1. Purpose
Our primary aim is to provide a standard format for protein-coding BioBricks intended for eukaryotes, especially in the nematode worm Caenorhabditis elegans, so that future teams will be able to correctly take advantage of the Kozak consensus sequence and its power to manipulate gene expression levels. There are currently, in fact, at least three protein-coding BioBricks with Kozak sequences that correctly conform to this convention, designed for yeast, (BBa_K165057, BBa_K165058, and BBa_K165059) but their format has not been documented or standardized. Since this is so important to full exploitation of current and future eukaryotic chassis, we feel it must be codified.

2. Relation to other BBF RFCs
Assumes use of BioBrick assembly standard 10. No known conflicts with other RFCs currently published.

3. Copyright Notice
Copyright (C) The BioBricks Foundation (2010). All Rights Reserved.

4. Background
Gene expression in eukaryotes as diverse as yeast and humans is greatly aided by the presence of a Kozak consensus sequence, named after Marilyn Kozak, who has extensively studied translational initiation since the 1980s. A Kozak sequence should not be thought of as analogous to the Shine-Dalgarno (SD) sequence, which is used by prokaryotic ribosomes as a binding site on the mRNA. This is because in eukaryotes, ribosomal binding occurs upstream of the Kozak sequence, at a protein complex bound to the 5' terminus of the transcript. [4] By contrast, since the prokaryotic ribosome binds directly at the SD sequence, it is not prevented from initiating translation by upstream barriers.
Our understanding of the Kozak sequence comes from induction: computational and comparative analyses of nucleotide conservation surrounding the start codon, and comparing the expression levels of modified Kozak sequences. [2] As such, it is not necessary for precise homology, and a wide variety of inferior Kozak sequences are possible, each with its own expression strength. Below is an example of a sequence often cited [6, 2] as the optimal Kozak sequence:

GCC GCC ACC AUG G

In Kozak’s studies of preproinsulin [2], the final G can improve expression by up to 1.5 times. We have not chosen to incorporate this into the standard Kozak sequence as it may not always be possible (if the protein coding sequence does not begin with a G). Previous recommendations have included the forcible inclusion of this G by placing the codon GAG after the start, but this is not advisable as the resultant glutamate residue could have significant impact on protein folding and functioning of the amino terminus. The three bricks mentioned in section 1 conform to this. Many proteins include this G anyway, as shown below in Figure 1.

![Figure 1. Sequence conservation around the start codon in human mRNAs. From Wikipedia [6].](image)

## 5. Relevance

As can be seen from Figure 1, the Kozak sequence is an environment for the start codon, and so cannot be separated from AUG. The typical BioBrick definition of a protein-coding brick [1], used for most prokaryotes, begins with the start codon and ends with the stop codon. This is less useful for C. elegans and other eukaryotes, as it wouldn’t allow for the installation of a Kozak sequence by three-way ligation. This is because the ligation assembly scar, from, for example assembly standard 10 [3], would separate the first portion of the Kozak sequence (hereafter simplified as ACC) from the second part of the Kozak sequence (AUGG) rendering it ineffectual. To permit the Kozak, we had to establish a new standard for the definition of a

---

1 The first A in the sequence is sometimes shown as “G or A”, but using an A can be up to 3 times the strength of a G in this position.
protein-coding BioBrick for use in *C. elegans*, which dictates that the part should start at the beginning of the Kozak sequence, and end with the stop codon.

The effect of a well-formed Kozak sequence can be dramatic. Kozak showed [2] in 1984 that even single nucleotide changes can cause a fifteenfold increase in gene expression in certain genes.

The Kozak sequence has not been well-understood in BioBrick-based synthetic biology. The Parts Registry incorrectly lists the Kozak sequence as a direct analog of the prokaryotic ribosomal binding site, and misleadingly includes separate BioBricks containing full Kozak sequences. If used as presented, these BioBricks would create a separate open reading frame (ORF) that starts in the Kozak BioBrick and ends immediately afterward in the ligation scar. This ORF would have extremely high translational efficiency and greatly reduce expression of the actual intended coding region.

In eukaryotes, instead of binding to a specific DNA sequence, the ribosome latches onto a structure of cofactor proteins that have bound to certain features at the start of the transcript, and then scans down the mRNA until it finds a start codon. [4] A ‘naked’ start codon is sufficient to start translation, but it is very weak, and it is likely that the ribosome will skip it over and not express the gene. This is exploited in naturally-occurring eukaryotic genomes to reduce expression of a protein by preceding it with short ORFs called uORFs, much like the situation outlined above.

### 6. Standard

We propose that future preparations of protein-coding BioBricks for use in eukaryotes **SHOULD** include the following sequence immediately preceding the start codon:

```
GCC GCC ACC
```

This sequence will provide strong gene expression, and is short enough for labs to modify it using PCR to reshape it into a weaker consensus sequence. Our second reference [2] gives some examples of how each nucleotide affects expression strength.

It should be noted that this alters the promoter as commonly used in *C. elegans* research. The start codon should be removed from the promoter segment. This is standard for BioBricks in the parts registry, but may appear unusual to researchers used to building fusions. The ligation scar ends up in the 5' untranslated region (UTR).

### 7. Authors’ Contact Information

Geoff Halliday: geoff.halliday@qgemteam.com
Chris Palmer: chrispalmer@qgemteam.com
8. References

   <http://partsregistry.org/Protein_coding_sequences>


   <http://en.wikipedia.org/wiki/Kozak_consensus_sequence>