Generative modeling of aptamers from Directed Evolution experiments

Stefano Crotti, Jorge Fernandez-de-Cossio-Diaz

Institut de Physique Théorique, CEA, CNRS, Université Paris-Saclay, France

Outline

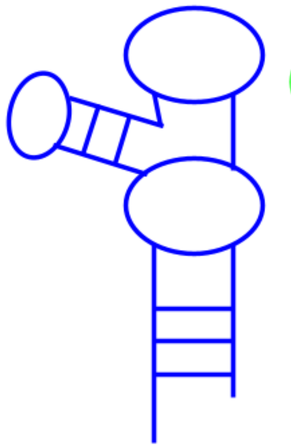
- Goal: **aptamer design**
- Experimental procedure: **Directed Evolution** (SELEX)
- Proposition: **complement** with **mathematical model**
- Results: lab validation
- **Challenges**: modeling, computation, noise

Introduction

Aptamers as ligands

Aptamer:

- Synthetic RNA/DNA strand (~100 bp)
- Used to bind a target (protein, virus, ...)



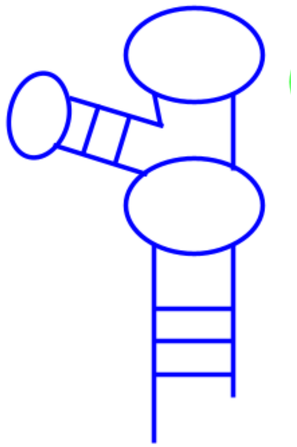
Aptamers as ligands

Aptamer:

- Synthetic RNA/DNA strand (~100 bp)
- Used to bind a target (protein, virus, ...)

Goal:

- Given target molecule, find good binder
- Multiple targets, different conditions, ...

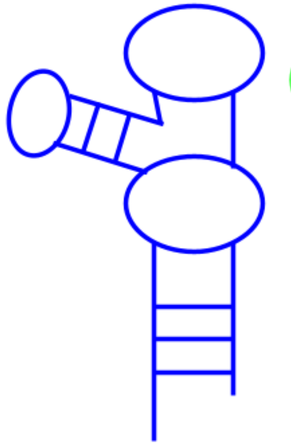


TARGET

Aptamers as ligands

Aptamer:

- Synthetic RNA/DNA strand (~100 bp)
- Used to bind a target (protein, virus, ...)



TARGET

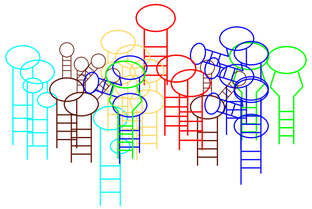
Goal:

- Given target molecule, find good binder
- Multiple targets, different conditions, ...

Applications:

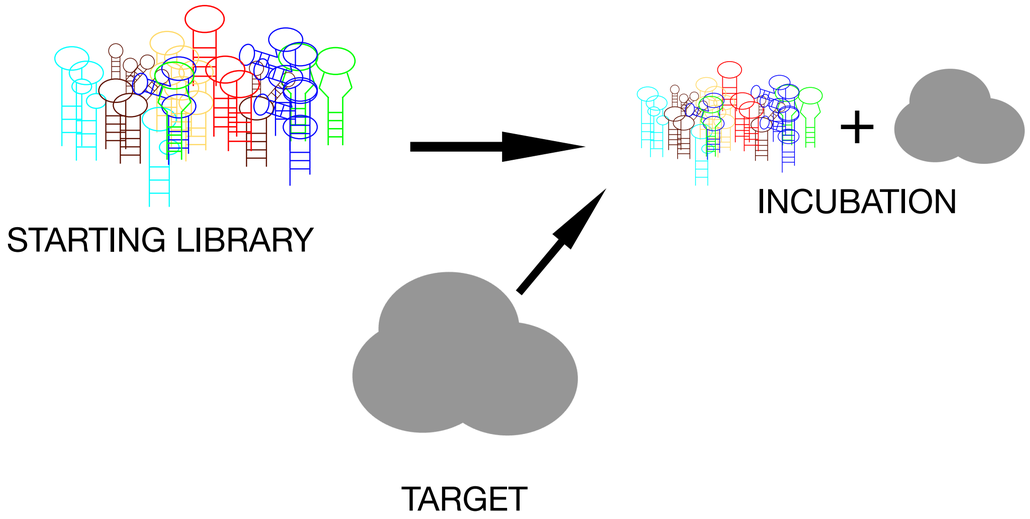
- Therapeutics, diagnostics, separation of compounds, ...

Directed Evolution Experiments (SELEX)

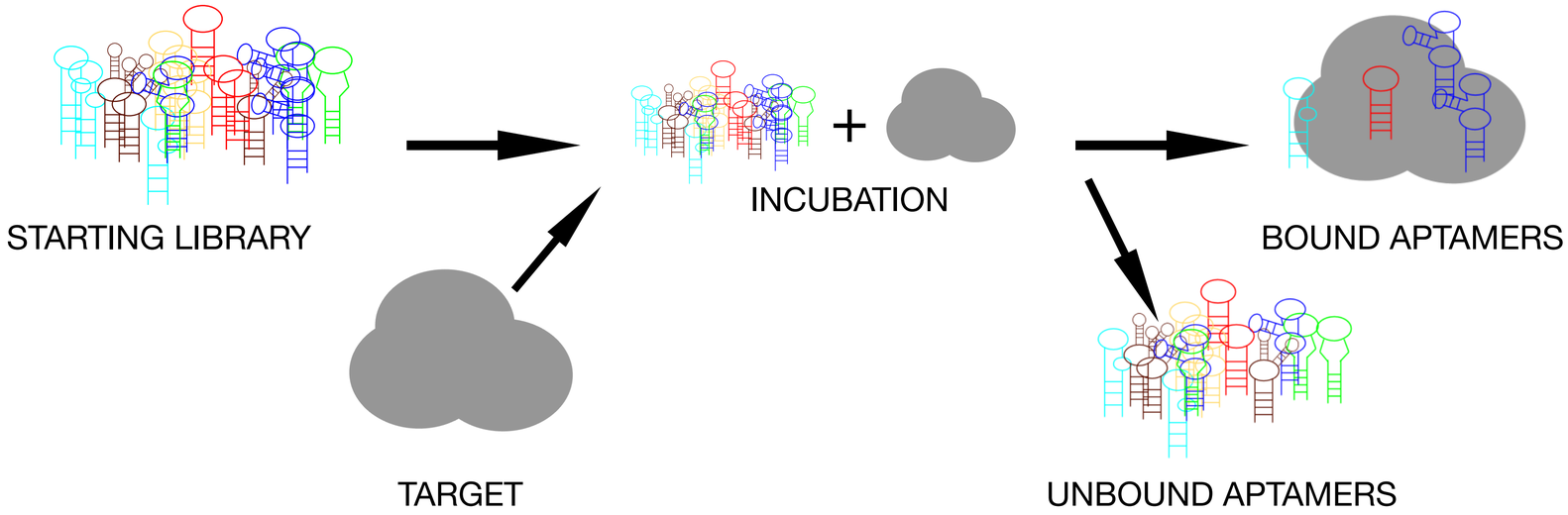


STARTING LIBRARY

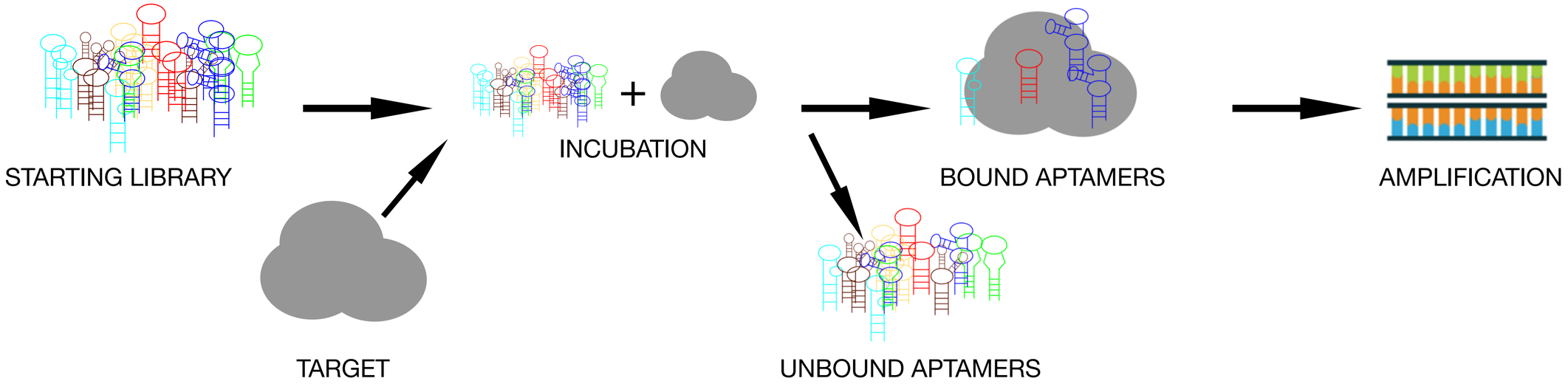
Directed Evolution Experiments (SELEX)



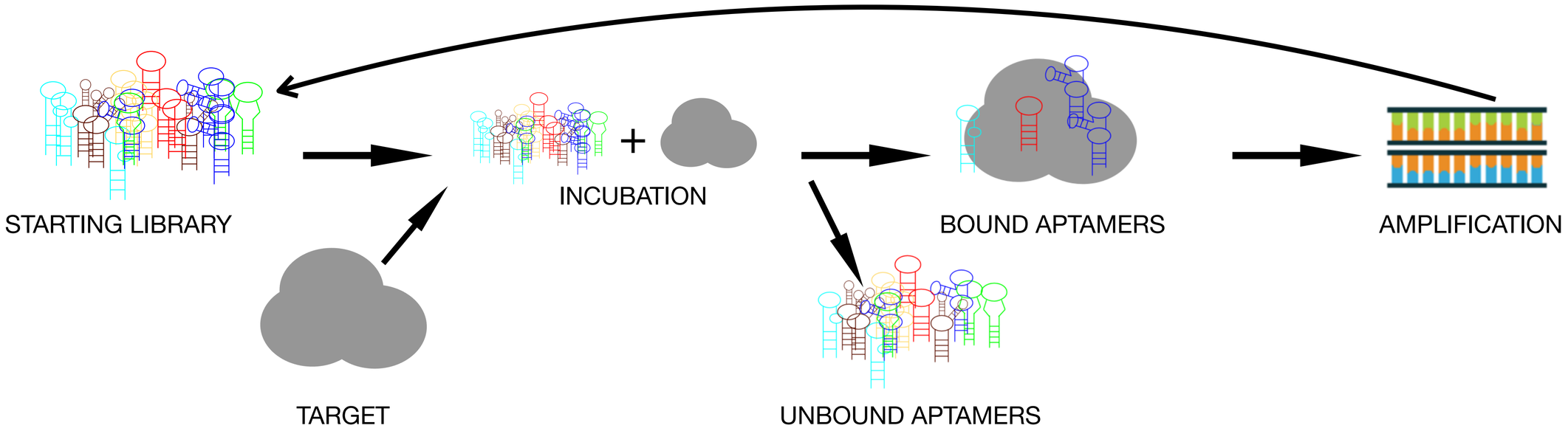
Directed Evolution Experiments (SELEX)



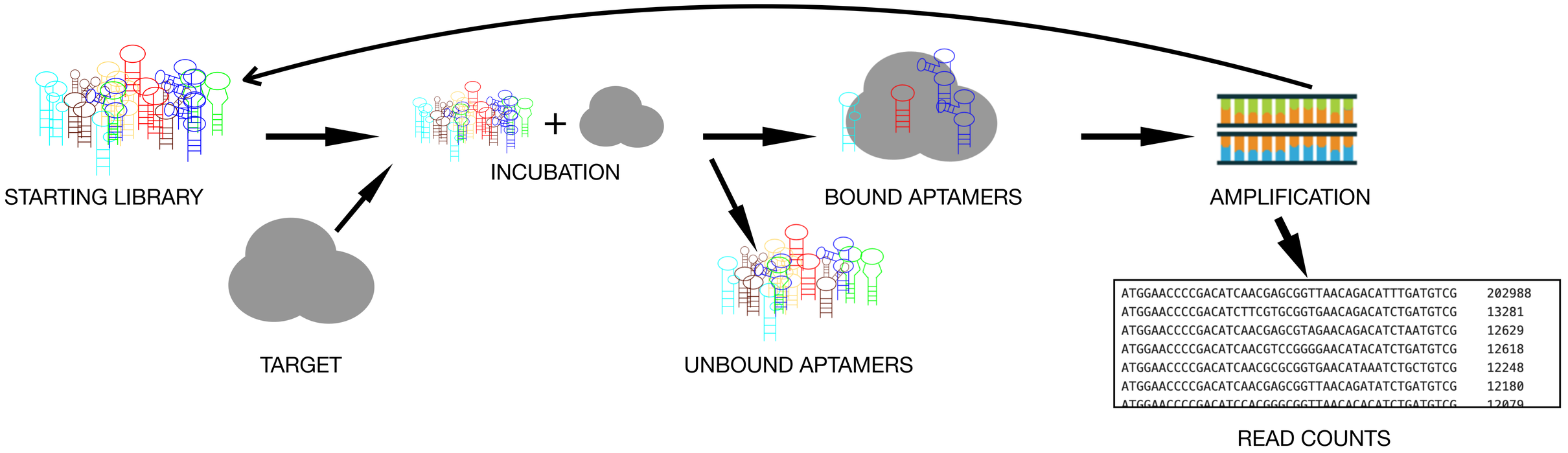
Directed Evolution Experiments (SELEX)



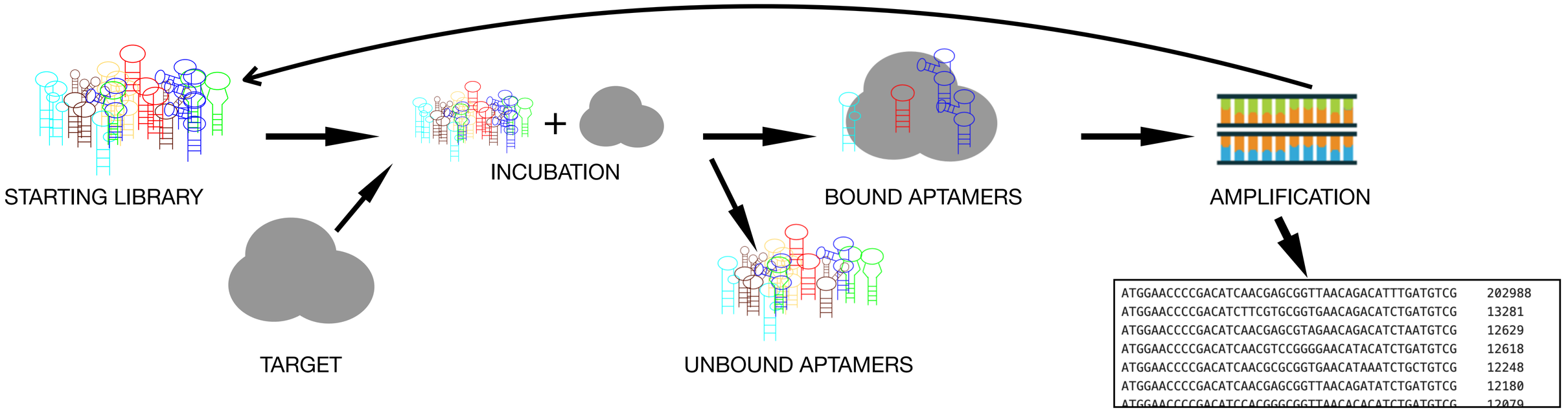
Directed Evolution Experiments (SELEX)



Directed Evolution Experiments (SELEX)



Directed Evolution Experiments (SELEX)

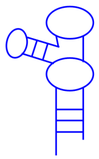


Limitations:

- Only explores a **subset** of the **combinatorially large** sequence **space**
- **No insight** about binding mechanism
- Hard for **multiple targets**

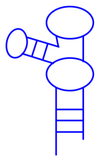
Mathematical model

Mathematical modeling

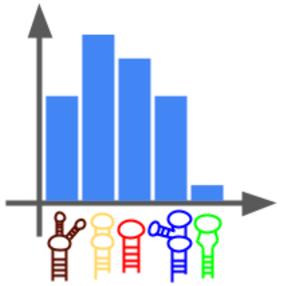
Work at the level of **sequences**:  $\leftrightarrow s \in \{A, C, G, U\}^L$

Otwinowski, *Molecular biology and evolution* 35.10 (2018)
Fernandez-de-Cossio-Diaz et al, *PLoS Computational Biology* 20.10 (2024)

Mathematical modeling

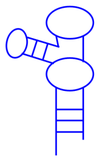
Work at the level of **sequences**:  $\leftrightarrow s \in \{A, C, G, U\}^L$

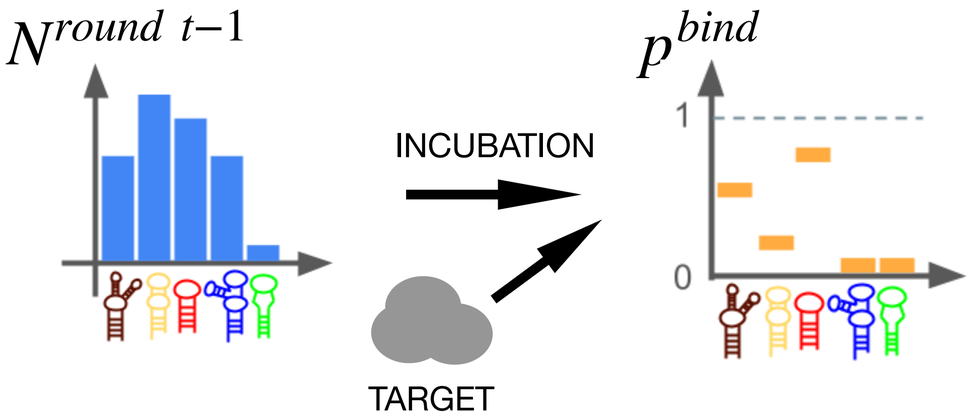
$N^{round\ t-1}$



Otwinowski, *Molecular biology and evolution* 35.10 (2018)
Fernandez-de-Cossio-Diaz et al, *PLoS Computational Biology* 20.10 (2024)

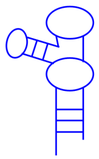
Mathematical modeling

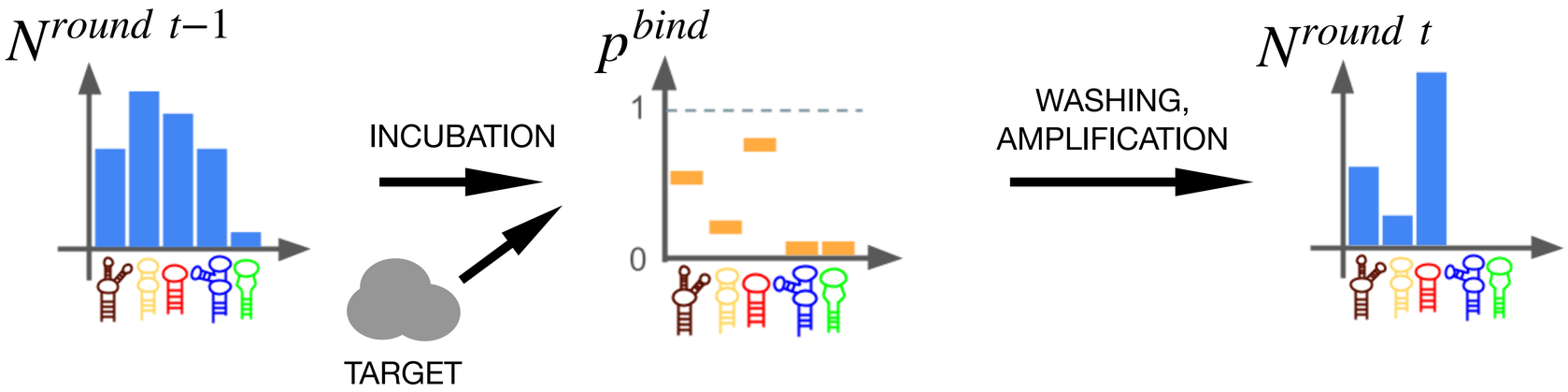
Work at the level of **sequences**:  $\leftrightarrow s \in \{A, C, G, U\}^L$



Otwinowski, *Molecular biology and evolution* 35.10 (2018)
Fernandez-de-Cossio-Diaz et al, *PLoS Computational Biology* 20.10 (2024)

Mathematical modeling

Work at the level of **sequences**:  $\leftrightarrow s \in \{A, C, G, U\}^L$

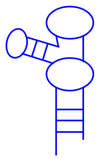


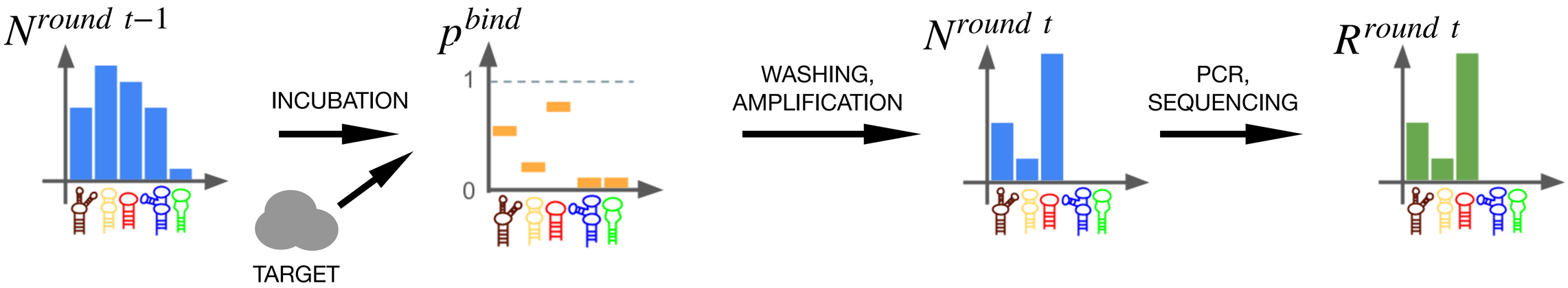
Otwinowski, *Molecular biology and evolution* 35.10 (2018)
Fernandez-de-Cossio-Diaz et al, *PLoS Computational Biology* 20.10 (2024)

Mathematical modeling

N: abundances (unobserved)

R: read counts (data)

Work at the level of **sequences**:  $\leftrightarrow s \in \{A, C, G, U\}^L$

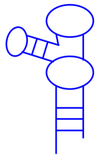


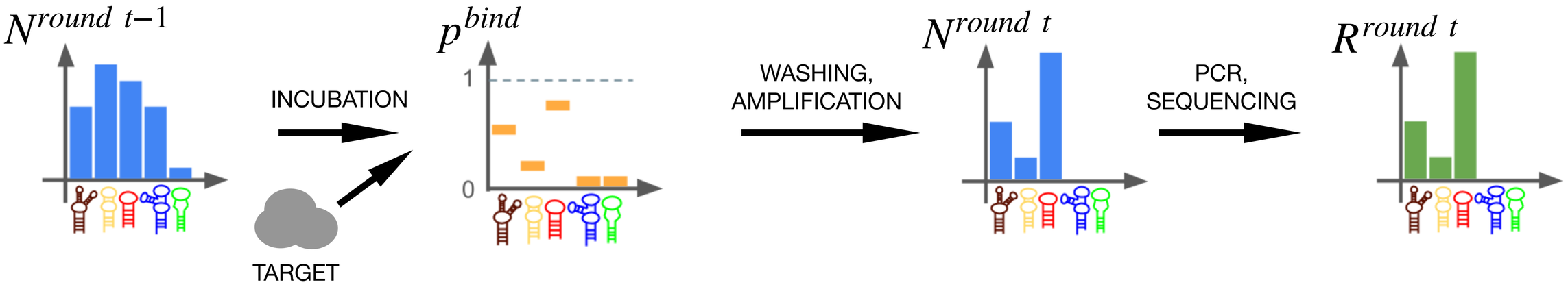
Otwinowski, *Molecular biology and evolution* 35.10 (2018)
Fernandez-de-Cossio-Diaz et al, *PLoS Computational Biology* 20.10 (2024)

Mathematical modeling

N: abundances (unobserved)

R: read counts (data)

Work at the level of **sequences**:  $\leftrightarrow s \in \{A, C, G, U\}^L$



$$\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right) \qquad \mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$$

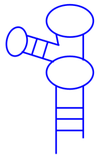
$$N^{round\ 0}(s) \qquad p^{bind}(s)$$

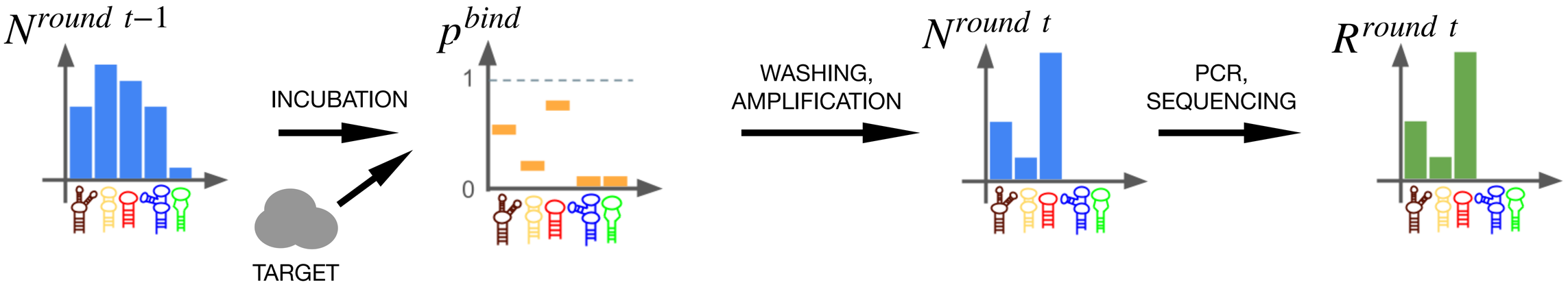
Otwinowski, *Molecular biology and evolution* 35.10 (2018)
Fernandez-de-Cossio-Diaz et al, *PLoS Computational Biology* 20.10 (2024)

Mathematical modeling

N: abundances (unobserved)

R: read counts (data)

Work at the level of **sequences**:  $\leftrightarrow s \in \{A, C, G, U\}^L$



$$\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right) \qquad \mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$$

Biology-informed

E.g. Poisson, Negative Binomial

$$N^{round\ 0}(s) \qquad p^{bind}(s)$$

Learned from data

E.g. Independent-sites, Pairwise interactions, Neural Network

Otwinowski, *Molecular biology and evolution* 35.10 (2018)
Fernandez-de-Cossio-Diaz et al, *PLoS Computational Biology* 20.10 (2024)

Generation of novel sequences

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$
2. Parametrize: $N_{\phi}^{round\ 0}(s),\ p_{\theta}^{bind}(s)$

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$
2. Parametrize: $N_{\phi}^{round\ 0}(s),\ p_{\theta}^{bind}(s)$
3. If multiple targets: $p_{\theta}^{bind,round\ t}(s)$

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$
2. Parametrize: $N_{\phi}^{round\ 0}(s),\ p_{\theta}^{bind}(s)$
3. If multiple targets: $p_{\theta}^{bind,round\ t}(s)$
4. Likelihood $\mathscr{L}$: probability of data given parameters

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$
2. Parametrize: $N_{\phi}^{round\ 0}(s),\ p_{\theta}^{bind}(s)$
3. If multiple targets: $p_{\theta}^{bind,round\ t}(s)$
4. Likelihood $\mathscr{L}$: probability of data given parameters
5. Maximize: $N_{\phi^*}^{round\ 0}(s),\ p_{\theta^*}^{bind}(s) = \max_{\phi,\theta} \mathscr{L}\left(N_{\phi}^{round\ 0}(s), p_{\theta}^{bind}(s)\right)$

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$
2. Parametrize: $N_{\phi}^{round\ 0}(s),\ p_{\theta}^{bind}(s)$
3. If multiple targets: $p_{\theta}^{bind,round\ t}(s)$
4. Likelihood $\mathscr{L}$: probability of data given parameters
5. Maximize: $N_{\phi^*}^{round\ 0}(s),\ p_{\theta^*}^{bind}(s) = \max_{\phi,\theta} \mathscr{L}\left(N_{\phi}^{round\ 0}(s), p_{\theta}^{bind}(s)\right)$
6. Look for sequences s for which $p_{\theta^*}^{bind}(s)$ is large

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$
2. Parametrize: $N_{\phi}^{round\ 0}(s),\ p_{\theta}^{bind}(s)$
3. If multiple targets: $p_{\theta}^{bind,round\ t}(s)$
4. Likelihood $\mathscr{L}$: probability of data given parameters
5. Maximize: $N_{\phi^*}^{round\ 0}(s),\ p_{\theta^*}^{bind}(s) = \max_{\phi,\theta} \mathscr{L}\left(N_{\phi}^{round\ 0}(s), p_{\theta}^{bind}(s)\right)$
6. Look for sequences s for which $p_{\theta^*}^{bind}(s)$ is large
7. Send them for lab validation

Generation of novel sequences

1. Pick $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$, $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$
2. Parametrize: $N_{\phi}^{round\ 0}(s),\ p_{\theta}^{bind}(s)$
3. If multiple targets: $p_{\theta}^{bind,round\ t}(s)$
4. Likelihood $\mathscr{L}$: probability of data given parameters
5. Maximize: $N_{\phi^*}^{round\ 0}(s),\ p_{\theta^*}^{bind}(s) = \max_{\phi,\theta} \mathscr{L}\left(N_{\phi}^{round\ 0}(s), p_{\theta}^{bind}(s)\right)$
6. Look for sequences s for which $p_{\theta^*}^{bind}(s)$ is large
7. Send them for lab validation

Challenges

Challenges

Modeling choices

Parametrization of initial library and selection: accuracy vs robustness.

- **Selection** mechanism $\mathbb{P}\left(N^{round\ t} \mid N^{round\ t-1}, p^{bind}\right)$: concentration around mean?
- **Noise** on **read counts** $\mathbb{P}\left(R^{round\ t} \mid N^{round\ t}\right)$: PCR amplification, nucleotide -> amino-acid translation. Poisson vs Negative Binomial.
- **Noise** from various **sub-sampling** stages: between preparation and initial library, before sequencing.

Computation

- Unsupervised learning of an **energy-based model**: needs MCMC for intractable partition functions.

Misc

- Hard to evaluate model goodness, Multiple targets, localized datasets, spurious selection for plastic support

To parametrize or not to parametrize

Initial library

To parametrize or not to parametrize

Initial library

Non-parametrized

$N^{round\ 0}(s) \in \mathbb{R}^{\#\ \text{seq}}$

- Most flexible
- #seq ~ 10^7 parameters
- Prone to overfitting

To parametrize or not to parametrize

Initial library

Non-parametrized

$$N^{round\ 0}(s) \in \mathbb{R}^{\#\ \text{seq}}$$

- Most flexible
- #seq ~ 10^7 parameters
- Prone to overfitting

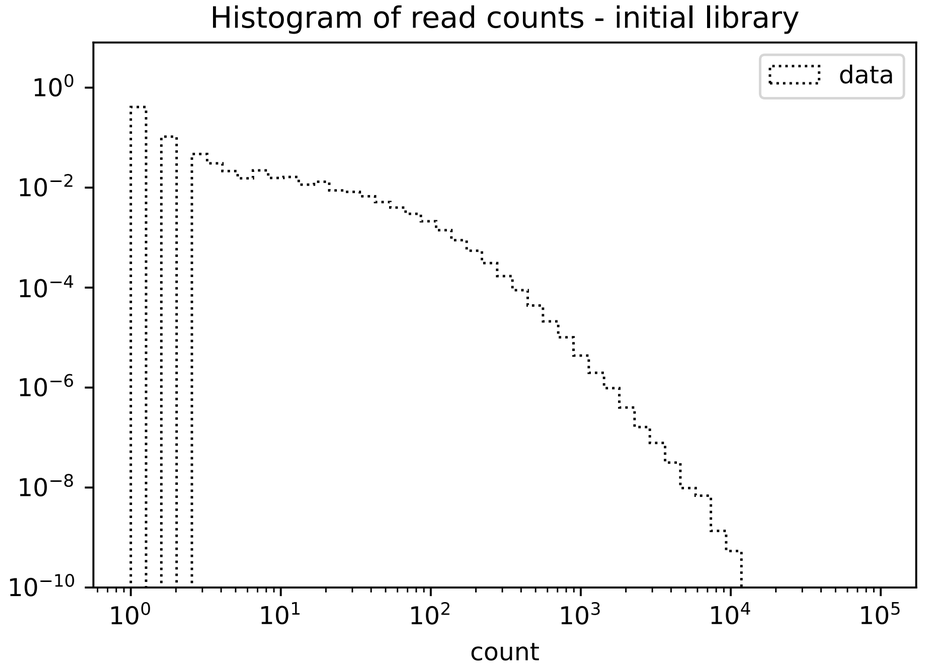
Parametrized

$$N_{indep}^{round\ 0}(s) = \prod_{i=1}^{L} f_i(s_i)$$

- Strong assumption
- 4L parameters
- If correct, more robust to noise

Initial library distribution

Checking model assumptions: fit and re-sample



Data from from phage display experiments, Fernandez-de-Cossio-Diaz et al. *PLoS Computational Biology* 20.10 (2024)

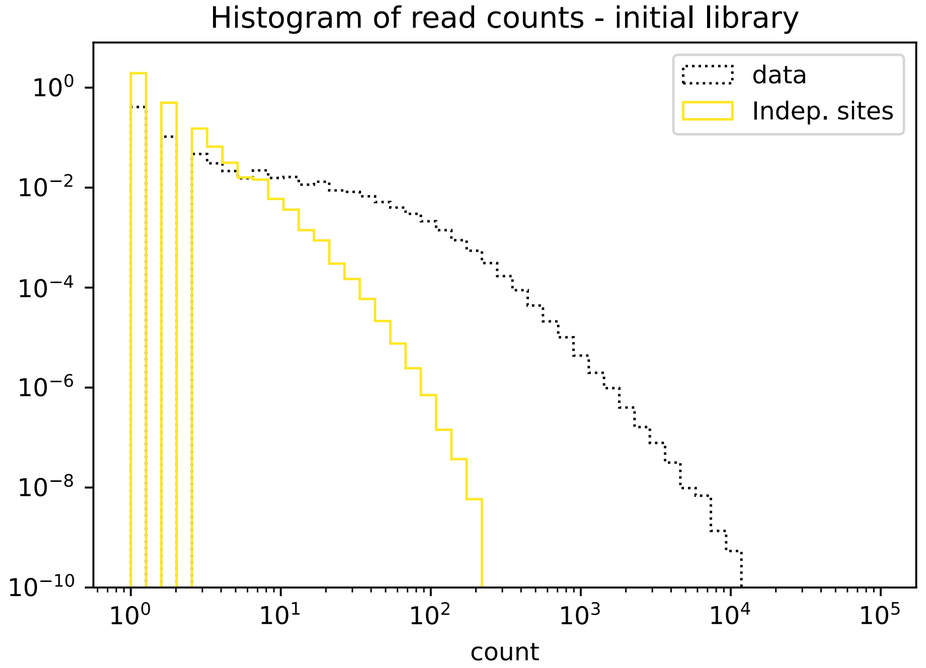
Initial library distribution

Checking model assumptions: fit and re-sample

Independent-sites

$$N_{indep}^{round\ 0}(s) = \prod_{i=1}^{L} f_i(s_i)$$

L: sequence length



Data from from phage display experiments, Fernandez-de-Cossio-Diaz et al. *PLoS Computational Biology* 20.10 (2024)

Initial library distribution

Checking model assumptions: fit and re-sample

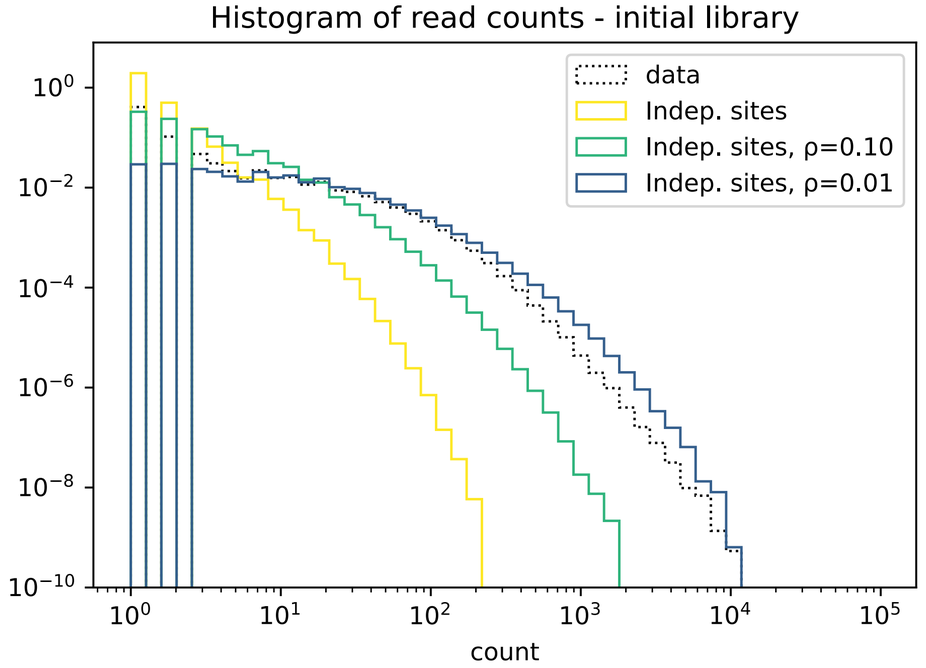
Independent-sites

$$N_{indep}^{round\ 0}(s) = \prod_{i=1}^{L} f_i(s_i)$$

L: sequence length

Truncated Independent-sites

$$N^{round\ 0}(s) \propto \begin{cases} N_{indep}^{round\ 0}(s) & \text{with prob } \rho \\ 0 & \text{with prob } 1-\rho \end{cases}$$



Data from from phage display experiments, Fernandez-de-Cossio-Diaz et al. *PLoS Computational Biology* 20.10 (2024)

Conclusions

Conclusions

- Directed Evolution experimentally finds **good aptamers for given target**

Conclusions

- Directed Evolution experimentally finds **good aptamers for given target**
- Mathematical model to **generate novel** ligands

Conclusions

- Directed Evolution experimentally finds **good aptamers for given target**
- Mathematical model to **generate novel** ligands
- Good model needs **good assumptions**

Conclusions

- Directed Evolution experimentally finds **good aptamers for given target**
- Mathematical model to **generate novel** ligands
- Good model needs **good assumptions**
- **Parametric** vs non-parametric, **initial library** distribution

Thank you!

stefano.crotti@ipht.fr

Collaborators

Jorge Fernandez-de-Cossio-Diaz

RNA aptamers for virus purification

Fréderic Ducongé

Lisa Le Dortz

Nicolas Destombes

Phage display for antibody design

Clément Nizak

Matilde Pattanaro

Olivier Rivoire

Guido Uguzzoni