# Codon Usage Bias in the fire ant, *Solenopsis invicta*

Background: Ants are important model species for sociobiology and behavioral ecology. Life in an ant colony is marked by cooperation, but it also entails conflicts. Both aspects have been studied extensively to understand the prerequisites for social behavior and to test the kin selection theory. New molecular and genomic techniques are making it possible to identify the genes underlying social behavior in ants and other social insects, as well as the mechanisms behind self-organization, life-history evolution, division of labor, and developmental plasticity.

UNIL researchers created the first genome-scale library of ant DNA by sequencing 21,715 cDNA clones from the fire ant *Solenopsis invicta* [Wang *et al*, 2007], which assemble to 12,000 putative messenger RNA sequences.

In this project, we propose to scan the entire library of ant sequences for Codon Usage Bias (CUB). CUB is the unequal use of synonymous codons in coding DNA sequence (if for example the amino-acid Tyrosine is frequently encoded by UAC and only rarely by UAU). CUB is linked to the relative abundance of tRNAs: codons for abundant tRNA are translated faster and more accurately than codons for rare tRNAs. Thus, natural selection acts on CUB to increase speed and accuracy. The strength of selection on any given gene can be measured using the Codon Adaptation Index (CAI) [Sharp & Li, 1987].
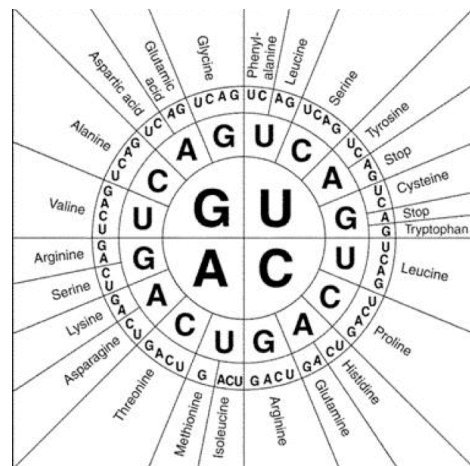
In this project, our aim is to identify the ant genes with the strongest and weakest CAI. Genes with strong CAI are under selection and may have been permissive in the evolution of fire ant-specific biology. On the other hand, genes with the weakest CAI may be "young" in fire ants.
Both categories of genes will subsequently be investigated (beyond the scope of this programming project).

Outline: First, the CUB of every fire ant gene will be calculated, requiring identification of the protein-coding sequence of each DNA sequence. Subsequently, the "reference" CUB for the fire ant genome can be calculated. Finally, using the above information, we will calculate the CAI for each fire ant gene.

Inputs:
    a. A FASTA file containing DNA sequences of assembled *Solenopsis invicta* cDNA sequences.
        o *Note: sequences may contain 3' and 5'UTRs (untranslated regions), and some will contain no protein-coding regions.*
    b. A BLASTX report file describing the best alignment ("hit") between each ant sequence from **a.** and the non-redundant protein database.

Outputs:
- A table summarizing "reference" CUB in fire ants
- A table indicating CAI for each fire ant gene

Detailed project: The project includes the following steps:
- Using **b.**: For ant sequences that have BLASTX hits, identify the reading frame.
- Using the sequence in **a.**, extend the reading frame in both the 3' and 5' directions, to respectively a start codon (AUG) and a stop codon (UAG, UGA, or UAA). Keep this "minimal open reading frame sequence (ORF)".
  - *Note: If the reading frame is negative, you will need to use the EST's reverse-complement.*
  - *Note: to obtain mRNA sequence, cDNAs were synthesized, thus Us are replaced by Ts.*
- *Optional: For ant sequences that have no BLASTX hit, identify the largest open reading frame (ORF), but only keep it if that ORF is longer than 50 codons.*
- Each ORF should be translated into the predicted protein sequence, and the codon usage counts recorded.
- Summarize CUB of all genes into a single reference table (similar to the RSCU column in Table 1 of Sharp & Li 1987).
- For each gene, calculate the Codon Adaptation Index [Sharp & Li 1987].

Supervisor: Yannick Wurm (Keller Lab, Department of Ecology and Evolution)

References:
Wang J, Jemielity S, Uva P, Wurm Y, Graff J, Keller L, 2007. **An annotated cDNA library and microarray for large-scale gene-expression studies in the ant *Solenopsis invicta*.** Genome Biology 8(1), R9. [Pubmed]

Vicario S, Mason C, White KP, Powell JR, 2008. **Developmental Stage and Level of Codon Usage Bias in *Drosophila*.** Molecular Biology and Evolution 25(11) [Pubmed]

Sharp PM, Li WH, 1987. **The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications.** Nucleic Acids Research 15(3) [Pubmed]