Development of an algorithm to deplete pseudo-ITR structures from AAV transgenes

Study Project Exposé

Jessica Kranz

December 2, 2020

1 Introduction

Gene therapy represents a comparably young discipline in medicine that treats diseases by replacing damaged genes with corrected genes [Panno, 2005, p. XV]. Synthetic recombinases such as Brec1 represent a tailored tool in the set of gene therapeutic applications [Sarkar et al., 2007]. Karpinsky et al. have shown that Brec1 is able to efficiently remove HIV-specific sequences from host cell genomes from infected liver and spleen using lentiviral vectors. In the same study, the authors plan to employ adeno-associated virus (AAV) vectors in the future instead of lentiviral vectors in order to expand the cell spectrum by T memory cells [Karpinski et al., 2016].

Transgenes of AAV vectors are flanked with inverted terminal repeats (ITR)s that are conserved packaging signals crucial for encapsidation of the DNA [Zhou et al., 2017]. These ITRs are composed of stem loop-like DNA secondary structures [Cataldi and McCarty, 2013]. One obstacle in the use of the AAV system is the truncation of transgenes due to ITR like secondary structures which can serve as aberant packaging signals [Zhou et al., 2017], [Krooss et al., 2020]. An empirical depletion of these structures led to optimized AAV-packaging performance and since the experimental synthesis of the manually corrected sequence has shown that little to no truncations occur, an algorithm is required that performs the optimization automatically [Krooss et al., 2020].

2 Improving the sequence

The DNA input sequence represents the starting state of the problem. For most codons there exists a synonym, another triplet encoding for the same amino acid. This phenomenon is called degeneration of the genetic code [Watson et al., 2011, p. 570]. Therefore, at any time, there is a defined set of possible operations on a

sequence but this set of operations can be empty, which is the case for the amino acids Methionin or Tryptophan that can only be encoded by a unique nucleotide triplet [Watson et al., 2011, p. 571]. To determine the possible operations, the nucleotides are analyzed triplet-wise [Watson et al., 2011, p. 579].

According to the proved working assumption of Krooss et al., supported by Huang and Nair, the following adaptions need to be applied to the transgene when packed in an AAV vector [Krooss et al., 2020], [Huang and Nair, 2017]. For all nucleotides at a distance of up to 30 nucleotides before and after the currently considered triplet, the following must apply: there are no inverted repetitions of the nucleotide sequence which could potentially form a stem loop with a stem longer than six nucleotides. For the whole sequence, the following must apply: no potential terminal resolution sites (5'-TTGGCC-3 'or 5'-GGCCAA-3') and no rep binding site sequences (5'-CTTTG-3') occur. The application of the specified conditions results in a sufficiently good sequence. A further reduction of the length of possible ITRs is not a further improvement, so that there is not one optimal solution, but a lot of sufficiently good solutions. Further, solving the decision problem "Is it possible to calculate a satisfying solution?" is not of significant relevance. Instead, the calculating power should be invested in finding a solution that satisfies the above mentioned requirements. If a sufficiently good solution cannot be found in the given time, the best solution up to that point should be returned together with an appropriate indication.

3 Related Work

The idea of modifying mRNA so that less secondary structures are formed is not new. Gaspar et al. describe an algorithm capable "[...] of avoiding stable secondary structures in mRNA molecules by means of maximizing the Maximum Free Energy (MFE) of the nucleotide sequences, without changing the resulting amino acid sequence." As Gaspar et al. claim, analysis of sample sequences, optimized with the developed algorithm, show an increase of $> 40\%$ in MFE, "[...] strongly reducing the strength of secondary structures, in only a few seconds" [Gaspar et al., 2013]. From a conceptual point of view, this existing algorithm pursues the same goal as the requiered algorithm should. Nevertheless, the achieved optimization does not satisfy the given requirements. The existing algorithm does not restrict the number of nucleotides in a secondary structure to a value below seven or does not restrict it at all. For this reason, there is a need for further investigations at this point.

4 Approach

For a first complexity assessment, the properties of the primary target sequence Brec1 are described. The open reading frame of Brec1 consists of 1032 nucleotides representing 344 amino acids [Bessen et al., 2019]. Brec1 contains 40 potential inverted terminal repeats with a stem length of at least 7 nucleotides

where the loop consists of 0-10 nucleotides, considering only those inverted repeats containing 0-1 missmatches (calculations based on [Bessen et al., 2019]]. Those are to be depleted using a skilful approach, so that few side effects occur. Concerning codon replacement it has to be considered, that on the one hand the replacement of codons of neighbouring secondary structures can cause consequential effects and on the other hand the neighbouring nucleotides can form a new secondary structure even without existing inverted repeats. This is exactly what happened in the manual correction process [Krooss et al., 2020].

There are essentially two different approaches to solve this. The first possibility is the deterministic calculation of all possible combinations with subsequent evaluation of each generated sequence. Another possibility is to change the sequence in a targeted manner, with reassessment of the sequence after each intermediate step. Based on the evaluation of the quality of a solution, it would then be possible to combine advantageous properties and, with additional variance introduced, to produce better solutions in almost every iteration.

The deterministic approach is supported by the fact that - if available - a sufficiently good sequence must be included in the set of generated sequences, since every possible combination is taken into account. This enables a clear statement to be made as to whether a sufficiently good sequence exists. In addition, this approach can be easily parallelized, since the calculation of the combinations is completely independent. Since this approach is untargeted, the amount of sequences to construct will be large, especially when there are little or no solutions.

The targeted replacement calculates fewer combinations and is therefore better suited if the number of possible combinations is too high to calculate every possible representation. Creating sequences without re-evaluation saves computational effort, but the re-evaluation after each step also needs to be considered. The possiblilty to parallelize is limited, since the next generation can only be calculated once the current sequence has been evaluated.

In the scope of this project, the deterministic solution is to be pursued first. From the perspective of the researchers involved, it is sufficient to be able to use the simplest possible algorithm. In order to gain experience with the properties of the data, it is advantageous to choose the approach that is best understandable and that carries out the same steps in each execution and comes to the same result. The calculation time is of secondary importance.

If it turns out that the deterministic approach is not well suited to solve the problem or that a targeted change can bring considerable advantages, then the approach can be developed step by step towards the variant of targeted replacement.

In both cases, built sequences must be evaluated against the defined criteria. This evaluation function should consider ITR-like structures, RBS and TRS sequences and weight them accordingly. The function may be calibrated using a fully optimized and a secondary structure-heavy sequence.

References

[Bessen et al., 2019] Bessen, J. L., Afeyan, L. K., Dančík, V., Koblan, L. W., Thompson, D. B., Leichner, C., Clemons, P. A., and Liu., D. R. (2019). High-resolution specificity profiling and off-target prediction for site-specific dna recombinases. Online available at https://www.nature.com/articles/s41467-019-09987-0; accessed on 02nd November 2020.

[Cataldi and McCarty, 2013] Cataldi, M. P. and McCarty, D. M. (2013). Hairpin end conformation of adeno-associated virus (aav) genome determines interactions with dna repair pathways. Online available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3578132/; accessed on 02nd November 2020.

[Gaspar et al., 2013] Gaspar, P., Moura, G., Santos, M. A. S., and Oliveira, J. L. (2013). mrna secondary structure optimization using a correlated stem–loop prediction. Online available at https://academic.oup.com/nar/article/41/6/e73/2902446; accessed on 5th October 2020.

[Huang and Nair, 2017] Huang, Z. and Nair, M. (2017). A crispr/cas9 guidance rna screen platform for hiv provirus disruption and hiv/aids gene therapy in astrocytes. Online available at https://www.nature.com/articles/s41598-017-06269-x; accessed on 21st April 2020.

[Karpinski et al., 2016] Karpinski, J., Hauber, I., Chemnitz, J., Schäfer, C., Paszkowski-Rogacz, M., Chakraborty, D., Beschorner, N., Hofmann-Sieber, H., Lange, U. C., Grundhoff, A., Hackmann, K., Schrock, E., Abi-Ghanem, J., Pisabarro, M. T., Surendranath, V., Schambach, A., Lindner, C., van Lunzen, J., Hauber, J., and Buchholz, F. (2016). Directed evolution of a recombinase that excises the provirus of most hiv-1 primary isolates with high specificity. Online available at https://www.nature.com/articles/nbt.3467; accessed on 29th October 2020.

[Krooss et al., 2020] Krooss, S., Vu, X.-K., Hauber, J., Bohne, J., Ott, M., and Büning, H. (2020). Generation of a computer-based algorithm for codon optimization to deplete pseudo-itr structures from aav transgenes 'itrex'. unpublished.

[Panno, 2005] Panno, J. (2005). *Gene Therapy*. The new biology, United States of America.

[Sarkar et al., 2007] Sarkar, I., Hauber, I., Hauber, J., and Buchholz, F. (2007). Hiv-1 proviral dna excision using an evolved recombinase. Online available at https://science.sciencemag.org/content/316/5833/1912; accessed on 02nd November 2020.

[Watson et al., 2011] Watson, J., Baker, T., Bell, S., Gann, A., Levine, M., and Losick, R. (2011). *Watson Molekularbiologie*. Pearson Studium, Munich, Germany.

[Zhou et al., 2017] Zhou, Q., Tian, W., Liu, C., Lian, Z., Dong, X., and Wu, X. (2017). Deletion of the b-b' and c-c' regions of inverted terminal repeats reduces raav productivity but increases transgene expression. Online available at https://www.nature.com/articles/s41598-017-04054-4; accessed on 02nd November 2020.

Materials

pCMV-Brec1 was a gift from David Liu (Addgene plasmid # 123135; http://n2t.net/addgene:123135; RRID:Addgene_123135)