Designing Highly Multiplex PCR Primer Sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE)

by **Nina G. Xie, Michael X. Wang, Ping Song, Shiqi Mao, Yifan Wang, Yuxia Yang, Junfeng Luo, Shengxiang Ren, David Yu Zhang**

Presenter: **Marie Hoffmann, M.Sc.**

published 2022/04/11 in Nature Communications, Volume 13, Issue 1

August 23, 2022

Abstract

Designing Highly Multiplex PCR Primer Sets with SADDLE

Problem tackled:

- Application: Multiplex-PCR
- Optimize composition of primer set S given pool of proto-primers w.r.t. low dimerization chance

Challenges:

- Dimerization – primer dimerization grows quadratically
- Combinatorial – exponentially many possibilities to form subset

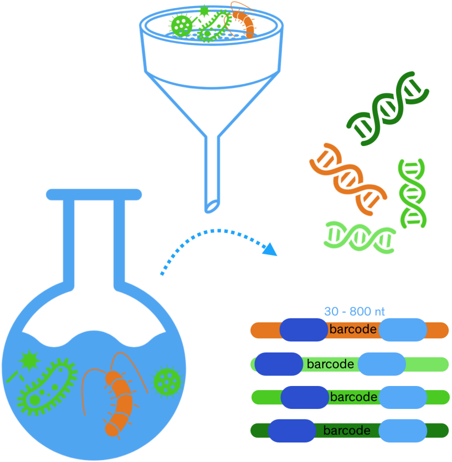
Results:

- In a 96-plex PCR primer set (192 primers), the fraction of primer dimers decreases from 90.7 % (naively designed) to 4.9%
- SADDLE-designed primer sets can be in NGS, but also qPCR

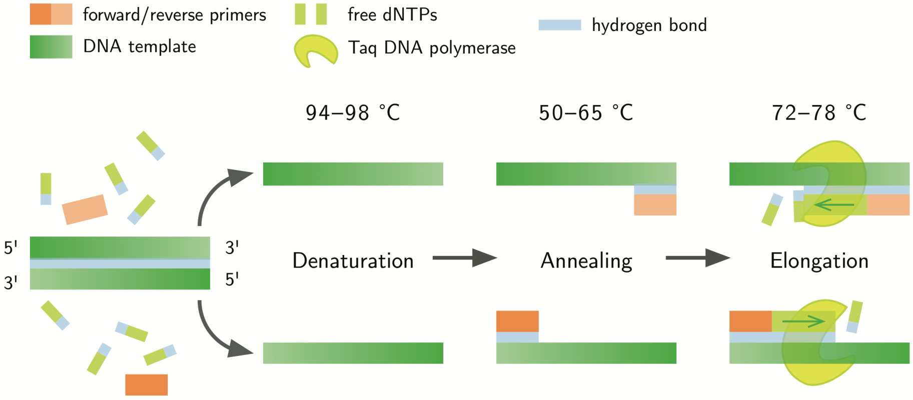
Field of Application: Metabarcoding

Technique

- Only technique capable of identifying up to thousands species/sample
- Affordable and well-established method
- Technique
  1. Batch-process DNA extracts
  2. Amplify barcode via PCR
  3. Sequence via NGS
  4. Identify via match against reference database



Polymerase Chain Reaction (PCR)



forward/reverse primers

DNA template

free dNTPs

Taq DNA polymerase

hydrogen bond

94–98 °C

50–65 °C

72–78 °C

5'

3'

3'

5'

Denaturation → Annealing → Elongation

Multiplex PCR

A Non-Convex Optimization Problem

Multiplexing: add multiple distinct primer pairs for simultaneous amplification of many regions

Challenges

1. Dimerization between P primer sequences

$$O(P^2)$$

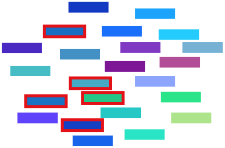
2. Sequence selection - exponentially many choices for a pool of N multiplex primers

$$\binom{N}{P}$$

3. Non-convex fitness landscape

```
  5-CGAAAGTCAGGGGATCG->
        ||||
5-CGAAAGTCAGGGGATCG->

  5-ACTTAGATGTACGTGG->
    || || || || ||
<-GGTGCATGTAGATTCA-5
```



Optimization Method: Simulated Annealing

via Stochastic Sampling

Goal: Minimize a energy E (or loss L) w.r.t. to a configuration θ

$$\widehat{\theta} = \arg\min_{\theta} E(\theta)$$

```
1: $S = S_0$
2: for $g = 1$ to $g_t$ do
3:     $T = \text{temperature}(1 - (g+1)/g_t)$
4:     $S_{\text{new}} = \text{neighbor}(S)$
5:     if $P(E(S), E(S_{new}), T) \geq \text{rand}(0,1)$ then
6:         $S = S_{\text{new}}$
7:     end if
8: end for
9: return $S$
```

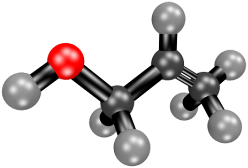


Figure 1: taken from Shutterstock

SADDLE: Loss Function and Initialization

1. Selection of an initial primer set S_0 from candidate pool
2. Evaluation of the Loss function $L(S_0)$

$$
\begin{aligned}
L(S) &= \sum_{b \geq a} Badness(p_a, p_b) && (1) \\
Badness(p_a, p_b) &= \sum_{q \in Q_a \cap Q_b} \frac{2^{|q|} 2^{\mathrm{GC}}}{(d_a + 1)(d_b + 1)} && (2) \\
Q &= \{q \in p : |q| \in [4; 8]\} && (3)
\end{aligned}
$$

Note: hashing of patterns reduces runtime to $\mathcal{O}(PN)$ per time step

SADDLE: Repeat Steps 3 and 4

3. Generate temporary primer set T based on set S_g (primer set from generation g) by randomly changing 1 or more primers
4. Evaluate $L(T)$, and set S_{g+1} to either S_g (no change) or T:

   case $L(T) < L(S_g)$ then $S_{g+1} = T$

   case $L(T) \geq L(S_g)$[1] then $S_{g+1} = T$ with $P = exp\{^{L(S_g) - L(T)}/_{C(g)}\}$

   otherwise stochastic gradient descent (SGD), i.e., $S_{g+1} = S_g$

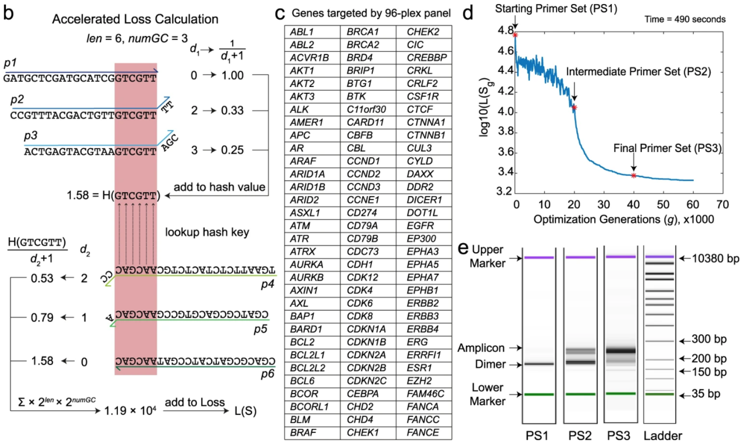
Notes: Probability $P(\cdot)$ depends on

- p depends on magnitude $L(S_g) - L(T)$
- generation-dependent and decreasing function $C(g)$

[1] and $g \leq g_T$

Evaluation I – Paper

96-plex primer set selected from cancer-related genes



ABL1	BRCA1	CHEK2
ABL2	BRCA2	CIC
ACVR1B	BRD4	CREBBP
AKT1	BRIP1	CRKL
AKT2	BTG1	CRLF2
AKT3	BTK	CSF1R
ALK	C11orf30	CTCF
AMER1	CARD11	CTNNA1
APC	CBFB	CTNNB1
AR	CBL	CUL3
ARAF	CCND1	CYLD
ARID1A	CCND2	DAXX
ARID1B	CCND3	DDR2
ARID2	CCNE1	DICER1
ASXL1	CD274	DOT1L
ATM	CD79A	EGFR
ATR	CD79B	EP300
ATRX	CDC73	EPHA3
AURKA	CDH1	EPHA5
AURKB	CDK12	EPHA7
AXIN1	CDK4	EPHB1
AXL	CDK6	ERBB2
BAP1	CDK8	ERBB3
BARD1	CDKN1A	ERBB4
BCL2	CDKN1B	ERG
BCL2L1	CDKN2A	ERRFI1
BCL2L2	CDKN2B	ESR1
BCL6	CDKN2C	EZH2
BCOR	CEBPA	FAM46C
BCORL1	CHD2	FANCA
BLM	CHD4	FANCC
BRAF	CHEK1	FANCE

Figure 2: Evaluated on 10 ng of NA18562 human genomic DNA.

Results I – Paper

96-plex primer set selected from cancer-related genes

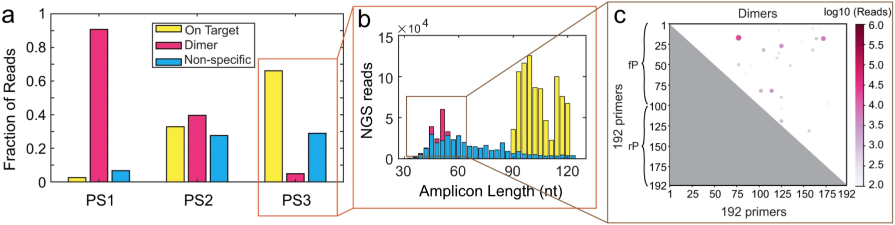


Figure 3: Read analysis. **On Target** – aligned to intended amplicon, **Dimer** – primer dimers, **Non-specific** – other

in silico upscale demo with 384 amplicon panel (768 primers) on 40 ng NA18562 genomic DNA:

- On-Target 43 %, Non-specific 56 %, and 1 % dimer amplicons

Evaluation II

Prediction Accuracy of the Badness function

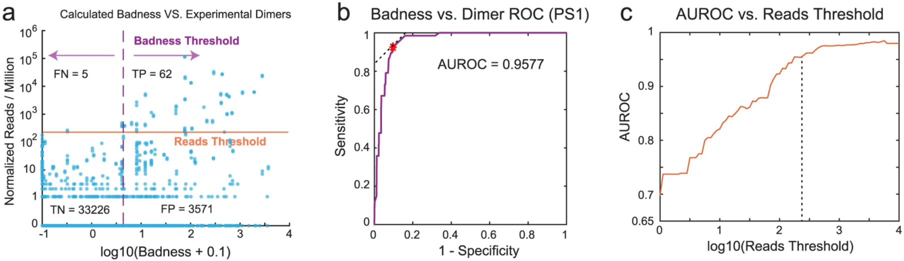


Figure 4: **a** Observed vs. predicted primer dimers. Reads threshold here: mean on-target read depth. **b** ROC Sensitivity vs. 1 - Specificity by shifting Badness threshold. **c** AUROC dependency of reads threshold.

Current setting: 92.5 % sensitivity and 90.3 % specificity

Evaluation II

Metabarcoding of Plankton

1. Compute primer sequences on 19 plankton clades [1]
2. Select proto-primers for head of Gibb's: $\Delta G \in [x_{\alpha_{5\%}} : x_{\alpha_{95\%}}]$ [2]
3. Use SADDLE to propose multiplex primers (iteration vs. loss plot)
4. Result: [Matlab Online]

Code: `https://github.com/mariehoffmann/PriSeT_X_SADDLE`

[2] $x_{\alpha_c} : cdf(x_{\alpha_c}) = c$

Evaluation II – Metabarcoding of Plankton

Tuning Acceptance Probability – Standard Error

$$\begin{aligned} P(S_g = T \mid L(T) > L(S_g)) &= \exp\left(L(S_g) - L(T)\right) \cdot \mathrm{stderr}^{-1} && (4) \\ \mathrm{stderr} &= \frac{\sigma}{\sqrt{p}} && (5) \end{aligned}$$

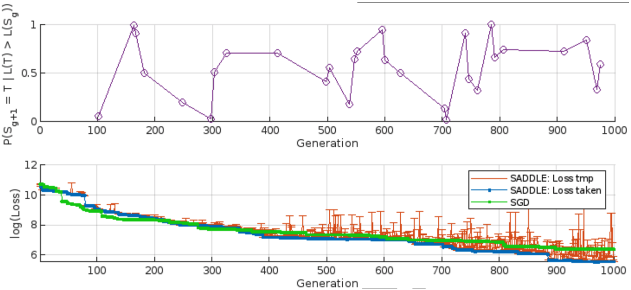


Figure 5: $N = 436$, $P = 64$, detriment normalized via stderr

Discussion

1. Baseline SGD can be better
   - Run with $P(\cdot)$, and SGD for comparison
   - Repeat with different seeds of RNG
   - Run sufficiently many iterations
2. Plenty of constraints unchecked
   - forward complementary, disconnected annealing patterns, $(A|T)$-tails, hairpins, etc.
   - Need better understanding of primer dimerization: high-ranked dimers from experiment had low badness score and vice versa
3. Parameter tuning C_g for acceptance probability difficult
   - Needs robustness based on Loss statistics
4. Control for proportion of fwd/rev primers

Conclusion

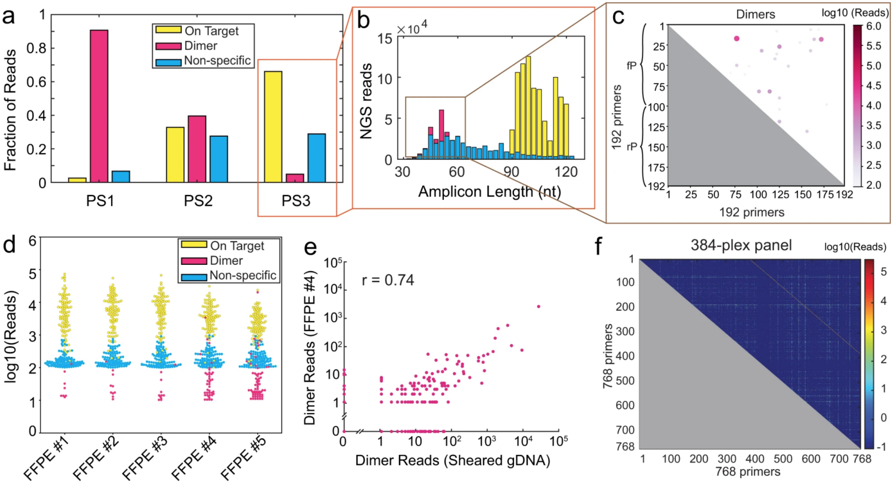
1. SADDLE reduces reagent costs while amplifying hundreds of target templates simultaneously
2. Broad Applicability of Framework
   - Different types of PCR
   - Amplification of multiple regions of same genome
   - Gene fusion detection
   - Amplification of similar regions in different genomes
3. Framework adjustable, e.g., Metabarcoding – identification of many phylogenetically diverse species
   1. Sample Clade index j
   2. Sample primer pair from Clade$_j$ for probabilistic exchange
4. More sequence checks can be easily added with low computational overhead (C++)

List of References

[1] Marie Hoffmann, Michael T. Monaghan, and Knut Reinert. “PriSeT: Efficient de Novo Primer Discovery”. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB ’21. Gainesville, Florida: Association for Computing Machinery, 2021. ISBN: 9781450384506. DOI: 10.1145/3459930.3469546. URL: https://doi.org/10.1145/3459930.3469546.

[2] Nina G Xie et al. “Designing highly multiplex PCR primer sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE)”. In: *Nature Communications* 13.1 (2022), p. 1881.

Appendix: Evaluation I



Appendix: Evaluation II

$$P(S_g = T | L(T) > L(S_g)) \quad = \quad \exp\left(L(S_g) - L(T)\right) \cdot \mathrm{stderr}^{-1} \qquad (6)$$

$$\mathrm{stderr} \quad = \quad \frac{\sigma}{\sqrt{p}} \qquad (7)$$

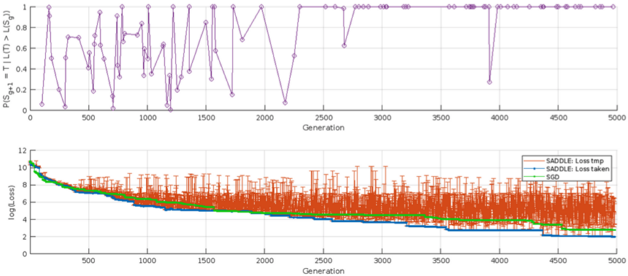


Figure 6: $N = 436$, $P = 64$, detriment normalized via stderr, $g_t = 5000$, $g_T = 5000$, $log(L^{\mathrm{SADDLE}}_{g_T}) = 7.3$, $log(L^{\mathrm{SADDLE}}_{g_T}) = 16.1$