Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Volume 39, Article 66, 2018

Bioinformatics of the Green Fluorescent Proteins

Alma E. Rodriguez Estrada

Aurora University, Biology Department, 347 S. Gladstone Ave, Aurora Illinois 60506 USA
(**arodriguezestrada@auroa.edu**)

Transformation of *Escherichia coli* with the Green Fluorescent Protein (GFP) is a laboratory activity that has become increasingly popular at diverse levels ranging from middle school to undergraduate courses. There are several mutant GFP genes that encode mutant proteins with small differences in their nucleotide and amino acid sequences. This bioinformatics activity was designed for an upper level course in molecular biology. Through this activity, students learn how to use GenBank® (the Basic Local Alignment Search Tool, BLAST®) and the Protein Data Bank (Benson et al. 2005, Berman et al. 2000) 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.

Keywords: Green Fluorescent Protein, GFP, bioinformatics, protein structure.

Introduction

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 GFP wild-type contains 238 amino acids and has a molecular weight of 26.9 kilodaltons (kDa). The protein is formed 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 open 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 the α-helix (Fig. 1). In the wild-type protein, the chromophore is composed of three amino acids: 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 long and it was the first cloned and expressed in other organisms (Zimmer 2002).

The green-fluorescent protein has been used in a wide variety of applications and to study protein properties and behaviors and cellular processes. Since this protein is found in jellyfish that inhibits 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 can be located close or far to the chromophore, buried or on the surface of the barrel structure (Zimmer 2002).

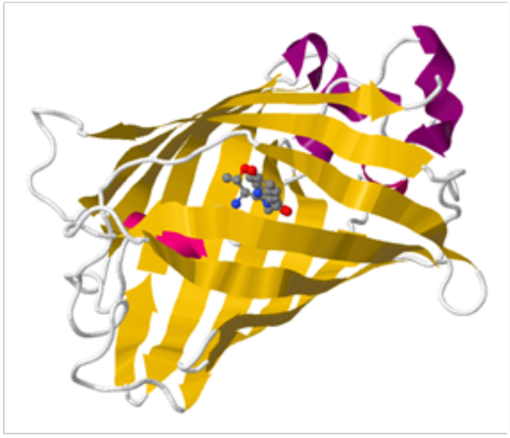


Figure 1. Structure of the wild-type GFP found in *Aequorea victoria*. Image from the RCSB PDB (www.rcsb.org) of PDB ID 1EMB (Brejc, K., Sixma, T. Green Fluorescent Protein (GFP) from *Aequorea victoria*, GLN 80 replaced with ARG).

Objectives*:*

1. Learn to use the Basic Local Alignment Search Tool (BLAST ®) and the Protein Data Bank
2. Determine the structure of the GFP gene
3. Characterize the GFP wild-type and two mutants

4. Review concepts related to the central dogma of molecular biology (transcription and translation)
5. Review the use of the genetic code table

Bioinformatics of the GFP

Although bacteria transformation with the GFP is a popular laboratory activity in lower and upper level courses at the college level and general biology courses in high schools (pGLO^{TM} Bacterial Transformation Kit, BIO-RAD), a deep understanding of the gene and protein structure is not necessarily addressed during the wet lab. 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 (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 to students accompanied by questions prompting students to explore the information available in the websites and to carefully analyze protein structure. For instance, students first explored the amino acid sequence, secondary and tertiary structure of the wild-type protein (1EMB) deposited in 1997 (Figs. 1 and 2).

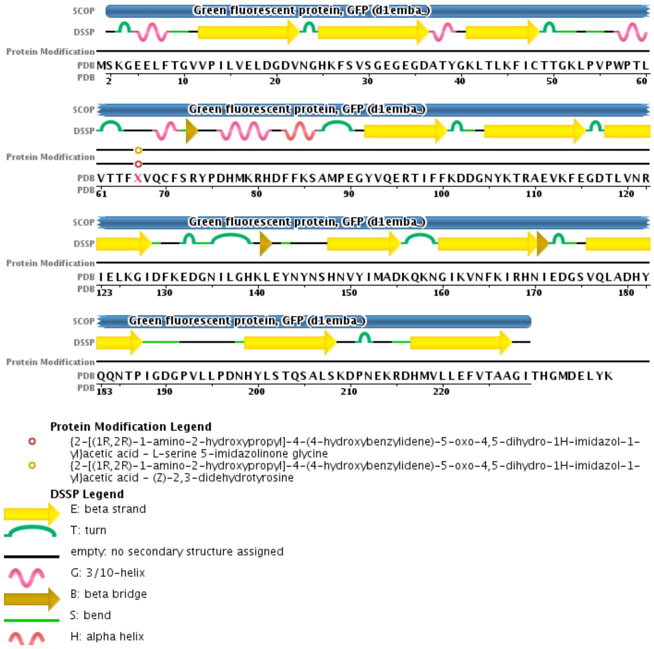

SCOP
DSSP
Protein Modification
PDB
Green fluorescent protein, GFP (d1emba_)
MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL
2 10 20 30 40 50 60
VTTFXVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNR
61 70 80 90 100 110 120
IELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHY
123 130 140 150 160 170 180
QQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK
183 190 200 210 220
Protein Modification Legend
{2-[(1R,2R)-1-amino-2-hydroxypropyl]-4-(4-hydroxybenzylidene)-5-oxo-4,5-dihydro-1H-imidazol-1-yl}acetic acid - L-serine 5-imidazolinone glycine
{2-[(1R,2R)-1-amino-2-hydroxypropyl]-4-(4-hydroxybenzylidene)-5-oxo-4,5-dihydro-1H-imidazol-1-yl}acetic acid - (Z)-2,3-didehydrotyrosine
DSSP Legend
E: beta strand
T: turn
empty: no secondary structure assigned
G: 3/10-helix
B: beta bridge
S: bend
H: alpha helix

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

From Figure 2, 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)
- Number of amino acids (238)

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^{TM} transformation experiment (BIO-RAD). The amino acid residues that are modified in the cycle 3 mutant can be identified using the "find" and "word count" function in word (Fig. 3).

Table 1. Comparison between the GFP wild-type found in *Aequorea victoria* and the cycle 3 mutant protein used in the pGLO^{TM} transformation experiment.

Position of the mutation	Amino acid in the wild type GFP protein	Amino acid in the cycle 3 mutant GFP
100	Phenylalanine	Serine
154	Methionine	Threonine
164	Valine	Alanine

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEG
DATYGKLTLKFICTTGKLPVPWPTLVTTFGYGVQC
FSRYPDHMKRHDFFKSAMPEGYVQERTIFFKDDG
NYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGH
KLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIED
GSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALS
KDPNEKRDHMVLLEFVTAAGITHGMDELYK

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

The National Center for Biotechnology Information (NCBI) website, Basic Local Alignment Tool (BLAST®) can then be used to find the gene sequence of the GFP. From the four tools available (blastn, blastp, blastx and tblastn), students were asked to select the one that is appropriate for that purpose. **Tblastn** allows finding the translated nucleotide form the amino acid sequence of a protein. The BLAST search can be limited 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
- **Amino acid residues forming the chromophore:** Gly-Tyr-Gly

AAGCTTTATTAAA**ATG**TCTAAAGGTGAAGAATTAT
TCACTGGTGTTGTCCCAATTTTGGTTGAATTAGATG
GTGATGTTAATGGTCACAAATTTTCTGTCTCCGGT
GAAGGTGAAGGTGATGCTACTTACGGTAAATTGAC
CTTAAAATTTATTTGTACTACTGGTAAATTGCCAGT
TCCATGGCCAACCTTAGTCACTACTTTC**GGTTATG**
GTGTTCAATGTTTTGCTAGATACCCAGATCATATG
AAACAACATGACTTTTTCAAGTCTGCCATGCCAGA
AGGTTATGTTCAAGAAAGAACTATTTTTTTCAAAG
ATGACGGTAACTACAAGACCAGAGCTGAAGTCAA
GTTTGAAGGTGATACCTTAGTTAATAGAATCGAAT
TAAAAGGTATTGATTTTAAAGAAGATGGTAACATT
TTAGGTCACAAATTGGAATACAACTATAACTCTCA
CAATGTTTACATCATGGCTGACAAACAAAAGAATG
GTATCAAAGTTAACTTCAAAATTAGACACAACATT
GAAGATGGTTCTGTTCAATTAGCTGACCATTATCA
ACAAAATACTCCAATTGGTGATGGTCCAGTCTTGT
TACCAGACAACCATTACTTATCCACTCAATCTGCC
TTATCCAAAGATCCAAACGAAAAGAGAGACCACA
TGGTCTTGTTAGAATTTGTTACTGCTGCTGGTATTA
CCCATGGTATGGATGAATTGTACAAA**TAA**CTGCAG

Figure 4. Nucleotide sequence of the GFP mutant 3 (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 sequence above (Fig. 4) 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 for 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 necessity for the development of mutant genes.

Conclusions

This bioinformatics activity can be completed in a single laboratory session of 170 minutes or less and could be completed 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 NCBI websites. A comparison of the wild-type protein, its mutants and other fluorescent proteins could be further performed using software such as MEGA.

Student Outline
Green Fluorescent Protein – Bioinformatics

Objectives

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

Introduction

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 green-fluorescent protein has been used in a wide variety of applications and to study protein properties and behaviors and cellular processes. Since this protein is found in jellyfish that inhibits 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 can be located close or far to the chromophore, buried or on the surface of the barrel structure (Zimmer 2002).

In this laboratory activity, you will use the Basic Local Alignment Search Tool (BLAST ®) and the Protein Data Bank in order to analyze the GFP protein and gene structure. This activity will also prompt you to review basic concepts related to gene expression and the genetic code.

Instructions

1. Visit the Protein Data Bank (PDB) website:
   http://www.rcsb.org/pdb/home/home.do

2. Briefly describe what this website is about (IN YOUR OWN WORDS):

3. In the search window type the following PDB code:

   1EMB

   and go!

4. What structure is this:

5. When was this structure deposited in the PDB?

6. Where was this protein expressed and isolated from?

7. Click the 3D view tab. Notice that you can rotate the image. Pick a view where you could clearly see the chromophore. Copy and paste that image in the space below?

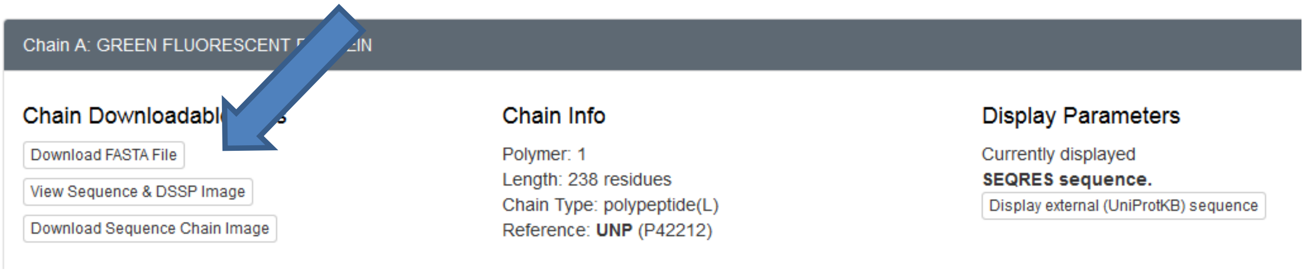
8. Click in the "sequence" tab. Copy and paste the image that shows the sequence and the secondary structures found in this protein:

9. How many beta sheets does the protein have (count from the image and number each sheet)?

10. How many helices (alpha and 3/10 helices) does the protein have (count from the image and number each)?

11. Identify the chromophore and mark it directly on the image.

12. Download the FASTA file:



Chain A: GREEN FLUORESCENT P EIN

Chain Downloadabl s

Download FASTA File

View Sequence & DSSP Image

Download Sequence Chain Image

Chain Info

Polymer: 1
Length: 238 residues
Chain Type: polypeptide(L)
Reference: **UNP** (P42212)

Display Parameters

Currently displayed
SEQRES sequence.

Display external (UniProtKB) sequence

13. The file will open in a notepad file. From that file, identify the amino acid sequence. Copy and paste it in the space below (there are many letters/words before the amino acid sequence actually starts… observe and figure out where is the start)

14. Use the "word count" and "find" functions in word to answer all the questions below.

    Number of amino acids in the amino acid sequence above: _________________

15. The sequence above and the sequence showed to you in the GFP laboratory manual have a slight discrepancy. Identify the discrepancy and briefly explain it below:

16. Identify the three amino acid changes described in page 4 of the GFP laboratory manual (BIO-RAD). Highlight those changes (yellow) in the sequence above (step 13).

17. Is the amino acid sequence above (1EMB, step 13) the wild type gene or the cycle 3 mutant (compare this sequence to the sequence in page 4 of your GFP laboratory manual).

    ____

18. Visit the National Center for Biotechnology Information (NCBI) website, Basic Local Alignment Tool https://blast.ncbi.nlm.nih.gov/Blast.cgi

19. Explore this website and circle out which tool would you use in order to find out the nucleotide sequence from which the amino acid sequence above (step 13) derives

    Nucleotide BLAST

    blastx

    tblastn

    Protein BLAST

20. Copy and paste the amino acid sequence (step 13) in the "Enter Query Sequence" window and scroll all the way down. Click BLAST
    From the list retrieved, search for the entry that has accession number GenBank: U73901.1. Click on that number. What kind of protein is this?

    Notice that you have now the nucleotide sequence!!! Click "FASTA". You will be prompted to a window with the DNA sequence. Copy and paste the nucleotide sequence in the space below:

21. In the sequence above, highlight in yellow the start and the stop codon. In the space below write the sequence of the start and stop codon identified above.

    Start codon: ____________

    Stop codon: ____________

22. What is the length of the gene's coding region?

23. Are there any introns in the EGFP gene?

24. Use the table that you created in your notebook as pre-lab last week. Identify the codons involved in the cycle 3 mutant. In order to facilitate your search, copy and paste below the sequence between the start and the stop codon. Remember; use the "word count" and "find" function in word to facilitate your search.

25. Keep in mind that the sequence above corresponds to the *egfp* gene mutant 3 (not the wild type or the cycle 3 mutant described in your GFP manual). In the sequence above, identify and highlight in blue the three codons involved in the formation of the chromophore.

26. What amino acid residues are involved in the formation of the chromophore in this mutant protein?

27. What amino acid residues are involved in the formation of the chromophore in the cycle 3 mutant.

28. Investigation:
    Many different GFP variants in nature exist and many different mutants have been developed in laboratories. Explain why.

Materials

Each student (or a pair of students) should have a computer with Internet access.

Cited 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^{TM} Protein Electrophoresis of GFP: A pGLO^{TM} Bacterial Transformation Kit Extension. BIO-RAD. Accessed on January 2017.

About the Authors

Alma Rodriguez is Associate Professor of Biology at Aurora University, where she teaches courses such as genetics, molecular biology, biochemistry, among others.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Rodriguez Estrada AE. 2018. Bioinformatics of the Green Fluorescent Proteins. Article 66 In: McMahon K, editor. Tested studies for laboratory teaching. Volume 39. Proceedings of the 39th Conference of the Association for Biology Laboratory Education (ABLE). **http://www.ableweb.org/volumes/vol-39/?art=66**