eurofins | Genomics

Eurofins' adaption and optimisation software "GENEius" in comparison to other optimisation algorithms

Uwe Köhler§, Sebastian Kubny, Julia Wiesemann
Gene Synthesis Department of Eurofins Genomics, Ebersberg

Eurofins Genomics' Gene Synthesis Service offers optimisation of coding gene sequences for improved expression in heterologous organisms. For this challenging task a tailor-made adaption and optimisation software has been developed in close cooperation with our bioinformatics partner BioLink GmbH. Over ten years of experience in gene synthesis and protein expression as well as valuable information from several publications was employed to develop our sequence optimisation algorithm. The resulting software is called "GENEius". It is an intelligent optimisation software with enhanced functions that enable us to design gene sequences for the best protein expression in different organisms.

Introduction

Most gene synthesis companies employ their own algorithm for codon usage adaption and we wanted to determine how well GENEius performs in comparison with these different software packages. We therefore asked five main competitors for optimisation of the jellyfish *Aequorea victoria* wild-type GFP for best expression results in *E.coli*. All but one competitor provided the optimised sequence and we then synthesised the genes ourselves. One competitor, however, did not provide the optimised sequence until we confirmed the gene synthesis order. We nevertheless synthesised the gene ourselves and were especially interested about expression results of this gene.

How does GENEius work?

During optimisation with GENEius the software randomly assembles the DNA sequence and then analyses it in relation to codon usage by comparing it to an input codon usage table. This input codon usage table is usually taken from the Kazusa Codon Usage Database (http://www.kazusa.or.jp/codon) but it can also be provided by the customer. Currently codon usage tables of over 35,000 organisms can be found in the Kazusa database. GENEius does not simply aim for a high codon adaption index (CAI), instead it harmonises the codons used. Frequently used codons from highly expressed genes are more often used in the resulting gene sequence than less frequently used codons. Very rare codons, however, will be completely avoided. During adaption GENEius also checks for "bad motifs" like restriction sites and avoids artificial splice sites, unspecific transcription factor binding sites, etc. Also, to minimise RNA structure direct and inverted repeats are avoided as they not only make synthesis more difficult, they can decrease DNA stability and reduce efficiency of transcription and translation in *E.coli*. And last but not least, the GC content is equally distributed to avoid GC-rich subsequences within in the gene. All these parameters are taken into account and a score is constantly being calculated. If this score falls below a certain threshold, the sequence is taken as the output. This procedure results in a different DNA sequence every time the optimisation is running. Therefore, THE best optimised sequence does not exist. As most amino acids are encoded by more than one triplett, the number of possible DNA sequence versions for a single protein sequence is very high. Two amino acids are encoded by only one codon (Met and Trp), all other amino acids are encoded by more than one codon. E.g. Cys, Lys or Asn are encoded by two triplets, Ile

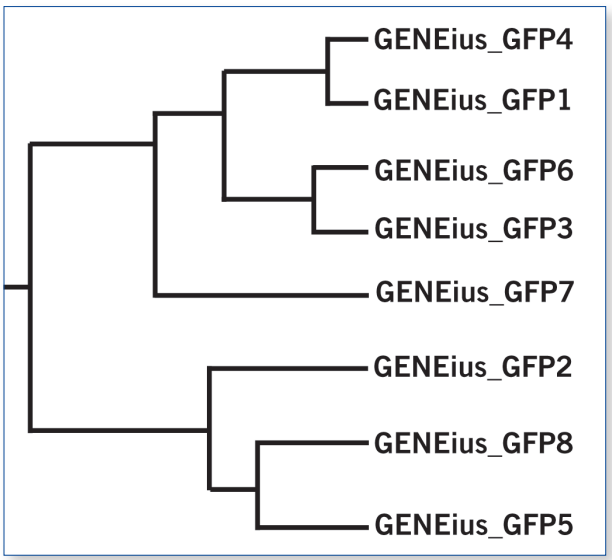

Fig. 1 Dendrogram with 8 different GENEius-optimised DNA sequence versions of GFP. Version GENEius_GFP4 and GENEius_GFP5 have been chosen for gene synthesis and were used in the expression experiments

is encoded by three triplets. Ala, Gly or Thr are encoded by four codons and there are three amino acids being encoded by six different triplets: Arg, Leu and Ser. On average an amino acid is encoded by 3 triplets, therefore a protein of 100 amino acids can be encoded by 3^{100} (= 5.2 x 10^{47}) different DNA versions!
We let GENEius run eight times to optimise the Aequorea victoria wild-type GFP and got eight different DNA sequences. With these eight sequences a dendrogram has been created using the programme CLUSTALW (http://www.genome.jp/tools/clustalw) and we chose the two most distantly "related" GFP versions to be synthesised. These were sequences GENEius_GFP4 and GENEius_GFP5 in Figure 1.

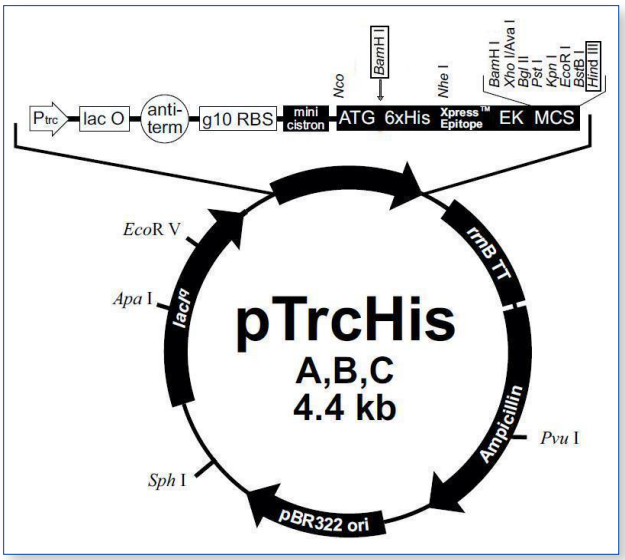

Fig. 2 Vector map of a modified pTrc that was used for expression of GFP versions

For subcloning a pTrc vector (Invitrogen) has been used for the two GENEius versions as well as for the 5 competitor's versions. We introduced a second BamHI site directly downstream of the ATG (see figure 2) and then subcloned all genes via BamHI and HindIII. Thereby the His-Tag, Xpress Epitope and enterokinase cleavage recognition sequence (EK) as well as the other restriction sites from the multiple cloning site of pTrc were removed.

Results

We asked competitors to adapt and optimise the wild-type GFP from jellyfish for optimal expression in *E.coli* avoiding cloning sites BamHI and HindIII. Once we had the DNA sequences we synthesised those genes and subcloned them into the modified pTrc. *E.coli* TOP10 cells were used in the expression experiments. Induction was done with 1 mM IPTG final concentration after growth at 37 °C with 150 rpm until OD_{600} was between 0.4-0.6. After induction, growth was continued for 6 hrs at 25 °C followed by 4 °C over night incubation. This was necessary for proper folding of the protein as wild-type jellyfish GFP has been used. Fluorescence of normalised *E.coli* cultures was measured with a Hitachi F2500 Fluorescence Spectrometer.

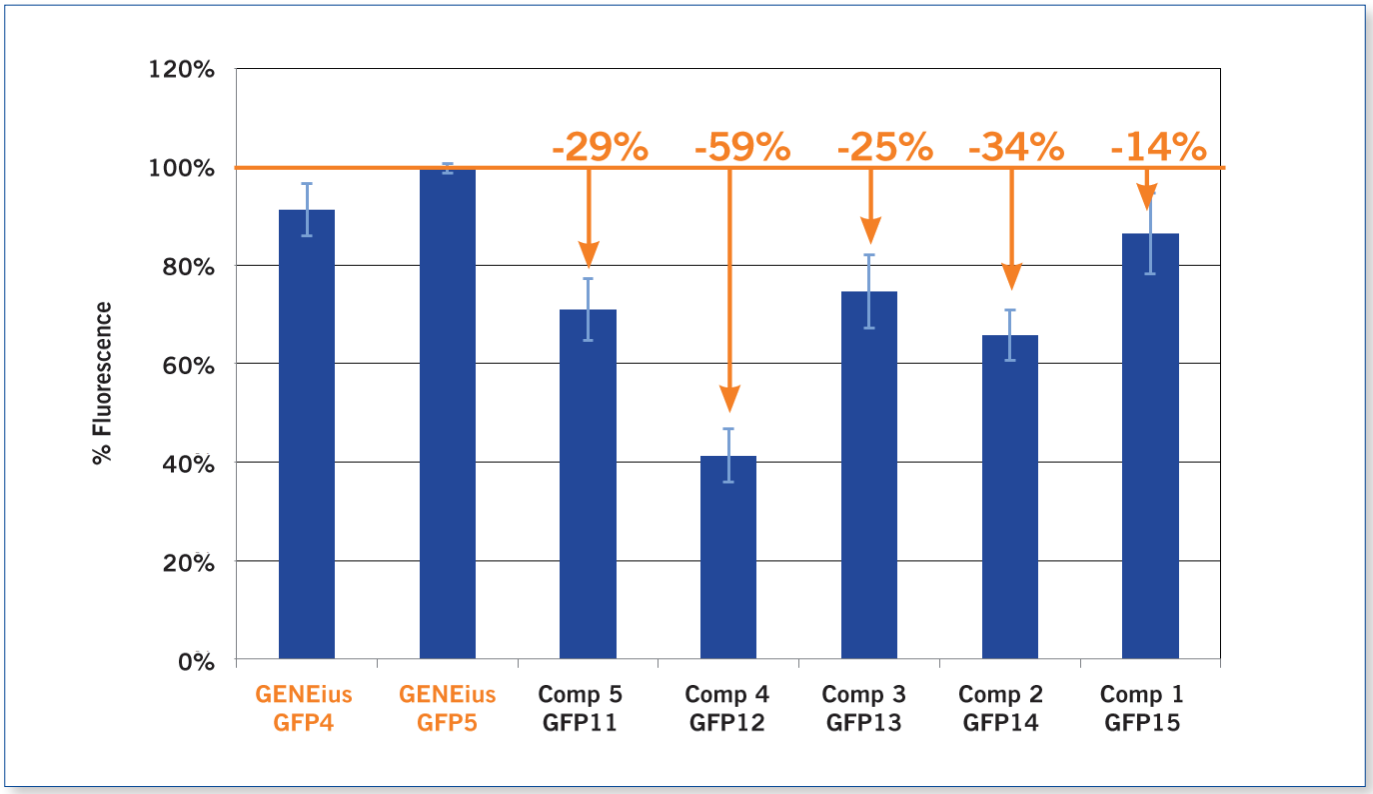

Fig. 3 GFP Fluorescence of optimised constructs

In earlier expression experiments the two GENEius optimised versions resulted in very similar expression levels (publication in preparation), therefore we did not expect to see huge differences between fluorescence of the GENEius versions and the optimised versions from the competitors. 10 independent expression experiments have been performed and average fluorescence values are shown in figure 3.

The construct with strongest expression was the GENEius optimised version GFP5. The fluorescence value of this construct was set to 100% and the other values were compared to this best expressing gene. The second GENEius version GFP4 resulted in the second strongest fluorescence. Fluorescence of competitor's constructs was between 14% and 59% lower.

Conclusion

This shows that our proprietary software GENEius is very well suited for codon usage adaption and optimisation of gene sequences to result in very high protein expression in *E.coli*. It is very likely that other genes optimised for expression in *E.coli* will also result in high protein expression, and we know from our customers that GENEius optimised genes express very well in other expression systems like mammalian cells, insect cells, yeast (S. cerevisiae and P. pastoris) and plants (monocots and dicots). It is also possible to optimise genes for expression in two different hosts. For this we employ proprietary codon usage tables e.g. for high expression of one gene in both mammalian and insect cells.

How to use GENEius for your project

GENEius is linked to our Ecom ordering system. When you choose codon usage adaption and optimisation of your gene sequences you can select the input codon usage table of your expression host and choose "bad motifs" like your cloning sites that will be excluded during adaption. You can even create your own bad motifs, e.g. transcription binding sites, artificial splice sites or polyadenylation signals. These sequence motifs will not be present in the optimised DNA sequence and therefore will not interfere with your downstream experiments

Contact (§)

Dr. Uwe Köhler
Eurofins Genomics
Head of Gene Synthesis
Email: uwekoehler@eurofins.com

GENEius software is designed and developed for Eurofins Genomics by BioLink Informationstechnologie GmbH.