Visualising DNA, RNA & Proteins in Jalview School Workbook
About this workbook:

The workbook contains 4 practical web-based bioinformatic projects:

**Project 1** views DNA and RNA sequences and their 3D structures.

**Project 2** views protein sequences and their 3D structures.

**Project 3** compares the sequence of human myoglobin protein with myoglobin sequences from other animals; the student then views the coding DNA sequence, and codons, alongside the protein.

**Project 4** views exons and introns in the *HBB* gene, which is involved in sickle cell anaemia. The genetic mutations linked to this disease is identified by viewing the coding DNA and its protein product in a split-screen view.

The workbook, the links for launching JalviewJS to run the exercises, as well as additional resources such as videos are available at the ‘Resources for Schools’ web page on the Jalview website: [http://www.jalview.org/school-resources](http://www.jalview.org/school-resources)

Who is this workbook for:

<table>
<thead>
<tr>
<th>Knowledge required:</th>
<th>Secondary school biology pupils (aged 16-18 years old)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment needed:</td>
<td>A desktop or laptop computer with a web browser and internet access</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Getting Started using JalviewJS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project 1-Viewing DNA &amp; RNA</strong></td>
<td></td>
</tr>
<tr>
<td>Exercise 1: Viewing DNA</td>
<td>2</td>
</tr>
<tr>
<td>Exercise 2: Viewing RNA</td>
<td>7</td>
</tr>
<tr>
<td><strong>Project 2-Viewing Proteins</strong></td>
<td></td>
</tr>
<tr>
<td>Exercise 1: Viewing different classes of proteins</td>
<td>8</td>
</tr>
<tr>
<td><strong>Project 3-Viewing Myoglobin Protein &amp; its coding DNA</strong></td>
<td></td>
</tr>
<tr>
<td>Exercise 1: Comparing myoglobin sequences</td>
<td>12</td>
</tr>
<tr>
<td>Exercise 2: Viewing myoglobin &amp; its coding DNA</td>
<td>13</td>
</tr>
<tr>
<td><strong>Project 4-Viewing the genetic mutation responsible for sickle cell anaemia</strong></td>
<td></td>
</tr>
<tr>
<td>Exercise 1: Viewing exons &amp; introns in the <em>HBB</em> gene</td>
<td>17</td>
</tr>
<tr>
<td>Exercise 2: Viewing coding DNA of the <em>HBB</em> gene &amp; its protein product</td>
<td>19</td>
</tr>
<tr>
<td><strong>Appendix</strong></td>
<td>23</td>
</tr>
<tr>
<td><strong>Glossary</strong></td>
<td>25</td>
</tr>
</tbody>
</table>

What is Jalview?

Jalview is free-to-use computer software developed at the University of Dundee. It is designed to assist research scientists to visualise and analyse DNA, RNA and proteins. It uses an interactive multi-window interface for viewing sequences, alignments, annotations and three-dimensional structures. Jalview can read files directly from public biological databases and has a number of analysis tools for aligning sequences, producing trees, measuring similarities and comparing structures.

[www.jalview.org](http://www.jalview.org)

The workbook was produced by Dr Suzanne Duce with help from Mungo Carstairs, Benedict Soares, Bob Hanson, Dmitry Finkelbergs, Charlotte Campbell, Jim Procter & Geoff Barton

We would like to acknowledge the BBSRC and Wellcome Trust who have funded this work.

Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, DD1 5EH

© 29th May 2020
Getting Started using JalviewJS

Learning Objectives:
• Access the Jalview schools web page
• Open the Jalview Schools Workbook
• Launch JalviewJS

1. Open a web browser such as Chrome or Firefox. We suggest you avoid Internet Explorer as JalviewJS does not always work in older browsers.

2. Search using keywords 'Jalview Schools'. From the list of results, select 'Resources for Schools' at http://www.jalview.org/school-resources. This will open the Jalview Schools web page.

3. Select the 'View the School Workbook' link on the Jalview Schools web page.

4. The Schools Workbook will open in an adjacent tab.

5. Return to Jalview Schools web page. Launch JalviewJS by clicking the link 'open JalviewJS'.

6. JalviewJS program opens in an adjacent window in the browser.

(i) To move the JalviewJS window, place the mouse on title panel on the top of the window, then click-and-drag.
(ii) To enlarge the window, place the mouse on the lower right hand corner of the window, then click-and-drag.
(iii) To close the window, click the ‘X’ in the top right hand corner of the window.
Project 1: 'Viewing DNA & RNA'
Project 1-Exercise 1: Viewing DNA

Learning Objectives:
• Open DNA sequences in Jalview
• Colour the nucleotide bases in the sequence
• View the 3D structure

1. **Click the link** named 'View DNA fragment' in Section (2): Project 1-Exercise 1 of the Schools web page.

2. JalviewJS with the DNA sequences opens in an adjacent window. Select the **Colour** menu in the alignment window. Select **Nucleotide**.

3. The Jalview nucleotide colour scheme is adenine bases are green, cytosine bases are yellow, guanine bases are red and thymine bases are blue.

4. This is the sequence from a fragment of B-DNA (PDB id 3BSE).

5. Select both the **A & B 3BSE** sequences named PDB/3bse/3BSE/... in the alignment window by **clicking and dragging** the mouse. **Right click** the mouse to open the pop-up menu. Select **3D Structure Data**.

6. In the 'Structure Chooser' box, select **both the 3BSE entries** by **clicking and dragging** the mouse or using the **Shift** key. Click **New View**.

7. A 3D structure window opens containing the molecule. Click this window with the mouse, then drag the mouse to change the view. Notice how the **adenine bases** align with the **thymine bases**. The **cytosine bases** align with the **guanine bases**. These are **complementary base pairings**.
### Project 1-Exercise 1: Viewing DNA

Q: What are the names of the 4 different DNA bases, and their single letters identifier?

<table>
<thead>
<tr>
<th>DNA Base</th>
<th>Single Letter Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenine</td>
<td>A</td>
</tr>
<tr>
<td>Thymine</td>
<td>T</td>
</tr>
<tr>
<td>Guanine</td>
<td>G</td>
</tr>
<tr>
<td>Cytosine</td>
<td>C</td>
</tr>
</tbody>
</table>

Q. How many nucleotide base pairs are there in this fragment of DNA? (Tip: count the base pairs in the 3D structure window)

8. Select the **Colours** menu in the Jmol structure window. Select the **Purine/Pyrimidine** colour scheme.

9. The **purines** (adenine and guanine bases) are coloured pink. The **pyrimidines** (thymine and cytosine bases) are coloured cyan.

10. In the alignment window, hover the mouse over the first base in the top strand (A) and view its location on the structure. Next, hover the mouse over the first base in the second strand (B) of the alignment and view its location on the structure.

Note where the first base in each sequence is located in the Jmol structure window (step 10).

Q. Are the DNA strands parallel or anti-parallel?

Q. Look at the shape of purine and pyrimidine, how are they different? (see Appendix Table 1)

11. In the Jmol structure window, select the **View** menu. Select **Show Chain**. Uncheck **3BSE:B**.

12. Drag the mouse to change the view, so you are looking **down** the spiral. (Check that first adenine base on the strand A is at the top by hovering the mouse over the base in the alignment). Check the direction in which the spiral turns.

Q. In Step 12, what direction is the strand turning, clockwise or anticlockwise?
For more information about the DNA fragment 3BSE visit the Protein Data Bank website at: www.rcsb.org/structure/3BSE

Change the appearance of the windows:
(i) To move the window, place the mouse on the title panel on the top of the window, then click-and-drag.
(ii) To enlarge the window, place the mouse on the lower right hand corner of the window, then click-and-drag.
(iii) To close the window, click the ‘X’ in the top right hand corner of the window.

Jmol window commands:
(i) To rotate the structure place mouse on the structure, then click-and-drag.
(ii) To zoom press the shift key, then click-and-drag.
For more information about Jmol visit: http://wiki.jmol.org/index.php/Main_Page

Introduction to JalviewJS’s multi-window interface
DNA (deoxyribonucleic acid) is made up of two strands in a double helix. The two DNA strands are anti-parallel with respect to each other.

A DNA strand is a polymer. Its sub-units are called nucleotides. A nucleotide consists of a phosphate group, a 5 ring sugar and a nitrogenous base (see Table 1 in the Appendix).

Each carbon in the sugar ring is assigned a number. The base is attached to the 1' carbon (reads 1 prime). The hydroxyl group is attached to the 3' carbon. The phosphate group is attached to the 5' carbon.

The backbone of a strand of DNA is made of deoxyribose sugars linked to phosphates by phosphodiester bonds. The bases are attached to the backbone as sidechains. There are four different bases: guanine (G), thymine (T), cytosine (C) and adenine (A).

The nucleotide bases between the two intertwining strands of DNA form weak hydrogen bonds. The adenine base aligns with the thymine base. They form 2 hydrogen bonds. The cytosine base aligns with the guanine base. They form 3 hydrogen bonds. These pairings are called complementary base pairings.

A DNA strand has polarity. One end is called the 3' end and other is 5' end. This relates to the position of the 3' and 5' sugar carbons.

In the figure opposite, the 3' carbons beside the hydroxyl groups are coloured cyan. The 5' carbons beside the phosphate groups are coloured purple.
Project 1-Exercise 2: Viewing RNA

Learning Objectives:
• Open an RNA sequence in Jalview
• Colour the nucleotide bases in the sequence
• View the 3D structure

1. **Click the link** named ‘View RNA’ in Section (2): Project 1-Exercise 2 on the Schools web page.

2. JalviewJS with the RNA sequence opens in an adjacent window. Select the **Colour** menu in the alignment window. Select **Nucleotide**.

3. The Jalview nucleotide colour scheme is adenine bases are green, cytosine bases are yellow, guanine bases are red and uracil bases are blue.

4. This is the SAM responsive riboswitch mRNA (PDB id 2GIS). For more information go to https://en.wikipedia.org/wiki/SAM_riboswitch_(S-box_leader).

5. Click the mouse on the sequence name **PDB/2gis/2GIS/....** in the alignment window. **Right click** the mouse to open the pop-up menu. Select **3D Structure Data**.

6. Select **2gis** in the ‘Structure Chooser’ box. **Click New View**.

7. A 3D structure window opens containing the RNA molecule. For instructions on how to change the appearance of the Jmol structure window go to page 5.

Q: What are the names of the four different RNA bases?
Q. How does RNA and its nucleotides differ from those of DNA? (see Appendix Table 1)
Project 2: 'Viewing proteins'
Project 2-Exercise 1: Viewing Proteins

Background: The amino acids and their order in a protein sequence determine the shape and chemical characteristics of a protein, this in turn, influences the function (role) of a protein.

Learning Objectives:
• Open the protein sequences in Jalview
• Colour the amino acid residues
• View the 3D structure

1. Click the link named 'View Myoglobin protein' in Section (2): Project 2-Exercise 1 on the Schools web page.

2. JalviewJS with the protein sequence opens in an adjacent window. Select the Colour menu in the alignment window. Select Taylor.

3. In the Taylor colour scheme, each amino acid residue has its own individual colour. (For the key to the 1-letter amino acid codes see Appendix Table 3).

4. Use the horizontal scroll bar to scroll to the end of the sequence. Click on the last residue and view the information in the status box in the lower left hand corner of the alignment window.

5. Click the mouse cursor on the myoglobin sequence name. Right click the mouse to open the pop-up menu. Select 3D Structure Data.

6. In the 'Structure Chooser' box, select PDB Id 3rgk entry. Click New View.

7. A 3D structure window opens containing the human myoglobin protein. (For instructions on how to change the appearance of the Jmol structure window go to page 5).

Q. How many amino acid residues are there in the human myoglobin protein? (see step 4)
Q. How many different protein amino acids are found in nature? (see Appendix Table 3)
Q. What is the name of the amino acid that is represented by the letter A?
Q. What is the chemical formula of the amino acid glycine (G)?
8 Repeat this process for the other proteins listed on the web page: 
collagen (structure - PDB id code 1cag) 
amylase (enzyme - PDB id code 1smd) 
antibody (defence - PDB id code 1igt) 
insulin (signalling - PDB id code 6bcx) 
ferritin (storage - PDB id code 5xb1) (this is a large file so it may take a while to download).
Click the appropriate link eg ‘View protein’ on the Schools web page.

9. Select the Colour menu. 
Select colour scheme eg Zappo.

10. Click and drag the mouse to select all the names of the sequences listed. 
Right click the mouse to open the pop-up menu. 
Select 3D Structure Data.

11. In the 'Structure Chooser' box, select PDB Id entries by clicking and dragging the mouse or using the Shift key. 
Click New View.

In the Jmol structure window, select the Colour menu. 
Select By Chain to view each of the chains in the 3D structure.

Q. In the 3D structure viewer, identify the alpha helix and beta sheet regions? 
(See the Appendix for more information about secondary structure).
Open sequence files by reading them from public databases

1. Go to the File menu in the desktop window. Select Fetch Sequences.

2. In the 'New Sequence Fetcher' box, select PDB from the 'Select Database' list. Select/Click OK.

3. Enter the PDB ID code eg 1smd in the 'PDB Sequences Fetcher' box. Press [return]. Select the PDB ID code from the list eg 1smd. Click OK. (Close the 'PDB Sequence Fetcher' box once the sequence has loaded).

4. Select the View menu in the alignment window. Uncheck Show Sequence Features. (Note: This needs to be toggled off otherwise it will mask the residue colour schemes).

<table>
<thead>
<tr>
<th>Protein type</th>
<th>Function</th>
<th>Example</th>
<th>PDB codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme</td>
<td>Catalysis of chemical reactions</td>
<td>Amylase [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=74">www.rcsb.org/pdb/101/motm.do?momID=74</a>]</td>
<td>1smd</td>
</tr>
<tr>
<td>Storage</td>
<td>Stores small molecules or ions</td>
<td>Ferritin [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=35">www.rcsb.org/pdb/101/motm.do?momID=35</a>]</td>
<td>5xb1</td>
</tr>
<tr>
<td>Transport</td>
<td>Carry substances around the body</td>
<td>Myoglobin [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=1">www.rcsb.org/pdb/101/motm.do?momID=1</a>]</td>
<td>1mbn</td>
</tr>
</tbody>
</table>

Alternative method to open files using public databases

Open sequence files by reading them from public databases
1. Go to the File menu in the desktop window. Select Fetch Sequences.

2. In the 'New Sequence Fetcher' box, select PDB from the 'Select Database' list. Select/Click OK.

3. Enter the PDB ID code eg 1smd in the 'PDB Sequences Fetcher' box. Press [return]. Select the PDB ID code from the list eg 1smd. Click OK. (Close the 'PDB Sequence Fetcher' box once the sequence has loaded).

4. Select the View menu in the alignment window. Uncheck Show Sequence Features. (Note: This needs to be toggled off otherwise it will mask the residue colour schemes).
Project 3:
'Viewing myoglobin protein & its coding DNA'
Project 3-Exercise 1: Comparing myoglobin sequences

Learning Objectives:
• Compare myoglobins from different animals
• Produce protein similarity tree

Background: Myoglobin is a protein located in muscle, it complexes with iron for oxygen storage. For example, the concentration of myoglobin in muscle cells affects how long an organism can hold its breath. In 1958, whale myoglobin was the first protein ever to have its 3D structure revealed by X-ray crystallography. Max Perutz and John Kendrew won a Nobel Prize in chemistry for this work.

1. Click the link named 'View Myoglobin Sequences and create a Tree' in Section (2): Project 3-Exercise 1 on the Schools web page.

2. JalviewJS with the protein alignment opens in an adjacent window. Select the Colour menu. Select Zappo.

3. In the Zappo colour scheme, the amino acids are coloured based on their physicochemical properties. (For the key to the 1-letter amino acid codes see Appendix Table 3).

4. Select the Calculate menu. Select Calculate Tree or PCA.

5. Click Calculate in the 'Choose Calculation' box. By default Neighbour Joining and Blosum62 should be selected, and the Principal Component Analysis option not selected. (Close the 'Choose Calculation' box once the tree has appeared by clicking the "X").

6. A tree window opens containing a tree generated from human, gorilla, whale, cow, mouse, dolphin, chicken, ostrich, dog, sheep and elephant myoglobin sequences. The pattern of branching in the tree reflects protein similarity. Move the tree window away from the alignment by placing the mouse on the top title bar, then click-and-drag the mouse.
7. Click the mouse at a different location on the tree, and a red vertical line appears. The red line groups sequences. Each group has its own randomly generated colour. Move the mouse to a different location on the tree, this will change the grouping and colours of groups. Each branch has a number, they can be added together to produce similarity scores.

8. Select the View menu in the tree window. Select Sort Alignment by Tree.

9. The sequences in the alignment window are reordered to reflect the tree. Note how the sequence names in the alignment window have the same colours as those in the tree window.

Q. Looking at the tree window, are the groupings what you might expect? (Tip: The shorter the length of the branches between species, the more similar they are. The length between each branch points is shown in brackets).

Q: From the tree (step 7), which animal has the most similar myoglobin to humans?

Q: From the tree (step 7), compared to humans which animals have the least similar myoglobin?

**Uniprot identifier codes used in this exercise**

<table>
<thead>
<tr>
<th>Uniprot ID</th>
<th>Species</th>
<th>Uniprot ID</th>
<th>Species</th>
<th>Uniprot ID</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>P02144</td>
<td>Human</td>
<td>P04247</td>
<td>Mouse</td>
<td>P63113</td>
<td>Dog</td>
</tr>
<tr>
<td>P02147</td>
<td>Gorilla</td>
<td>P68276</td>
<td>Dolphin</td>
<td>P02190</td>
<td>Sheep</td>
</tr>
<tr>
<td>P02185</td>
<td>Whale</td>
<td>P02197</td>
<td>Chicken</td>
<td>P02187</td>
<td>Elephant</td>
</tr>
<tr>
<td>P02192</td>
<td>Cow</td>
<td>P85077</td>
<td>Ostrich</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P02144; P02147; P02185; P02192; P04247; P68276; P02197; P85077; P63113; P02190; P02187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Project 3-Exercise 2: Viewing myoglobin & its coding DNA

Learning Objectives:
• View the myoglobin protein sequence and its coding DNA sequence in a split-screen window
• View the DNA codons

1. Click the link named 'View Myoglobin protein & its coding DNA' in Section (2) Project 3-Exercise 2 on the Schools web page.

2. In the JalviewJS desktop window, select the File menu. Select Fetch Sequences.

3. Select EMBLCDS from the 'Select Database' list in the 'New Sequence Fetcher' box.

4. Enter the ID CAG46747 in the 'New Sequence Fetcher' box. Click OK.
(Close the 'New Sequence Fetcher' box once the sequence has loaded).

5. Select the Calculate menu in the alignment window. Select Get Cross References. Select Uniprot.

6. Note: It may take a little while for Jalview to retrieve the files depending on wi-fi speeds.

7. A split-screen window opens containing the coding DNA sequence (upper panel) and the myoglobin protein sequence (lower panel).
Project 3-Exercise 2: Viewing myoglobin & its coding DNA

8. The DNA and protein sequence panels are linked. Place the mouse over an amino acid and the associated DNA triplet (codon) is highlighted in a black box.

Note: There are 3 DNA bases for each amino acid residue (see the codon code in Appendix Table 2).

9. Place the mouse on the leucine amino acid L at residue 3, and note the associated 3 DNA bases (codon) highlighted in the black box. Repeat for leucines at residue 10, and for leucines at residue 12.

Q. When an amino acid residue is selected in the protein sequence, why are three DNA bases highlighted by a black box in the DNA sequence?
Q. What DNA triplet bases are associated with leucines at residue 3, 10 and 12?
Q. Would you expect them to be the same? (see Appendix Table 2)
Project 4: 'Viewing the genetic mutation responsible for sickle cell anaemia'
A chromosome is made up of DNA tightly coiled around proteins called histones. Each human cell normally contains 23 pairs of chromosomes, i.e. a total of 46 chromosomes. This is an image of a set of 46 stained metaphase chromosomes from a male human. [Image courtesy of National Human Genome Research Institute]

Chromosomes are often represented in an idealized arrangement with their centromeres aligned. In the image both the male (XY) and female (XX) versions of the 23rd chromosome are shown. [Image courtesy of NCBI Genome Decoration Page]

This is a graphical representation of human chromosome 11. Each chromosome contains several genes. Genes are regions of DNA that produce proteins during translation. [Image courtesy of National Library of Medicine]

The location of the haemoglobin-beta (HBB) gene, responsible for sickle cell disease is marked with an arrow. [Image courtesy of National Library of Medicine]

Genes are made up of protein coding regions called exons, separated by regions called introns. Introns do not encode protein products but appear to play a role in regulation and gene expression. [Image courtesy of National Human Genome Research Institute]

The figure showing the relationship between the DNA in chromosomes, genes, exons and introns. [Image courtesy of Thomas Splettstoesser]
1. Click the link named ‘View Exons & Introns in the HBB gene’ in Section (2): Project 4-Exercise 1 on the Schools web page.

2. JalviewJS with the HBB gene opens in an adjacent window. The pink DNA regions are introns. The green DNA regions are exons.

3. Move the overview window and increase its size. The red box shows the part of the sequence visible in the alignment window. Use the mouse to drag the red box to the right.

4. Note: The bases in the haemoglobin beta (HBB) gene are labelled with the 1-letter identifier: G (guanine), T (thymine), C (cytosine) and A (adenine).

5. To toggle the Overview window on and off, go to the View menu in the alignment window. Select the Overview Window.

Q. What is the main difference between exons and introns? (see page 18)
1. Click the link named 'View the coding DNA & its protein product' in Section (2): Project 4-Exercise 2 on the Schools web page.

2. In the empty JalviewJS desktop window, select the File menu. Select Fetch Sequences.

3. Select EMBLCDS from the 'Select Database' list in the 'New Sequence Fetcher' box.

4. Enter the ID CAG46711 in the 'New Sequence Fetcher' box. Click OK to retrieve the coding DNA sequence. (Close the 'New Sequence Fetcher' box once the sequence has loaded).

5. Select the Calculate menu in the alignment window. Select Get Cross References. Select Uniprot. (Note: You may have to wait a little whilst Jalview fetches the data).

6. A split-screen window opens that contains the HBB coding DNA sequence (top panel) and HBB protein (lower panel). The panels are inter-linked. Place the mouse over an amino acid in the protein and view the associated codon in the DNA sequence.

7. There are two links on the top left hand corner of the JalviewJS web page. Click-and-drag the top link Drag 'this link on DNA sequence to add features' onto the HBB gene sequence panel in the upper DNA alignment window. This opens the features file and colours the DNA sequence.
8. Click-and-drag the second link Drag 'this link on protein sequence to add features' onto the protein sequence panel in the lower protein alignment window. This opens the features file and colours the protein sequence.

9. The sickle cell mutation is highlighted in red on both the DNA and protein sequences. The HBB gene contains 3 exons (one green and two purple).

10. In the upper DNA panel, select the View menu. Select the Overview Window to view the exons.

11. In the upper panel, place the mouse over the adenine (A) at base number 20 in the DNA sequence. A tooltip opens, note information. In the lower protein panel, repeat with the glutamic acid amino acid (E) at residue number 7. A tooltip opens, note information.

   Note: If the tooltip doesn't open, right click the mouse and select Feature Details instead. Additional information is available in Status Bar in lower left hand corner.

Q. What nucleotide triplet codon code produces the glutamic acid at residue 7?
Q. From the information in the tooltip, what is the change in the nucleotide base at base number 20 that is responsible for sickle cell anaemia?
Q. In the mutated HBB protein, what amino acid replaces the glutamic acid at residue 7?

12. Click the mouse cursor on the protein sequence name. Right click the mouse to open the pop-up menu. Select 3D Structure Data.

13. In the 'Structure Chooser' box, select PDB id 3nmm entry (the first entry in the Best Quality list). Select New View.
14. A 3D structure window opens containing the HBB protein. (For instructions on how to change the appearance of the Jmol structure window go to page 5).

15. In the protein sequence panel, select the View menu in the alignment window. Select Features Settings...

16. In the 'Sequence Feature Settings' box, click the mouse on the red Sickle_cell_variant feature name and drag it to the top of the list, above the green RESNUM feature name. Click OK.

17. Rotate the 3D structure to locate residue 7 (coloured red) that is mutated in sickle cell anaemia. Mouse over the protein residue 7 in the protein sequence (lower panel of the split-screen viewer). See the effect in the 3D viewer.

Q. How does a single mutation at base 20 in the HBB gene result in a person getting sickle cell disease?

Q. Do you know the evolutionary advantage that heterozygote sickle cell mutation infers (only 1 of the HBB genes has the mutation), that has resulted in an increased prevalence of sickle cell anaemia in Africa? (see https://en.wikipedia.org/wiki/Heterozygote_advantage).

If you enjoyed the exercises in this workbook, we would appreciate it if you would post us a message on Twitter or Facebook. Please include @Jalview, #STEM, #genetics

Additional Help: Help can be accessed from the Help menu in the Jalview's desktop window, and select Documentation. If you are interested in learning more, the Jalview manual contains several hands-on exercises, available at www.jalview.org/about/documentation. Or visit the Jalview YouTube training channel at https://www.youtube.com/channel/UCjpnvZB770yz7tbrJ0tw and view the videos and playlists.
Table 1: Composition of DNA & RNA subunits

DNA and RNA are polymers made up of nucleotide sub-units. The nucleotide consists of a phosphate group, a 5 ring sugar and a nitrogenous base. DNA contains a deoxyribose sugar and has a thymine base. RNAs contains a ribose sugar and has an uracil base. RNAs are usually single-stranded.

Table 2: Codon Table

A codon is a set of three nucleotides, or triplet, that code for a specific amino acid residue during protein synthesis.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>F</td>
</tr>
<tr>
<td>TTC</td>
<td>L</td>
</tr>
<tr>
<td>TTA</td>
<td>S</td>
</tr>
<tr>
<td>TTG</td>
<td></td>
</tr>
<tr>
<td>TCT</td>
<td>P</td>
</tr>
<tr>
<td>TCC</td>
<td></td>
</tr>
<tr>
<td>TCA</td>
<td></td>
</tr>
<tr>
<td>TCG</td>
<td></td>
</tr>
<tr>
<td>TAT</td>
<td>H</td>
</tr>
<tr>
<td>TAC</td>
<td></td>
</tr>
<tr>
<td>TAA</td>
<td>stop</td>
</tr>
<tr>
<td>TAG</td>
<td></td>
</tr>
<tr>
<td>TGT</td>
<td>C</td>
</tr>
<tr>
<td>TGG</td>
<td>stop</td>
</tr>
<tr>
<td>CTG</td>
<td></td>
</tr>
<tr>
<td>CTA</td>
<td></td>
</tr>
<tr>
<td>CTC</td>
<td></td>
</tr>
<tr>
<td>CTG</td>
<td></td>
</tr>
<tr>
<td>CCT</td>
<td></td>
</tr>
<tr>
<td>CCC</td>
<td></td>
</tr>
<tr>
<td>CCA</td>
<td></td>
</tr>
<tr>
<td>CGG</td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td></td>
</tr>
<tr>
<td>CAC</td>
<td></td>
</tr>
<tr>
<td>CAA</td>
<td></td>
</tr>
<tr>
<td>CAG</td>
<td></td>
</tr>
<tr>
<td>CGT</td>
<td></td>
</tr>
<tr>
<td>CGC</td>
<td></td>
</tr>
<tr>
<td>CGA</td>
<td></td>
</tr>
<tr>
<td>CGG</td>
<td></td>
</tr>
<tr>
<td>GCT</td>
<td></td>
</tr>
<tr>
<td>GCC</td>
<td></td>
</tr>
<tr>
<td>GCA</td>
<td></td>
</tr>
<tr>
<td>GCG</td>
<td></td>
</tr>
<tr>
<td>GAT</td>
<td></td>
</tr>
<tr>
<td>GAC</td>
<td></td>
</tr>
<tr>
<td>GAA</td>
<td></td>
</tr>
<tr>
<td>GAG</td>
<td></td>
</tr>
<tr>
<td>GGT</td>
<td></td>
</tr>
<tr>
<td>GGC</td>
<td></td>
</tr>
<tr>
<td>GGA</td>
<td></td>
</tr>
<tr>
<td>GGG</td>
<td></td>
</tr>
</tbody>
</table>

Proteins

Biologists use four terms to describe the structure of proteins:

- **Primary structure** is the sequence of the amino acids present in the polypeptide chain.
- **Secondary structure** are two particularly stable subunit structures. An α (alpha) helix is a spiral, with the R side chain groups sticking outwards. A β (beta) sheet is a corrugated sheet, with the R groups lying above and below the sheet. The structures are stabilised by hydrogen bonds.
- **Tertiary structure** describes the 3-dimensional folded shape of the protein. The tertiary structure is stabilised by interactions between groups, including hydrogen bonds, ionic bonds, disulphide bridges, hydrophobic interaction and Van der Waals interactions.
- **Quaternary structure** refers to the structure that forms when several individual polypeptides subunits or proteins link together to form a larger aggregate protein complex.
Amino Acids

Amino acids are the building blocks (sub-units) of proteins. Attached to the central carbon is a hydrogen, a NH$_2$ amino group (this can have a positive charge depending on pH), a COOH carboxylic group (this can have a negative charge depending on pH) and variable side group R. The R sidechain influences whether an amino acid is polar or non-polar, acidic or basic.

Amino acids link together to form proteins by forming peptide bonds between the COOH and NH$_2$ of neighbouring amino acids during a condensation reaction that releases water.

### Table 3: Amino acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3 letter code</th>
<th>1 letter code</th>
<th>Sidechain polarity</th>
<th>Linear</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>nonpolar</td>
<td>CH$_3$-CH(NH$_2$)-COOH</td>
<td>C$_3$H$_7$NO$_2$</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>basic polar</td>
<td>HN=C(NH$_2$)-NH-(CH$_2$)$_2$-CH(NH$_2$)-COOH</td>
<td>C$_6$H$_9$N$_4$O$_2$</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>polar</td>
<td>H$_2$N-CO-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_4$H$_7$N$_3$O$_2$</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
<td>acidic polar</td>
<td>HOOC-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_4$H$_7$NO$_4$</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>nonpolar</td>
<td>HS-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_3$H$_7$NO$_2$</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
<td>acidic polar</td>
<td>HOOC-(CH$_2$)$_2$-CH(NH$_2$)-COOH</td>
<td>C$_5$H$_9$NO$_4$</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
<td>polar</td>
<td>H$_2$N-CO-(CH$_2$)$_2$-CH(NH$_2$)-COOH</td>
<td>C$_5$H$_9$N$_3$O$_3$</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td>nonpolar</td>
<td>H-CH(NH$_2$)-COOH</td>
<td>C$_2$H$_5$NO$_2$</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>basic polar</td>
<td>NH-CH=NH-CH=CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_6$H$_9$N$_2$O$_2$</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>nonpolar</td>
<td>CH$_3$-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_6$H$_13$NO$_2$</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td>nonpolar</td>
<td>(CH$_2$)$_2$-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_6$H$_13$NO$_2$</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td>basic polar</td>
<td>H$_2$N-(CH$_2$)$_2$-CH(NH$_2$)-COOH</td>
<td>C$_6$H$_14$N$_2$O$_2$</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
<td>nonpolar</td>
<td>CH$_3$-S-(CH$_2$)$_2$-CH(NH$_2$)-COOH</td>
<td>C$_5$H$_11$NO$_2$S</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
<td>nonpolar</td>
<td>Ph-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_6$H$_11$NO$_2$</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
<td>nonpolar</td>
<td>-NH-(CH$_2$)$_2$-CH=COOH</td>
<td>C$_4$H$_7$NO$_2$</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
<td>polar</td>
<td>HO-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_3$H$_7$NO$_3$</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
<td>polar</td>
<td>CH$_2$-CH(OH)-CH(NH$_2$)-COOH</td>
<td>C$_4$H$_9$NO$_4$</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
<td>nonpolar</td>
<td>Ph-NH-CH=CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$<em>{11}$H$</em>{12}$N$_2$O$_2$</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
<td>polar</td>
<td>HO-Ph-CH$_2$-CH(NH$_2$)-COOH</td>
<td>C$_5$H$_11$NO$_3$</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
<td>nonpolar</td>
<td>(CH$_2$)$_2$-CH-CH(NH$_2$)-COOH</td>
<td>C$_5$H$_11$NO$_2$</td>
</tr>
</tbody>
</table>
Glossary

**Amino acid:** molecular sub-units of peptides and proteins.

**Bioinformatics:** the application of computer and statistical techniques to the management of biological data.

**cDNA (complementary DNA):** cDNA sequence is synthesized from an RNA template by reverse transcription. It contains 5' and 3' untranslated regions (UTRs) as well as coding regions.

**CDS (protein-coding sequence):** the portion of the mRNA transcript that is translated by ribosomes into proteins.

**Chromosome:** located in the cell nucleus, it contains the cellular DNA along with a number of proteins (e.g., histones) that compact and package the DNA.

**Codon:** a set of three adjacent nucleotides (triplet) that code for a specific amino acid residue during protein synthesis.

**DNA (deoxyribonucleic acid):** the molecule that encodes genetic information. It carries the instructions for all aspects of an organism's functions such as growth, metabolism and reproduction. These chains can be over 100,000,000 molecules in length.

**Exon:** the sections of a gene that are translated into proteins, they remain in the transcript (mRNA) after introns have been spliced out of the genomic sequence.

**Gene:** a region of DNA that encodes a specific protein or protein subunit.

**Genetic code:** sets of triplet nucleotides letters that encodes specific amino acids.

**Genome:** all of the genetic material in the chromosomes of a particular organism.

**Genomic DNA (gDNA):** all the DNA residing in the chromosomes.

**Genotype:** all the genes in a particular individual.

**Intron:** the noncoding part of the genome that is transcribed then spliced out of the RNA.

**Phenotype:** the observable characteristics or features of a living organism.

**Phylogenetic tree:** an evolutionary tree for organismal species or cellular macromolecules that is built using inheritance or molecular sequence information.

**Protein:** a biological macro-molecule composed of a string of amino acids joined together by peptide bonds.

**Protein sequence:** the sequence of amino acids in a protein.

**Nucleoside:** nucleotides without a phosphate group.

**Nucleotide:** building blocks of RNA and DNA made up of a nitrogenous base, a molecule of sugar and phosphoric acid.

**Multiple sequence alignment:** an alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column.

**RNA (ribonucleic acid):** RNA are similar to DNA but containing the ribose sugar rather than deoxyribose sugar and the base uracil (U) rather than thymine (T). Typically they are single-stranded.

**Replication:** process by which DNA makes a copy of itself during cell division.

**Sequence alignment:** arranging the sequences of protein, RNA or DNA to identify regions of similarity. The similarity could be a consequence of functional, structural, or evolutionary relationships.

**Translation:** process where mRNA is decoded by ribosomes to produce specific amino acids and polypeptides.

**Transcription:** process where a segment of DNA is copied into RNA by the enzyme RNA polymerase.

---

**Free Public Biological Databases:**

- **UniProt** is a database of protein sequences ([http://www.uniprot.org/](http://www.uniprot.org/)).
- **Protein Data Bank (PDB)** is a database of crystallographic, three-dimensional structural data of large biological molecules ([http://www.rcsb.org/](http://www.rcsb.org/)).
- **Ensembl** is a genomic database ([http://ensemblgenomes.org/](http://ensemblgenomes.org/)).
- **EMBL (CDS)** data originates from the European Nucleotide Archive (ENA) database of annotated DNA and RNA sequences ([https://www.ebi.ac.uk/ena](https://www.ebi.ac.uk/ena)).