Structured Discovery: Learning about Gene Structure and Function through Analysis of *Arabidopsis* Genes

Andrew W. Woodward
The University of Mary Hardin-Baylor
Belton, Texas, USA

Undergraduate biology majors must memorize basic aspects of gene structure and function. However, students often have a shallow understanding of basic gene and protein features. I developed activities in which students analyze DNA and protein sequences to look for patterns. Students use NCBI[1] and The Arabidopsis Information Resource[2] (TAIR) to search for sequences. Students summarize intron/exon boundaries in an *Arabidopsis* gene to look for patterns. In their own words, students write an apparent rule for intron/exon boundary sequences based on data from one large gene. Then, they examine other genes from *Arabidopsis* and other creatures to determine whether their rule seems generally correct. Next, students read about snRNP complexes to learn how the splicing process relates to the sequence rule they devised. In a related protein module, students align protein sequences and make hypotheses about active-site and mutant amino acids. Finally, they observe the three-dimensional shape of the protein using UCSF Chimera[3] software, and they evaluate the accuracy of their hypotheses. Through data collection, hypothesis development, and model refinement, these *in silico* activities offer students a deeper experience of gene structure and function than is often possible in the classroom.

In the first phase of the activity, students analyze the alpha-amylase gene that encodes an enzyme we have already discussed during lecture. Students collect data to identify a pattern in the intron/exon boundaries:

1. Navigate to the UMHB Library website.
2. Under "Research," choose "Databases A-Z."
3. Choose "T" to choose the link "TAIR - The Arabidopsis Information Resource."
   a. Note that this database is UMHB login protected, but available to us off campus.
4. Hover over "Tools" and choose "SeqViewer."
5. *Await verbal instruction to proceed.*
6. In the search box, write the gene name for one of the three *Arabidopsis* α-amylase genes, "At1g69830" and click "Submit."
7. Click the red location marker on Chromosome 1.
8. Navigate left and right if necessary until a gene highlighted in yellow appears in the "Gene" portion of the viewer.
9. *Await verbal instruction to proceed.*
10. Hover over the highlighted gene. Read the summary information in the pop-up box, then click the blue "nucleotide seq view" link at the bottom of the pop-up box.
11. *Await verbal instruction to proceed.*
12. Reverse the sequence orientation so that you are viewing the coding strand for this gene rather than the template strand.
13. *Await verbal instruction to proceed.*
14. Now, begin examination of the twelve(!) splice-site boundaries in this large gene. Summarize the sequences below:

					5' Splice site						*Ignore the middle of the intron*							3' Splice site					
-5	-4	-3	-2	-1	1	2	3	4	5	-5	-4	-3	-2	-1	1	2	3	4	5

After writing nucleotides into the boxes, students make a hypothesis about the important nucleotides. Then, they look up another plant gene and a human gene to test whether their hypothesis holds with these other genes. Many students' hypotheses are simple and clear enough to apply. Other students write detailed hypotheses that do not apply outside the context of the *Arabidopsis* alpha-amylase gene.

Thus, phase one of the project is to construct a hand-made nucleotide sequence alignment and analyze the results. Next, students use a computer program to align amino acids.

In the second activity, students analyze protein sequences. Specifically, they construct a multiple-sequence alignment using the freely-available MEGA7[4] program.

1. Google "Pubmed" and follow the link to the Pubmed Home site.
2. In the drop-down box next to the search bar, the default is PubMed. Switch to Protein.
3. Enter the search term "PEX4" and click enter.
4. Follow the first link to investigate the PEX4 from *Saccharomyces cerevisiae* (bread yeast).
   a. A great deal of information is show. Scroll to the bottom of the page to glance at the protein sequence (shown using the one-letter abbreviations that you have memorized!).
5. Scroll back to the top and click the link to access the protein sequence in "FASTA" format.
6. Now click "Send to" (top right of the screen), select "File," and click "Create file."
   a. Save the file to the desktop using a code for the genus and species: "ScPEX4"
7. Follow steps 3-6 above for each of the following organisms (it would be fastest to use the organism name and PEX4 as search terms). Each time, be sure to follow careful file naming, using the genus and species letters before PEX4.
   a. *Cladophialophora carrionii* (a fungus that eats decaying plant material)
   b. *Arabidopsis thaliana* (a plant that needs no introduction)
   c. *Candida albicans* (a fungus, a yeast that lives in the human mouth)
   d. *Nicotiana attenuata* (a wild tobacco plant)
8. Before comparing these sequences, which organism on the list do you expect to have the most similar PEX4 protein sequence to *Arabidopsis,* and why?

The instructions continue guiding students through the construction of a multiple-sequence alignment. Then, students hypothesize where the active site of this enzyme might be (it is within the most conserved block of the sequences). Finally, they notice that an amino acid that is defective in the *pex4-1* mutant[5] of *Arabidopsis* appears to be distant from the active site based on the alignment.

Finally, in part three of the activity, students analyze the three-dimensional structure of the PEX4 protein using freely-available UCSF Chimera[3] software. In the process, they discover that the mutant amino acid is physically near the active site of the enzyme when viewed three dimensionally. This contrasts with its apparent distance from the multiple-sequence alignment in lesson two, in which the mutant amino acid and the active site appeared to be distant. Thus, the third phase of the activity highlights that higher order structure is a critical consideration when analyzing proteins.

References:

[1] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2020 Apr 17]. Available from: https://www.ncbi.nlm.nih.gov/

[2] The Arabidopsis Information Resource (TAIR), www.arabidopsis.org, [cited 2020 Apr 17].

[3] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25(13):1605-1612 (2004).

[4] Kumar, S., Stecher, G., and Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33(7):1870-1874 (2016).

[5] Zolman, B.K., Monroe-Augustus, M., Silva, I.D., and Bartel, B. Identification and functional characterization of Arabidopsis PEROXIN4 and the interacting protein PEROXIN22. *Plant Cell* 17(12):3422-35 (2005).

Notes:
I regret not including more screen shots of websites and student work, but I thought it inappropriate to post such items on a public site. I have attempted to format this poster for viewing on small screens.