How Combinatorial Optimization Enables Rapid Discovery of High-Performance Variants

In the forward engineering approach to optimizing biological systems, diversity generation is followed by a combinatorial optimization phase. The process involves the discovery of single edit hit variants from the diversity generation phase and partnering them in novel configurations through combinatorial libraries.

But creating combinatorial libraries is not always straightforward. A key challenge is identifying the best way to generate libraries and to search through them quickly to identify improved performance variants.

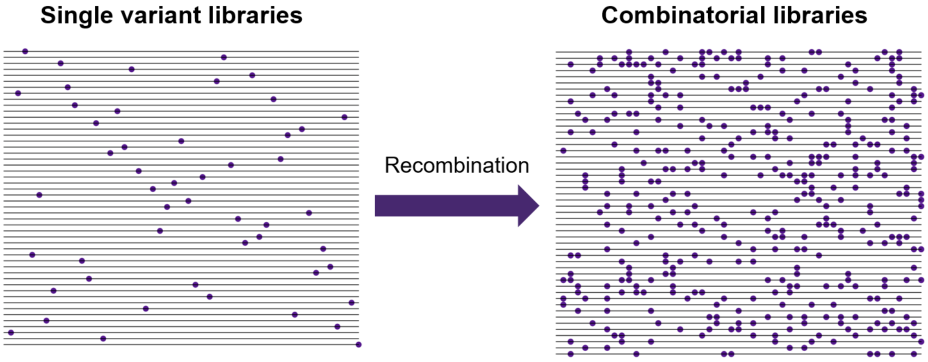


Figure 1: Combinatorial libraries (right) allow for important optimization of biological variant performance, significantly improving on the gains obtained from in single-variant libraries (left). Each line represents a single genetic sequence in a population of related variants, where dots represent mutations.

A research approach from the mid-'90s focused on recombination at a sub-genomic scale, using DNA shuffling as one mechanism for achieving high yield variation. The method involved selecting a set of related gene variants and shuffling them together in a hypersexual way to recombine fragments of different genetic material from various parent sequences to produce a progeny population for screening. Repeating this process through several rounds of shuffling and screening can simultaneously weed out and accumulate diversity that is detrimental or beneficial to the trait of interest and produce novel variants with improved capabilities. This process of DNA shuffling is comparable to decades-old computational methods employing evolutionary algorithms for *in silico* optimization problems.

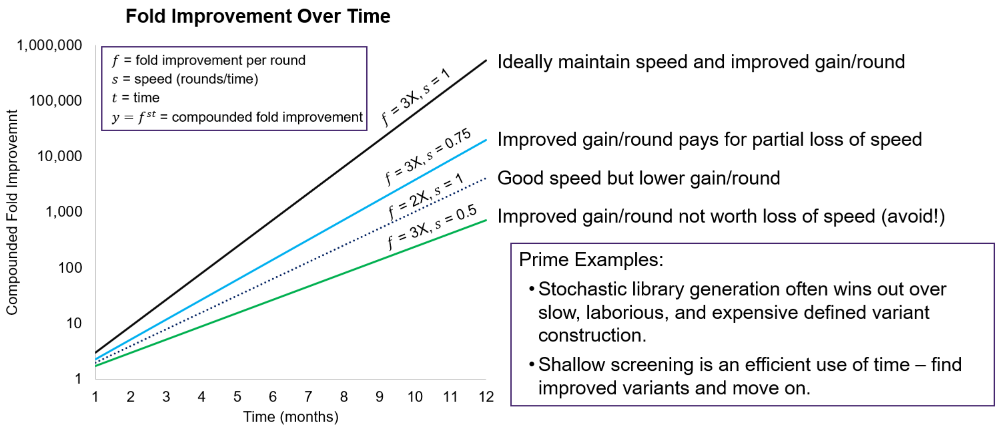
In the wet lab context, asexual DNA mutagenesis involves introducing mutations in parental sequences to generate diversity and using high-throughput phenotyping to select an optimal strain that serves as the input for the next round of mutagenesis. Although asexual approaches work, they fail to take advantage of a wealth of available beneficial genetic diversity in the rest of the population that can be recombined and shuffled to create variants with improved function.

A combinatorial optimization strategy that leverages existing beneficial diversity facilitates a more rapid exploration of sequence space compared to asexual approaches. The benefits of such a strategy were demonstrated in an experiment where a starting strain that produced a molecule of interest was subjected to 20 years of classical mutagenesis, effectively asexual reproduction, to obtain higher-yield progeny strains. Although the program resulted in high titer mutants, the approach consumed many years and resources. Alternatively, by adopting an approach consisting of an initial diversity generation phase followed by only two rounds of recombination of beneficial mutations from across the genomes of several individually improved strain variants, a new production strain comparable to the one obtained from the 20-year process was delivered in a single year and with significantly fewer resources. The use of more recently developed genome engineering methods on a strain engineering project of this kind can reduce the time and resource requirements even further [(1)].

The Importance of Fuel and Speed for Evolution

In many experiments, the desired combinations of mutations that yield the most optimal results are quite rare. Given a complex combinatorial library, strategies that focus on deep screening tend to offer minimal return on investment. Once the first few hundred or few thousand variants have been sampled, a deep screening approach will return very few improved variants from the population (over what has already been discovered) and will almost certainly fail to explore all possible variation, including the most optimal variants. A more effective strategy is to screen fewer variants from these libraries, select the best ones, and then recombine them to create novel genetic configurations that further improve performance. Repeating this process over several rounds ultimately homes in on many of the rarer configurations of beneficial mutations that the deep screening approach is likely to miss.

Operational speed and large reserves of genetic fuel for evolution are two key parameters that govern fitness gains in performance over repeated rounds of combinatorial optimization. After multiple rounds of evolution, the phenotypic variance in the population starts to drop off as the genotypic variance is purified out. This results in less diversity with each generation and affects the ability to gain greater fitness with continued rounds of recombination. Avoiding this issue requires stockpiling large quantities of starting beneficial diversity. This provides more evolutionary fuel for the combinatorial optimization step and profoundly improves performance over successive rounds of evolution.



Prime Examples:

- Stochastic library generation often wins out over slow, laborious, and expensive defined variant construction.
- Shallow screening is an efficient use of time – find improved variants and move on.

Figure 2: Slower strategies sacrifice rateof performance gains, while faster versions – even imperfect ones – consistently perform better.

Additionally, it is important to avoid strategies that improve the fitness gain per round at the expense of operational speed. In both diversity generation and combinatorial optimization, stochastic library generation wins out over slow, laborious, and expensive defined variant construction approaches. Even though stochastic libraries are not perfect in their representation of different mutations, they are very robust and fast to create and phenotype — making them a far more efficient use of resources.

Machine Learning-Guided Models for Improved Optimization

Combinatorial optimization produces valuable diversity by selecting the best options and randomly recombining their genetic material. Approaches guided by machine learning can be used to speed up the process by mining sequence data to establish genotype-phenotype relationships, and then constructing *in silico* models that guide future rounds of evolution.

An optimization experiment to improve the catalytic performance of an enzyme offers a representative example of using machine learning in this way(2). The experiment began with a diversity generation step using the wild type enzyme. This resulted in 171 beneficial mutations as potential inputs for the optimization procedure. The best mutations from the diversity generation round were selected and then screened using a combinatorial library of several thousand variants. The genotype-phenotype information from this set of mutations was used to build a statistical model that could infer from sequence data which mutations were beneficial, neutral, or deleterious for the property of interest. The best variant from the second round served as the input for generating the library that was used in the third round.

In this project, researchers used the statistical model to make additional changes to the variant, including 18 mutations from diversity stockpiled in the first round. They then screened the library to identify new hits. The outcome from the third round was a variant that showed two orders of magnitude improvement compared to the starting enzyme.

In the first large scale example of this method(3), researchers used machine learning guided evolution to create an enzyme that showed a 4,000-fold improvement in volumetric productivity over the starting material. The final production variant had 51 codon changes, nearly 40 of which were coding variants. It is important to note that these beneficial mutations were discovered across the enzyme. By way of comparison, the researchers also performed classical DNA shuffling in parallel to the machine learning approach. Comparing the results of the two indicated that in 14 out of the first 15 rounds of evolution, the machine learning approach outperformed the DNA shuffling approach at discovering improved variants. The team also found that many of the beneficial mutations identified through the statistical model would have been missed by traditional DNA shuffling since they were not present in the top variants the first time they appeared.

Summary

Recombination is an extremely powerful tool for optimization. It facilitates the rapid accumulation of beneficial diversity and achieves much higher fitness levels faster than standard asexual or rational engineering methods. Making the most of a combinatorial optimization program requires the design of large, stochastic libraries where theoretical sizes exceed the traditional capacity to generate and test variants. Combinatorial optimization is most efficient when coupled with shallow sampling of these large libraries and with iterative Design-Generate-Test-Learn (DGTL) rounds for achieving the largest gains in performance most rapidly. When the empirical approach is paired with machine learning models, the process of identifying optimal variants becomes even faster.

References

[1] Zhang Y., Perry K., Vinci A.V., Powell K., Stemmer W.P.C., Del Cardayré S.B. (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature 45 (6872):644-6.

[2] Fox J.R. (2013). IBC 5. Applications for Enzyme Technologies.

[3] Fox J.R. , Davis S.C., Mundorff E.C., Newman L.M., Gavrilovic V., Ma S.K., Chung L.M., Ching C., Tam S., Muley S., Grate J., Gruber J., Whitman J.C., Sheldon R.A., Huisman G.W. (2007). Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol. 25(3): 338-44.

Learn more at **INSCRIPTA.COM**

INSCRIPTA®

INSCRIPTA.COM