Ultra High Diversity Factorizable Libraries for Efficient Therapeutic Discovery

Zheng Dai[1(✉)], Sachit D. Saksena[1(✉)], Geraldine Horny[2], Christine Banholzer[2], Stefan Ewert[2], and David K. Gifford[1(✉)]

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
{zhengdai,sachit,gifford}@mit.edu
[2] Novartis Institutes for BioMedical Research (NIBR), Basel, Switzerland

The successful discovery of novel biological therapeutics by selection requires highly diverse libraries of candidate sequences that contain a high proportion of desirable candidates. Here we propose the use of computationally designed *factorizable libraries*, whose sequences are made of concatenated segments from smaller *segment libraries*, as a method of creating large libraries that meet an objective function at low cost.

Designing segment libraries that result in a factorizable library that meets an objective function is a computationally difficult task. We present a computational method we call Stochastically Annealed Product Spaces (SAPS), which optimizes segment libraries though iterative improvements with respect to an objective function to design a full length factorizable library. Key to our method is the *reverse kernel trick*, which allows us to efficiently evaluate an objective over the full factorizable library by casting the objective function as an inner product of feature vectors (see Fig. 1).

We show that SAPS outperforms five different benchmark sampling approaches on simulated datasets. We next apply SAPS to design factorizable libraries of the third complementarity determining region of antibody heavy chains (CDR-H3s). We show that this framework can generate factorized CDR-H3 segment libraries that, when joined combinatorially, contain $\sim 10^9$ unique sequences with highly specific and flexible design parameters. We compare these libraries to a randomized library and show that SAPS designed libraries are more diverse and more enriched in desirable sequences.

Applications of factorizable libraries include the discovery of biologics such as monoclonal antibody therapeutics [5], discovery of adeno-associated vectors (AAV) for gene therapy [1,8], T-cell receptor (TCR) discovery [2,4,7], and aptamer libraries [3,6].

Full Text Preprint: https://www.biorxiv.org/content/10.1101/2022.01.17.476670v1.

Data Availability: https://github.com/gifford-lab/FactorizableLibrary.

Z. Dai, S. D. Saksena and D. K. Gifford: Equal contribution.

I. Pe'er (Ed.): RECOMB 2022, LNBI 13278, pp. 390–392, 2022.
https://doi.org/10.1007/978-3-031-04749-7_40

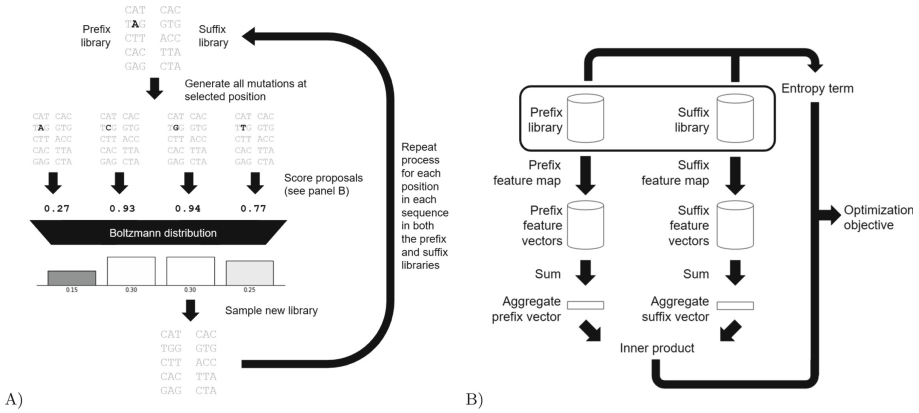

Fig. 1. Factorizable library evaluation and optimization. A) Optimization is achieved through iterative stochastic updates. An update step is performed by selecting a position in a sequence in one of the libraries and generating all possible mutations for that position. The mutated libraries are then scored, and then a Boltzmann distribution over the libraries is generated using the negated scores as energy values. The update is then sampled from the distribution. A full update sweep performs this for all positions in all sequences in both libraries. Multiple sweeps are done and the temperature of the Boltzmann distribution is lowered over time. For simplicity, the figure depicts this optimization on small DNA libraries. B) The score of the factorizable library is evaluated by mapping all the sequences in its prefix and suffix libraries to feature spaces. The feature vectors are then aggregated, and an inner product is taken between them, which by the distributive property produces the total score for the whole factorizable library. We refer to this as the reverse kernel trick, since this optimization requires expressing a "kernel function" that maps prefix suffix pairs to real values as an inner product. A position based entropy term is evaluated to quantify the diversity of sequences in the library, and a weighted sum of the two is then used to guide optimization.

Acknowledgements. This work was funded by NIH Grant R01 CA218094, and a gift from Schmidt Futures to D.K.G. The experimental work was funded by Novartis.

References

1. Bryant, D.H., et al.: Deep diversification of an AAV capsid protein by machine learning. Nat. Biotechnol. **39**, 691–696 (2021)
2. Holler, P.D., Holman, P.O., Shusta, E.V., O'Herrin, S., Wittrup, K.D., Kranz, D.M.: In vitro evolution of a t cell receptor with high affinity for peptide/mhc. Proc. Nat. Acad. Sci. **97**(10), 5387–5392 (2000). https://doi.org/10.1073/pnas.080078297, https://www.pnas.org/content/97/10/5387
3. Keefe, A.D., Pai, S., Ellington, A.: Aptamers as therapeutics. Nat. Rev. Drug Discov. **9**(7), 537–550 (2010)
4. Li, Y., et al.: Directed evolution of human t-cell receptors with picomolar affinities by phage display. Nat. Biotechnol. **23**(3), 349–354 (2005)
5. Liu, G., et al.: Antibody complementarity determining region design using high-capacity machine learning. Bioinformatics **36**(7), 2126–2133 (2020)

6. Maier, K.E., Levy, M.: From selection hits to clinical leads: progress in aptamer discovery. Mol. Ther. Methods Clin. Dev. **5**, 16014 (2016)
7. Smith, S.N., Harris, D.T., Kranz, D.M.: T cell receptor engineering and analysis using the yeast display platform. Methods Mol. Biol. **1319**, 95–141 (2015)
8. Wang, D., Tai, P.W.L., Gao, G.: Adeno-associated virus vector as a platform for gene therapy delivery. Nat. Rev. Drug Discov. **18**(5), 358–378 (2019)