

Recursive Ensemble Mutagenesis Plus 32 Years

Douglas C. Youvan

doug@youvan.com

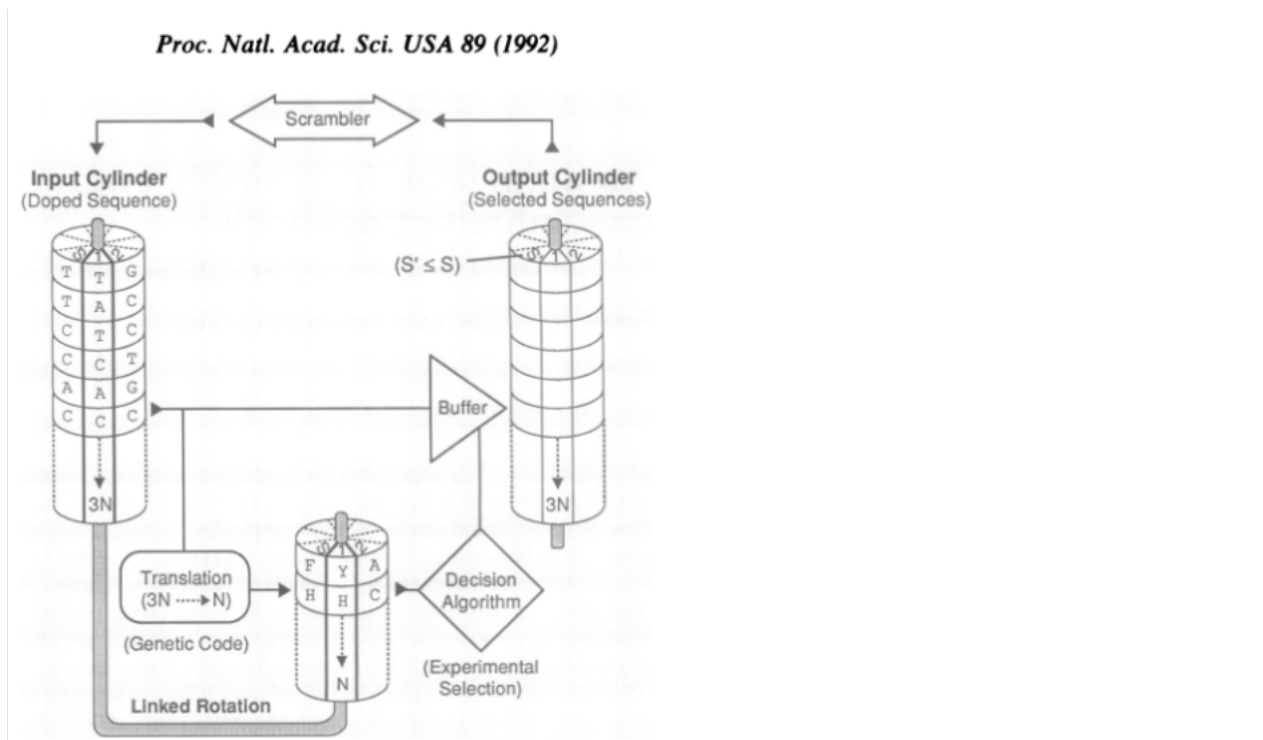
March 23, 2024

Since its inception in 1992, Recursive Ensemble Mutagenesis (REM) has catalyzed a paradigm shift in protein engineering, laying the groundwork for the iterative, computational approaches that have come to define the field. This method, pioneering at its time for employing a feedback-driven cycle of mutagenesis to explore protein sequence space, presaged the fusion of computational biology with experimental techniques that is now standard. As we mark 32 years since REM's introduction, it is imperative to review the trajectory of protein engineering and in vitro evolution, noting how REM's principles have permeated and propelled these disciplines forward. This paper, "Recursive Ensemble Mutagenesis Plus 32 Years," aims to chronicle the advancements since REM, highlighting the integration of computational models with experimental mutagenesis and the transformative impact on biotechnology. Reflecting on the legacy and evolution of REM provides a lens through which to appreciate the current state of protein engineering and envision the future directions of this ever-evolving field.

Keywords: Recursive Ensemble Mutagenesis, protein engineering, in vitro evolution, computational biology, experimental mutagenesis, bioinformatics, directed evolution, synthetic biology, machine learning, artificial intelligence, high-throughput sequencing, biotechnology advancements.

Original: Arkin and Youvan, PNAS (1992):

<https://www.pnas.org/doi/epdf/10.1073/pnas.89.16.7811>



Introduction

In the landscape of protein engineering, the 1992 publication of the Recursive Ensemble Mutagenesis (REM) paper marked a pivotal moment, introducing an algorithmic approach to the iterative refinement of protein sequences. The primary objective of this paper was to present a novel strategy for protein engineering—one that leveraged a feedback mechanism from successive rounds of mutagenesis to effectively search and select for proteins with desired characteristics within the vastness of sequence space. This approach was grounded in the principles of combinatorial optimization and artificial intelligence, which at the time were novel applications within the field of biochemistry.

The findings from Arkin and Youvan laid a conceptual foundation for what would evolve into an entirely new methodology for protein engineering. The REM algorithm was not merely a theoretical construct; it was a feasible

proposition for guiding the laborious process of protein optimization, with the potential to significantly streamline the discovery of proteins with improved or new functions. It was an ambitious fusion of the burgeoning fields of genetic algorithms and protein design, envisioned during an era when the capabilities for gene synthesis and sequencing were just beginning to burgeon.

The historical context in which REM was introduced cannot be overstated. The early '90s witnessed the birth of the Human Genome Project and the burgeoning of recombinant DNA technologies. This era was ripe for innovation, with molecular biology tools advancing rapidly and computational power increasing exponentially. The REM paper was not only a product of its time but also a harbinger for a future where these two domains would intersect with remarkable synergy.

Since the publication of the REM paper, the advancements in the field of protein engineering have been exponential. The tools for gene synthesis and mutagenesis have become more accessible and robust, enabling the creation of vast libraries of protein variants. High-throughput screening methods have developed to a point where millions of variants can be assayed for desired functions in a relatively short period. Perhaps most crucially, the computational algorithms, once nascent and simplistic, have grown in sophistication to match the complexity of biological systems. Machine learning, for example, has emerged as a powerful tool for predicting protein structure and function, far exceeding the capabilities that were available to Arkin and Youvan at the time of their research.

In this introduction, we set the stage for a comprehensive review that not only reflects on the legacy of REM but also charts the course of the subsequent three decades of innovation. The evolution of in vitro protein evolution, sparked in part by the foresight embedded in the REM paper, has led to significant scientific and therapeutic achievements. As we delve into this rich history, we underscore the undeniable impact of REM on the trajectory of protein engineering, shaping a future that continues to expand the boundaries of what is possible in the biosciences.

REM: The Concept Revisited

The Recursive Ensemble Mutagenesis (REM) process, as conceived in 1992, represented a groundbreaking approach to protein engineering, built on the idea that the evolutionary process could be harnessed and accelerated in the laboratory setting. The REM algorithm sought to iteratively generate and refine a diverse population of protein variants, using a feedback loop to hone in on sequences with optimal characteristics. This concept was an early forerunner to what is now commonly referred to as directed evolution, though at the time it stood at the confluence of cutting-edge bioinformatics and practical molecular biology.

The REM process begins with a population of protein variants—initially quite diverse but somewhat related to a known wild-type sequence. Variants were generated through mutagenesis, introducing alterations at the DNA level to be translated into changes in the amino acid sequence of the resultant proteins. These variants were then screened based on a set of predefined criteria, such as activity, stability, or binding affinity. Sequences that met these criteria were carried forward to the next round of mutagenesis and screening, while those that did not were discarded. The iterative nature of this process, akin to natural selection, was key to its function: each cycle was intended to increase the prevalence of beneficial mutations and to decrease the frequency of deleterious or neutral ones.

The original algorithm put forth a number of theoretical predictions about the efficiency of this method. It proposed that REM would be capable of finding sequences with desired properties more efficiently than random mutagenesis because it could 'learn' from each iteration, becoming more focused in its search as it progressed. The algorithm was also designed to be flexible, allowing it to be applied under various selection criteria and experimental conditions, reflecting the diverse needs of protein engineering projects.

The figure provided in the original paper serves as a conceptual representation of this process. It illustrates the translation of a doped DNA

sequence (the input cylinder) into an ensemble of protein variants. These variants pass through a 'scrambler'—analogous to mutagenesis—and are then evaluated by the decision algorithm (the experimental selection), which selects the most promising candidates for the next round (the output cylinder). This diagram encapsulates the cyclic nature of the REM process, from initial sequence input through to the selection of refined outputs.

Revisiting this concept today, it is clear that while the specifics of the REM algorithm have been surpassed by more sophisticated computational tools and methods, the core principles remain relevant. The idea of using iterative, feedback-driven cycles to evolve proteins towards a desired function is now a cornerstone of in vitro evolution. The legacy of REM is evident in the way we approach protein engineering challenges—always building upon the rounds of selection and screening that enable us to climb peaks in the fitness landscape more effectively than ever before.

Evolution of Protein Engineering

Since the inception of Recursive Ensemble Mutagenesis (REM) in 1992, the field of protein engineering has witnessed an array of transformative milestones, each contributing to the current state-of-the-art practices and technologies.

Chronological Milestones in Protein Engineering:

- **Mid-Late 1990s:** The rise of phage display technology, which allowed for the screening of vast libraries of peptides and proteins, came on the heels of REM. This era saw the development of new selection techniques that could identify strong binders from a sea of candidates.
- **Early 2000s:** Advancements in high-throughput screening and the maturation of yeast display methods provided researchers with the tools to screen protein libraries at a throughput that was previously

unimaginable, leading to an explosion in the number of potential drug candidates.

- **2003:** The Human Genome Project's completion gave researchers a blueprint of human genetic material, offering unprecedented insights into protein functions and facilitating targeted protein design.
- **2010s:** The advent of next-generation sequencing (NGS) technology revolutionized protein engineering by allowing the rapid sequencing of entire libraries, providing detailed feedback on the effects of mutations.
- **2018:** The Nobel Prize in Chemistry was awarded to Frances H. Arnold for her work on the directed evolution of enzymes, highlighting the significance and potential of in vitro evolution in protein engineering.

Case Studies of Successful In Vitro Evolution:

- **Thermostable Enzymes:** Researchers have used directed evolution to create enzymes that can withstand high temperatures, which are crucial in industrial processes like biofuel production.
- **Antibody Development:** In vitro evolution has been critical in developing highly specific antibodies for therapeutics, including those used in the treatment of diseases such as cancer and autoimmune disorders.
- **Biosensors:** Through iterative cycles of mutation and selection, proteins have been engineered to function as biosensors, capable of detecting environmental pollutants or disease biomarkers with high sensitivity.

Anticipating Modern Workflows:

The REM algorithm anticipated many features of modern protein engineering workflows. The concept of a feedback loop, integral to REM, is now commonplace in directed evolution experiments where the results of one round of selection inform the design of the next. REM's use of computational algorithms to manage and guide the selection process presaged the current integration of machine learning and AI in protein

design. These computational tools can now predict the effects of mutations, design mutagenesis strategies, and even simulate the evolutionary trajectories of proteins.

Moreover, REM's approach to mutagenesis—targeting specific sites for modification and using a decision algorithm to select the most promising candidates—has been refined and expanded upon in modern methods. For example, techniques like error-prone PCR, DNA shuffling, and CRISPR-based editing have enhanced the precision and effectiveness of mutagenesis.

The overarching principle introduced by REM, that of searching sequence space in a structured, efficient manner, remains central to protein engineering today. It has influenced the development of systematic approaches to library construction and screening, ensuring that protein engineering remains both a science and an art—combining empirical data with the strategic exploration of the unknown. As a result, the capabilities for designing proteins with bespoke properties have advanced from a theoretical possibility to routine practice, creating an era of biotechnological innovation built on the shoulders of foundational concepts like those of REM.

Computational and Experimental Synergy

The symbiotic relationship between computational models and experimental mutagenesis has become a defining characteristic of contemporary protein engineering, setting the stage for rapid advances and innovative applications.

Synergistic Relationship:

The iterative loops of mutagenesis, selection, and screening are now closely interwoven with computational predictions and analyses. This synergy

allows for the refinement of experimental strategies based on computational outputs and, conversely, the improvement of computational models using experimental data. Computational models can simulate the effects of mutations on protein structure and function, which informs the design of mutagenesis experiments. The resulting experimental data, in turn, validate and refine the predictive power of these models, leading to a continually evolving and improving cycle of protein engineering.

Modern Software and Algorithms:

Several software and algorithms stand out for their contributions to accelerating the pace of protein engineering:

- **Rosetta**: A comprehensive suite for protein modeling, Rosetta has been instrumental in predicting protein structures, designing protein folds, and simulating protein-protein interactions.
- **FoldX**: An algorithm used to predict the impact of mutations on the stability of proteins, FoldX has facilitated the design of proteins with enhanced stability or altered binding properties.
- **DeepMind's AlphaFold**: Representing a quantum leap in protein structure prediction, AlphaFold's ability to predict protein structures with near-experimental accuracy has significantly reduced the time and resources required for structure determination.
- **Machine Learning Platforms**: Various machine learning approaches, such as neural networks and deep learning, have been developed to predict the outcomes of mutagenesis, identify likely candidates for successful protein evolution, and even suggest novel protein sequences that might not be readily apparent from the existing sequence space.

Realization and Expansion of REM's Vision:

The computational advancements have not only realized but significantly expanded upon REM's original vision. Where REM proposed an algorithmic approach to searching protein sequence space, current computational tools

have taken this to a new level of precision and complexity. The ability to rapidly model thousands of mutations, understand the relationships between sequence, structure, and function, and predict the results of mutagenesis has transformed REM's foundational concepts into a robust, high-throughput reality.

Modern computational tools have also enabled the exploration of more extensive and complex sequence spaces than REM's initial algorithm could have managed, allowing for the discovery of proteins with properties that may not be found in nature. The development of in silico directed evolution approaches, where computational models simulate multiple rounds of evolution, reflects the ultimate realization of REM's iterative, feedback-driven strategy.

In summary, the fusion of computational and experimental methods has created a dynamic and potent interface that has propelled protein engineering into a new era. This interface has allowed researchers to conduct virtual screening of protein libraries, predict the results of mutagenesis with greater accuracy, and achieve levels of protein function and diversity that were once only theoretical. This synergy has made the lofty goals of REM more achievable than ever, showcasing the power of computational biology to catalyze real-world biological innovation.

Current State of In Vitro Evolution

In vitro evolution has evolved dramatically since the 1990s, with advances across a wide range of technologies enhancing our ability to evolve biomolecules in a laboratory setting.

Current Methodologies and Technologies:

- **Directed Evolution:** This process mimics natural selection to evolve proteins or nucleic acids toward a user-defined goal. It has become

more sophisticated, with methodologies now incorporating multiplex automated genome engineering (MAGE), CRISPR-based techniques, and other forms of targeted mutagenesis.

- **Display Technologies:** Phage display, yeast display, and mRNA display techniques have been refined, allowing for the rapid screening of peptide and protein libraries against a vast array of targets.
- **Synthetic Biology:** Synthetic biology has contributed new tools for in vitro evolution, such as the creation of synthetic gene circuits and the use of unnatural amino acids to expand the genetic code.

Impact of High-Throughput Sequencing and Machine Learning:

- **High-Throughput Sequencing:** Next-generation sequencing (NGS) technologies have had a profound impact on in vitro evolution, allowing for the rapid analysis of mutant libraries. This technology has made deep mutational scanning possible, providing insights into sequence-function relationships on a scale not previously possible.
- **Machine Learning:** Machine learning and AI are being used to predict the outcomes of mutations, design proteins, and even create novel protein sequences. These computational approaches can analyze vast datasets produced by high-throughput sequencing, extracting patterns and predicting the fitness landscapes of proteins with increasing accuracy.

Comparative Analysis of Efficiencies:

- **Increased Efficiency:** In vitro evolution has become more efficient over the decades. Early techniques were limited by the size of libraries that could be screened and by the fidelity of the methods used to introduce mutations. Today, vast libraries can be constructed and screened, and specific mutations can be introduced with precision.
- **Speed of Evolution:** The speed at which proteins can be evolved has also increased. What might have taken years of work can now often be achieved in a matter of weeks or months due to the automation and parallelization of both the mutagenesis and screening processes.

- **Quality of Evolution:** The quality of evolved proteins has improved as well. Early in vitro evolution often produced proteins that were marginally better than the starting material, whereas current methods can yield proteins that are substantially improved in terms of stability, activity, and specificity.

In summary, the current state of in vitro evolution is characterized by a sophisticated suite of tools and techniques that allow for the rapid and efficient development of biomolecules with desirable properties. The integration of high-throughput sequencing and machine learning has not only increased the speed and efficiency of these processes but also the quality of the resulting proteins. The field has progressed from simple mutation and selection to a nuanced orchestration of biological systems, where the ability to read, write, and edit biological information has unlocked new horizons in protein engineering.

Case Studies: REM's Legacy in Action

The foundational principles of Recursive Ensemble Mutagenesis (REM) have been applied, either directly or in spirit, across various fields within and beyond protein engineering. Here are some specific examples where REM-like strategies have been instrumental:

Examples of REM-like Strategies in Practice:

- **Enzyme Development for Industrial Processes:** Inspired by REM's iterative approach, researchers have successfully developed enzymes with enhanced catalytic properties suitable for industrial applications, such as biofuel production. For example, the enzyme Cel7A, crucial in the conversion of cellulose to glucose, has been iteratively mutated and selected for improved performance at industrial conditions.
- **Antibody Engineering for Therapeutics:** The development of adalimumab, a monoclonal antibody used in the treatment of

rheumatoid arthritis, involved processes similar to REM. Iterative rounds of mutagenesis and selection were used to increase its affinity for TNF- α , demonstrating how REM-like strategies facilitate the fine-tuning of therapeutic molecules for specificity and efficacy.

- **Biosynthesis Pathway Optimization:** Synthetic biologists have employed REM-like methods to optimize metabolic pathways. Through iterative engineering and selection, strains of microorganisms have been developed that can produce pharmaceuticals, such as artemisinin, more efficiently and cost-effectively.

Iterative Feedback in Current Platforms:

Iterative feedback is a cornerstone of modern protein engineering platforms. Directed evolution workflows, for example, rely on feedback from screening results to guide subsequent rounds of mutagenesis. High-throughput screening technologies have enabled this feedback to be processed rapidly, allowing for the accelerated evolution of proteins with enhanced or novel functions.

Principles of REM Embedded in Modern Approaches:

- **Algorithmic Selection:** Modern bioinformatics tools incorporate algorithmic selection principles akin to those outlined in REM. They use machine learning to predict which mutations are likely to yield functional proteins, thereby guiding the design of mutagenesis strategies.
- **Adaptive Learning:** REM's adaptive learning—where the process 'learns' from previous iterations—is mirrored in the use of adaptive laboratory evolution (ALE). In ALE, populations of organisms are iteratively cultured under selective pressures, with beneficial mutations naturally accumulating over time.
- **Recursive Improvement:** The principle of recursive improvement—central to REM—is now seen in the use of recursive DNA shuffling, where shuffled genes from one round of selection are recombined for

further rounds, each time with an increased likelihood of finding improved variants.

In sum, REM's legacy endures through its conceptual contributions to these areas. Its core principles have permeated modern methodologies, shaping the landscape of protein engineering and biotechnology. The iterative, selection-driven approach of REM has found resonance in a multitude of successful applications, proving its lasting value in the ongoing quest to understand and harness the power of biological systems.

Future Directions and Applications

The future of protein engineering and in vitro evolution appears poised to transcend current boundaries, leveraging the foundational principles of REM to achieve unprecedented innovations.

Predictions for Protein Engineering and In Vitro Evolution:

- **Integrated Systems Biology:** We foresee a future where in vitro evolution is tightly integrated with systems biology, creating a holistic approach that considers entire biological systems for the optimization of proteins, rather than focusing on individual proteins in isolation.
- **Advanced Computational Predictive Models:** As computational power and algorithmic sophistication continue to grow, predictive models will become even more accurate, enabling the precise design of proteins with desired functions before any wet lab work is initiated.
- **AI-Driven Design:** Artificial Intelligence (AI) is expected to take a front seat in the design and evolution of proteins, with systems capable of autonomously planning and executing iterative cycles of evolution with minimal human intervention.

Potential Applications of Evolved Proteins:

- **Therapeutic Proteins:** The pharmaceutical industry will likely see an expansion in the variety and efficacy of therapeutic proteins, including enzymes, antibodies, and novel biologics for the treatment of complex diseases like cancer, Alzheimer's, and autoimmune disorders.
- **Environmental Remediation:** Proteins evolved to degrade pollutants or convert waste into valuable products could revolutionize environmental remediation and waste management.
- **Personalized Medicine:** Tailored enzymes that can specifically target disease markers or modulate individual metabolic pathways could usher in an era of personalized medical treatments.

Integration of REM Principles in Emerging Fields:

- **Synthetic Biology:** In synthetic biology, REM's principles are being integrated to construct entirely new biological circuits and pathways. Recursive optimization of these circuits could lead to robust and efficient systems with applications in bio-manufacturing, diagnostics, and therapeutics.
- **Nanotechnology:** Protein engineering will increasingly intersect with nanotechnology, where evolved proteins may serve as components of nanoscale devices or systems, offering precise control at the molecular level for applications in computing, material science, and medical devices.
- **Gene Therapy and Editing:** As gene editing technologies like CRISPR become more sophisticated, the ability to introduce beneficial proteins directly into the genome could be enhanced by REM-like approaches, enabling more effective gene therapies for a range of genetic disorders.

In conclusion, the legacy of REM is set to expand into new realms, where its recursive, iterative approach to mutation and selection will be applied across the spectrum of biotechnology and beyond. The fusion of this

approach with emerging disciplines promises not only to push the envelope in terms of what is possible in protein engineering but also to address some of the most pressing challenges faced by humanity. Whether it be combating disease, protecting the environment, or creating new materials, the principles of REM are expected to play a critical role in the innovative solutions of tomorrow.

Conclusion

The introduction of Recursive Ensemble Mutagenesis (REM) in 1992 represented a seminal moment in the field of protein engineering, marking the beginning of a new era where computational insights and experimental practice began to intertwine with unprecedented synergy. The impact of REM has reverberated through the decades, influencing not only the methodologies of protein engineering but also the philosophy underlying modern molecular evolution.

REM brought forth the concept that the vastness of protein sequence space could be navigated efficiently and systematically through recursive, feedback-driven processes. It anticipated the integration of computational models to direct and streamline experimental mutagenesis, a principle that has been central to the numerous successes of recursive approaches in molecular evolution. From the enhancement of enzymes for industrial use to the refinement of antibodies for therapeutic applications, the iterative principles of REM have underpinned advancements across the spectrum of biotechnology.

Reflecting on the success of these recursive approaches, it is evident that the iterative nature of evolution, harnessed and accelerated in the lab, has not only enabled the fine-tuning of biomolecules but also facilitated a deeper understanding of the fundamental processes that guide biological development. The successes seen in the laboratory mirror the power of

natural selection and demonstrate the potential for human-directed evolution to yield solutions to complex problems.

As we look forward, the integration of historical concepts, like those introduced by REM, into future innovations remains a guiding light. The future of protein engineering is poised to be driven by a combination of advanced computational techniques, such as machine learning and AI, alongside increasingly sophisticated experimental methods. The convergence of these tools will allow us to explore new dimensions of biological possibility, from the design of proteins with novel functions to the creation of entirely new forms of life through synthetic biology.

In essence, REM's legacy is not confined to the specific techniques or experiments it initially inspired. Rather, its enduring impact lies in the demonstration of a powerful idea: that through recursive improvement and the judicious application of selection pressure, both guided by intelligent design and analysis, we can coax the vast potential of biological systems into yielding profound and beneficial innovations. This legacy ensures that as we move into the future, the principles of REM will continue to influence the trajectory of protein engineering, shaping the endless frontier of biological discovery and application.