Abstract

The genetic toolkit at our disposal is rapidly expanding, yet the application of such tools is greatly limited to a small set of genetically tractable model organisms. One of the most significant barriers to genetic tractability in prokaryotes is the Restriction Modification (R-M) system, an innate immune system that functions to selectively degrade foreign DNA [1]. Overcoming these barriers will advance the potential for synthetic biology to address real-world problems. We developed Chameleon tools pStealth to provide an integrative solution to addressing the issue of R-M evasion by improving upon and leveraging the Stealth bioinformatic program, capable of predicting R-M target motifs in genomic sequence, with a codon optimization algorithm that takes pattern constraints (under-represented sequences) into account as to minimize occurrences in the optimized plasmid. The primary aim of Chameleon tools is to lower the overhead required for researchers to perform genetic modification experiments outside the limited scope of current model organisms, and to help enable the engineering of non-model species with desirable phenotypes.

Introduction

The presence of Restriction-Modification (R-M) systems is widespread in prokaryotes and presents one of the most significant barriers impeding genetic tractability in non-model organisms [1]. Evasion of the R-M system may be achieved by ensuring that synthetic DNA does not contain active restriction targeting sites. One approach to R-M evasion called "mimicry-by-methylation" often yields marginal improvements in transformation efficiency. The goal of this approach is to replicate the methylation profile of the target bacteria in synthetic DNA, rendering it inert to the R-M system, via the *in vitro* pretreatment of synthetic DNA with a methylase cocktail. Another approach, termed "Stealth by engineering", is achieved by synthesizing DNA engineered to be void of restriction targets, and has been shown to be highly efficient.

Nonetheless, these existing approaches hinge on the characterization of the target bacteria's R-M system- a requirement that remains unfulfilled for many non-model prokaryotes by definition. Established approaches to characterize R-M targeting profiles include the use of specialized sequencing technologies, such as Single-molecule real-time (SMRT), that reveal methylomic insight. Such technologies are not accessible to many researchers, drastically limiting the application of modern genetics in non-model organisms and presenting a potentially insurmountable road-block.

Dr. Bernick introduced us to the Stealth program, which capitalizes on the hypothesis that restriction motifs become under-represented in the genome over time. Due to the semi-conservative nature of DNA replication, there is a temporary lapse in the protective methylation of endogenous DNA as nascent DNA strands have not yet interacted with the modification component of the R-M system. This renders the unmethylated DNA vulnerable to interaction with the restriction component of the R-M system, an interaction that will induce a costly DNA damage response in the host bacterium. This results in a selective pressure against the presence of restriction motifs, favoring variants with fewer occurrences of such motifs and leading to their under-representation in the genome over time.

The original version of Stealth assesses the expected frequency of all possible motifs, using a hidden markov model based approach, and compares these values to the actual counts of motif within genomic sequence. The program then outputs a list of under-represented motifs. The researchers behind Stealth have shown an increase in transformation of over 40,000X by employing the stealth^2 by engineering approach in H. Pylori, an organism shown to resist plasmid transformation [4].

The Stealth framework provides a strong foundation for the evasion of the R-M system by revealing targeting DNA motifs. We set out to develop the Chameleon pipeline as an end-to-end solution to barriers imposed by the R-M system of a target bacteria, dependent on a basic genomic sequence of a target bacteria. The Chameleon pipeline offers sequence optimization by combining established codon optimization methodologies with the consideration of negatively weighted motifs, yielded from the Stealth program, to determine the optimal sequence. This avoidance of underrepresented motifs should increase the genetic tractability of non-model bacteria with extensive R-M systems, and in turn increase the transformation efficiency of synthetic plasmid into the live cell. Another key consideration in the avoidance of such sequences is an integrative approach to sequence optimization that considers other factors such as codon usage.

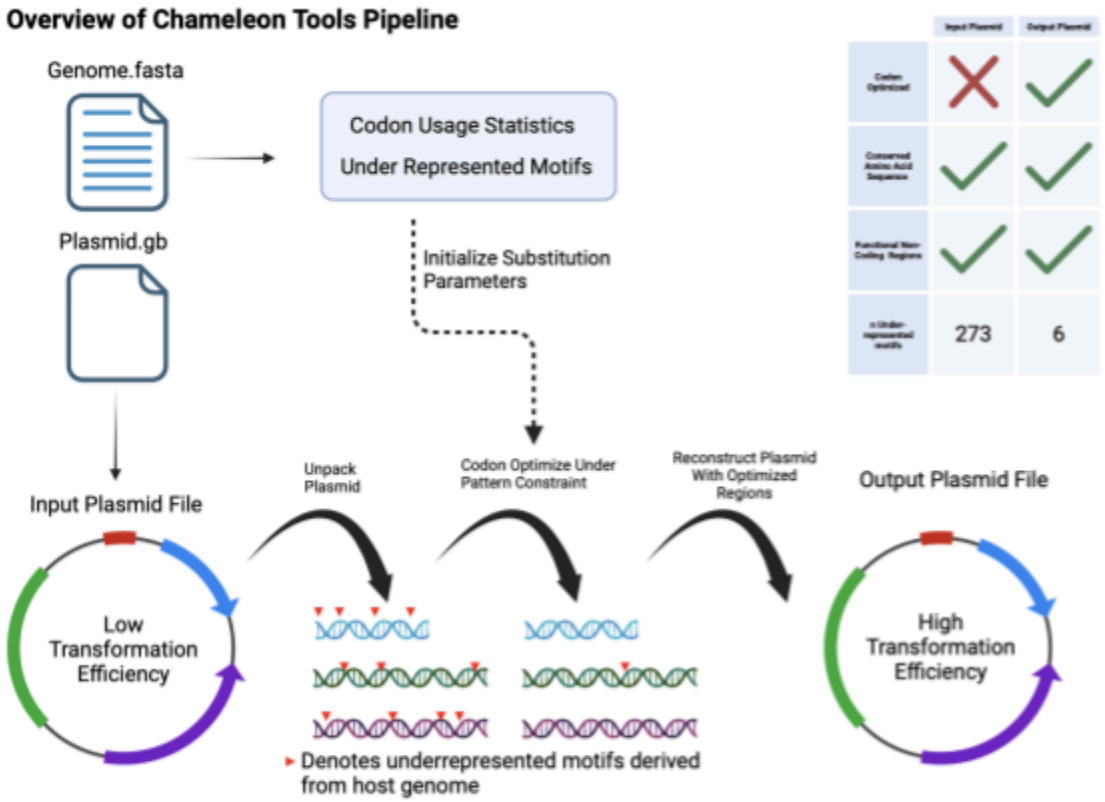

Fig. 1. An overview of the Chameleon Tools Plasmid Stealth (pStealth) Pipeline. Required input consists of a fastA format genomic sequence file and a genbank format plasmid file with annotated features. Codon usage statistics and under-represented genomic motifs are identified and collected from the genome file, and genbank annotations are used to deconstruct the plasmid so that protein coding regions can be optimized. The output consists of an optimized plasmid file that closely aligns to the hosts codon usage profile while containing a minimal number of under-represented motifs.

Significant variation in codon usage is observed between different biological species, and even strains. This leads to characteristic codon usage bias seen in specific biological contexts, which is known to play a significant role in the efficacy of heterologous protein expression. The expression levels of protein coding genes taken from one biological context and applied to another are subject to discrimination based on the degree of concordance in codon usage with the host's baseline codon usage profile. Thanks to developments in DNA synthesis, researchers are now able to apply the principles of rational design to synthetic DNA constructs in order to optimize exogenous genes for application in specific host contexts.

The engineering of a DNA sequence to shift its codon usage profile for compatibility with a specific host is known as codon optimization and has been extensively studied.

Chameleon Tools

We began development of the Chameleon toolkit to complement a synthetic biology research project after existing solutions to codon optimization under pattern constraints proved limited and difficult to integrate into an open source and cohesive pipeline flow. We wanted a comprehensive and flexible solution to DNA construct optimization, to not only enable our engineering project in non-model prokaryotes with extensive R-M systems but to build a tool that is simple and straightforward for anyone to use. This open source toolkit simplifies the task of enhancing plasmid sequences for efficient transformation and expression in non-model prokaryotes, complete with a straightforward command line interface (CLI) and accessory modules.

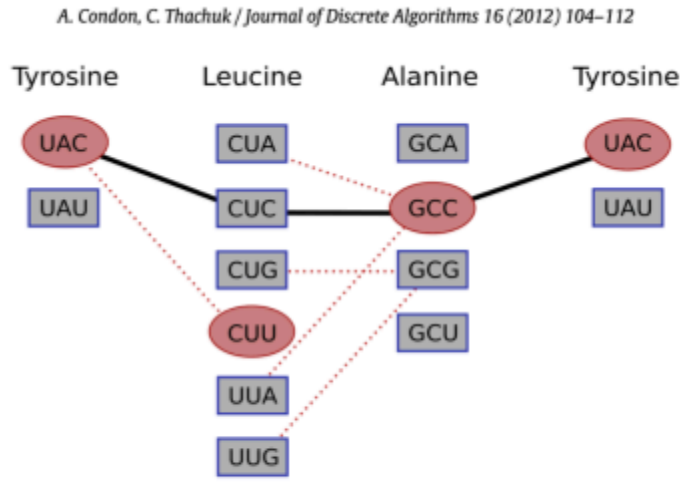

Fig. 2. Graphical overview of joint codon optimization and motif avoidance selection process. Preferred codons are indicated as red ovals, while secondary synonymous codons are shown in grey boxes. Consider a set of motifs to avoid {CCUU, AGC, UGGC}. Dashed lines indicate paths that result in occurrences of motifs, and the solid black lines represent a path that avoids all specified motifs by substituting with a less preferred codon. Note that the total occurrences of forbidden motifs in a permutation cannot be determined without a full traversal of each path. While a branch and bound approach could reduce the number of full traversals, this is negated by the application of our algorithm to a fixed window size, and the practical application of only short forbidden motifs. Graphic is repurposed from Condon et al. (2011), where they employed an alternative approach to a similar problem [3].

Several distinct codon optimization methodologies are well established and supported by extensive experimental data. While there are varying nuances in these approaches, a core foundation is to engineer synthetic constructs that reflect the hosts tRNA population in codon usage [2]. Chameleon tools identifies putative genes in the genome file by using an orf finder, however, we are working on an option that

identifies genes based on sequence annotations when available. ORFS are then assayed for codon usage statistics which are available in GCG wisconsin format for user reference. Protein coding genes are then unpacked from the plasmid file (per genbank style annotations) and passed to the codon optimization model. Our basic codon optimization methodology is designed to reconstruct protein coding genes to reflect the hosts codon usage biased through a stochastic substitution methodology where synonymous codons are selected at a rate that matches the target usage frequencies, subject to an optional minimum frequency parameter so that low frequency codons can be excluded if desired (Fig. 2.).

Results/Methods

Since the release of Chameleon Tools V1, Dr. David Bernick recognized statistical limitations associated with the hidden Markov model approach of the original Stealth program. StealthV3 was developed, integrating bootstrapping alongside a chi-squared test of independence and a false discovery rate (FDR) adjustment to address multiple hypotheses. While these statistical advancements should enhance data accuracy, their effectiveness remains unverified by empirical studies unlike the original Stealth V0. Additionally, in order to accommodate the more rigorous statistical analysis conducted in Stealth V3, two manual calibration procedures absent in Stealth V0 were introduced. These procedures involve determining the number of samples to be drawn from random sampling with replacement, and establishing an appropriate "Bootstrap Score" adjusted to a specified FDR. The bootstrapping score aggregates each motif and its frequency over 100 iterations of random sampling with replacement. Both calibration steps rely on user input, facilitated through a graphical plot generated by the program with specified parameters. Notably, the runtime of StealthV3 exhibited a substantial increase compared to StealthV0, nearly 25 times longer. The integration of our pipeline with StealthV3 was done under the guidance of Dr. David Bernick.

In addition to establishing the pipeline flow (**Fig. 1.**), two core functional requirements of the pipeline included integrating the Stealth program with our pipeline, and designing an algorithm to address codon optimization while considering the pattern constraints aspect of optimizing protein coding sequences (**Fig. 2**.).

Motif Avoidance Algorithm - ChromatoSeq

To evaluate the performance of our motif substitution and codon optimization algorithm during development, we primarily focused on comparing the performance of our algorithm with existing tools such as DNAworks. It should be noted that the complete avoidance of forbidden motifs through synonymous mutation is not always possible, and the exclusion of low-frequency codons (parameter of pStealth) provides further constraints. Our current algorithm identifies regions of protein coding genes that contain forbidden motifs through a sliding window approach, and scores each possible alternative permutation before selecting the optimal substitutions and continuing.

Motif Presence Score (M):
This is the total count of motif occurrences in a permutation.

$$M(P) = \sum_{m \in P} 1_{m \in \text{motifs}}$$

where $1_{m \in motifs}$ is an indicator function that is 1 if m is in motifs and 0 otherwise.

Entropy Score (S):
This is the adherence of a permutation to the reference sequence.

$$S(P) = \frac{1}{|P|} \sum_{i=1}^{|P|} 1_{P_i = \text{refSeq}_i}$$

where $1_{P_i = refSeq_i}$ is an indicator function that is 1 if the i-th codon in P matches the i-th codon in refSeq and 0 otherwise.

Preferential Codon Score (F):
This is the average frequency of the codons in a permutation.

$$F(P) = \frac{1}{|P|} \sum_{i=1}^{|P|} \text{frequencies}_{P_i}$$

where frequencies_{P_i} is the frequency of the i-th codon in P.

The final rank of a permutation P is then given by:

$$\text{Rank}(P) = (w_2 \cdot S(P) + w_3 \cdot F(P) - M(P))$$

where w2 and w3 are weights that variably shift the impact of each secondary criterion.

ChromatoSeq processively assays codon optimized protein coding sequences in overlapping 4 codon windows for occurrences of forbidden motifs (4-9bp). When a forbidden motif is identified, alternative synonymous permutations of the 4 codon window are constructed and scored. Because each window is a part of the whole sequence, adapter sequences are needed in order to consider edge-case motifs that are impacted by the surrounding sequence, however, these adapter regions are static for each window.

Synonymous permutations generated using all codons above the minimum usage frequency threshold and are scored based on three metrics: Motif presence score (M), Entropy score (S), and Preferential codon score (F). M provides a measure of motif occurrences and is always an integer above or equal to zero. As our primary goal is to avoid forbidden motifs, M carries the most significant weight. S is used to minimize disturbances to the previously applied stochastic codon optimization algorithm. F is similar to codon adaptation index (CAI), as its score is proportional to usage of preferred codons.

The scoring metrics are combined to rank synonymous codon permutations, the permutation with the highest rank is selected for substitution before the next 4 codon window is examined. Our current implementation has a runtime of O(m!) with m representing the number of forbidden motifs.

To further evaluate ChromatoSeq, we made use of the Codon Adaptation Index (CAI) metric. CAI is the geometric mean of the relative frequencies of codons within a gene, and is an established metric of gene fitness relating to the usage of preferred codons. There is an established correlation between relatively high CAI values and genes that are likely to be highly expressed [5]. Additionally, we use the gross number of occurrences of forbidden motifs to evaluate the success of our algorithm, as the balance between codon optimization and motif avoidance is critical. We successfully reduce occurrences of forbidden motifs in protein coding genes while simultaneously optimizing codon usage (**Fig. 3.).** Furthermore, we successfully applied an early version of the pipeline to help researchers working on engineering the non-model cyanobacteria Microcystis Aeruginosa (**Fig. 4.).**

Fig. 3. Scatter plot illustrating the change in Codon Adaptation Index (CAI) versus the percentage of motifs avoided before and after optimization. Sequence test data was generated using 500 randomly generated protein coding DNA sequences, and a list of 50 forbidden motifs were considered. The 'before' data points are represented in black while the 'after' data points are depicted in red. The shift from black to red points demonstrates the improvement in CAI and motif avoidance due to the optimization process. Note that in practice, the pStealth pipeline pre-optimizes sequences before handling motif avoidance in order to improve background codon usage (baseline codon usage in regions not impacted by forbidden motifs). Here we show that our motif removal algorithm successfully avoids forbidden motifs while improving CAI.

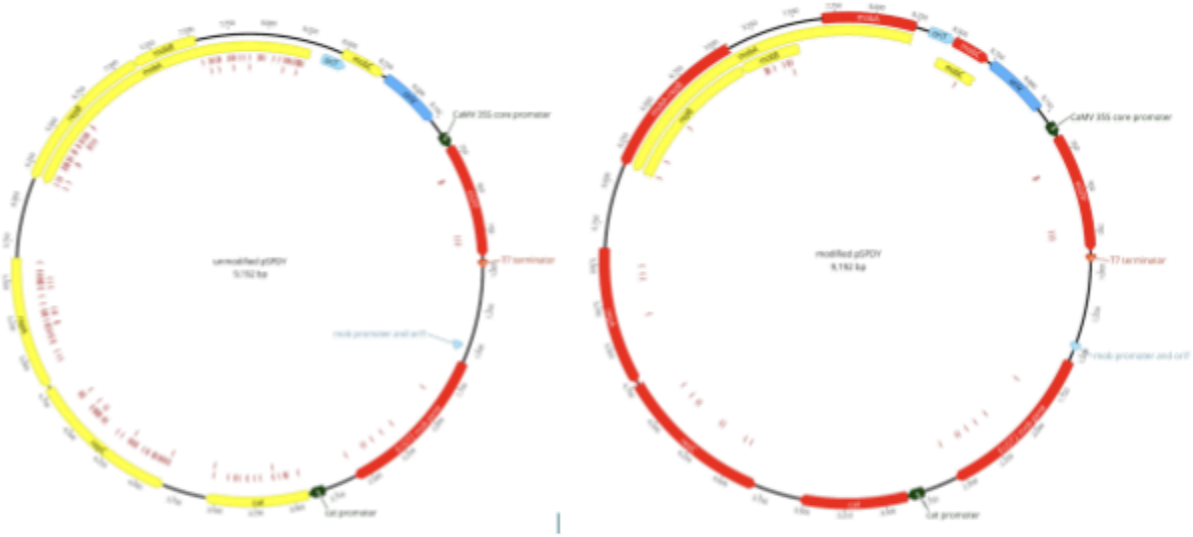

Fig. 4. Results of pStealth pipeline on pSPDY plasmid with putative restriction sites annotated. Each short pink annotation corresponds to a restriction site identified by Stealth that was matched to a sequence on the modified plasmid. After processing of the unmodified plasmid with Chameleon, the majority of restriction sites on coding regions were removed without altering corresponding amino acid sequences. 40 putative restriction sites were retained on the modified pSPDY plasmid. Annotated regions colored in red were passed through Chameleon to remove putative restriction sites. Annotated regions colored in yellow were not passed through Chameleon to remove putative restriction sites. Image made using Geneious Prime.

Stealth Optimization

Fig 5. Runtime of an unmodified StealthV3 (left) and runtime of a modified StealthV3 (right). A run-specific 36x improvement by removing overhead.

Close examination of the StealthV3 runtime during the bootstrapping process identified an expensive runtime process that could be moved to a one time overhead without affecting any final results. This change was made in coordination with Dr. David Bernick who confirmed the importance of the edit. The resulting speedup made possible by that change resulted in roughly a 10x average speedup over the original runtime, greatly improving the fluidity of Pstealth runtime with Stealth V3 (Fig 2).

FDR adjustments are now automatically calibrated, eliminating the need for the manual calibration of sampling parameters in specific genomic context. During this integration we primarily focused on automating the calibration procedure with respect to *Microcystis Aeruginosa*, and later found drops in performance on other genomes. Previously, FDR calibration involved analyzing an elbow curve to select an acceptable FDR threshold, allowing users to determine the bootstrapping cutoff score. Taking inspiration from the 'elbow-method' in K-means clustering, Dr. David Bernick suggested seeking the sampling point where diminishing returns are seen, such that increasing the bootstrapping score no longer significantly improves the FDR. This was achieved by processing a 1D numpy array containing the FDR values per bootstrap iteration to identify the local minimum of the overall curve, thereby determining the bootstrapping score at the point of diminishing returns.

Discussion

We set out to develop Chameleon Tools as a generalizable solution to restriction site avoidance and codon optimization for researchers working in non-model prokaryotes. In the broader scope, this project opens the design space for work with numerous other non-model bacteria species that would otherwise be unfeasible to engineer; Stealth and Chameleon are highly generalizable programs that could be applied in any case of an R-M system challenging transformation efficiency, which are ubiquitous among bacteria.

Our work to automate the sampling calibration of Stealth V3 has highlighted the nuances of distinct genomic context and underscores the need for adaptive sampling strategies. In the future, we aim to develop a neural network to oversee sample calibration. While a sampling size of 500,000 adequately identified motifs of 4 to 6bp in M.aeruginosa, the same sampling rate failed to produce accurate results for E.coli. Given the challenge of identifying an appropriate sampling size for each genome, the Pstealth pipeline was streamlined to focus primarily on assisting genetic modification experiments in M.aeruginosa. Additionally, attempts to explore more optimal algorithms for codon optimization while avoiding underrepresented motifs revealed that the problem is NP-complete. This complexity is illustrated in figure 2, where each codon is depicted as a node in a graph with optional edges leading to synonymous codon nodes. Akin to the NP-complete "Decision Path" problem, every path must be explored before

determining the most optimal path. This expense is negated by the fixed window size employed in ChromatoSeq, but further investigation will be done.

A primary goal of our work is to reduce experimental barriers of entry for researchers working with non-model species, and we are currently recruiting collaborators willing to try our tool. Further experimental evaluation of the underlying Stealth program is currently underway, as is that of the complete pStealth pipeline.

Bibliography

[1] Johnston CD, Cotton SL, Rittling SR, Starr JR, Borisy GG, Dewhirst FE, Lemon KP. Systematic evasion of the restriction-modification barrier in bacteria. Proc Natl Acad Sci U S A. 2019 Jun 4;116(23):11454-11459. doi: 10.1073/pnas.1820256116. Epub 2019 May 16. PMID: 31097593; PMCID: PMC6561282.

[2] 1.Al-Hawash, A. B., Zhang, X. & Ma, F. Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems. *Gene Reports* **9,** 46–53 (2017).

[3] Condon, A., Thachuk, C. (2011). Efficient Codon Optimization with Motif Engineering. In: Iliopoulos, C.S., Smyth, W.F. (eds) Combinatorial Algorithms. IWOCA 2011. Lecture Notes in Computer Science, vol 7056. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-25011-8_27

[4] Hu S, Giacopazzi S, Modlin R, Karplus K, Bernick DL, Ottemann KM. Altering under-represented DNA sequences elevates bacterial transformation efficiency. mBio. 2023 Oct 31;14(6):e0210523. doi: 10.1128/mbio.02105-23. Epub ahead of print. PMID: 37905805; PMCID: PMC10746208.

[5] Lee S, Weon S, Lee S, Kang C. Relative codon adaptation index, a sensitive measure of codon usage bias. Evol Bioinform Online. 2010 May 5;6:47-55. doi: 10.4137/ebo.s4608. PMID: 20535230; PMCID: PMC2880845.