Online Instruction – Bioinformatics Lesson for a COVID-19 Vaccine

NADJA ANDERSON, MARGARET WILCH

Abstract

In the spring of 2020, remote learning was implemented in schools throughout the world due to the pandemic of SARS-CoV-2, the novel coronavirus that causes the disease COVID-19. Thrust into online instruction, many science teachers scrambled during this transition, and classes were severely hampered by a lack of hands-on investigations involving critical thinking and problem-solving skills. In response to a need for online experimentation, bioinformatics lessons centered around SARS-CoV-2 were developed. This article presents a multipart bioinformatics lesson that allows students to (1) compare spike protein sequences from the database portal NCBI Virus, to investigate whether this protein would be a good target for a vaccine against COVID-19; and (2) create phylogenetic trees and demonstrate evolutionary relatedness of human coronaviruses. This lesson allows for instruction in molecular biology, virology, immunology, bioinformatics, and phylogenetics, as well as analysis of scientific data. It is appropriate for high school AP Biology and biotechnology courses and can be taught entirely online.

Key Words: alignment; bioinformatics; COVID-19; online; phylogeny; SARS-CoV-2; spike protein.

Introduction

The SARS-CoV-2 pandemic created a need for online high school instruction throughout the world. Major school districts, including San Diego Unified School District, Los Angeles Unified School District, and Chicago Unified School District, postponed the return to in-person classrooms and instead used online instruction in the fall of 2020 (Stieb, 2020). As long as states are encountering increasingly widespread pandemic case numbers, it would be irresponsible and dangerous to students and staff if they were to open their classrooms (LaFrance, 2020). Here, we describe a multipart bioinformatics lesson that can easily be taught online, requiring only a computer and an internet connection. The lesson is appropriate for high school AP Biology and biotechnology courses.

Advances in biology and biotechnology classes require that students interact with current, relevant concepts and master the tools used to answer authentic questions in molecular biology. Most secondary advanced biology coursework standards include hands-on laboratory work with an emphasis on inquiry-based investigations. In the classroom, students can gain science practices by utilizing tools such as pipetters, electrophoresis equipment, and thermal cyclers to better understand proteins and DNA. In addition to bench laboratory techniques, state standards require that bioinformatics be included in biotechnology curricula. Bioinformatics is a growing field and a much-utilized research tool that is applied across a broad spectrum of life sciences (Wefer, 2008; Machluf et al., 2017). In the current climate of the COVID-19 pandemic, remote classrooms can take advantage of lessons utilizing bioinformatics.

The lesson described here utilizes sequences that have been uploaded to GenBank (Mizrachi, 2002), a public database made available through the National Center for Biotechnology Information (NCBI) (Cooper et al., 2010). The resources and tools provided through GenBank and NCBI are used by most life science researchers and are essential to the understanding of current biological and health sciences. Utilizing these resources, students compare the spike protein, a viral surface protein (Figure 1), from several related strains of coronaviruses to determine whether the SARS-CoV-2 spike protein would be a target for a vaccine against COVID-19.

Prior to this lesson, students should have some knowledge of how the immune system functions in response to both infections and vaccines. In preparation for this lesson, students can explore the mechanism of different types of vaccines such as the whole pathogen vaccine (inactivated or live-attenuated), subunit vaccine, and nucleic acid vaccines (DNA or RNA). This activity focuses on the comparison of spike protein amino acid sequences using the multiple alignment and phylogenetic tools in DNA Subway (Hilgert et al., 2014). Students will retrieve protein sequences from NCBI and align sequences in DNA
Subway, providing them with similarities and differences between spike proteins of SARS-CoV-2 and other coronaviruses. The lesson begins with the comparison of spike proteins from the seven coronaviruses known to infect humans. Each part of the lesson will hone the comparison to more closely related viruses and, presumably, greater sequence similarity of spike protein (Figure 2). In the end, students will have tools to assess whether spike protein is a good target for a vaccine against COVID-19.

Materials for this lesson, as well as others, can be downloaded from the University of Arizona BIOTECH Project website (http://biotech.bio5.org/activities/covid19) at no cost for educators to use and adapt.

Bioinformatics Lesson – Developing a Vaccine against COVID-19

The goal of a vaccine is to safely create an immune response in an individual so that they are protected against an infectious disease or pathogen. An effective vaccine needs to elicit an immune response, the target for which is a part of the virus that is exposed (a surface protein) and therefore visible to the immune system. The vaccine target also needs to not change (mutate) and therefore elude the immune system. For example, HIV, the virus that causes AIDS, mutates rapidly, so attempts to create a vaccine against it have been unsuccessful. Mutations have created ~200 different strains of influenza, which necessitates new vaccines every year.

Students will need a DNA Subway account and will be using the Blue Line (Determining Sequence Relationships) for this analysis (Figure 3). Students will analyze the sequences of the spike protein, which is on the outside (surface) of all coronaviruses and is therefore exposed to the immune system, thus fulfilling one of the requirements for an effective vaccine target. There are seven different coronaviruses known to infect humans:
• SARS-CoV-2, which causes COVID-19 (Beta-CoV)
• SARS-CoV, the original SARS from 2003 (Beta-CoV)
• MERS (Middle East Respiratory Syndrome) outbreak in 2012 (Beta-CoV)
• OC43, which causes common cold symptoms (Beta-CoV)
• HKU1, which causes common cold symptoms (Beta-CoV)
• 229E, which causes common cold symptoms (Alpha-CoV)
• NL63, which causes common cold symptoms (Alpha-CoV)

To address the question of whether the spike protein is a good target for a COVID-19 vaccine, students will conduct five alignments using the spike proteins of the other coronaviruses that infect humans and ascertain the sequence similarities by comparing the amino acid sequences.

Part 1: How Similar Are the Spike Proteins between All of the Seven Coronaviruses?

Section A – Identifying and copying the spike protein translations for all seven coronaviruses that infect humans

In order to compare the spike proteins of these seven coronaviruses, students conduct an internet search for NCBI Virus, and click on the blue box Search by virus (Figure 4A). You can then input information by name or ID number, or select Up-to-date-SARS-CoV-2 (Figure 4B).

The taxid for all SARS-CoV-2 sequences in the NCBI database is 2697049 (Figure 5). The first result(s) shown are called RefSeq (reference sequence). Note that the first SARS-CoV-2 genome sequence was added on January 13, 2020, and has an accession number (NC_045512). Every gene sequence uploaded to NCBI is given a unique accession number that serves as a catalog number in the genomic archive GeneBank. After the RefSeq, the SARS-CoV-2 sequences are listed in chronological order, starting with the most recent deposit. The number of deposits is growing daily, with a total

Figure 4. (A) NCBI Virus search window: search for SARS-CoV-2 under Search by virus. (B) Select Up-to-date SARS-CoV-2. Alternatively, type in taxid in search window; the taxid is 2697049.
of 61,835 on February 5, 2021. As a comparison, that number was 17,855 on August 31, 2020, when this paper was first submitted.

When NC_045512 is selected, a Nucleotide Details window opens, with attributes about that sequence, such as its length (Figure 5). Be sure the sequence you are looking at is the entire genome, approximately 30,000 base pairs.

The newly opened GenBank record lists the accession number, NC_45512, as well as the genome size in base pairs and notes that the genome is linear single-stranded RNA. Scroll down the page to see the FEATURES section, which details the translated sequences of open reading frames (ORFlab) and genes associated with the virus; and, at the very bottom of the page, the ORIGIN section provides the nucleotide sequence, which is presented as DNA, even though this is an RNA genome. In the middle of the FEATURES section, students will find the translated amino acid sequence of the spike protein, labeled “S” for gene and “surface glycoprotein” for product (protein product) (Figure 6).

Students will copy the entire translated amino acid sequence from MFV…HYT, and paste this sequence into a Word document and save it in FASTA format, in which each unique sequence is identified with a > followed by a defining name. For example:

```text
>refseq_SARS-CoV-2_NC_045512_2020Jan13
MFVFVLPLPSQCVNLTTQPLPPATNS...HYT
```

Save the SARS-CoV-2 sequence in FASTA format, which renders it ready for DNA Subway analysis. Add the other six coronaviruses’ spike protein sequences in FASTA format to the Word document. To add the sequence from the original SARS, from the 2003 outbreak, search SARS in NCBI Virus and select the first result, Severe acute respiratory syndrome-related CoV taxid:694009 (Figure 7). On February 5, 2021, there were 63,342 results, including both the original SARS-CoV and the current SARS CoV-2; almost 62,000 of these were SARS-CoV-2 sequences. The first result is the SARS-CoV-2 reference sequence and the second is the 2003 SARS reference sequence (accession NC_004718). Click the accession number, copy the spike protein translated sequence, and add the FASTA format sequence to your Word document. Be sure to create a new name using the > and an identifying name for each new virus.

The sequence for the other five viruses can be found using the same instructions. MERS taxid is 1335626, OC43 taxid is 31631, HKU1 taxid is 290028, 229E taxid is 11137, and NL63 taxid is...
The FASTA-formatted spike protein translations for all seven coronaviruses that infect humans can then be added to DNA Subway for alignment of the sequences.

Section B – Aligning coronavirus spike protein sequences

Open DNA Subway, Blue Line (Figure 8). Select project type protein, select enter sequences in FASTA format, and copy your seven FASTA-formatted sequences in your Word doc and paste them in the window (or upload the file). Give your project a name and continue.

View the sequences in sequence viewer. Align the sequences using MUSCLE, selecting all of the sequences to align the seven CoVs. A barcode alignment will be displayed (Figure 9) and compared to the “consensus sequence.” The consensus sequence is somewhat arbitrary, in that it represents whichever amino acid is found at that position most often out of the sequences you have selected. One additional sequence could significantly change the consensus. Students can look at the differences between the CoV spike protein sequences and determine whether there are any patterns of similarities between any of the viruses. Select SEQUENCE SIMILARITY % to see the similarity between any two sequences.

Section C – Building a phylogenetic tree

To determine evolutionary relationships between different groups, the analysis requires a rooting or outgroup. This is an organism that is closely related to the study group but not part of it. Students can choose a variety of viruses as the outgroup, but a virus with a spike protein would allow the analysis to be a direct comparison to the sequences in the investigation. There are many coronaviruses that do not infect humans; choosing one in another genus would provide a good outgroup for evolutionary comparison of the seven coronaviruses that infect humans. Any Gamma-CoV would be a good choice to serve as a rooting outgroup for the phylogenetic tree. Go back to NCBI Virus, Search by virus and type in Gamma-CoV, then select Gamma-CoV, taxid:694013. Copy the spike protein sequence and add the FASTA format to the DNA Subway project.

Select data, including the Gamma-CoV, align with MUSCLE, and then build a tree using PHYLIP ML (Figure 10).

Students will be able to assess the evolutionary relationships of the viruses based on the spike protein. The length of the tree branch represents the number of amino acid differences between the sequences, which can be used as a correlation of time since the most recent common ancestor between the viruses. Students should be able to determine that the two SARS-CoV have a common ancestor in the recent past, as do the two “common cold” Beta-CoVs. Although this tree put MERS more closely related to the SARS viruses, the common ancestor between these three is almost as distant as the common ancestor for all of the beta coronaviruses. Given the limitations of these data, the question of whether the beta coronaviruses had a more recent ancestor to the alpha coronaviruses or the gamma coronavirus was left unresolved.

The next two parts of the lesson will utilize the sequences added during part 1.

Part 2: How Similar Are the Five Beta-CoVs’ Spike Proteins?

Using the sequences in DNA Subway, select all the Beta-CoVs: OC43, HKU1, MERS, SARS, and SARS-CoV-2, and align using MUSCLE (Figure 11). Students will be able to look for similarities in the alignment. They can begin to look at the regions of similarity – which, if the region is similar to the consensus sequence, is designated by the whitish bands. In the end, these regions could be important targets for vaccine development.

Students will be able to build a phylogenetic tree using the Gamma-CoV as the outgroup and compare the tree with all seven human coronaviruses to the tree with just the five beta coronaviruses. This analysis may allow them to see that addition or removal of data can change the representation of the relationships. Another exercise to demonstrate this effect is through building the tree with different outgroups.

Part 3: How Similar Are the Two SARS-CoV (SARS CoV, SARS CoV-2) Spike Proteins to Each Other?

Select sequences for SARS CoV, SARS-CoV-2 and use MUSCLE for the alignment. Students will notice an even greater amount of similarity between these two sequences, getting them even closer to finding regions of the spike protein that could be a good target for a vaccine. However, even these two sequences have significant differences.
Part 4: What Is the Similarity among Different SARS-CoV-2 Spike Proteins?

Students will now have the experience to understand the importance of similarity in alignment of sequences. Our question as to whether the spike protein could be a good target for a vaccine can be further addressed by analyzing different SARS-CoV-2 spike proteins and seeing if mutations have accumulated in these sequences. To do this, students will need to return to NCBI Virus and search for several SARS-CoV-2 sequences. Search for the virus by typing in SARS-CoV-2 (taxid is 2697049) in the NCBI Virus search window (see instructions above). Select at least 10 other SARS-CoV-2 spike protein sequences to add to the DNA Subway project to compare with your reference sequence (RefSeq), which was collected in December 2019. Make sure to search for full-length SARS-CoV-2 sequence, look at the length of the sequence (top light blue row of results), and only select sequences with close to 30,000 nucleotides, which represents the entire SARS-CoV-2 genome.

To capture the scope of mutations or differences in SARS-CoV-2, select sequences based on time, particularly the date collected, and on location. Notice that after the RefSeq, the sequences are in order of most recently released. Some of those may be from a much earlier date; for example, accession no. MW570899 was released on February 5, 2021, from a sample that was collected on August 19, 2020 (Figure 12). Scroll right in the results window to see the collection date.
The main object of this lesson is to introduce students to computational analysis for biological questions. In this case, the question is: "Would SARS-CoV-2 spike protein be a good choice as a target for vaccine development?" Students will have the opportunity to learn about immunology and development of different kinds of vaccines, as well as whether the spike protein is a good vaccine target candidate. Students will also experience navigating NCBI's GenBank genome submissions and manipulate these sequences in DNA Subway. Students will compare diverse coronaviruses and successively make the comparisons more specific, beginning with the seven CoVs that infect humans, to the five Beta-CoVs, to the two SARS-CoVs, and, finally, compare the amino acid sequence of spike protein from several SARS-CoV-2. These selected data will be imported into an alignment tool in DNA Subway and compared to each other. In addition, DNA Subway offers students the opportunity to see the evolutionary relationships of the viruses in the form of phylogenetic trees. One of the highlights of this lesson is that the alignment and tree-building exercises can yield different results depending on the data used. In the end, students will be able to see that the SARS-CoV-2 spike protein fits the criteria of a part of the virus that is exposed and a protein that does not appear to be changing rapidly, and therefore offers the majority of the protein as potential targets for vaccine development.

**Conclusion**

The main object of this lesson is to introduce students to computational analysis for biological questions. In this case, the question is:
LaFrance, A. (2020). This push to open schools is guaranteed to fail. *Atlantic*, August 2.

Machluf, Y., Galbart, H., Ben-Dor, S. & Yarden, A. (2017). Making authentic science accessible – the benefits and challenges of integrating bioinformatics into a high-school science curriculum. *Briefings in Bioinformatics, 18*, 145–159

Mizrachi, I. (2002) [updated 2007]. GenBank: the nucleotide sequence database. In J. McEntyre & J. Ostell (Eds.), *The NCBI Handbook* (pp. 1–17) Bethesda, MD: National Center for Biotechnology Information

Stieb, M. (2020). The major school districts that will remain online-only this fall. *New York Magazine*, August 4.

Wefer, S. (2008). Bioinformatics in high school biology curricula: a study of state science standards. *CBE–Life Sciences Education, 7*, 155–162.

NADJA ANDERSON (nadja@bio5.org) is the BIOTECH Project Director at the University of Arizona, Tucson, AZ 85721. MARGARET WILCH (margaret@sarsef.org) is the Director of Research at SARSEF, Tucson, AZ 85711.
Article Categories

A note about article word count: Please recognize that tables, figures, and photographs add to the overall length of the article. One page of text has approximately 1,000 words, therefore a 1/4-page graphic will count for 250 words. More extensive graphics should be budgeted accordingly. References are also included in the final article word count.

**Feature Article** (up to 4,500 words) includes topics of general interest to readers of *ABT*. Consider the following examples of content that would be suitable for the feature article category:

- Research on teaching alternatives, including evaluation of a new method, cooperative learning, concept maps, learning contracts, investigative experiences, educational technology, simulations and games, and biology and life science education standards
- Social and ethical implications of biology and how to teach such issues as genetic modification, energy production, agriculture, climate change, health care, nutrition, and cultural responsiveness
- Reviews and updates of recent advances in the life sciences in the form of an “Instant Update” that brings readers up-to-date in a specific area
- Imaginative views of the future of biology education and suggestions for adjusting to changes in schools, classrooms, and student populations
- Other timely, relevant, and interesting content such as discussions of the role of the Next Generation Science Standards in biology teaching, considerations of the nature of science with implications for the classroom, considerations of the continuum of biology instruction from K–12 to post-secondary teaching environments, or contributions that consider the likely/ideal future of science and biology instruction

**Research on Learning** (up to 4,500 words) includes reports of original research on innovative teaching strategies, learning methods, or curriculum comparisons. Studies should be based on sound research questions, hypotheses, discussion of appropriate design and procedures, data and analysis, discussion on study limitations, and recommendations for improved learning outcomes.

**Inquiry and Investigations** (up to 3,500 words) is the section of *ABT* that features discussion of innovative laboratory and field-based strategies. Strategies in this section should be original, engaging, practical, and related to either a particular program such as AP and/or linked to standards such as NGSS. Submissions should also be focused at a particular grade/age level of student and must include all necessary instructions, materials list, worksheets, and assessment tools. Other appropriate contributions in this category are laboratory experiences that engage students in inquiry.

**Tips, Tricks and Techniques** (up to 1,500 words but may be much shorter) features a range of suggestions useful for teachers including laboratory, field, and classroom activities; motivational strategies to assist students in learning specific concepts; modifications of traditional activities; new ways to prepare some aspect of laboratory instruction; etc.

Writing & Style Guidelines

The *Chicago Manual of Style, 17th Edition* is the guide for questions of punctuation, abbreviation, and style. List all references in alphabetical order on a separate page at the end of the manuscript. Please review a past issue for examples. Use first person and a friendly tone whenever appropriate. Use concise words to emphasize your point rather than capitalization, underlining, italics, or boldface. Use the SI (metric) system for all weights and measures.

While calls for specific themed issues of *ABT* are infrequent, February and April are traditionally themed editions on Evolution and the Environment, respectively.

Preparing Tables, Figures, and Photographs

**General Requirements**

- When your article is accepted, we will require that figures be submitted as individual figure files in higher resolution format. See below for file format and resolution requirements.
- Authors should be aware that color is limited within the journal. All artwork, figures, tables, etc. must be legible in black and white. If color is important to understanding your figures, please consider alternative ways of conveying the information.

**Article Photographs**

Digital files must meet the following guidelines:

- Minimum resolution of 300 DPI, 600 DPI is preferred
- Acceptable file formats are TIFF and JPEG
- Set to one-column (3.5” wide) or two-column size (7” wide)
- If figure originates from a website, please include the URL in the figure caption. Please note that screen captures of figures from a website are normally too low in resolution for use.

**Tables and Figures**

- Minimum resolution of 600 DPI, 1200 DPI is preferred
- Acceptable file formats are TIFF, BMP, and EPS
- Set to one-column (3.5” wide) or two-column size (7” wide)

If you have any questions, contact Valerie Haff at managingeditor@nabt.org.
Submission Guidelines

NOTE: All authors must be current members of NABT or a charge of $100 per page is due before publication.

All manuscripts must be submitted online at http://mc.manuscriptcentral.com/u PRESS-ABT

- Authors will be asked to register the first time they enter the site. Upon receiving a password, authors can proceed to upload their manuscripts through a step-by-step process. Assistance is available in the “Author Help” link found in the menu on the left side of the page. Additional assistance is available from the Managing Editor (managingeditor@nabt.org).
- Manuscripts must be submitted as Word or WordPerfect files.
- Format manuscripts for 8.5 x 11-inch paper, 12-point font, double-spaced throughout, including tables, figure legends, and references.
- Please place figures (including photos) and tables where they are first cited in the text along with appropriate labels. Make sure to include figure and table citations in the text, as it is not always obvious where they should be placed. At the time of initial submission, figures, tables and images should be low resolution so that the final file size remains manageable.
- If your article is accepted, the editors will require that figures be submitted as figure files in higher resolution form. See section on Preparing Tables, Figures, and Photographs.

Supplemental Materials

In order to maintain the word count for individual articles, we are pleased to facilitate publication of supplemental materials accompanying the online issue. If authors have materials (figures, examples, worksheets, appendices, multimedia files, etc.) that support but are not essential to the printed text of the article, authors can include those as separate files with their article submission.

Editorial Procedures

- Communications will be directed only to the first author of multiple-authored articles.
- Typically, three individuals who have expertise in the respective content area will review each article.
- The editors attempt to make decisions on articles as soon as possible after receipt, but the process can take six to eight months, with the actual date of publication to follow. Authors will be emailed editorial decisions as soon as they are available.
- Accepted manuscripts will be forwarded to the Copy Editor for editing. This process may involve making changes in style and content. However, the author is ultimately responsible for scientific and technical accuracy. Page proofs will be sent to authors for final review before publication at which time only minor changes can be made.

Submitting Images

Cover Images

Submissions of cover photographs from NABT members are strongly encouraged. Covers are selected based on the quality of the image, originality, composition, and overall interest to life science educators. ABT has high standards for cover image requirements and it is important for potential photographers to understand that the required size of the cover image generally precludes images taken with cell phones, point-and-shoot cameras, and even some older model digital SLR cameras.

Please follow the requirements listed below.

- Email possible cover images to Kathleen Westrich at kmwestrich@yahoo.com.
- ABT covers feature an almost-square image with a slight vertical orientation.
- Choose an image with a good story to tell. Do not crop the subject too tightly. It is best to provide an area of background around the subject.
- Include a brief description of the image, details of the shot (i.e., circumstances, time of day, location, type of camera, camera settings, etc.), and your biographical information in an email message.
- Include your name, home and email addresses, and phone numbers.
- Please ensure that the image meets the minimum standards for publication listed below and has not been edited or enhanced in any way. The digital file must meet the minimum resolution of 300 pixels per inch (PPI)—preferred is 400 PPI—and a size of 8.5 x 11.25". We accept TIFF or JPEG images only.

BioMystery Images

Authors who have a photo that does not meet the high resolution required for a cover photo can consider submission as the “BioMystery” image. These submissions can be sent to ABTEditor@nabt.org with a description of the subject of the image.

Thank you for your interest in The American Biology Teacher. We look forward to receiving your manuscripts.

William McComas, Editor-in-Chief
ABTEditor@nabt.org
Valerie Haff, Managing Editor
managingeditor@nabt.org