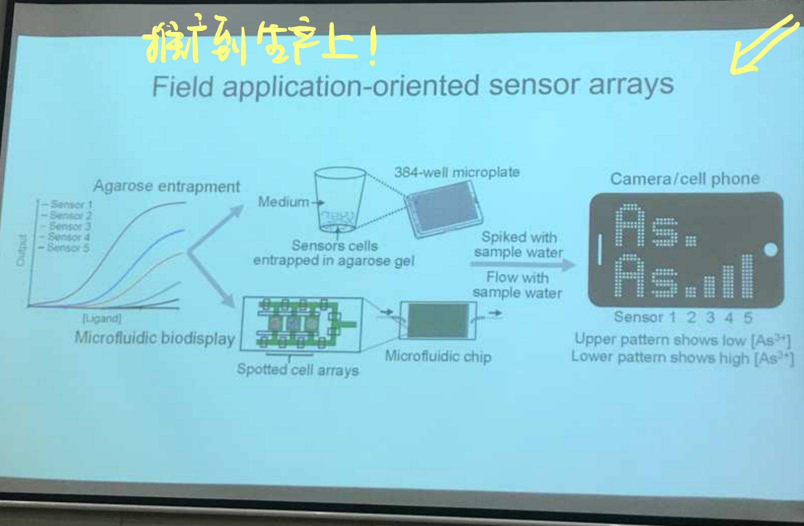


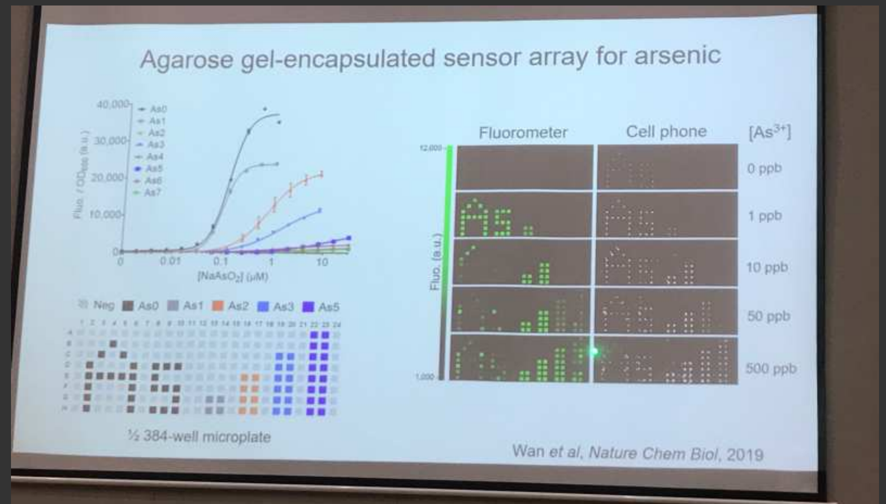
灵敏度提高3个数量级

它篇文章除As外、还有Hg

Wan *et al*, *Nature Chem Biol* 2019

搬到生产上!



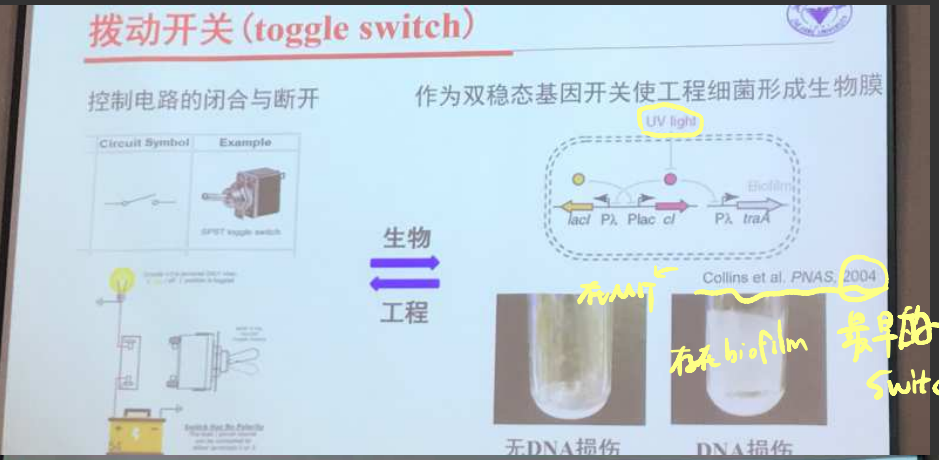


孟加拉国超一半以上人的饮用水中

As^{3+}超标

"Nat chem bio"

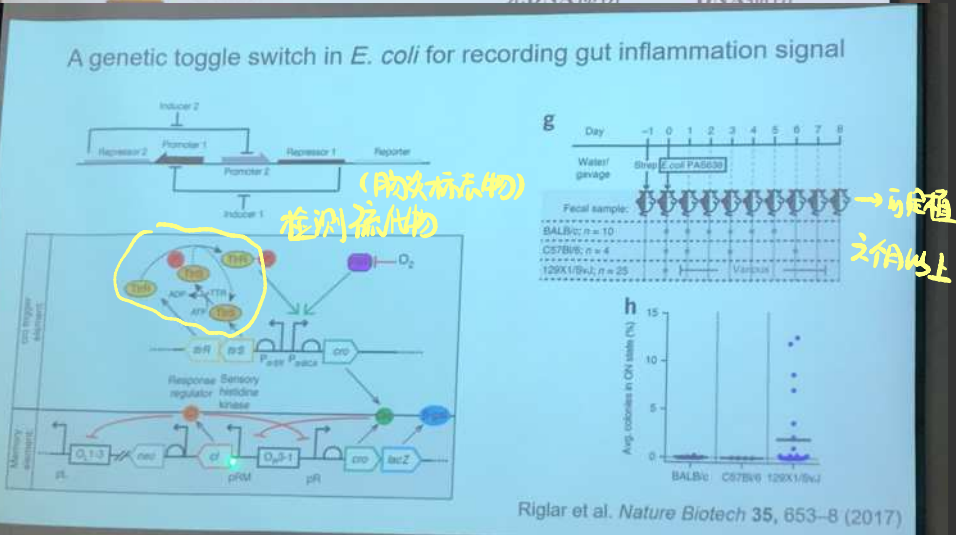
Engineer memory circuits —— data storage in living cells



Q: 可逆的toggle switch?

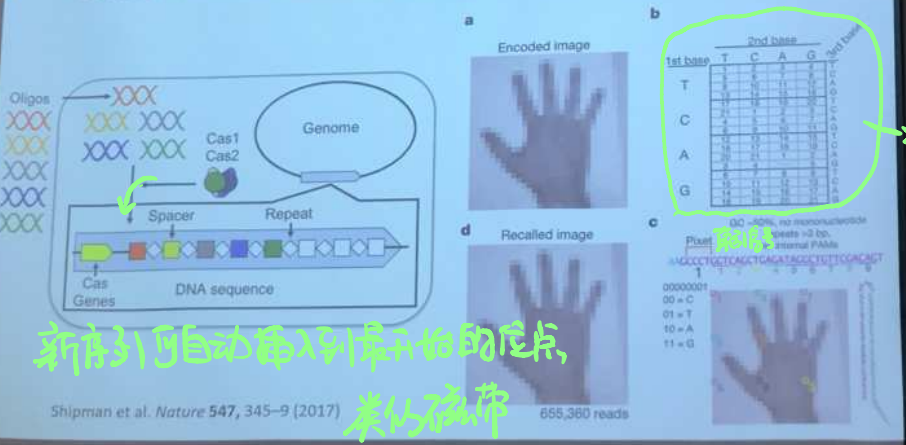
Angel的报告有提到(后面PPT)

基因存在用相同信号调控状态切换的toggle switch.



Riglar et al. *Nature Biotech* **35**, 653–8 (2017)

CRISPR-based memory storage in the genome of cells



Shipman et al. *Nature* **547**, 345–9 (2017)

三个碱基有一个冗度

新序列可自动插入到最初的位点，类似磁带

Cello (Nielsen 2016 Science)

自动化设计软件

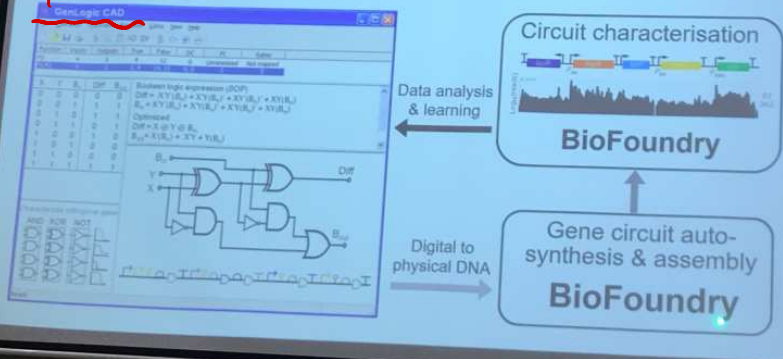
SynBio trends in the next decade

Present	Future 2021-2030
• Trial-and-error	• Predictive & automated
• Building - limiting step	• Data interpretation -limiting step
• Labour intensive	• Machine intensive
• Small scale systems	• Large scale complex systems
• Single cell organisms	• Multicellular organisms
• Static, module focused	• Dynamic, system focused
• Lab proofs-of-concept	• Real-world applications

假想的模式图，不存在真实工具

基因线路设计自动化

Design program to automate matching input-output responses of multi-layered logic gates



Analogy of Bio-IC design to ARM processor design

ARM develop and license CPU cores

\+

ARM development tools (EDA software)

IDEs DS-5 | Keil MDK
Compilers Arm Compiler 6
Boards Juno Arm Platform

→

ARM core designs licensees

Apple (A6 – A12 chips)
Huawei (HiSilicon)
Samsung, Intel …

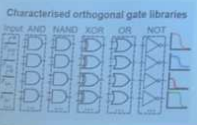
Bio-IC design IP cores
Libraries of genetic logic blocks

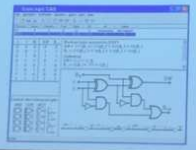
\+

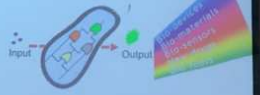
Bio-CAD software

→

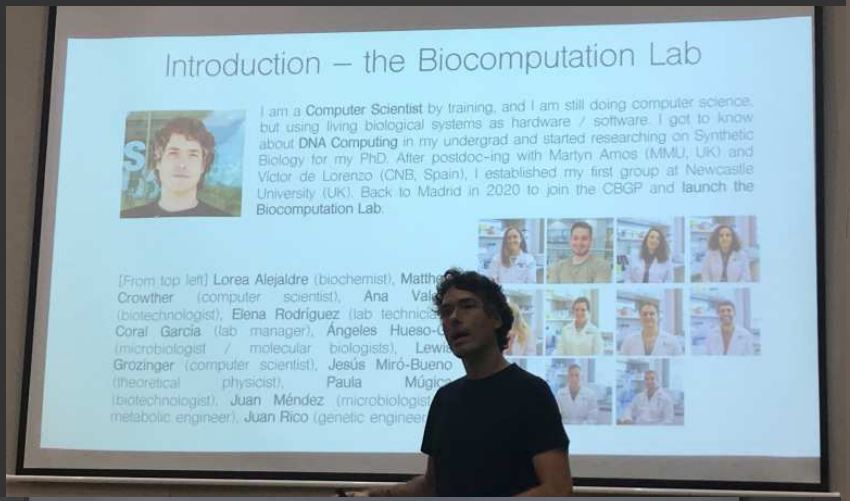
Bio-circuits/ASICs for bio-computing/control applications

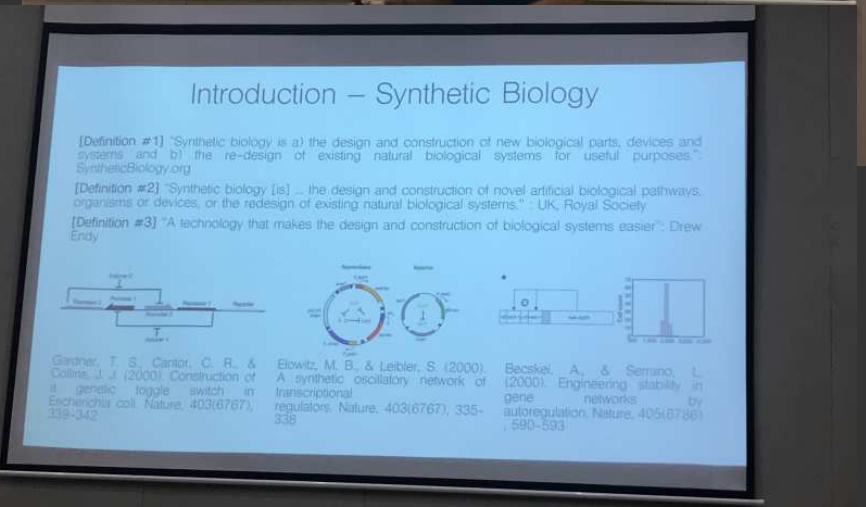


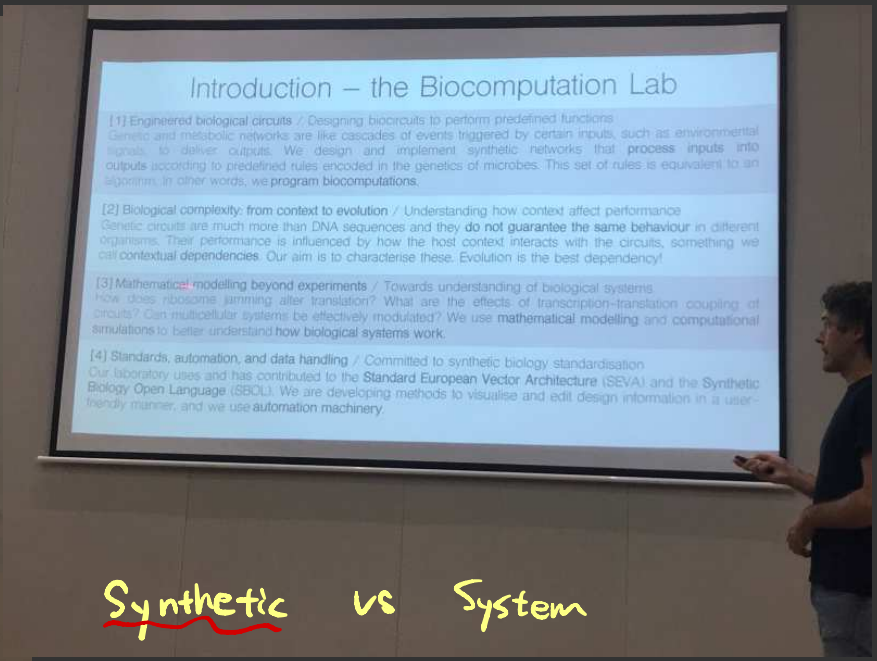




0824 Angel







Synthetic vs System

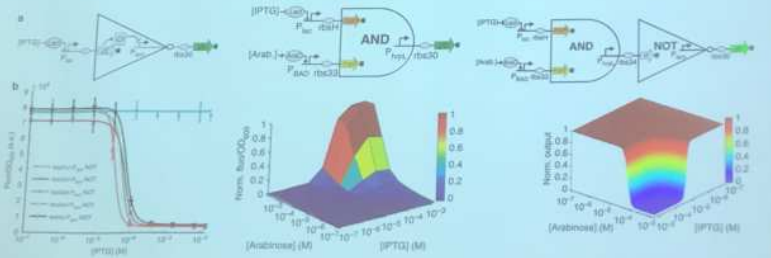
Bottom - Up

Top- Down

Introduction – Biocomputing / programming cells

Synthetic biology: Engineering Escherichia coli to see light. Levskaya et al. Nature 438, 441-442 (2005)

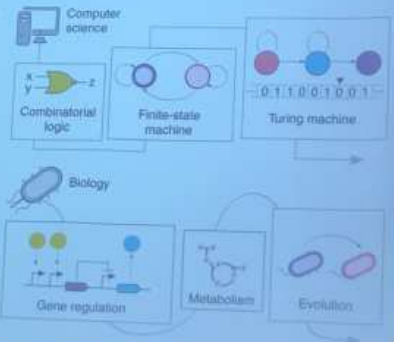
NOT / AND / NAND logic gates



Wang, B. Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. Nature Communications (2011). DOI: 10.1038/ncomms1516

Introduction – Biocomputing

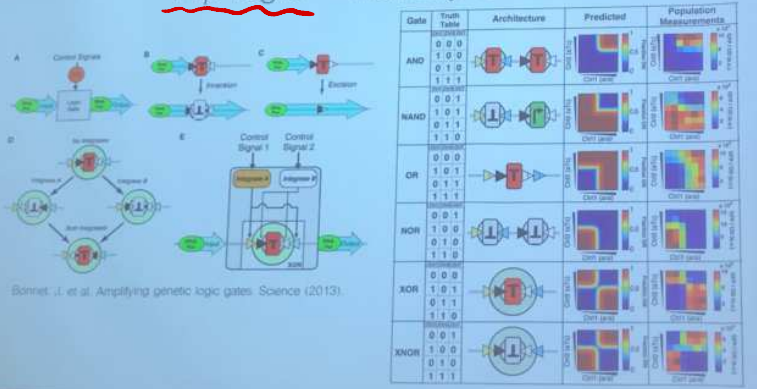
- Not all models of computation are equally powerful, that is, they are not equivalent in terms of the range of computations they can describe.
- Combinatorial logic is extremely weak.
- Two extremely powerful models are the Lambda calculus and the Turing Machine.



Goñi-Moreno, A., Amos, M., Gorochowski, T. E., Carbonell, P., Oyarzún, D. A., Stan, G., & Goñi-Moreno, A. (2019). Formality to cellular supremacy in biocomputing. Nature communications, 10(1), 1-11.

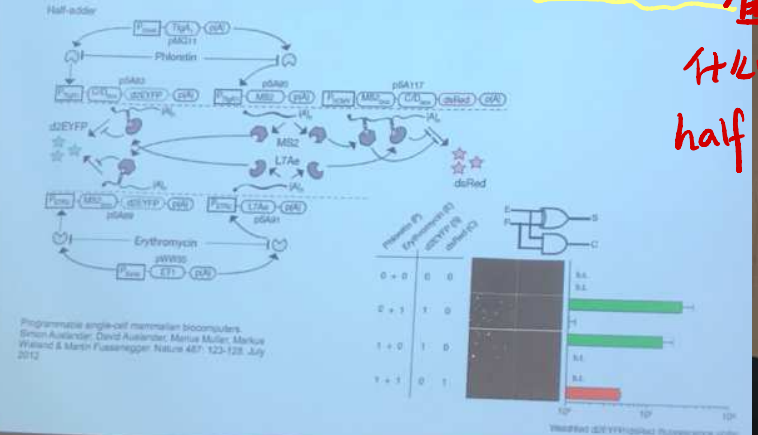
DNA로 XOR Logic Gate!

Flip logic – memory circuits



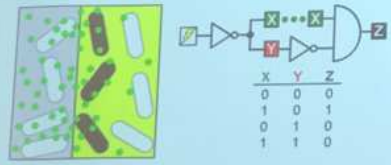
Bonnet, J. et al. Amplifying genetic logic gates. Science (2013).

Mammalian logic circuits – half adder

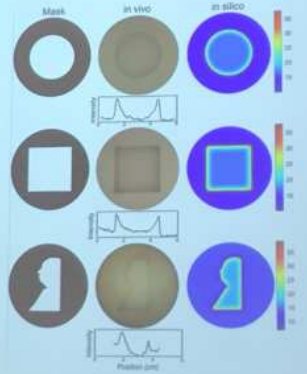


查一下什么叫做 half adder

The edge detector (single cell + populations)

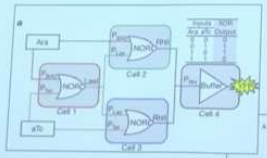


A synthetic genetic edge detection program. Jeffrey J. Tabor, Howard M. Salis, Zachary Booth Simpson, Aaron A. Chevalier, Anselm Levskaya, Edward M. Marcotte, Christopher A. Voigt and Andrew D. Ellington. Cell, 137: 1272–1281 (2009)

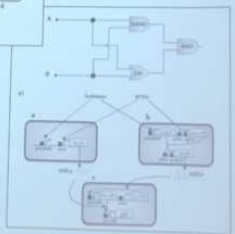


Multicellular distributed computations

However, for certain computations, algorithms exist in which the ordering of computational steps may be relaxed, or abandoned entirely. These algorithms display a level of concurrency.

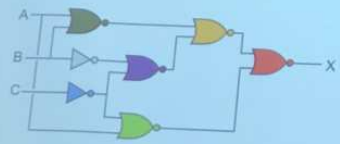


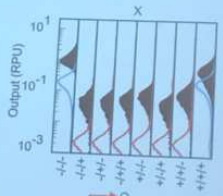
Tamsir, A. et al. Robust multicellular computing using genetically encoded NOR gates and chemical wires. Nature (2011)



Goni-Moreno, A., et al. Biocircuit design through engineering bacterial logic gates. Natural Computing (2010)

Large-scale logic circuits





55 parts

Nielsen, A. et al. Genetic circuit design automation. Science (2016)

Predictive Biology

繁杂的复杂

体系混乱的复杂

Complicated is different than complex

An engine is complicated. A complicated system has a direct and clear **cause and effect relationship**. Its elements interact in a **predictable** way. For each action, there is a proportional reaction. Its problems are often difficult to solve but can be solved with rules and processes after a rational and specialized assessment.

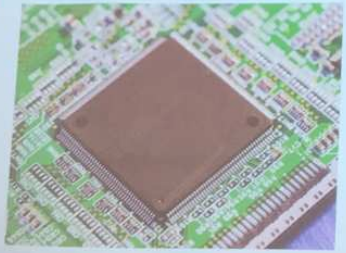
A complex system, in turn, is composed of **elements that interact with each other** exhibiting a dynamic and adaptive behavior. The relationships are more important than the elements themselves. Interactions are what matter. These interactions can give rise to **unpredictable** behaviors, with no single identifiable cause. Small actions can lead to huge reactions, while great interventions may prove ineffective.

Complicated is different than complex

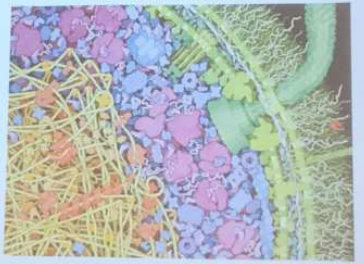
We can't control [complex] systems [permanently] or figure them out. But we can dance with them!

–Donella Meadows

Chips are **complicated**



Living systems are **complex**

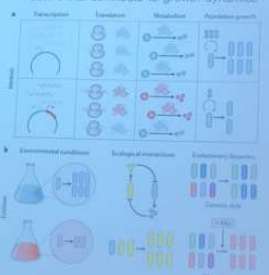


Predictive biology as a challenge

Predictive biology is **the next great chapter in synthetic and systems biology**, particularly for microorganisms. Tasks that once seemed infeasible are increasingly being realized such as designing and implementing intricate synthetic gene circuits that perform complex sensing and actuation functions, and assembling multi-species bacterial communities with specific, predefined compositions. These **achievements have been made possible by the integration of diverse expertise** across biology, physics and engineering, resulting in an emerging, quantitative understanding of biological design. As ever-expanding multi-omic data sets become available, their potential utility in transforming theory into practice remains firmly rooted in the underlying quantitative principles that govern biological systems.

Lopatkin, A. J. & Collins, J. J. (2020). Predictive biology: modelling, understanding and harnessing microbial complexity. *Nature Reviews Microbiology*, 18(9), 507–520.

Factors that contribute to growth dynamics



Predictive biology as a challenge

Complex dynamics in engineered and natural populations
- Systematic characterization of defined ecological modules tested in diverse environments
- Implement high-throughput experimental techniques to increase the parameter space of every isolate-by-environment interaction

Increasingly large data sets
- Minimize 'fishing expeditions', where appropriate, by validating insights from large quantitative data sets with specific testable hypotheses through mathematical models and controlled experiments
- Computational tools that facilitate integrating diverse levels of information into predictive models
- Centralized and accessible parameter reporting

Accurate parameter estimates
- Establish an appropriate level of abstractness with which to characterize a given system
- Match parameter definitions with experimental estimation
- Use conditions that more closely mimic the process of interest to improve model accuracy
- Improved technical tools to standardize parameter estimates
- Efforts to use consistent terminology to facilitate cross-literature data compilation and review

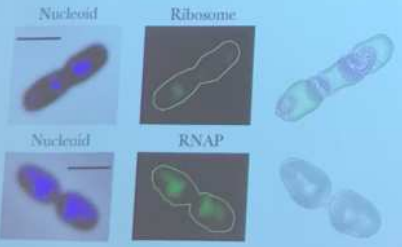
Accounting for stochastic and deterministic evolution in complex populations
- Integration with machine learning and systematic analyses into evolutionary constraints
- Expand sequencing breadth and depth to explore evolutionary responses at lower detection limits

Translating in vitro predictions to in vivo outcomes
- Implement experimental conditions that are biologically relevant
- Inclusion of animal models and other experimental platforms that can better simulate natural environments of interest
- Incorporate physiological conditions into modelling analysis
- Advances in culturing techniques
- Establish in situ quantification methods

Lopatkin, A. J. & Collins, J. J. (2020). Predictive biology: modelling, understanding and harnessing microbial complexity. *Nature Reviews Microbiology*, 18(9), 507–520.

Measure with **image analysis**

- Using an image analysis software (CellShape) we measured the correlation between nucleoid, ribosome and RNAP areas.
- Prediction: localization important & RNA/TF diffusion should play a role

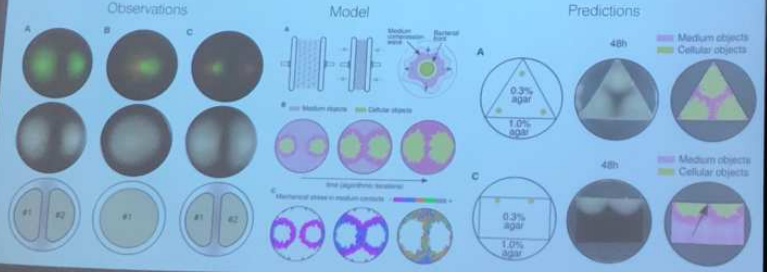


Kim, J., Gorfi-Moreno, A., Calles, B. and de Lorenzo, V. (2019). Spatial organization of the gene expression hardware in *Pseudomonas putida*. Environ Microbiol, 21: 1645–1658.

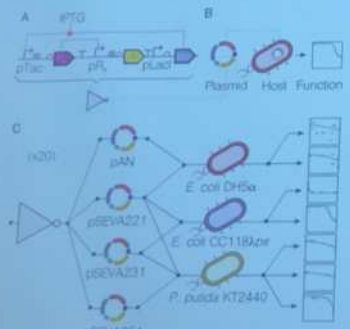
Predicting populations with **agent-based models** (I)

Predictions suggested that the medium is compressed in the direction of the bacterial front motion. This phenomenon generates what was termed a compression wave that goes through the medium preceding the swimming population and that determines the visible high-level pattern.

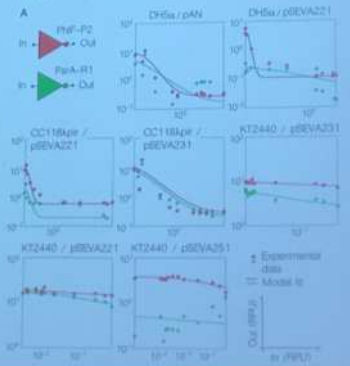
Espeso, D. R., Martínez-García, E., De Lorenzo, V., & Gorfi-Moreno, Á. (2016). Physical forces shape group identity of swimming *Pseudomonas putida* cells. *Frontiers in microbiology*, *7*, 1437.



Predictions are stopped by **nonlinearities**

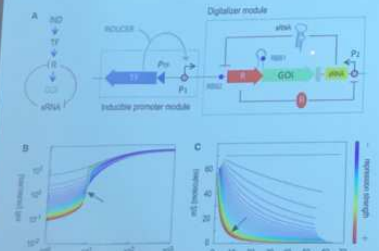


Two gates change performance in different ways. Very difficult to build predictions!



Nuoyan Tao et al. Contextual dependencies expand the reusability of genetic inverters. Nature Communications (2021).

Predicting dynamics with **ODEs** (the *digitalizer*)

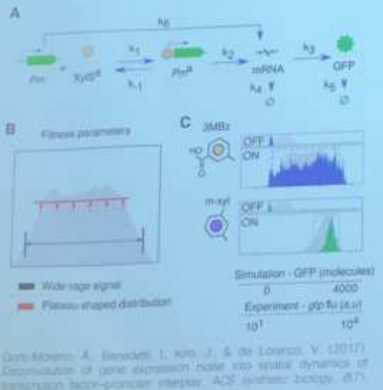


Bottom left graph. Each line is a single simulation that measures the level of mR (y axis) while the concentration of TF molecules increases (x axis). The colour of the lines goes from dark red (both repressions very strong*) to dark blue (both repressions very weak).

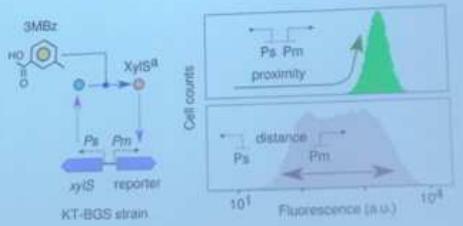
Bottom right graph. Same as before, but plotting mS vs mR. We see the relationship between the two mRNA species in the system. The stronger the repression forces (both) are, the more exclusive this relationship is.

Calles, B., Gorfi-Moreno, Á., & de Lorenzo, V. (2019). Digitalizing heterologous gene expression in Gram-negative bacteria with a portable ON/OFF module. *Molecular systems biology*, *15*(12), e8877.

Predicting dynamics with SSA

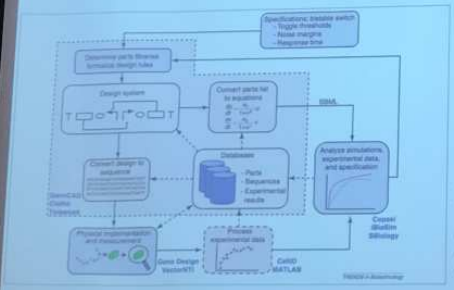


Transcriptional noise was predicted to depend on the intracellular physical distance between regulator source (where XylS is produced) and the target promoter. Experiments with engineered bacteria in which this distance is minimized or enlarged confirmed the predicted effects of source/ target proximity on noise patterns.



Quo-Mowrin, A., Benedetti, I., Kim, J., & de Lorenzo, V. (2017). Deconvolution of gene expression noise into spatial dynamics of transcription factor–promoter interplay. *ACS synthetic biology*, *6*(7), 1359-1369.

Computer Aided Design (CAD) / Aided Modelling (CAM)



"The aim of synthetic biology is to make genetic systems more amenable to engineering, which has naturally led to the development of computer-aided design (CAD) tools. Experimentalists still primarily rely on project-specific ad hocworkflows instead of domain-specific tools, which suggests that CAD tools are lagging behind the front line of the field. Here, we discuss the scientific hurdles that have limited the productivity gains anticipated from existing tools. We argue that the real value of efforts to develop CAD tools is the formalization of genetic design rules that determine the complex relationships between genotype and phenotype"

Lux, M. W., Bramlett, B. W., Bull, D. A., & Peccoud, J. (2012). Genetic design automation: engineering fantasy or scientific renewal?. *Trends in biotechnology*, *30*(2), 120-126.

The Synthetic Biology Open Language (SBOL)

The SBOL data standard is a data exchange representation for synthetic biology designs. Its goal is to improve the efficiency of data exchange and reproducibility of synthetic biology research. SBOL introduces a standardized format for the electronic exchange of information on the structural and functional aspects of biological designs. The standard has been designed to support the explicit and unambiguous description of biological designs by means of a well defined data model. The standard further describes the rules and best practices on how to use this data model and populate it with relevant design details. SBOL uses existing Semantic Web practices and resources, such as Uniform Resource Identifiers (URIs) and ontologies, to unambiguously identify and define genetic design elements. The definition of the data model and associated format, the rules on the addition of data within the format and the representation of this in electronic data files are intended to make the SBOL standard a useful means of promoting global data exchange between laboratories and between software programs.

SBOL

https://sbolstandard.org

A community effort. Current specification version 3.0.1. You can find it here:

https://sbolstandard.org/docs/SBOL3.0.1.pdf

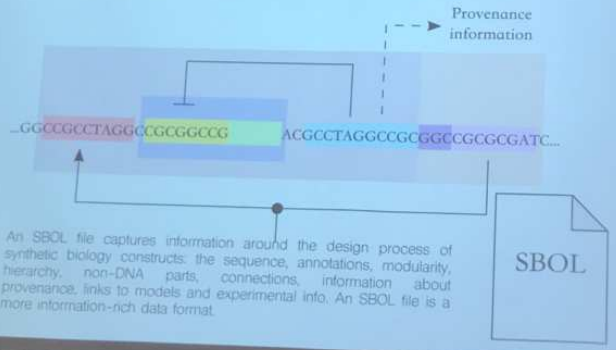
The Synthetic Biology Open Language (SBOL)

...GGCCGCCTAGGCCGCGGCCGCGCGAACGCCTAGGCCGCGGCCGCGCGATC...

A GenBank file encodes a sequence with annotations. That is, it captures information on what are specific regions on that sequence and other information the user wants to input concerning sequence annotations. This type of format is much more useful if you need to identify different regions, parts or sub-sequences.

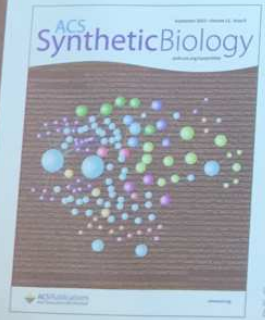
GenBank

The Synthetic Biology Open Language (SBOL)



An SBOL file captures information around the design process of synthetic biology constructs: the sequence, annotations, modularity, hierarchy, non-DNA parts, connections, information about provenance, links to models and experimental info. An SBOL file is a more information-rich data format.

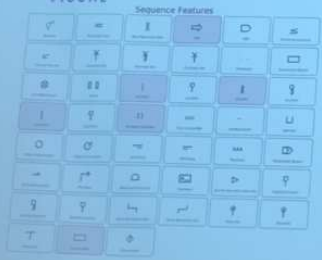
Networks to visualise/analyse SBOL information



"As genetic circuits become more sophisticated, the **size and complexity of data** about their designs increase. The data captured goes beyond genetic sequences alone; information about circuit modularity and functional details improves comprehension, performance analysis, and design automation techniques. However, new data types expose new challenges around the accessibility, visualization, and usability of design data (and metadata). Here, we present a method to **transform circuit designs into networks** and showcase its potential to enhance the utility of design data. Since networks are dynamic structures, initial **graphs can be interactively shaped into subnetworks of relevant information based on requirements** such as the hierarchy of biological parts or interactions between entities. Additionally, several visual changes can be applied, such as coloring or clustering nodes based on types (e.g., genes or promoters), resulting in easier comprehension from a user perspective."

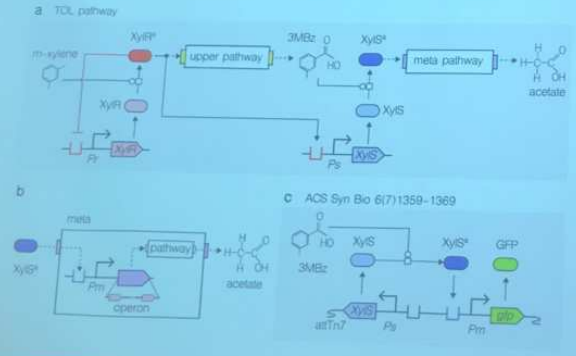
Crowther, M., Wipat, A., & Goñi-Moreno, Á. (2022). A network approach to genetic circuit designs. *ACS Synthetic Biology*, *11*(9), 3058–3066.

SBOL Visual



People who are engineering biological organisms often find it useful to **communicate in diagrams**, both about the structure of the nucleic acid sequences that they are engineering and about the functional relationships between sequence features and other molecular species. Some **typical practices and conventions** have begun to emerge for such diagrams. SBOL Visual aims to organize and systematize such conventions in order to produce a coherent language for expressing the structure and function of genetic designs. At the same time, **we aim to make this language simple and easy to use**, allowing a high degree of flexibility and freedom in how such diagrams are organized, presented, and styled—in particular, it should be readily possible to create diagrams both by hand and with a wide variety of software programs.

SBOL Visual – example I



Beal, J., Nguyen, T., Gorochowski, T. E., Goñi-Moreno, Á., Scott-Brown, J., McLaughlin, J. A., & Wipat, A. (2018). Communicating structure and function in synthetic biology diagrams. *ACS synthetic biology*, *8*(8), 1818–1825.

The Systems Biology Markup Language (SBML)

https://sbml.org

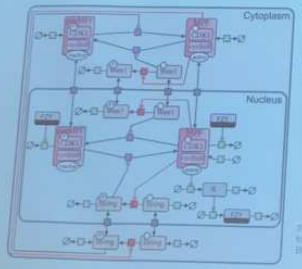
SBML

The starting point is an appreciation that **computational modeling of biological systems is no longer a fringe activity—it's a *requirement* for us to make sense of our vast and ever-expanding quantities of data**. This reality is acknowledged and reinforced by a vast increase over the past two decades in the number of journals, books and articles having computational and systems biology emphases.

At the level of software, a different format is needed for quantifying a model to the point where it can be simulated and analyzed. That's where the Systems Biology Markup Language (SBML) comes in.

Simply put, **SBML is a machine-readable format for representing models**. It's oriented towards describing systems where biological entities are involved in, and modified by, processes that occur over time. An example of this is a network of biochemical reactions. SBML's framework is suitable for representing models commonly found in research on a number of topics, including cell signaling pathways, metabolic pathways, biochemical reactions, gene regulation, and many others.

The Systems Biology Graphical Notation (SBGN)



*"Welcome to the global portal for documentation, news, and other information about the **Systems Biology Graphical Notation** (SBGN) project, an effort to standardise the graphical notation used in maps of biological processes."*

https://sbgn.github.io

Touré, V., Le Novère, N., Waltemath, D., & Wolkenhauer, O. (2018). Quick tips for creating effective and impactful biological pathways using the Systems Biology Graphical Notation. PLoS computational biology, 14(2), e1005740.

The Standard European Vector Architecture (SEVA)

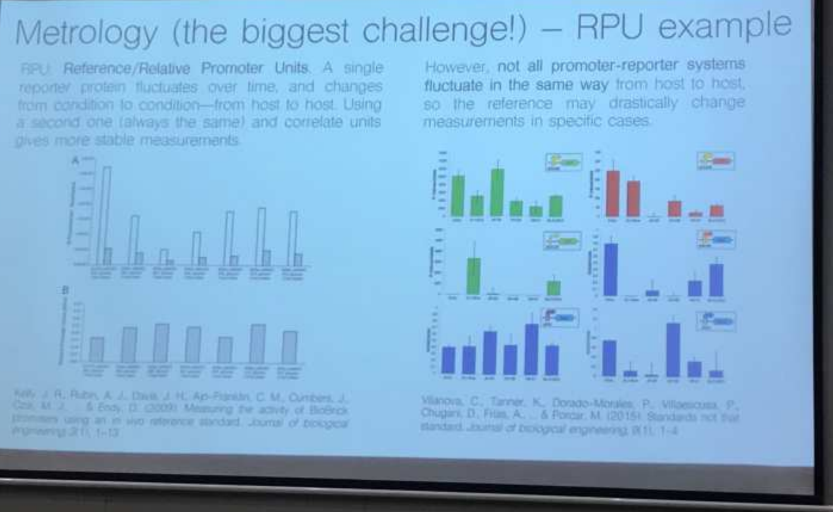
https://seva-plasmids.com

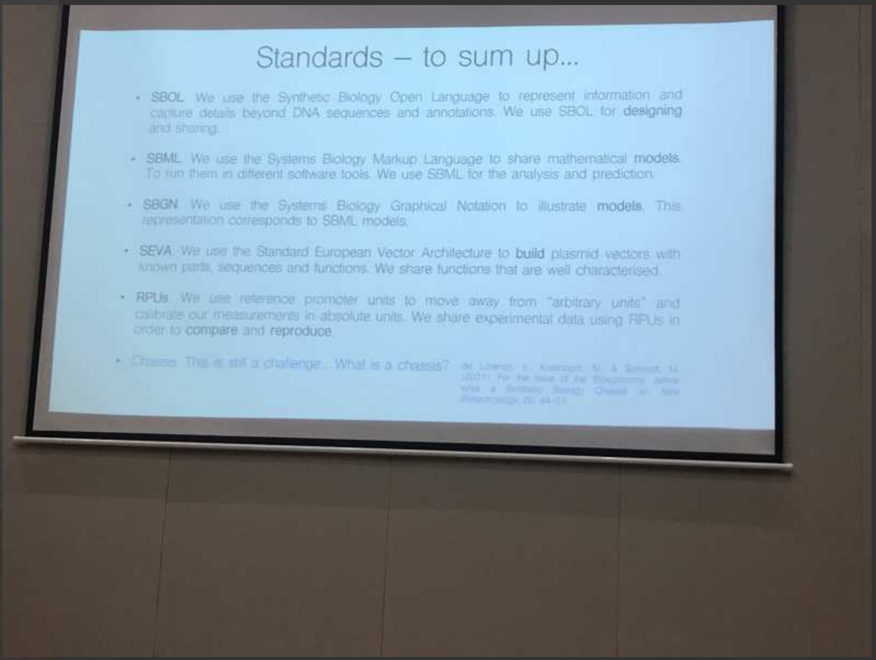
The Standard European Vector Architecture (SEVA) platform is a **web-based resource** *and a material clone* **repository** *to assist the choice of optimal plasmid vectors for de-constructing and re-constructing complex prokaryotic phenotypes.*

The SEVA database (**SEVA-DB**) is a resource for implementation of a standard for physical assembly of vector plasmids and for their non-ambiguous nomenclature as well as the index for a repository of functional sequences and actual constructs available to the community. The database was designed to simplify the choice of a given vector for the sake of specific applications, in such a way the user can easily decide the best configuration of replication origins, antibiotic resistance and business segments.

The Standard European Vector Architecture (SEVA)

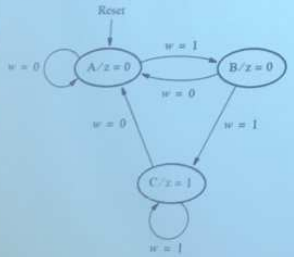
Rules for naming each of the 4 positions available for composing a complete SEVA code. The first position identifies the antibiotic resistance. The second position is the origin of replication (a sole numeric code 1 to 9 and then a capital letter is added). In case of either variants or addition(s) of either the antibiotic marker or a second replication origin, a lower case letter is then inserted next. The third position is the cargo, which can be mono-function (named 1 to n) or variants/multi-function thereof (number followed by a capital letter). Finally, the fourth position is for the gadgets, which are designated by lowercase Greek letters (α to ω).


Metrology (the biggest challenge!) – RPU example
RPU: Reference/Relative Promoter Units. A single reporter protein fluctuates over time, and changes from condition to condition—from host to host. Using a second one (always the same) and correlate units gives more stable measurements.
However, not all promoter-reporter systems fluctuate in the same way from host to host, so the reference may drastically change measurements in specific cases.
A
B
Kelly, J. R., Rubin, A. J., Davis, J. H., Ajo-Franklin, C. M., Cumbers, J., Czar, M. J., ... & Endy, D. (2009). Measuring the activity of BioBrick promoters using an in vivo reference standard. Journal of biological engineering, 3(1), 1-13.
Vilanova, C., Tanner, K., Dorado-Morales, P., Villaescusa, P., Chugani, D., Frías, A., ... & Porcar, M. (2015). Standards not that standard. Journal of biological engineering, 9(1), 1-4.

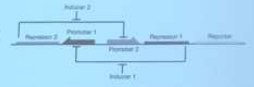

Standards – to sum up...
SBOL: We use the Synthetic Biology Open Language to represent information and capture details beyond DNA sequences and annotations. We use SBOL for designing and sharing.
SBML: We use the Systems Biology Markup Language to share mathematical models. To run them in different software tools. We use SBML for the analysis and prediction.
SBGN: We use the Systems Biology Graphical Notation to illustrate models. This representation corresponds to SBML models.
SEVA: We use the Standard European Vector Architecture to build plasmid vectors with known parts, sequences and functions. We share functions that are well characterised.
RPUs: We use reference promoter units to move away from "arbitrary units" and calibrate our measurements in absolute units. We share experimental data using RPUs in order to compare and reproduce.
Chassis: This is still a challenge... What is a chassis?

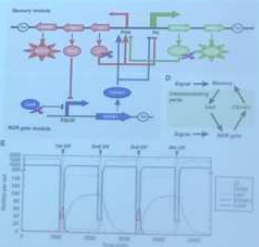
Sequential Logic – States

The output of stateful computations depends not only on the current input, but also on the current state. State Machines are more powerful than Combinatorial Logic.



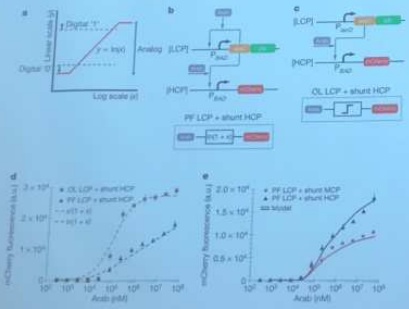
Gardner, T., Cantor, C., Collins, J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* (2000).





Lou, C., Liu, X., Ni, M., Huang, Y., Huang, Q., Huang, L., ... & Ouyang, Q. (2010). Synthesizing a novel genetic sequential logic circuit: a push-on push-off switch. *Molecular systems biology*, *6*(1), 350.

Analogue Biocomputing (in/out range pop. level)



This work shows variations in dynamic ranges in relation to network motif (PF vs OL).

PF returns wider dynamic range (i.e., more values at the output). Here, the OL returns a more digital-like behaviour (at the population level).

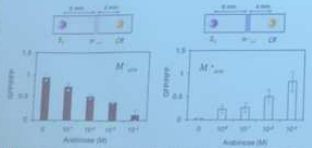
Computing here is done by turning inputs into outputs, in the analogue domain!

Daniel, R., Rubens, J. R., Sarpeshkar, R., & Lu, T. K. (2013). Synthetic analog computation in living cells. *Nature*, *497*(7451), 619-623.

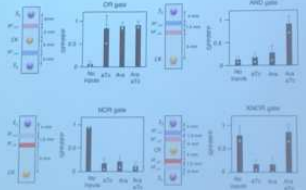
Analogue Biocomputing (in/out range pop. level)

By introducing (extracellular) **space** as another parameter, analogue signalling is used to process information. S: sender strain; CR: reporter strain, M: modulation cells; CS: carrying signal.

M- The more Ara, the less CS goes through (block).
M+ The more Ara, the more CS goes through.



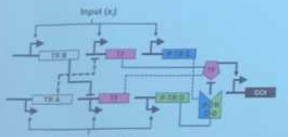
Modulating space in between S, CR and M cells, the reporter performed different functions. **The genetic programs are the same, but processing is analogue.**



Rubio-Diez, S., Gonzalez-Flo, E., & Macia, J. (2021). 2D printed multicellular devices performing digital and analogue computation. *Nature communs*, *12*(1), 1-10.

Reprogrammable biocomputers

Individual cells (left) can let inputs (x) go through (ID) if the reprogramming signal (p) is 0. They invert (NOT) the input if the reprogramming signal is 1. By mixing different cells in a chamber, the program is reconfigured dynamically (right): same cells, same genetic circuits, different reprogramming signals to build a NOR an N-IMPLY or an AND function. Key concept: **there are inputs, outputs and reprogramming signals (*three different types!*).**



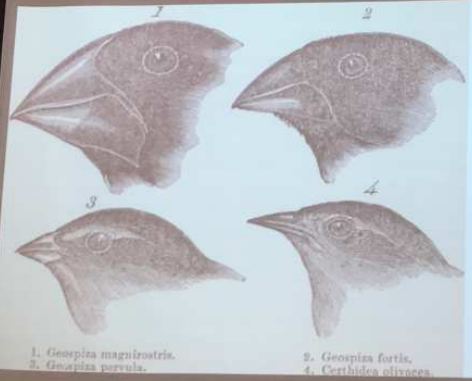
p	x	TR A	TR B	P-TR C	P-TR D	P-TR	TF	GOI
0	0	0	0	0	0	0	0	0
0	1	0	1	1	0	0	1	1
1	0	1	0	0	1	0	1	1
1	1	1	1	1	1	1	0	0

Canadell, D., Diez-Vaquerizas, N., Mogas-Diez, S., de Nadal, E., Macia, J., & Posas, F. (2022). Implementing reconfigurable biological computation with distributed multicellular consortia. *Nucleic Acids Research*.

Biocomputing beyond Turing?

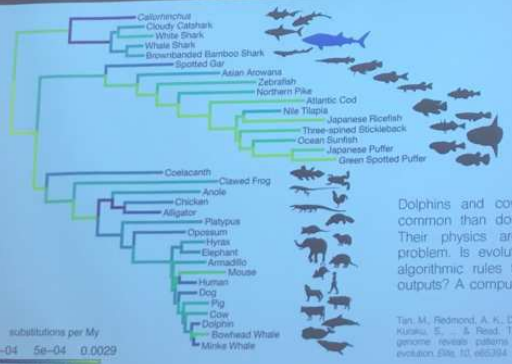
"A priori it is not obvious that every function which we would intuitively regard as computable by an algorithm can be computed using a Turing Machine. Church, Turing and many other people have spent a great deal of time gathering evidence for the Church–Turing thesis, and in sixty years no evidence to the contrary has been found. Nevertheless it is possible that in the future we will discover in Nature a process that computes a function not computable on a Turing Machine. It would be wonderful if that ever happened, because we could then harness that process to help us perform new computations which could not be performed before. Of course we would also need to overhaul the definition of computability, and with it, computer science."

Nielsen, M. A., & Chuang, I. (2002). Quantum computation and quantum information.



Is this not the solution to a problem? (no matter how the solution is found). Is this not a computation? There are inputs, outputs and rules. What about those rules?

Darwin's finches or Galapagos finches. Darwin, 1845. Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N. 2d edition. 1

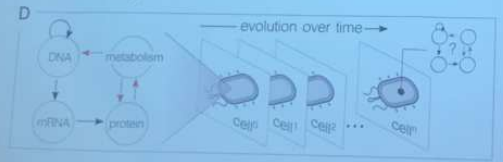


Dolphins and cows have more in common than dolphins and sharks. Their physics are solutions to a problem. Is evolution then a set of algorithmic rules to turn inputs into outputs? A computation then?

Tan, M., Redmond, A. K., Dooley, H., Nozu, R., Saito, K., Kuraku, S., ... & Read, T. (2021). The whale shark genome reveals patterns of vertebrate gene family evolution. *Elife*, *10*, e65394.

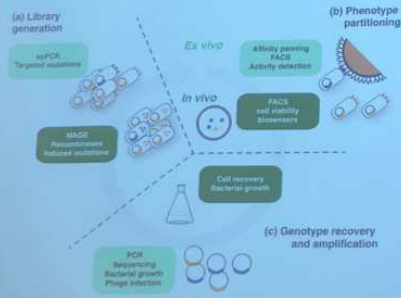
Motivation: Evolutionary Computing vs. Evolution

- Evolutionary computing is a field within computer science that develops and applies algorithms inspired by Darwinian evolution.
- Living systems evolve.
- Both computer scientists and living systems use evolutionary algorithms to generate algorithmic solutions.
- However, the rational engineering of autonomous evolutionary computing in living cells is still an overarching challenge.
- A potential initial step would be to expand on the standard representation of the central dogma of molecular biology.



Grozinger, L., Amos, M., Gorochowski, T. E., Carbonell, P., Oyarzún, D. A., Stoof, R., ... & Corti-Moreno, A. (2019). Pathways to cellular supremacy in biocomputing. *Nature communications*, *10*(1), 1-11.

Side [2]: using natural evolution



- **Directed evolution** for optimizing and engineering novel functions in both nucleic acids and proteins.
- Similarities with Darwinian evolution: genetic and phenotypic diversity, selection...
- Those simple steps hide a huge number of methodologies and procedures: library quality, evolutionary landscapes...
- Rational design vs. directed evolution (full knowledge of the system vs. gaps)
- More than 20 years of successful directed evolution (> than synbio)

Pedro Tizei et al. Selection platforms for directed evolution in synthetic biology. Biochem. Soc. Trans. (2016) 44, 1165–1175

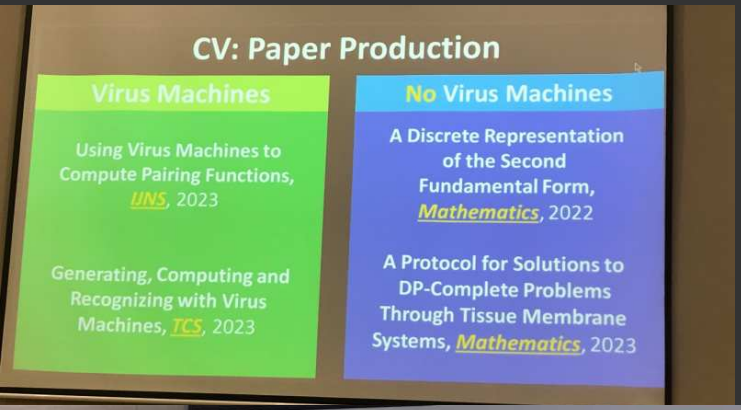
Side [2]: using natural evolution (Logic example)

Evolved the connection between gates (the cI repressor in this example). Warning: directed evolution needs manual steps (i.e., selection).

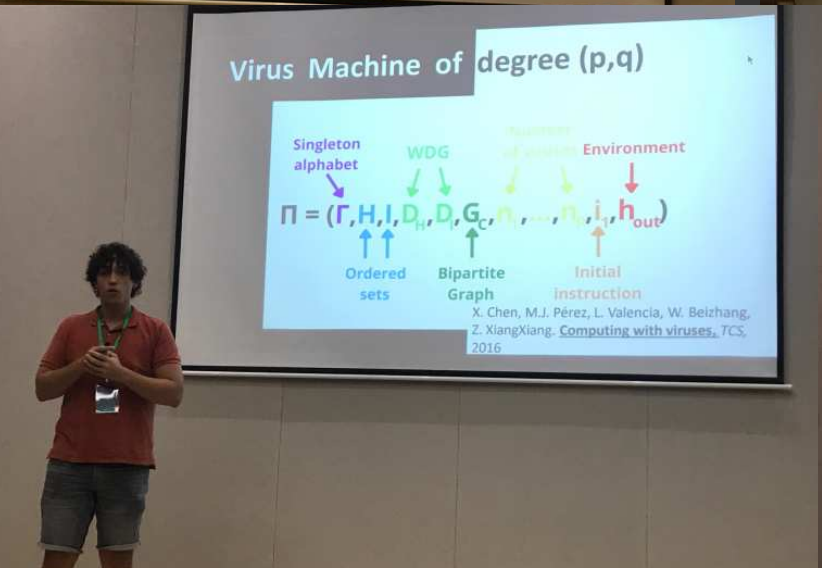
"We propose a combined rational and evolutionary design strategy for constructing genetic regulatory circuits, an approach that allows the engineer to fine-tune the biochemical parameters of the networks experimentally in vivo. By applying directed evolution to genes comprising a simple genetic circuit, we demonstrate that a nonfunctional circuit containing improperly matched components can evolve rapidly into a functional one"

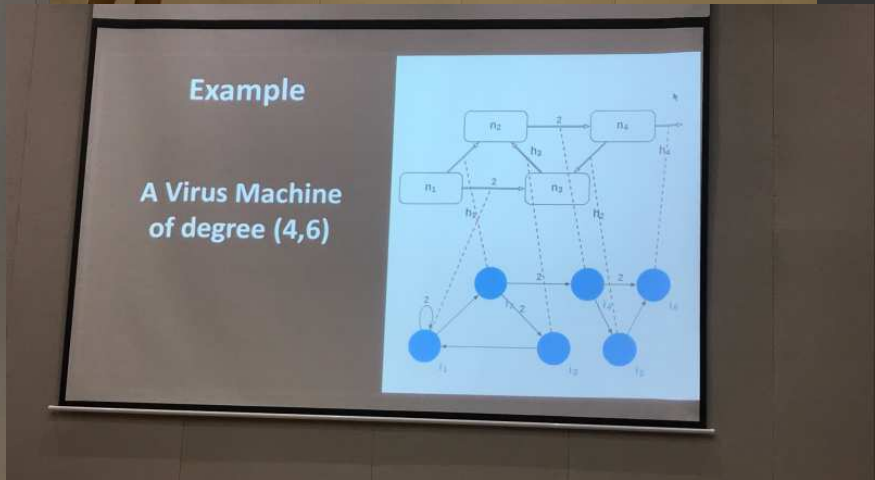
lacI	IPTG	cI	EYFP
0	0	1	0
0	1	1	0
1	0	0	1
1	1	1	0

Yokobayashi, Y., Weiss, R., & Arnold, F. H. (2002). Directed evolution of a genetic circuit. *Proceedings of the National Academy of Sciences*, *99*(26), 16587–16591.


CV: Paper Production
Virus Machines
Using Virus Machines to Compute Pairing Functions, IJNS, 2023
Generating, Computing and Recognizing with Virus Machines, TCS, 2023
No Virus Machines
A Discrete Representation of the Second Fundamental Form, Mathematics, 2022
A Protocol for Solutions to DP-Complete Problems Through Tissue Membrane Systems, Mathematics, 2023


CV: Presentations
Conferences
IWINAC 2022 Spain
CMC2022 Italy
(A)CMC2023 China
Best Paper Award


Virus Machine of degree (p,q)
Singleton alphabet
WDG
Environment
Ordered sets
Bipartite Graph
Initial instruction
X. Chen, M.J. Pérez, L. Valencia, W. Beizhang, Z. XiangXiang. Computing with viruses, TCS, 2016


Example
A Virus Machine of degree (4,6)

Power	Turing Universality
Formal Framework	Formal Verification
Simulation	Software
Problems	Cryptography
Efficiency	Sequential

Parallelism	• Instructions • Channels
Bio-extensions	• Host excitation • Mutation
Real-World	• Image processing • Time series

2

My Research Contents：研究背景

脉冲神经P系统（Spiking neural P systems, SNP systems）

- 脉冲神经P系统是神经型膜计算模型的主要形式，是一类受生物神经元以脉冲的方式传递信息的机制启发设计的神经计算模型，抽象于生物神经元发放脉冲的现象。

2

My Research Contents：研究动机

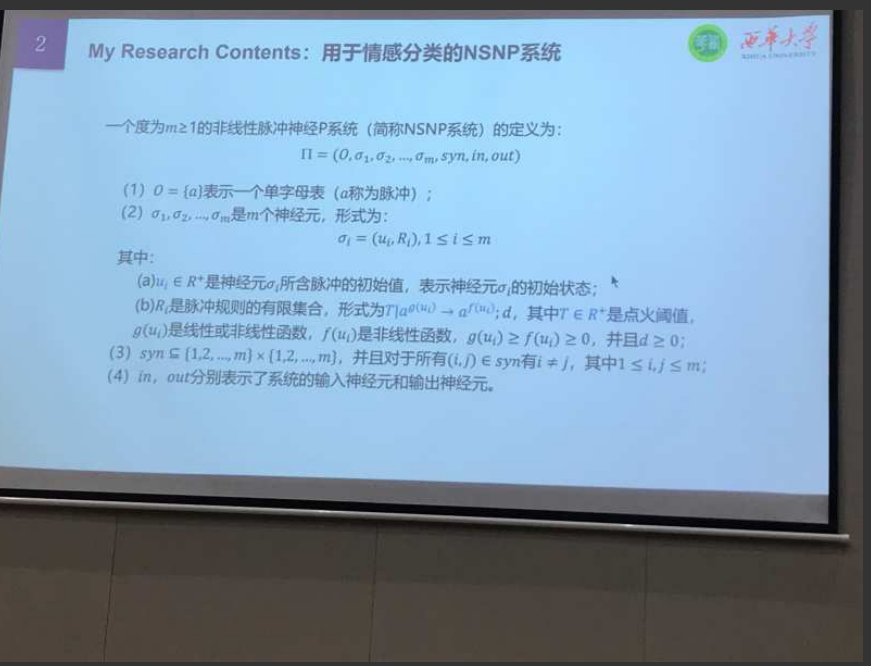
Peng et al. Nonlinear Spiking Neural P Systems, International Journal of Neural Systems, 2020.

非线性脉冲神经P系统（Nonlinear spiking neural P systems, NSNP systems）
（为脉冲神经P系统引入非线性机制构建了非线性脉冲神经P系统）

Motivaton：利用SNP systems的机制开发一个循环神经网络(RNN)模型

从NSNP systems出发构建这个RNN模型

RNN – Recurrent neural networks (循环神经网络)



彭宏教授　罗晓晖教授　西华大学

先荣康　袁艳萍

图像　NLP

“非线性脉冲P系统”

演化计算方法

（DE　PSO等）

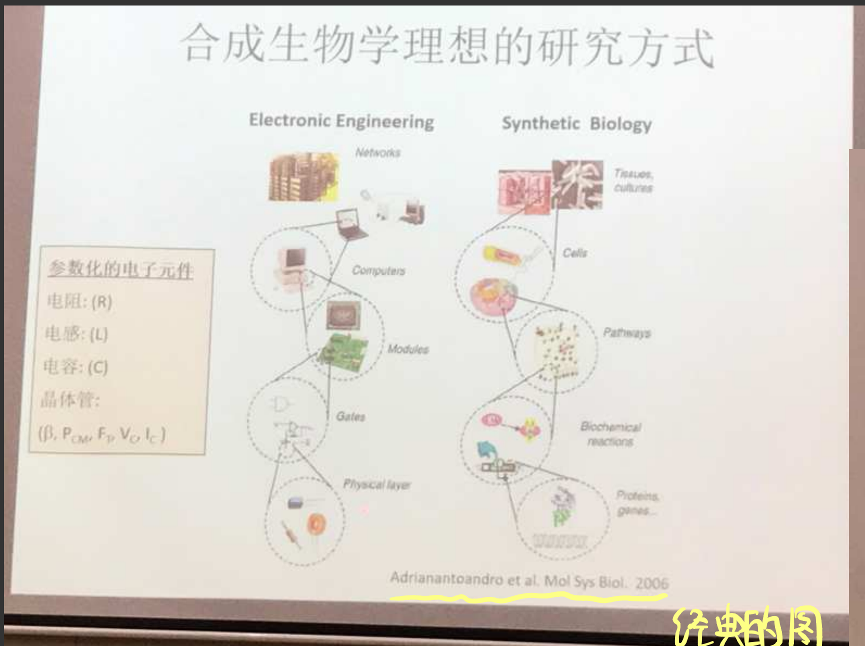
0825 娄春波 cb.lou@siat.ac.cn

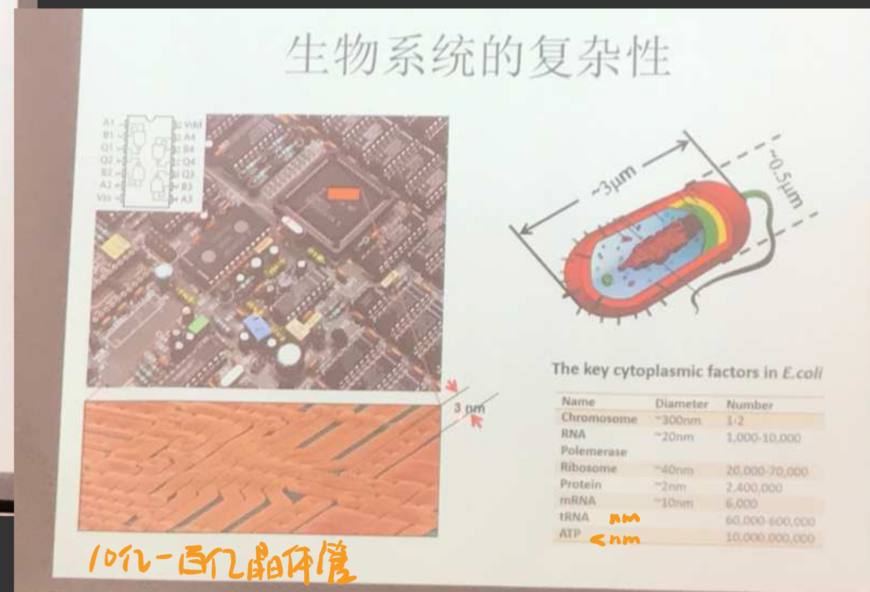
中科院深圳先进院合成所

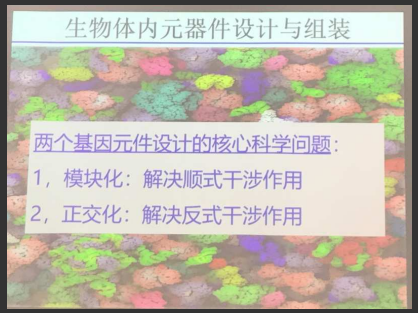
物理（北大本博 理论生物学中心 CQB）

很好的夏令营 { 院议、施一公　Bio2000
清华 暑期学校

设计
↕ 在电子领域可分开
制造





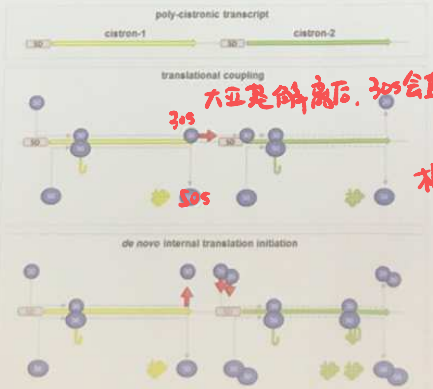


复杂的体系：——→ 解偶合、绝缘化 → 模块化

SD序列（RBS）

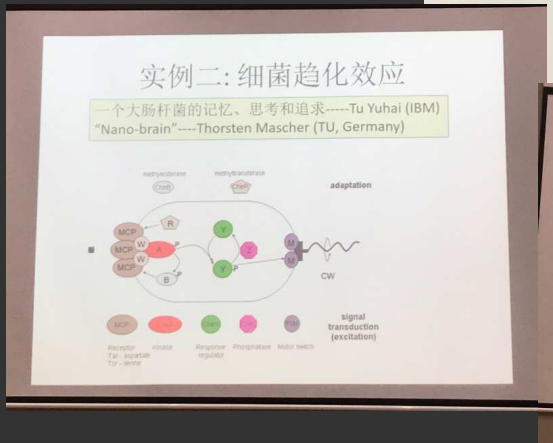
生物复杂性实例一：翻译偶联

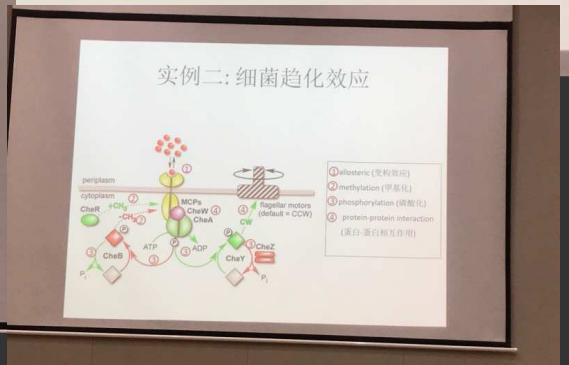
- Translational Coupling in prokaryotic operon



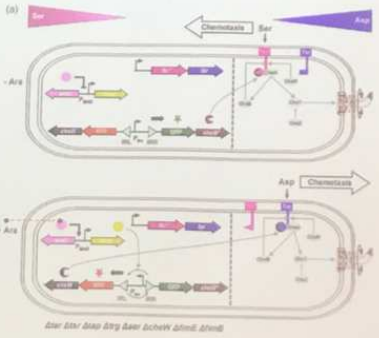
大亚是解离后，30S会直接结合下一个RBS

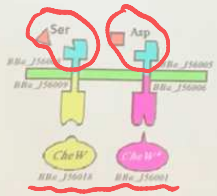
相偶联





2006年UCSF的iGEM项目





J. Mol. Biol. (2011), 406, 215-227

20年前的工作

工程化改造 CheW 的信号接收模块

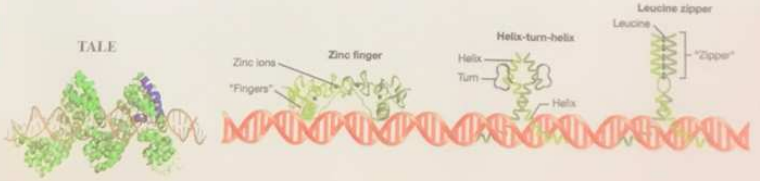
在生物中的 Sensor

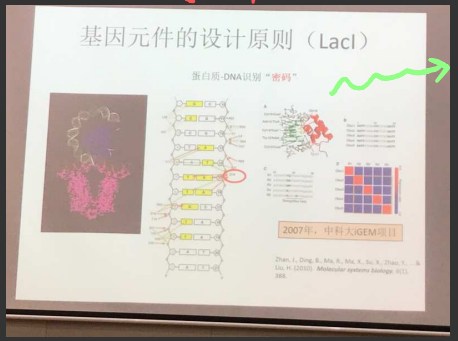
基因元件的设计原则

设计原则

- 简化原则（模块化）

（以转录因子(TF)与operators相互识别为例）

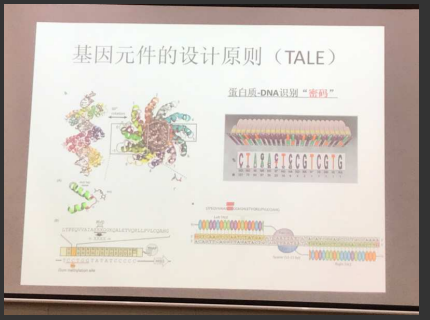




一个aa对应多个碱基

Zinc Finger

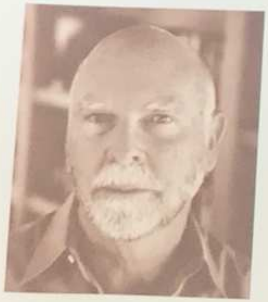
3个半碱基



一个aa ⟺ 一个碱基

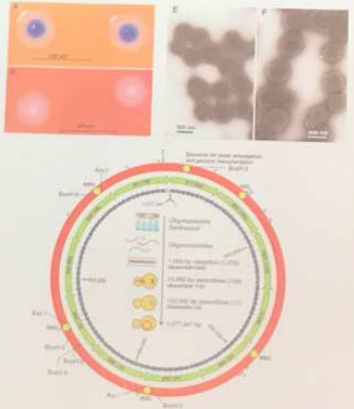
逐步模块化

合成生物学发展的关键问题



J. Craig Venter

科学家



四种基本密码子(A, T, C, G)

2008年 合成第一个基因组 2010年文章

14年采访：合成生物学最大的挑战？

不仅是DNA合成价格，更是在不清楚功能时做设计。 DOI: 10.1089

本课程的主要内容：

1，iGEM等工程化标准 "Tom Nedd" 计算机 "Jun Andy" 建筑

2，DNA组装技术

3，基因调控线路设计

4，次级代谢基因簇的设计

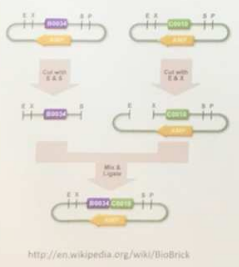
5，DNA折纸的设计

6，信号转导网络的设计

BioBricks-I

iGEM Synthetic Biology based on standard parts

Biobrick assembly standard



BioBrick standard assembly prefix and suffix sequences

Prefix 5′ GAATTC GCGGCCGC T TCTAGAG 3′
EcoRI NotI XbaI

Suffix: 5′ T ACTAGT A GCGGCCG CTGCAG 3′
SpeI NotI PstI

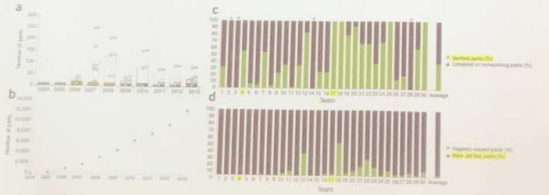
XbaI/SpeI scar

切割1后末端相同

http://en.wikipedia.org/wiki/BioBrick

Registry of Standard Biological Parts

http://parts.igem.org/Catalog



Cumulative number of iGEM parts in the registry

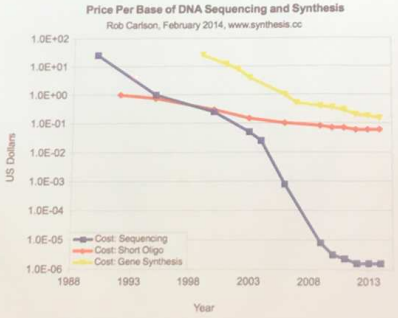
Nature Biotechnology **32**, 420–424 (2014) doi:10.1038/nbt.289

gBlock 是一个基因合成的公司

DNA组装技术与基因组合成

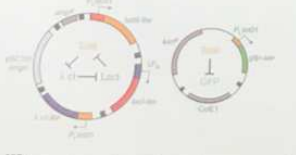
Method	Mechanism	PCR-free	Scar-free	parallelity	Size (bp)
BioBricks	RE	Yes	NO	single	10-10k
BglBricks	RE	Yes	NO	single	10-10k
Golden Gate	Type IIs RE	Yes	NO	multiple	10-10k
Infusion	Overlap	NO	Yes	multiple	10-10K
SLIC	Overlap	NO	Yes	multiple	1k-10K
USER	Overlap	NO	Yes	multiple	10-10K
Gibson	Overlap	NO	Yes	multiple	1-500K
RedET	Recombination	NO	Yes	single	1k-250k
Yeast TAR	Recombination	NO	Yes	multiple	1k-1M
OLMA	Type IIs RE /others	YES	YES	multiple	10-250K

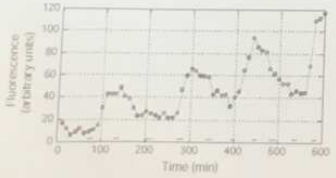
基因“读”和“写”的价格趋势



基因线路的标志性工作

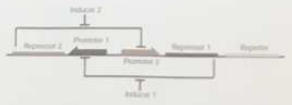
Repressilator（振荡器）

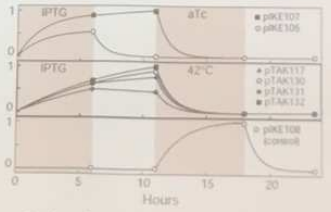




Elowitz & Leibler Nature (2000) 403:335-338

Toggle switch（双稳开关）

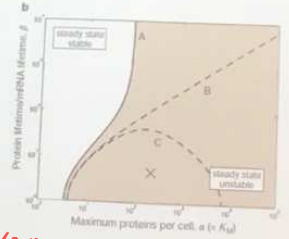




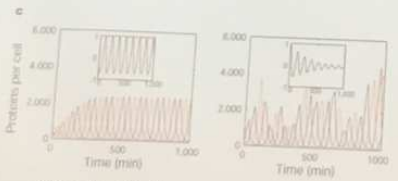
Gardner & Collins Nature (2000) 403:339-342

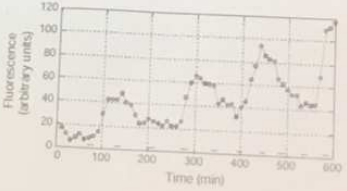
实例1：生物振荡器设计

Elowitz and Leibler, Nature, 2000, 403:335-338



降解快 SsrA

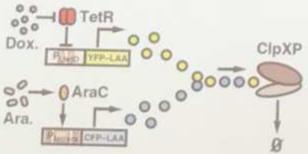




Elowitz & Leibler Nature (2000) 403:335-338

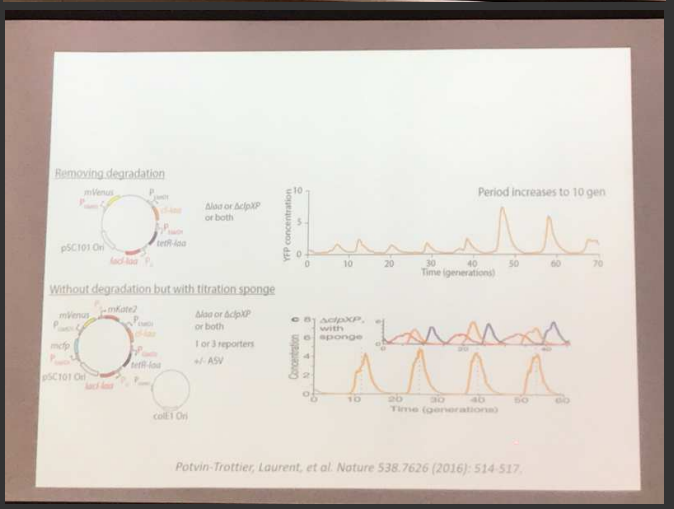
ssrA tag的解耦合与“未解耦合”

Correlated signaling through coupled degradation



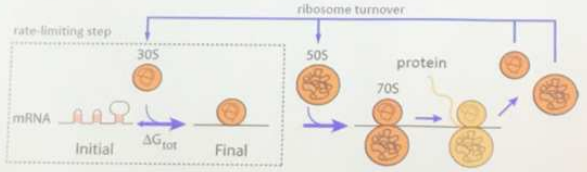
Cookson et al., Mol. Syst. Biol., 2011, 7:1-11

2016年改进



RBS 计算器的生物物理模型

Translation is a multi-step process
Translation initiation is often the rate-limiting step



$$m + R \leftrightarrow m::R$$

- m: 30S subunits
- R: RBS sequence on mRNA
- m::R: bound 30S subunits

$$\frac{m::R}{m\,R} = \exp\left(-\beta \Delta G_{tot}\right)$$

- β: Boltzmann constant
- ΔG_{tot}: Gibbs free energy change of ribosome binding

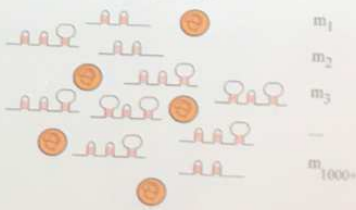
RBS 计算器的热力学模型（续）

$$m_i + R \leftrightarrow m_i::R$$

$$\frac{m_i R}{m_i::R} = \exp\left(\beta \Delta G_{tot}^{(i)}\right)$$

$$r_i \propto \frac{m_i \exp\left(-\beta \Delta G_{tot}^{(i)}\right)}{1 + \sum_j m_j \exp\left(-\beta \Delta G_{tot}^{(j)}\right)}$$

The "Ribosome Ensemble"
many mRNAs



$$r_{translation} \propto \exp\left(-\beta \Delta G_{tot}\right)$$

物理假设：能量最低状态是唯一状态

150+ Predictions Tested

RBS Calculator v1.1
tunable control of the translation initiation rate

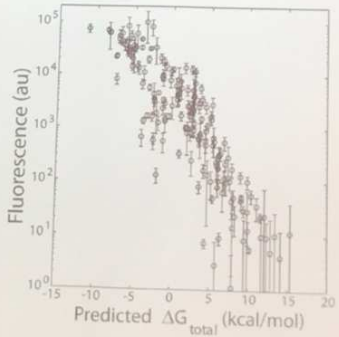
$$r_{translation} \propto \exp\left(-\beta \Delta G_{tot}\right)$$

Rational control of translation initiation over a 100,000-fold scale

$R^2 = 0.80$

$\beta^{-1} = 0.45 \pm 0.05$ mol/kcal

v1.1: improved thermodynamic model



Salis, Methods in Enzymology, 2011
Salis, Mirsky, & Voigt, *Nature Biotechnology*, 2009

正向设计的RBS元件更准确

Systematic metabolic pathway optimization requires three ingredients:

1. the ability to quantitatively control enzyme expression

We've developed a way to **control bacterial enzyme expression** across a 100,000-fold scale

RBS Calculator v1.1
rational control over the translation initiation rate

AGAGTACTACAGGAGCAAAAAGGCCCCAGAGGCGTATGCAGCGATCGACGGACGACTATCTACGAGCGAGCA

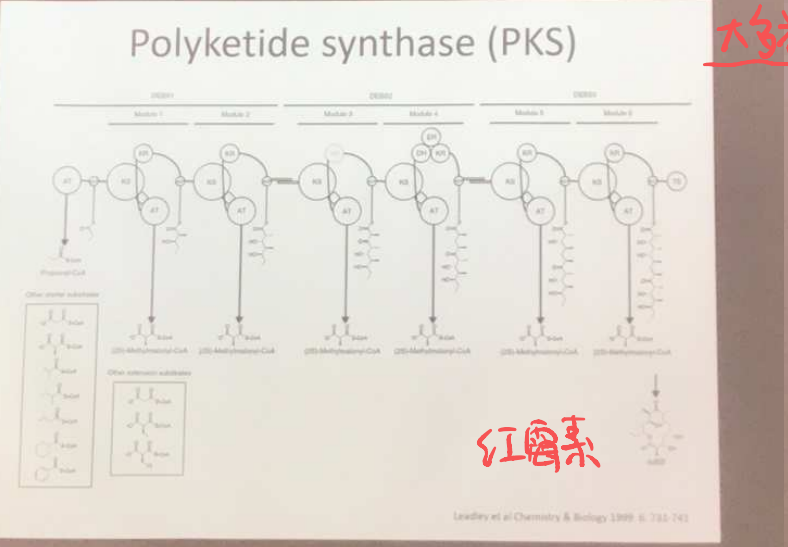
ribosome binding site — protein coding sequence

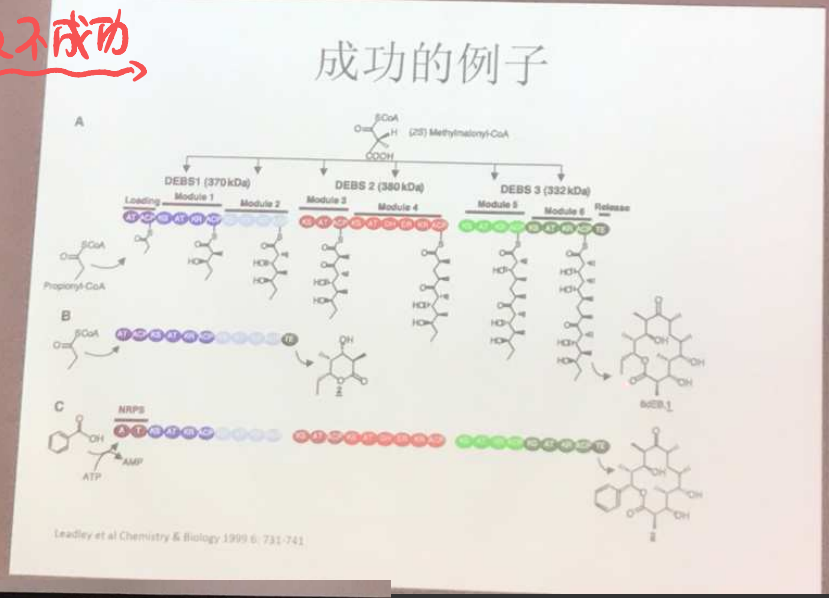
Thermodynamic Model

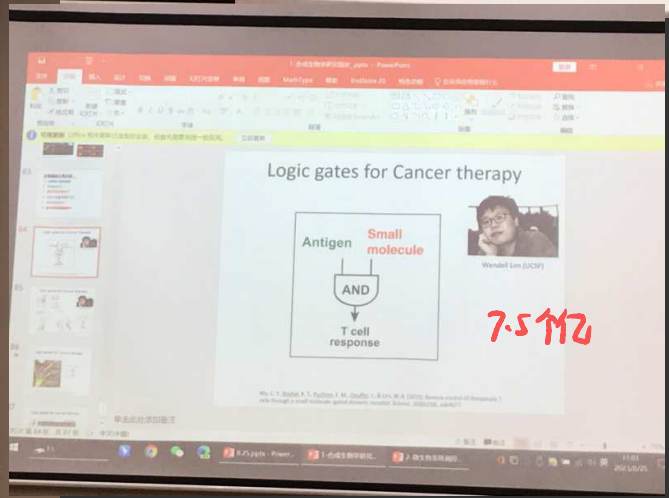
1064 au (proportional scale from 1 to 100,000)

Salis, Methods in Enzymology, 2011
Salis, Mirsky, & Voigt, *Nature Biotechnology*, 2009

为数不多的靠谱的预测软件



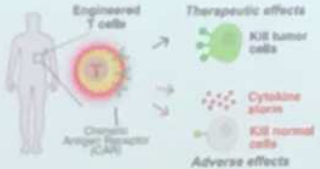




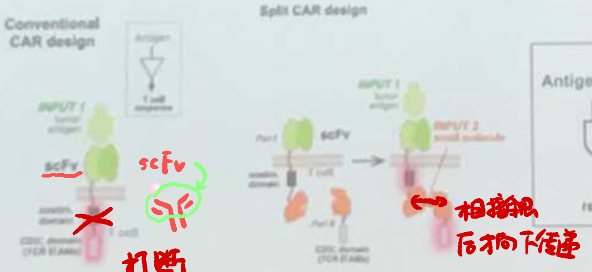
魏泽西

CarT 免疫风暴

Logic gates for Cancer therapy



Wendell Lim (UCSF)



scFv

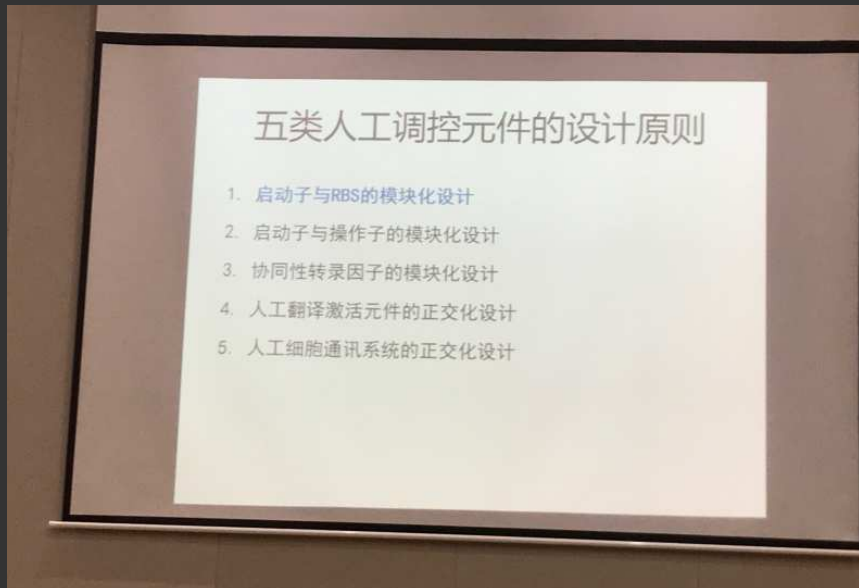
打断

相接触后才向下传递

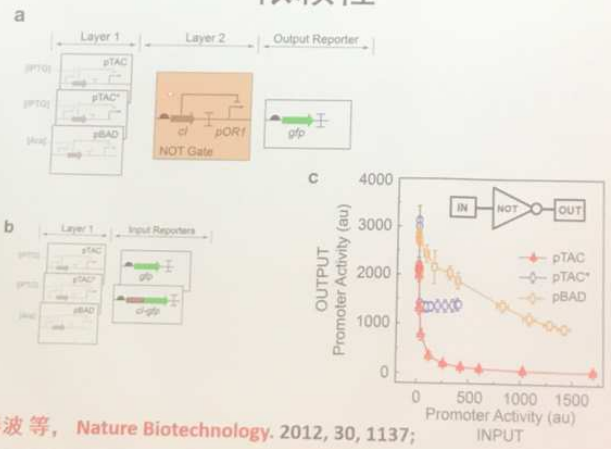
Wu, C. Y., Roybal, K. T., Puchner, E. M., Onuffer, J., & Lim, W. A. (2015). Remote control of therapeutic T cells through a small molecule-gated chimeric receptor. *Science*, *350*(6258), aab4077.

NEXT

到你自己的工作.

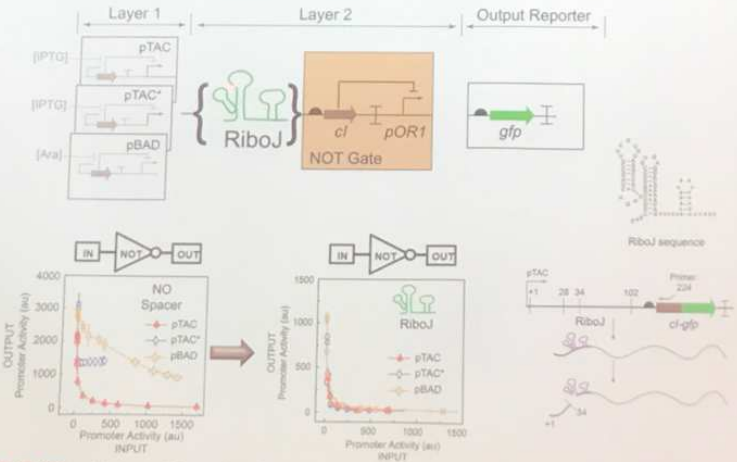


生物信号“转换函数”的周边序列依赖性



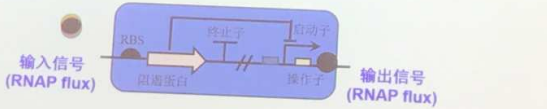
娄春波 等，Nature Biotechnology. 2012, 30, 1137;

Predictable NOT gate transfer function



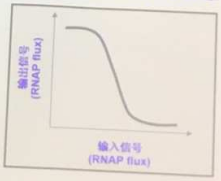
娄春波 等，Nature Biotechnology. 2012, 30, 1137;

➢ 唯一信号载体：RNA聚合酶“流”（RNAP flux）



- ✓ 可定量
- ✓ 可预测
- ✓ 可微调

（生物元件的模型化、参数化）



细胞重编程的前提条件

Nielsen, A. A., Der, B. S., Shin, J., ... & Voigt, C. A. (2016). Genetic circuit design automation. Science, 352(6281).

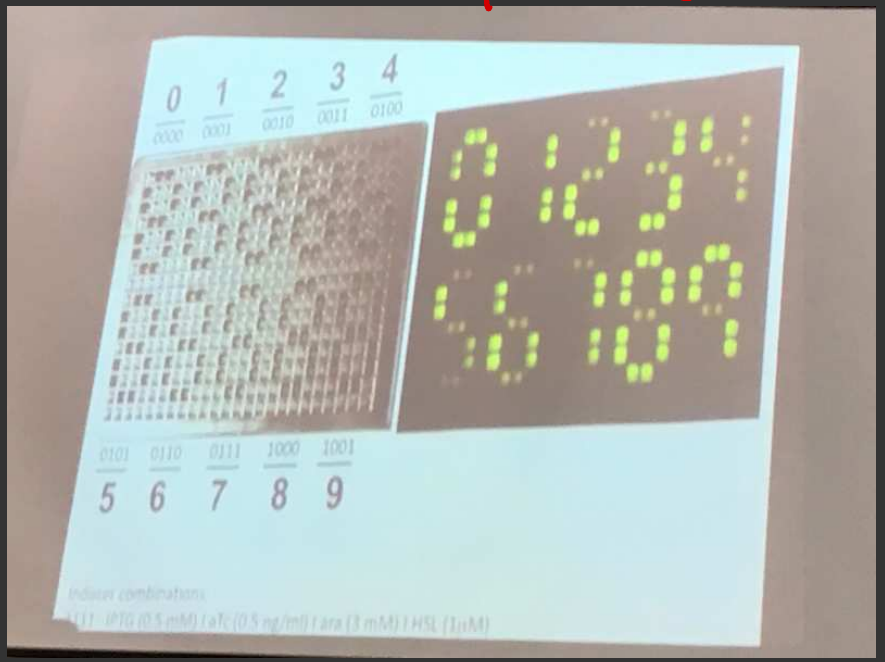
Voigt 等，Science. 2016, 352, 6281;

cello 公司

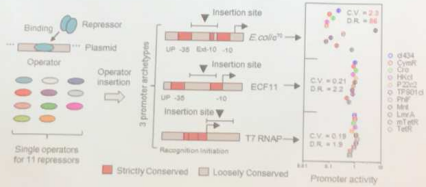
~$2.6亿
~¥10+亿
} 融资

7段式数字显示

Q：单cell or 多cell？

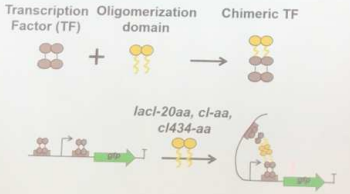


启动子"核心区"模块化



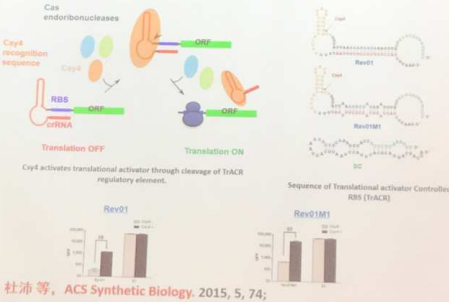
Zong, ..., Lou*, *Nature Communications*, 2017, 8, 8101

协同性转录因子的模块化设计



侯君然 等，*ACS Synthetic Biology*. 2017, 7, 1188;

基于Cas6 RNA内切酶的人工翻译激活元件

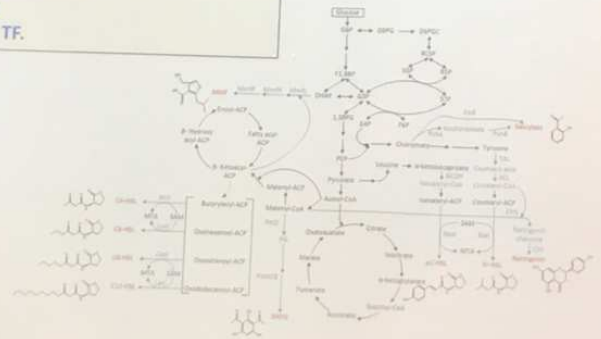


杜沛 等，*ACS Synthetic Biology*. 2015, 5, 74;

更多正交型细胞间通讯系统

设计原则:

1. diffusible molecules across membrane
2. Synthesized from CCM
3. <5 enzymes
4. sensed by a TF.



元器件设计原则：正交化

- 已完成的元器件：
  细胞通讯系统（10对）；Csy4翻译激活元件（3对）；T7-family RNAP（6个）；LacI阻遏元件（5对）；TALE（26个）；TetR家族阻遏元件（20对）；ECF sigma激活元件（20对）；sRNA（14对）；toehold-RNA（26对）；CRISPRi（任意多），NF-κB（6对）等等

- 尚待完成的元器件：
  原核调控元件：双组分调控元件；磷酸化修饰元件；翻译机器元件；蛋白降解元件；DNA复制调控元件；感应蛋白元件；能量供应元件；RNA结合蛋白；RNA降解元件；。。。

  真核调控元件（几乎是空白！！）：启动子；激活因子；增强因子；泛素化修饰因子；磷酸化修饰元件；蛋白降解元件；RNA剪切元件；RNA结合元件；。。。

Ribo profiling

翻译速率 ∽ 表达量

复合体中的蛋白比例固定

理想的务实主义者.

Science partner 期刊 ←——

23082班 宋理宸 DNA信息存储技术原理和应用

基因存储很厉害的单位

DNA存储优势：处理并行度高，常温保存、功耗低、时间久

电磁特性不敏感 → 数据安全性高

三维存储、信息密度好.

DNA信息存储的里程碑工作

Church *et al.* 2012
Harvard University

AT - 0
GC - 1

由于没有采用纠错码，解码数据发现少量的错误

问题：
- 连续碱基
- 极端GC含量序列
- 特殊结构序列


BREVIA

Next-Generation Digital Information Storage in DNA

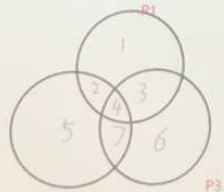
George M. Church,[1,2] Yuan Gao,[1] Sriram Kosuri[1,2,*]

首次商业存储、读取（几百kb）

冗余纠错基本原理

至少3个copy才能纠错

只能检错



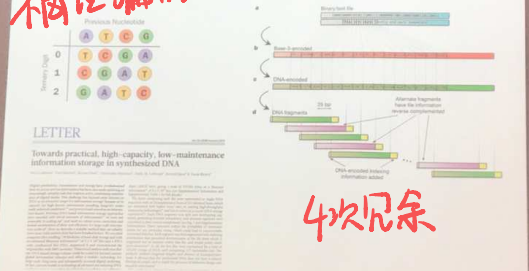
汉明码

P1组：1,2,3,4
P2组：2,4,5,7
P3组：3,4,6,7

不检错，也可以纠错

DNA信息存储研究进展介绍 2

不同位编码冗余

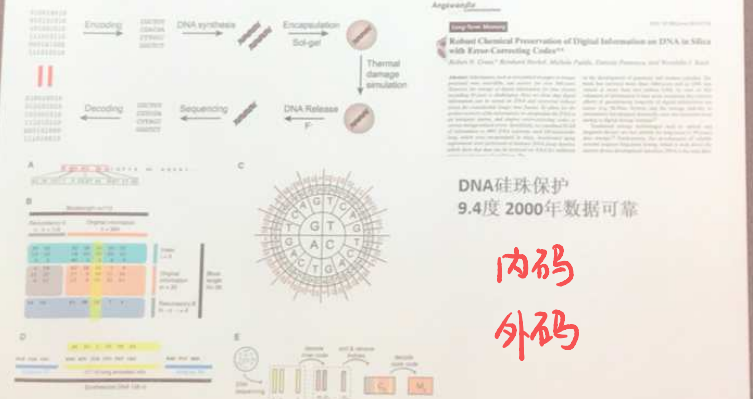

LETTER

Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

Goldman et al. 2013
European Bioinformatics Institute

4次冗余

DNA信息存储研究进展介绍 3



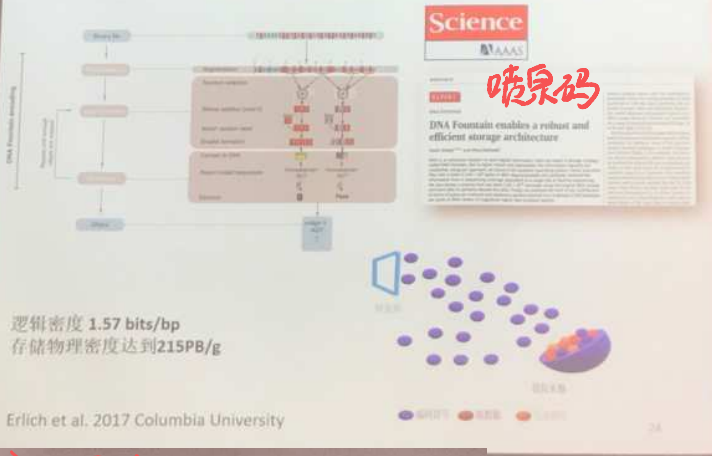
DNA硅珠保护
9.4度 2000年数据可靠

内码

外码

Grass et al. 2015 ETH Zurich

DNA信息存储研究进展介绍 4



喷泉码

逻辑密度 1.57 bits/bp
存储物理密度达到215PB/g

Erlich et al. 2017 Columbia University

DNA信息存储研究进展介绍 5

200 MB 数据规模随机访问

ARTICLES

Random access in large-scale DNA data storage

数据规模↑

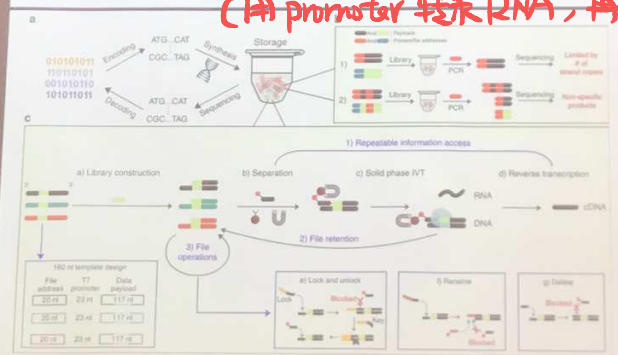
随机访问(引物)

Organick et al. 2018 Nature Biotechnology

测序时不消耗原DNA

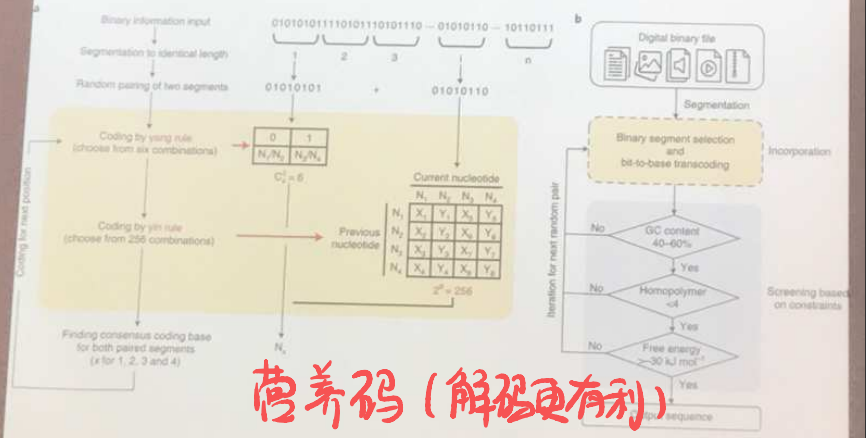
(用 promoter 转录 RNA，再cDNA后测序)

DNA信息存储研究进展介绍 6



Lin, K.N., Volkel, K., Tuck, J.M. *et al.* Dynamic and scalable DNA-based information storage. *Nat Commun* **11**, 2981 (2020). https://doi.org/10.1038/s41467-020-16797-2

DNA信息存储研究进展介绍 7



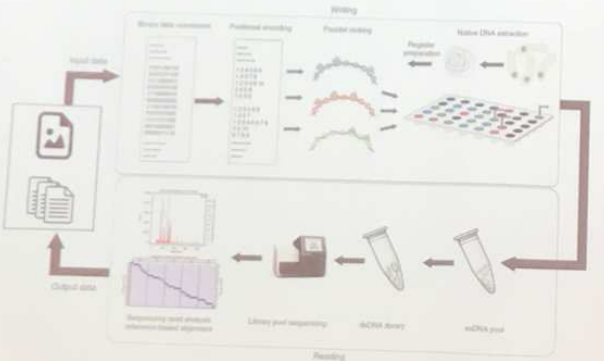
简并码（解码更有利）

Ping, Z., Chen, S., Zhou, G. *et al.* Towards practical and robust DNA-based data archiving using the yin–yang codec system. *Nat Comput Sci* 2, 234–242 (2022). https://doi.org/10.1038/s43588-022-00231-2

顺序码 解码纠错研有块.

DNA信息存储研究进展介绍 8

DNA打孔卡技术

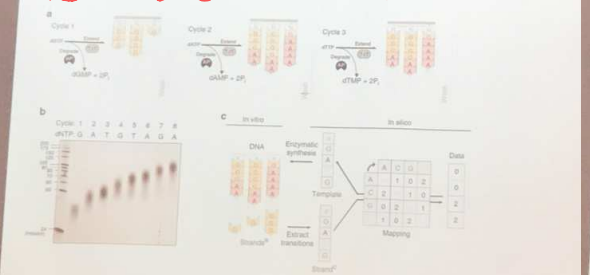


Tabatabaei, S. Kasra et al. (2020): DNA punch cards for storing data on native DNA sequences via enzymatic nicking. *Nature communications* 11 (1), p. 1742.

DNA信息存储研究进展介绍 9

非阻断型TdT酶法DNA合成的DNA信息存储

不好说



Church* et al. (2019): Terminator-free template-independent enzymatic DNA synthesis for digital information storage. In *Nature communications* 10 (1), p. 2383. DOI: 10.1038/s41467-019-10258-1.

DNA信息存储研究进展介绍 10

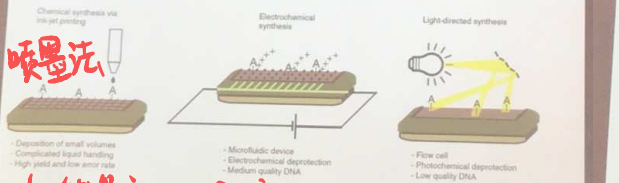
高错误率的DNA合成技术
用于DNA信息存储

错误率高 通量低

喷墨法

电化学

光控

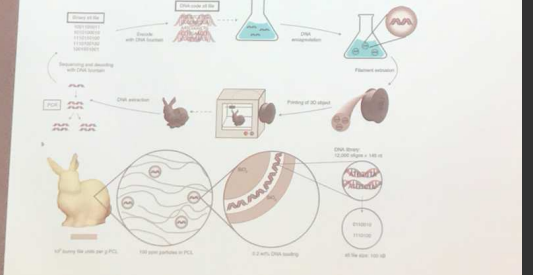


↳ 错误率低、成本高

Grass, Robert N*. et al(2020): Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nature communications* 11 (1), p. 5345.

下一代 DNA存储

DNA信息存储研究进展介绍 11

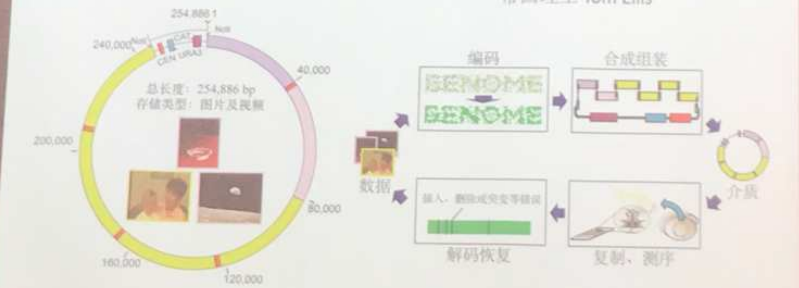


A DNA-of-things storage architecture to create materials with embedded memory

DNA信息存储研究进展介绍 12

“人工染色体数据存储的关键概念验证”

--帝国理工 Tom Ellis



提出并验证了“人工染色体如何以一种稳定且廉价复制的方式用于数据存储

帝国理工 Tom Ellis “Self-replicating digital data storage with synthetic chromosomes” (FEMS Yeast Rev 2021; 8: nwab086)

An artificial chromosome for data storage

Wenjang Chen[1,2,†], Mingzhe Han[1,2,†], Jianting Zhou[1,2,†], Qi Ge[1], Panpan Wang[1], Xinchen Zhang[1,2], Siyu Zhu[1,2], Lifu Song[1,2] and Yingjin Yuan[1,2,*]

细胞膜阻碍

质写

进展介绍13

清华大学

镜像DNA存储.

天然中无降解酶 更稳定

DNA链式存储信道的错误类型重新定义

Original DNA sequence

G-T-A-A-G-A-C-G-T-C-T

Type 1. Substitution A to G

S1-G-T-A-A-G-G-C-G-T-C-T

Type 2. Indel

insertion

S2-G-T-A-A-T-G-A-C-G-T-C-T

S3-G-T-A-A-A-C-G-T-C-T

G deletion

Type 3. DNA breaks

S4-G-T-A-A G-A-C-G-T-C-T

S5-G-T-A-A-G-A-C-G T-C-T

S6-G-T-A-A-G-A C-G-T-C-T

Type 4. DNA rearrangements

S7-G-T-A-A←T-C-T

S8-G-T-A-A-G-A-C-G←C-G-T-C-T

S9-G-T-A-A-G-A←G-A-C-G-T-C-T

DNA存储中可能的错误

断裂和重排是DNA链式存储最基本的错误类型

插入删除和替换错误也是特殊的“断裂重排”错误

DNA信息存储信道的多拷贝特征

写入—DNA合成

复制—PCR扩增

读取—测序读取

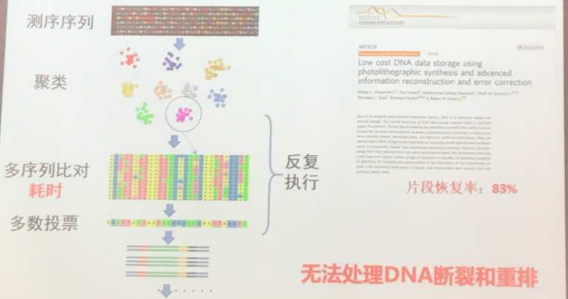
多拷贝不可避免

多拷贝特性是关键！

百万年猛犸象基因组

有大量尝试将拷贝数将为1的努力

传统基于聚类-多序列比对的内码解码过程



片段恢复率：83%

无法处理DNA断裂和重排

极度耗时

De Bruijn Graph 德布莱英图的基本原理

sequence ATGGAAGTCGCGGAATC

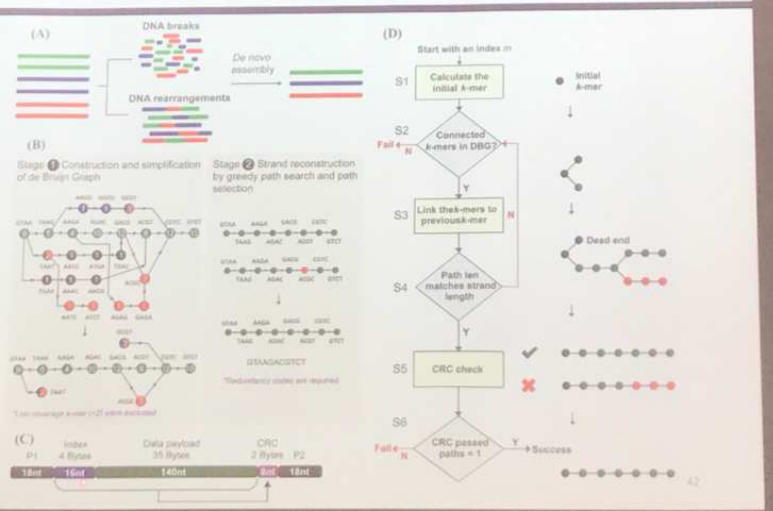
7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph

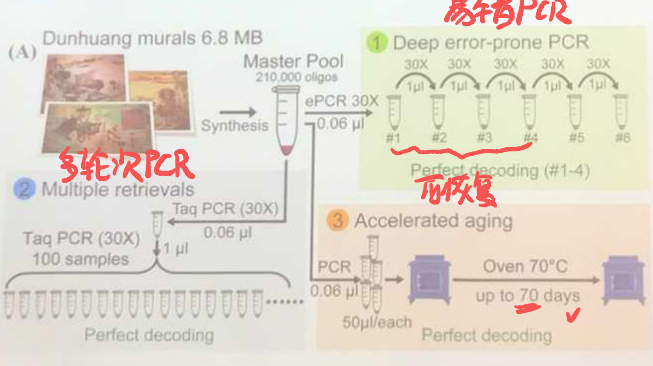
ATGGAAG → TGGAAGT → GGAAGTC → GAAGTCG → AAGTCGC → AGTCGCG → GTCGCGG → TCGCGGA → CGCGGAA → GCGGAAT → CGGAATC

基于图论的组装内码算法



DNA信息存储的数据鲁棒技术

DNA信息存储的数据鲁棒技术



NC文章. 不保护可2W年 DNA断裂可恢复

1）基于DNA链置换的高精度随机访问技术

普通引物

5'端区分碱基

DNA片段池

PCR选择扩增

PCR产物完全无法区分

引物不结合时不影响PCR

链置换引物

DNA片段池

增强引物

磁珠

目标DNA片段

主要步骤1 链置换选择

5‘端选择性强

磁珠

主要步骤2 PCR选择

3‘端选择性强

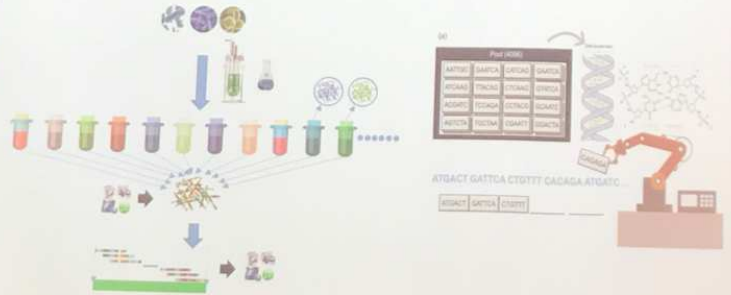
高纯度PCR产物

CTCAACTTTCC

CTCAACTTTCC

双碱基条件选择性提高90%以上

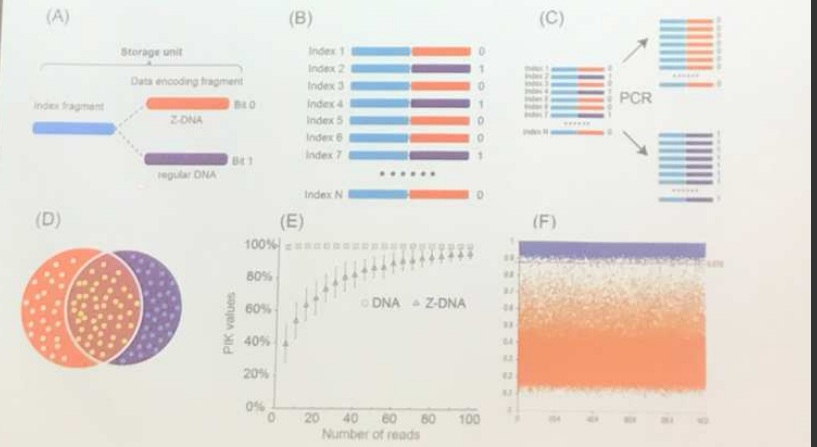
2）“DNA活字存储”写入和编码技术



Engineering DNA materials for sustainable data storage by a DNA movable-type system

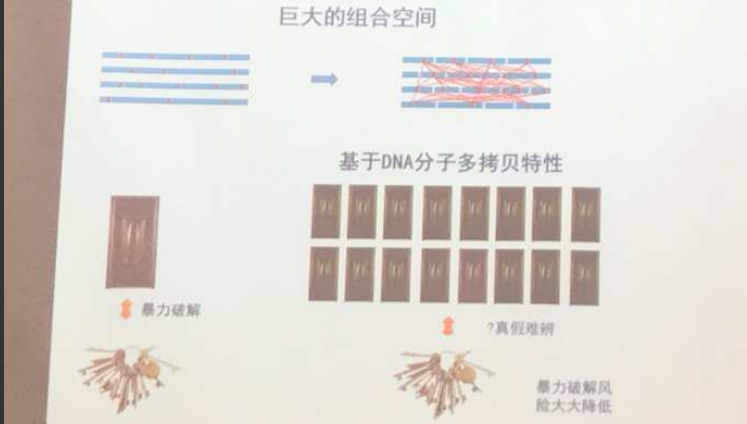
Engineering 工程院院刊 Accepted

3）基于Z-DNA的密钥存储技术



Z-DNA与普通DNA结构相同 PCR不可避免污染

4）“影子保护”技术



5）数据安全：“影子保护”技术