Bioinformatics of the Green Fluorescent Proteins

Dr. Alma Rodriguez-Estrada (arodriguezestrada@aurora.edu), Department of Biology, Aurora University

ABSTRACT

Transformation of *Escherichia coli* with the Green Fluorescent Protein (GFP) is a laboratory activity that has become increasingly popular at different levels ranging from middle school to undergraduate courses (pGLO™ Bacterial Transformation, BIO-RAD). There are several mutant GFP genes that encode mutant proteins with small differences in their nucleotide and amino acid composition. This bioinformatics activity was designed for an upper level course in molecular biology. Through this activity, students learn how to use the GenBank® (Basic Local Alignment Search Tool, BLAST®) and the Protein Data Bank (PDB) to analyze the gene and amino acid sequences of different GFP variants while reviewing general concepts of gene and protein structure in addition to primary literature.

THE GREEN FLUORESCENT PROTEIN (GFP)

The Green Fluorescent Protein (GFP) naturally found in *Aequorea victoria* (jellyfish) is a protein that produces glowing light in the umbrella margin of jellyfish. The GFP emits light in response to energy transferred by aequorin, a calcium-activated protein also found in this organism. The wild-type GFP contains 238 amino acids and has a molecular weight of 26.9 kilodaltons (kDa). The protein is composed of 11 β-sheets that form a barrel-like structure (24 Å diameter and 42 Å height) and an α helix that runs diagonally through the barrel (Zimmer 2002). Additional short helical sections form lids on both ends of the barrel. The chromophore is the structure that confers fluorescence to the protein. This structure is buried in the center of the barrel and is joined to the barrel by an α-helix (Fig. 1). The chromophore is composed of three amino acids. In the wild-type, the amino acid composition is 65Ser-Tyr-Gly67. The chromophore forms through an internal posttranslational autocatalytic cyclization where no cofactors are needed and only the presence of oxygen is necessary. The gene that encodes for the wild-type GFP is 966 bp (Zimmer 2002).

The green-fluorescent protein has been used in a wide variety of applications and to study protein properties, behaviors and cellular processes. Since this protein is found in jellyfish that lives in the cold Pacific Northwest, the native GFP efficiently folds and produces luminescence at temperatures lower than 37 °C. Thus, mutant proteins that efficiently fold at higher temperatures have been developed and are called folding mutants. Mutations locate close or far from the chromophore, buried or on the surface of the barrel structure (Zimmer 2002).

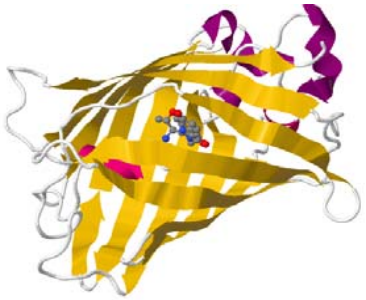

Figure 1. Structure of the wild-type GFP found in *Aequorea victoria*. Image from the RCSB PDB (www.rcsb.org) ID 1EMB (Brejc, K., Sixma, T. Green Fluorescent Protein from *Aequorea victoria*, GLN 80 replaced with ARG).

OBJECTIVES:

1. Learn to use the Basic Local Alignment Search Tool (BLAST®) and the Protein Data Bank to determine the GFP gene sequence and protein structure
2. Characterize the wild-type GFP and two mutants
3. Review concepts related to the central dogma of molecular biology (transcription and translation)
4. Review the use of the genetic code table

BIOINFORMATICS OF THE GFP

The laboratory activity "Bioinformatics of the Green Fluorescent Proteins" was designed for an upper level course (molecular biology) where students learned how to use the Basic Local Alignment Search Tool (GenBank, BLAST®) and the Protein Data Bank (PDB) to analyze the gene and amino acid sequences of different GFP variants (Benson et al. 2005, Berman et al. 2000). Step by step instructions and related questions were provided prompting students to explore the information available in the websites and to carefully analyze protein structure. For instance, students first explored the primary, secondary and tertiary structure of the wild-type protein (1EMB) (Fig. 1 and 2). From figure 1, the following information can be determined: number of beta-sheets (11) and short helices (6), position of the chromophore (65-67, represented with an X) and number of amino acids residues (238).

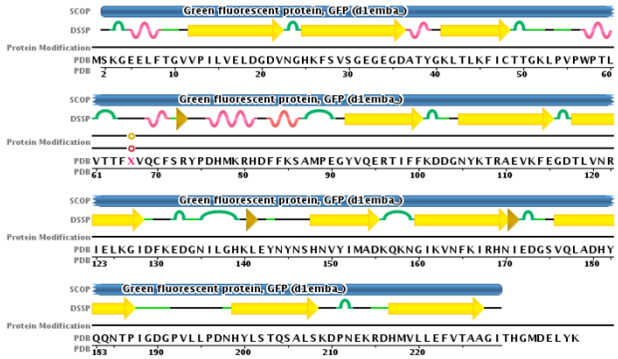

Figure 2. Amino acid sequence and secondary structure of the wild-type GFP. Image from the RCSB PDB (www.rcsb.org) ID 1EMB (Brejc, K., Sixma, T. GFP from *Aequorea victoria*, GLN 80 replaced with ARG).

The amino acid sequence was then downloaded as a FASTA file for further analysis (FASTA files open in Notepad and can be exported in word). Table 1 shows a simple comparison between the amino acid sequence between the wild-type and the mutant protein (cycle 3 mutant) used in the pGLO™ transformation experiment (BIO-RAD). The amino acid residues that are modified in the cycle 3 mutant can be identified in the amino acid sequence of the protein using the "find" and "word count" function in word (Fig. 3).

Table 1. Comparison between the wild-type GFP and the cycle 3 mutant protein.

Mutation Positions	Amino acid in the wild type GFP	Amino acid in the cycle 3 mutant GFP
100	Phenylalanine	Serine
154	Methionine	Threonine
164	Valine	Alanine

MSK…..TIFFK….YIMAD….IKVNF…..ELYK

Figure 3. Amino acid sequence of the wild-type GFP. The amino acid residues substituted in the cycle 3 mutant are highlighted in yellow.

BLAST can then be used to find the gene sequence of the GFP. **Tblastn** is used to find the translated nucleotide sequence from the amino acid composition of a protein (limit the BLAST search to the genus *Aequorea*). Accession U73901.1 shares 99% of sequence identity and the lowest E value. Thus, this entry can be used for further analysis. From the nucleotide sequence (downloaded as FASTA file), the following can be identified: **start codon** (ATG), **stop codon** (TAA), **length of the coding region** (717 bp), **chromophore codons** (GGT TAT GGT) and the **amino acid residues forming the chromophore** (Gly-Tyr-Gly).

AAGCTTTATTAAAATGTCTA……..TTTCGGTTATGGTGTTC……..AAATAACTGCAG

Figure 4. Nucleotide sequence of the mutant 3 GFP (GenBank: U73901.1). The start and stop codons are in blue and red, respectively. The codons for the amino acid residues that form the chromophore are shown in green.

The nucleotide and amino acid sequences above do not correspond to the wild-type where the chromophore is formed by Ser-Tyr-Gly. Accession U73901.1 corresponds to mutant 3, a yeast-enhanced GFP that fluoresces at 488 nm rather than 398 nm as the wild-type (Cormack et al. 1996). A wrap up activity included the review of primary literature in order to elucidate the full range of diversity and applications of the fluorescent proteins and the logic behind the need to develop mutants.

CONCLUSIONS

This bioinformatics activity can be completed in a single laboratory session of 170 minutes or less before or after the wet laboratory (bacteria transformation and electrophoresis of the GFP). If done before any of the wet labs, this activity is an excellent opportunity for inquiry since no background information will prompt students to characterize the GFP based on pure exploration of the nucleotide and amino acid sequences available in the PDB and GenBank websites. A comparison of the wild-type protein, its mutants and other fluorescent proteins could be performed using specialized software such as MEGA.

References:

- Benson D.A., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D.L. 2005. GenBank. Nucleic Acids Research 33:34-38 (GenBank; https://www.ncbi.nlm.nih.gov/genbank/)
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N. Weissig, H., Shindyalov, I.N. and Bourne, P.E. 2000. The Protein Data Bank. Nucleic Acids Research, 28: 235-242 (PDB; http://www.rcsb.org/pdb/)
- Cormack, B. P., Valdivia, R. H. and Falkow, 5. 1996. FACS-optimized mutants of the green fluorescent protein (GFP). Gene 173:33-38
- Zimmer, M. 2002. Green Fluorescent Protein (GFP): Applications, Structure and Related Photophysical Behavior. Chem. Rev. 102: 759-781.
- No Author. BioRad. Biotechnology Explorer™ Protein Electrophoresis of GFP: A pGLO™ Bacterial Transformation Kit Extension. BIO-RAD. Accessed on January 2017.